www.ajocict.net

African Journal of Computing & ICTs

Volume 6. No. 2. June, 2013

June, 2013
www.ajocict.net

African Journal of Computing & ICTs

Volume 6. No. 2. June, 2013

June, 2013
www.ajocict.net

A Journal of the Institute of Electrical & Electronics Engineers (IEEE)
Computer Chapter Nigeria Section

ISSN- 2006-1781

www.ajocict.net

Contents

www.ajocict.net

www.ajocict.net

Editorial Board

www.ajocict.net

Foreward

The African Journal of Computing & ICT remains at the nexus of providing a platform for contributions to discourses, developments, growth and implementation of Computing and ICT initiatives by providing an avenue for scholars from the developing countries and other nations across the world to contribute to the solution paradigm through timely dissemination of research findings as well as new insights into how to identify and mitigate possible unintended consequences of ICTs. Published papers presented in this voume provide distinctive perspective on practical issues, opportunities and dimensions to the possibilities that ICTs offer the African Society and humanity at large. Of note are the increasing multi-disciplinary flavours now being demonstrated by authors collaborating to publish papers that reflect the beauty of synergistic academic and purpose-driven research. Obviously, these developments will drive growth and development in ICTs in Africa.

This issue of the African Journal of Computing & ICTs contains journal articles with a variety of perspective on theoretical and practical research conducted by well-grounded scholars within the sphere of computer science, information systems, computer engineering, electronic and communication, information technology and allied fields across the globe. While welcoming you to peruse this volume of the African Journal of Computing and ICTs, we encourage you to submit your manuscript for consideration in future issues of the Journal

Wishing you a productive reading

Thank you

Longe Olumide Babatope PhD
Managing Editor
Afr J Comp & ICTs

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

A Metric for Evaluating Information System Management

[1]M.C. Okoronkwo & [2]M.N. Agu
Department of Computer Science
Univerisity of Nigeria
Nsukka Nigeria
[1]mcjemeka@yahoo.com, matthew.okoronkwo@unn.edu.ng [2]monica.agu@unn.edu.ng

ABSTRACT
Everything that is created including Information systems IS, starts to age from the moment of birth. The ageing progresses gradually or rapidly depending on management, control and maintenance (MCM) strategy adopted. The MCM affects the ageing and determines how fast IS becomes outdated and seizes to promote or realize the service goals that it's aimed at providing. This paper provides a framework for evaluating the MCM processes of IS by providing a metric for effective and efficient measurement of IS management. Regardless of the technical or organisational environment of the IS, its primary objective is to serve the organisation it is situated. But the world around it is in constant change, eventually causing changes in the technical environment and the overall effectiveness and efficiency of the systems. The framework that is presented in this paper is conceptual in nature; it provides a facility for understanding, explanation and anticipation the state of IS in any organisation. It uses IS entity: functions and technology to describe and qualify the past/present and to anticipate the future development path of an IS, so as to maintain or increase the return on investment while reducing its ageing and eventual death process lag.

Keywords: Evaluation, Information Systems, Management, Situation, Metrics, Description and Qualification**.**

African Journal of Computing & ICT Reference Format:
M.C. Okoronkwo & M.N. Agu (2013). A Metric for Evaluating Information System Management.
Afr J. of Comp & ICTs. Vol 6, No. 2. pp 1-10

1. INTRODUCTION

During the development of information systems or any other type of complex application, most of the effort expended is on getting the system to run, incorporating novel features, and allocating resources to accomplish project goals in a timely fashion. Many times evaluation of these systems is viewed as something that can be postponed until the end of the process, but all too frequently there is no time to do the needed testing then; at other times, evaluation is not even factored into the goals of the development effort. If convenient, easy to use methods were available in an environment in which evaluation were being fostered, then evaluation might be an activity that serves as an end-point for development [2].

Even in ideal cases where all requirements are met by perfect design, the system starts to outdate from the very day of its completion, due to changes in its use environment, that is, the domain area the system was designed to cover changes after the system was implemented. When people come and go through organization, the rules provide a constant for handling routine situations. Thus rules not only transfer past learning, they also control behaviour within the organization [8] **.**

There is the need according to [10] for Metrics; measurable standards that monitor the effectiveness of goals and objectives to be established for IS. The metrics measure the implementation of IS policy, the results of services and the impact of management on an enterprise's mission. Victor et al stated that IS metrics can be obtained at different levels within an organization and that detailed metrics collected at the system and network level, can be aggregated and rolled up to progressively higher levels, depending on the size and complexity of an organization. If such measurements are instantaneous snapshots of a particular measurable parameters, then the metrics are more complete pictures, typically comprised of several measurements, baselines, and other supporting information that provide context for interpreting the measurements. Good metrics are *goal-oriented* and should have the following features: *specific, measurable, comparable, attainable, repeatable,* and *time dependent*.

In this framework for evaluating IS, the entire organisation management processes are divided into it primary entities and examine from management perspective. The step of evaluating an organisation IS starts from describing the present situation (PS) of the IS entities: the Real System (RS), Information Technology (IT) and the Management, Control and Maintenance (MCM). Further a description of their Relationships and Influences were outlined.

These steps are necessary if a get clear picture of the PS is to emerge. We need to distinguish between issues that are critical for IS from those issues that are non-critical. The next stage is the qualification of each entity; here the issues that we need to qualify and how to qualify them in the PS situation are discussed. The procedure is to determine what issues to qualify in each entity and the next step is to state how to qualify the issues, in the respective entities.

Having described and qualified the entities, the next is to apply the metrics, using the weighting process attributes approach [11] to differentiate process attributes that are considered more important from those that are less important according to the perceptions of the person(s) conducting the qualification in line with organisation goals. The proposed metrics format presented, with which one can apply various uses statistics to determine the status of management at each entity and decide on lines of actions to address the outcome, which if implemented will take the IS to the desired situation.

2. THE DESCRIPTION OF THE ENTITIES IN PRESENT SITUATION (PS)

2.1 Real System (RS):

In the RS entity key issues that should be of concern to the evaluator include: user requirements and preconditions imposed by Information technology (IT) utilization and the management. User requirements to consider are availability, flexibility, maintainability, performance, reliability, and security of IT. Other issue include preconditions imposed by utilization and management, information policy and planning (IPP), centralization and de-centralization of activities, concentration and de-concentration of IT, financial resources, personnel allocation, safety of staff and users, standardization of IT, and service level agreements concerning IT services.

These preconditions make up the initial framework from which other factors contingent upon the case study is examined. Other concepts related to IT in the entity RS include the situational or contingency factors, which impact on the utilization and management of IT. These include size and location [6]**,** the technology environment, organizational culture, and the general level of communication infrastructure in and around the organisation.

Information Technology (IT):

In the IT entity, the issues considered vital for inclusion are the hardware, software and the network components based on their demand, common utility and usage. The user requirements for IS, i.e. availability, maintainability, performance, reliability, and security of IS, are included to correspond to IT service support and service delivery processes.

A service is defined herein as an essential intangible set of benefits or activities that are sold or provided by one party to another. Further the IT issues may be depicted to determine their states in the Extended Model. In such a case the focus include development, utilization, exploitation and maintenance of IT as these concern users more directly than other states. In addition to the above, complexity factors associated with IT may be examined. It is desirable that for less costly IS management, the level of complexity be reduced, however, not all complexity factors need be to reduced, but with knowledgeable understanding, proper assessment of the complexity factors becomes easier to handle.

2.1 Management, Control and Maintenance (MCM)

In MCM entity, functional management (FM), application management (AM), and technical management (TM), at Strategic Level (SL), Tactical Level (TL) and Operational Level (OL), form an important part of management of information systems. They are included to emphasize the important role that they play in IT service provision; however, discussion of tasks, task areas and task fields at the three levels is implicitly assumed within MCM processes. It is considered desirable that the goals of the MCM entity should be aligned to the goals of the entity RS for effective and efficient management and utilization of ICT to achieve the overall mission and objectives of the organization. In summary, the following relationships between the entities are defined (the arrow → indicate the relationship between entities):

RS→ICT RS *exploits* ICT	users in the entity RS require resources in the entity ICT to enable users to perform their day-to-day activities
ICT → RS ICT *supports* RS	Resources in the entity ICT support the activities in the entity RS through enhancing effectiveness and efficiency
ICT→MCM (ICT *supports* MCM)	Resources in the entity ICT provide useful information required for ICT management to the personnel staff in the entity MCM
MCM→ICT (MCM *manages* ICT)	Personnel in the entity MCM manage the resources in the entity ICT
RS→MCM; (RS employs MCM)	The entity RS *employs* the entity MCM to manage its information and communication technology resources.
MCM→RS:	technical staff members in the entity MCM *respond to* requests in RS

3. THE EVALUATION METHODOLOGY

3.1 Describing the PS of IS

The second step in the description function is to determine how the PS is depicted. The criterion used for description is *existence*. This criterion is predicated on the premise that before we can determine what issues within the IS framework are required or should be improved upon; there is the need to determine if the issues do indeed exist. If the issue *exists*, then a value 1 is assigned and if it does *not exist* a value *0* is assigned. A checklist of all the above IS related issues in the framework is then created and the values are tallied accordingly. The overall result indicates the number of issues that exist in the PS IS. This depiction, which includes verification of the issues, forms the basis upon which qualification of the PS, definition of the desired situation DS and transformation from PS to DS state depend.

In summary each entity is examined from the perspective of the entire IS State in as follows:

RS: - User requirements, Preconditions and Situational factors

IT: - Hardware, Software, Network components, Extended State Model (ESM) and Complexity Factors (CFs).

MCM: - Functional Management, Application Management and Technical Management at Strategic, Tactical and Operational levels.

Service support processes: - Change, Configuration, Help desk and Problem Managements.

Service delivery processes: - Availability, Capacity, Contingency, Cost and Service level managements

Relationships between entities: - RS & IT, MCM & IT and RS & MCM

Influences: - Managerial, Donor, Technological, economic and Cultural.

3.2 Qualification of the PS

Here we seek to qualify the issues depicted above. In doing so, we isolate the issues in each of the entities. In the entity RS, the main focus is on the *actors (*or players), who make demands, referred to as requirements, for IT to enable them perform their activities. Hence, in *user requirements*, the focus is on whether the users (staff) are aware of their own requirements. Also, we seek to determine whether user requirements have been formulated.

In *preconditions*, we seek the levels of user awareness and the formalization of user requirements in a framework, called preconditions, by senior management officials, that has been achieved. These include IPP, centralization/decentralization of activities, concentration of IT, de-concentration of IT, financial resources, personnel allocation, safety of staff, standardization of IT, and service level agreements. In *situational factors*, we seek an indication of whether the factors have had an impact on the utilization, exploitation and maintenance of IT, in the organisation.

In the entity IT, the focus is on the quality and states of the *objects* or IT resources, i.e. if and to what extent the hardware, software, and network components are utilized, exploited and maintained by the organization. In considering the availability of hardware resources, i.e. basic application, under the prevailing situational factors as depicted in the PS situation. It is critical to reflect on whether these resources are effectively and efficiently utilized, exploited and maintained or not. In the entity MCM, the primary focus is on the forms of management and processes**.** In determining what to qualify in this entity, the capability maturity model approach by [12] is adopted. Using this approach, each process is qualified according to whether the processes are: recognized, planned and practiced, the technical staff are trained, enforced and measured, performed rationally, cost effective and on schedule. Furthermore, whether before any process commences, there are any preparedness criteria, documented, standardized, peer reviews are conducted. On completion of any process, are: there any criteria for its completion, the services provided predictable, structured actions and organized as a service.

In the relationships between entities, we seek to determine the strength of the relationships between entities. Clearly, strong and positive relationships between entities are more desirable than weak or negative ones. Strong relationships indicate greater awareness of IS issues in the organization as opposed to weak or negative relationships, which indicate a lack of awareness and thus possible source of problems in the organization with regard to IS issues. It is, therefore, worth investigating and designating values to the relationships with a view later to understanding the level of involvement of various entities in the support and delivery of IS services in the organization.

In examining the relationship between entities, the key issues are whether:

- the Organisation *exploits* IT for its own benefit and the IT *supports* the activities in the Organisation
- the IT *supports* management activities and the management in the Organisation *employs* and offer support to technical staff
- the technical staff members have the required skills and knowledge to *manage* IT
- the technical staff *respond to* user requirements and demands
- the influences impact on IT utilization, exploitation and maintenance

The discussion on what to qualify in the PS covered all the five entities of the management paradigm of looijen (2001). Rather than qualifying individual aspects separately, the procedure examined the issues in relation to their impact on the utilization, exploitation and management of IT in an organization. In effect, this approach yields more informative qualification than the sum of individual qualification taken separately. Next, we examine the manner in which the qualification of the same issues is carried out.

3.3 Qualifying the IS PS

Different design approaches for standardization are proposed in the qualification of the PS for different entities. Standardization provides a benchmark against which measurements or comparisons can be made. This is necessary for the fact that issues in different entities differ remarkably and can also differ within the same entities. These issues, therefore, require different frameworks for qualification.

Since standards of comparison of IT related issues have not been developed to a significant degree, for issues in the entity RS, a normative approach to qualification is proposed [1]**.** This implies that, in the absence of an existing standard by which to compare the present situation, we base the qualification on what, in our opinion, it should be which we take to be the norm for that particular issue. The norm becomes the standard measure, and according to the issues in the entity RS, the norm is assigned the values High (or 3), Medium (or 2), and Low (or 1), relative to the norm. In the absence of the issue in question, the value Absent/Missing (or 0) is assigned.

3,4 The Qualification of RS issues

ABSENT implies -No user requirements, no user awareness, no preconditions, no positive impact of situational factors on utilization, exploitation and maintenance of IT with no operational plans to deal with them. No financial resources, no personnel allocation of any kind and the overall contribution of the issue to organizational effectiveness and efficiency is *zero*.

LOW implies- low level of user awareness, unformulated preconditions, low positive impact of situational factors on utilization, exploitation and maintenance of IT with limited or no operational plans to deal with them. Inadequate financial resources, low-skilled personnel with irrelevant (non-IT) qualifications and in insufficient allocations are made. Here the overall contribution of the issue to organizational effectiveness and efficiency is *low.*

MEDIUM implies- limited definition of user requirements, moderate level of user awareness, limited formulated preconditions, and moderate positive impact of situational factors on IT utilization, exploitation, and maintenance with tactical or operational plans to deal with them. Barely adequate financial resources, moderately skilled personnel in IT limited numbers. Here the overall contribution of the issue to organizational effectiveness and efficiency is *medium.*

HIGH - This implies well defined user requirements, high level of user awareness, well formulated preconditions, and high positive impact of situational factors on utilization, exploitation and maintenance of IT with strategic, tactical and operational plans to deal with them. Adequate financial resources, highly skilled personnel in IT fields in sufficient numbers are allocated. Here the overall contribution of the issue to organizational effectiveness and efficiency is high.

In the case of IT, we seek to specify the level of hardware/software technology in use, the extent to which IT is utilized, exploited and maintained to support the Organisation in its activities. Therefore, we need to determine:

- The quality and standard of the available IT resources - hardware, software and network components
- The stage of implemented IT resources and how well defined the management tasks are
- The impact of complexity factors on utilization, exploitation and maintenance of IT

3.5 The qualification of IT issues

ABSENT implies:
- no hardware, software and network components - only manual system
- no utilization, exploitation or maintenance of any kind
- no presence of complexity factors, i.e. quantity, diversity, distribution, dynamics, cohesion, ownership, utilization, and functionality

Here the overall contribution of IT issue to organizational effectiveness and efficiency is *zero.*

LOW implies:
- outdated non-standard hardware, software and network components by more than 10 years functions are not well utilized, exploited and maintained
- poor IT utilization, exploitation or maintenance and management tasks are not defined
- low levels of complexity factors - quantity, distribution, diversity, dynamics, utilization, ownership, cohesion between IT components, and functionality - all with low negative impact on IT utilization, exploitation and maintenance.

The overall contribution of IT issue to organizational effectiveness and efficiency is *low.*

MEDIUM:
- lagging state-of-the-art the resources are partially utilized, exploited and maintained
- limited IT utilization and tasks are barely defined
- moderate levels of complexity factors

The overall contribution of IT issue to organizational effectiveness and efficiency is *medium.*

HIGH:
- up-to-date state-of-the-art standardized IT components by less than 5 years across organization whose functions are well utilized, exploited and maintained
- enhanced IT utilization and tasks are well defined
- high levels of complexity factors.

The overall contribution of IT issues to organizational effectiveness and efficiency is *high.*

3.6 The qualification of MCM

In the entity MCM, the approach used in the qualification of the PS is based on the concept of capability maturity model (CMM) levels. While the concept of CMM is defined in terms of software development, in this paper, the first three capability maturity model levels are defined in terms of IT management according to [4] as follows:

Level 1(Initial): Reactions to events happen in an *ad hoc* manner at this level, and there are no management processes. The management is completely in the hands of operational "authorities" who decide what has to be done. The workload is most of the time high, and the work is carried out in an uncoordinated manner. This doesn't necessarily mean that all efficiency is ignored. The point is that it cannot be measured, because each separate realization is strongly influenced by the individuals who carry it out.

Level 2 (Repeatable)*:* At this level the work is done in a processes-like manner, but there are no formal process descriptions. The management recognizes the importance of processes like incident management, problem management, change management, configuration management and software control and distribution, and carries them out in a pragmatic way. The decision-making with regards to this is mostly inspired by the fact that similar processes in other organizations or in other situations seemed useful and looked susceptible to repetition, hence the name *repeatable.*

Level 3 (Defined): At this level the management processes are documented and standardized, they are related to Service Level Agreements (SLA), which are established between the management and the users of the services. This implies that the services of the management are more predictable and actions are taken in a structured way on issues like performance improvement, problem resolution, transmission of data and data storage capacity. Here the service users consider IS management to be a service organization.

3.7 The qualification criteria

ABSENT- No processes, no authorities, no work, no IT management. No tasks. The contribution of process to organizational effectiveness and efficiency is *zero.*

LOW- No processes, ad hoc and chaotic situation, 'authorities' decide, uncoordinated work, efficiency not measured, personalized command/control and bureaucratic form of IT management. No tasks defined. The overall contribution of process to organizational effectiveness and efficiency is *low.*

MEDIUM - Processes are on schedule, cost-effective, planned, practiced, trained, enforced, measured, process-like, recognized, pragmatic processes which repeat earlier successes. Tasks here are Moderately-defined. Overall contribution of process to organizational effectiveness and efficiency is *medium.*

HIGH- Preparedness criteria, documented, standardized across organization, peer-reviewed, completion criteria related to SLAs are predictable, structured actions, and tasks well-defined. The overall contribution of the Process to organizational effectiveness and efficiency is *high.*

These attributes constitute the process parameters required to qualify individual processes. We then use the following four numerical values to qualify attributes within the maturity levels: 2 and 3 for the purposes of computing capability maturity levels for the MCM processes, based on effectiveness and efficiency criteria:

0 if the attribute is absent
1 if the attribute is rated LOW
2 if the attribute is rated MEDIUM
3 if the attribute is rated HIGH.

Effectiveness of a purposeful entity is conceptualized in terms of the degree to which it attains its end-objectives and that for most entities having multiple end-objectives; overall measures of effectiveness must invariably be multi-dimensional [9]. The criterion *effectiveness* here implies the extent to which the goals and objectives of an organization are being achieved. It addresses the outputs from the organization or its units.

To be effective, therefore, management of entities must be able to address the needs of their user systematically, and in a participative and responsive manner. In [7] *efficiency* is defined as putting emphasis on reducing the labour required to achieve a certain objective, with the result of reducing routine and tedious work. The benefits are greater speed in output, greater accuracy, better customer service, and greatly reduced costs

4. THE EVALUATION METRIC

The evaluation uses attributes weighting of process, which is necessary to differentiate process attributes that are considered more important from those that are less important according to the perceptions of the person(s) conducting the qualification. There is no fixed rule governing the weighting (Wi) hence it can be varied as desired. The method involves multiplying process attribute weights by the values entered (Xi) and summed up to Σ(Wi* Xi), for all the attributes at both Level 2 and Level 3. The process capability level, L, is computed from the weights according to the equation below in line with [11]**.**

$$\mathbf{L = \Sigma(Wi \,.\, Xi) / \Sigma(Wi)}$$

The values for each process are added up and the average value for each process is determined. The value indicates the capability level for the specific process. These values are rational and range from 1.0 to 3.0 inclusive. The standard deviations from the mean are also calculated. In principle, all capability maturity level values below 1.0 are regarded as being at Level 1. These are the values that are entered in the qualification sheet as LOW, if less than 2.0, as MEDIUM, if 2.0 or more but less than 3.0 and HIGH if 3.0.

The format for capability maturity levels, CPL, is presented in table1. In the table, the attributes of the management processes are listed in the first column, each within its level. The second column consists of weighted values of the attributes, which may be adjusted. In this example, the weights take integral values 1, 2 and 3. The management forms FM, AM, and TM are treated as management processes for purposes of evaluation. The other processes are Service Support processes and Service Delivery processes.

Entries are made in the table for each process. If a process is absent, a value 0 is entered in each of the attributes. Otherwise, values 1, 2, or 3 are entered according to whether they are judged to be LOW, MEDIUM or HIGH, respectively.

TABLE 1: Format For IT Management Processes

	WT	Func. Mgt			Appl Mgt			Tech Mgt			Service Support					Service Delivery mgt				
Level 3		SM	TM	OM	SM	TM	OM	SM	TM	OM	chg	conf	hlp	prb	scd	av	ap	cf	ct	Sl
Seen as service	3																			
Structured action	3																			
predictable service	3																			
related to SLA	3																			
completion criteria	2																			
peer reviewed	2																			
standardized	2																			
documented	1																			
preparedness criteria	1																			
Level 2																				
Repeats earlier successes	3																			
performed rationally	3																			
Process is recognised	3																			
Activities are performed processes-like	3																			
measured	2																			
enforced	2																			
Trained	2																			
Practiced	2																			
planned	2																			
cost-effective	1																			
on schedule	1																			
Mean Process Level																				
Mean overall level																				
Standard dev.																				

Table Key: *Function, Application and Technical: SM=Strategic mgt, TM=Tactical mgt, OM=Operational*
Service Support Process: Change, Conflict, Help desk, Problem, SW control & Distribution
Service Delivery Process: Availability, Capability, Contingency plan, Service Level Agreement

5. DISCUSSIONS

The calculations for CPL for each process are made and entered at the bottom of each column. An overall average value for the capability maturity model level for all the processes is also calculated to give an indication of the general level of IS management. The average CPL for all the processes may be obtained to show, in general terms, the extent to which the organization has implemented the processes. This will provide some information about what processes require improvement and the extent of the improvements required.

The main objective of this function is to specify the future; the Desired situation DS for actors, objects or processes, in effect, what needs to be done to get to the DS. To achieve this, it is necessary to examine each issue and identify the aspects that are missing. For example, if the PS for a given issue is LOW, then the DS for that issue should be at least MEDIUM, if PS qualification is MEDIUM, then the DS situation should be defined as HIGH. In the event that the PS is HIGH, then the DS is defined as HIGH. This procedure applies to the issues that have positive impact on IT.

In case of those issues that have negative impact on management of IT such as preconditions (centralization of activities, concentration of IT), situational factors, and external influences, the reverse is true, in which case the DS should aim at minimizing their impact. These concepts are illustrated in Tables 2 and 3 below.

Table 2: Definition of DS (positive impact)

Qualification of PS	Definition of DS
Low	Medium
Medium	High
High	High

Table 3: Definition of DS (negative impact)

Qualification of PS	Definition of DS
Low	Low
Medium	Low
High	Medium

Using this process for each phase, the transformation of the PS to DS becomes realistic and attainable. The following steps are recommended:

Step 1: Establish the objectives for improvement and set up a team to oversee it. The team should be composed of a senior management official, IT management staff, and a trained capability maturity assessor.

Step 2: Analyze and reach a consensus regarding the gaps between the PS and the DS, that is, between where the State is today and where it wishes to be in future. Use the next higher capability maturity level as a reference land mark to identify what needs to be improved.

Step 3: Outline the specific activities required to make the transition between PS and DS. In particular, clearly defined remedial tasks which must be assigned, these are:

- deliverables or expected results
- responsibility/ownership of improvement processes
- timeframes for the accomplishment of tasks
- resources needed to perform the tasks
- risks involved in carrying out the tasks
- measurements of the achievements

Step 4: Set criterion to evaluate success of attainment bearing in mind the goals/objectives in step 1. This step also serves as a means to determine how and why some objectives are not being achieved.

Step 5: Devise ways to maximizing the desirable activities (enablers) and minimizing inhibitors (disablers). This step is the most difficult to achieve and requires strict discipline, and commitment on the part of implementers.

6. SUMMARY AND CONCLUSION

The aim of this paper is partly to answer the following research question. – How to develop a model of IS support that can describe and qualify PS, define DS for an achievable transform from PS to DS. For this an assessment process based on the management paradigm was applied to the functional specifications – description and qualification of PS, definition of the DS, and the transformation from PS to DS. Within the management paradigm, the issues in the entities RS, IT, MCM, and relationships between entities, were identified. The criteria for depiction and qualification of the PS were given, and the manner of qualification of the issues was also presented. One of the aims of the metric is to make the model relevant to the environments in which it will eventually be used. The other aim is to pursue the objectives of this paper – provision of a metric for the evaluation of IS against organization existing situations.

Change is necessary if any improvement of the PS is to be realized, but with every change there is an adjunct of and requirement for new concepts, new skills, new patterns, new commitments, new strategies and new resources in terms of training, financial resources and time needed to control and manage the realization of new situation and to prevent it from getting out of hand. Skills and financial commitments are needed to stabilize and sustain the new situation once it has been realized so that it can last for as long as the users and the organization require. But the identification of changes that are needed to improve the situation is not sufficient.

To be able to transform from PS to DS, we must examine the benefits that justify the costs associated with changing the situation. In this respect, studies have shown that, economic gains alone may neither be necessary nor sufficient to judge institutional performance or to justify investment in changes aimed at improvement [3]**.** A process improvement procedure is required to transform the PS to the DS. It is essential that the process improvement model be practical, feasible, cost effective and within the financial and technical abilities of the organization to implement.

REFERENCES

[1] Checkland, P.B (1981): Systems Thinking Practice, John Wiley & Sons, Chichester, p.3.

[2] Emil L Morse(2002): Evaluation Methodologies for Information Management Systems D-Lib Magazine September 2002, Volume 8 Number 9, ISSN 1082-9873 Emile L. Morse

[3] Kaplan, R.S., and D.P. Norton,(1992): The Balance Scorecard-Measures that Drive Performanc

[4] Looijen Wanyembi G. (2000): A Model for Improving ICT Management, Proceedings of the 2000 IEEE International Conference on Management of Innovation and Technology, Singapore, 12-15

[5] Looijen M. (2001): Information Systems Management and Culture Experiences from a Chinese perspective Information Systems, Management, Control and Maintenance P.37, **http://alexandria.tue.nl/extra2/200410651.pdf**

[6] Mintzberg 1979: The Structuring of Organizations: A Synthesis of the Research, Englewood Cliffs, N.J.: Prentice-Hall

[7] Remenyi, D.S.J., A. Money, A. Twite, A, (1993): Guide to Measuring and Managing IT Benefits, 2/e, NCC Blackwell

[8] Saastamoinen Heikki (2005): Exception Based Approche for Information Systems Evaluation: The Methods and its Benefits to Information System Management" The Electronic Journal of Information systems Evaluation, Vol. 8, Iss 1 pp 51-60.

[9] Sudit Ephraim (1996): Effectiveness, Quality and Efficiency: A Management Oriented Approach, Kluwer Academic Publishers.

[10] Victor- Valeriu P, Lustin P. ,Sabastian Nicolaescu(2006): Security metrics for Enterprise Systems. Journal of Applied Quantitative Methods

[11] Wanyembi G. N. (2002): Improving ICT Management in Public Universities in Kenya – PhD Thesis

[12] Whitten J. L, Lonnie D.s Bentley and Kevin C. Dittman: Systems Analysis and Design Methods 5th Edition. McGraw-Hill, Irwin, pp.76.

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Costs of Information Overload to Organisations – An Information Technology Perspective

G.K.S. Akorfu
Computing Technology Department
Wisconsin International University College – Ghana,
P.O. Box Lg 751, Legon, Accra, Ghana
Tel: +233 0244520207
George.Akorfu@Wiuc-Ghana.Edu.Gh

ABSTRACT

This paper illustrates how information overload costs organisations. Information overload describes an excess of information that results in the loss of ability to make decisions, process information, and prioritize tasks. Organisations need to cost Information Overload because Information Overload is dulling employees' senses and limiting their ability to absorb more in-depth and complex thoughts and content in this technological world. This is causing organisations to lose huge sums of money due to the unproductive nature exhibited by employees because of information overload. The method involves distributing questionnaire to 200 employees of private tertiary institutions. Questions were asked on the time they spent a day on performing unproductive functions on information overload related issues due to information communication technology in their institutions during working hours. This unproductive time is then converted into cash based on the wages/salaries the employees receive in the institution. The unproductive costs are the cost of information overload to the institution. Given that any of the 38 items is used in any institution by junior staff, the institution will lose on average GH¢4,428.49 annually, while by senior staff, the institution will lose GH¢414.51 annually. Collaborative technologies are becoming more and more integrated into how staff works. As these become more pervasive within institutions, staff will have more interruptions, instant messages, e-mails, social networking tools to interfere with their work, thus affecting institutions' profitability. But still, staff can be their own worst enemy if decisive precautionary measures are not taken in handling information overload related items.

Keywords: Information overload, Costs, Wages/Salaries, Information technology, Social networking.

African Journal of Computing & ICT Reference Format:
G.K.S. Akorfu (2013). Costs of Information Overload to Organisations – An Information Technology Perspective
Afr J. of Comp & ICTs. Vol 6, No. 2. pp 11- 30.

1. INTRODUCTION

According to the latest research by Spira [28], Information Overload costs the U.S. economy a minimum of $900 billion per year in lowered employee productivity and reduced innovation. Despite its heft, this is a fairly conservative number and reflects the loss of 25% of the employee's day to the problem. The total could be as high as $1 trillion. [28]. Research shows that only 5% of the employees' day is available for thought and reflection. The largest single block of time in the average day (25%) is spent dealing with Information Overload-related issues, such as interruptions, excessive e-mails, and failed searches. [28].

Indeed, in order to remain competitive in the 21st century, companies will need to begin an information overload bailout, i.e. taking active countermeasures, in order to remain competitive. Nothing is more disruptive to the way employees work than information overload and the trend needs to be reversed as quickly as possible.

There is the need to cost Information Overload because the problem is getting worse, not better. For institutions to even begin to regain their lost productivity, scattered focus, and decimated work/life balance, the first thing that must happen is for everyone in the institution to acknowledge the problem, and then take action.

Individual workers, teams, and entire organizations suffer diminished productivity and the loss of the ability to make sound decisions, process information, and prioritize tasks because of information overload [28]. The cost of Information overload is an opportunity for workers and organizations to take stock of the cost that this serious problem is having on their productivity and work life balance, not to mention on their organizations' bottom line - profit.

One issue about Information Overload is how easy it is to take steps to lessen its impact by simply raising awareness about the cost to an organization of the problem. For instance, reading and processing just 100 e-mail messages can occupy over half of a worker's day and for every 100 people who are unnecessarily copied on an e-mail, eight hours are lost. [28].

One would think that the world has reached a point where systems and computers should work flawlessly but that is less and less the case every day. On the one hand, the Information Revolution of the late 20th century has resulted in an anywhere, anytime information society that has become accustomed to boundless gobs of information on demand. From a technical standpoint, the advent of true ubiquitous computing has markedly changed employees' attitude towards and interactions with information. Employees' constant exposure to information leads them to have the expectation that it will be shared across systems, accurately and quickly. Unfortunately, information does not always get to where it needs to be.

Today, despite tremendous advances in technology over the course of 50 years, information often fails organisations and society. Calls to customer service representatives at call centres asking the same or similar questions yield widely disparate answers, despite the fact that the agent is being guided by the system. [28].

From endless e-mail, social media, and texting, to poor search tools and a dramatic increase in information generation, Information Overload is stretching the bandwidth of businesses and employees at unprecedented levels. Revealing how the very tools deployed to make employees more efficient have in turn bogged productivity down. Information Overload explores the many ways today's tidal wave of information has bombarded and dulled employees' senses as well as hampered their ability to innovate and produce.

2. PROBLEM STATEMENT

Spira, [28] examines the staggering statistics of time and money lost due to Information Overload, including the facts that:

- Information Overload cost the U.S. economy almost $1 trillion in 2010.
- A minimum of 28 billion hours is lost each year to Information Overload in the United States.
- Reading and processing just 100 e-mail messages can occupy over half of a knowledge worker's day.
- It takes five minutes to get back on track after a 30-second interruption.
- For every 100 people who are unnecessarily copied on an e-mail, eight hours are lost.
- 58 percent of government workers spend half the workday filing, deleting, or sorting information, at a cost of almost $31 billion dollars.
- 66 percent of knowledge workers feel they don't have enough time to get all of their work done.
- One major Fortune 500 company estimates that Information Overload impacts its bottom line to the tune of $1 billion per year.
- Information Overload has caused people to lose their ability to manage thoughts and ideas, contemplate, and even reason and think.
- The reality that many e-mail exchanges which go on for days and weeks at a time could be resolved with a five-minute phone call.
- Employees use 15% of working time searching for relevant information (and an estimated 50% of searches fail).
- Employees use 28% of working time on unnecessary interruptions followed by "recovery time" to get back on track after thinking and reflecting.

The question then is: "What is Information Overload costing your organization?" Information Overload is a problem because it creates a bottleneck that stops employees from absorbing all of the information being thrust at them. Clearly, some information is left behind. Some of it might even be useful or important. What's worse is that employees don't generally know what they don't need, so they may make decisions based on the information they have available to them, even though they are overlooking some information (that may or may not be critical) simply because they are unaware of its existence.

The problem of information overload is not just time spent on low-value activities. Interrupting employees is costly. Research has shown that every time an employee must refocus his or her attention, it takes time and energy to mentally switch from Topic A to Topic B and back again [3]. Beyond the time spent processing the high volume of messages lies another problem. Employees spend a huge amount of their time looking for information they can't find – or recreating information that already exists [3].Recent research on knowledge work shows that knowledge workers spend more time recreating existing information than they do turning out information that does not already exist. Some studies suggest that 90% of the time that knowledge workers spend in creating new reports or other products is spent in recreating information that already exists [3]. Tangri says stress is responsible for 19 percent of absenteeism, 40 percent of turnover, 55 percent of employee-assistance program costs – and much more [30]. Stressed-out employees are not likely to contribute the creativity, innovation, imagination and energy their organizations need to compete [3].

Costs of information overload such as lost productivity, diminished quality of thought, increased level of stress and so on are direct costs. However, there is another category of cost which is often overlooked because it is not that obvious. These are the ***indirect costs*** of information overload, i.e. the missed opportunities due to the "solutions" that employee is forced to apply when dealing with Information Overload**.**

Employees need mental breaks and down time from their work during the course of the day, but they need to be at a time and place of the employees' choosing, not when an interruption breaks their concentration. Reducing Information Overload is about increasing productivity and the Return on Investment (ROI) for the organization's employees. Understanding the potential impact, i.e. the amount of money that it takes to pay employees during time that they may not be at their peak efficiency, allows organizations to conceptualize the problem and begin to take action.

2.1 Definition of Information Overload.

At the most basic state, information overload refers to the simple notion of receiving or having to go through too much information [7]; [26]. In order to go to a deeper definition of information overload, researchers have taken two major paths: **objective sense** and **subjective sense**.

Objectively, information overload can be defined based on the information processing view that information overload occurs when the information processing requirements (IPR) exceed the information processing capacities (IPC) of an individual (IPR>IPC) [31].

Subjectively, information overload has been investigated by researchers who believe that information overload cannot be investigated under experimental conditions as time constraints and forced absorption set in; experimental conditions do not apply in most real life situations [20]. These researchers define information overload as being burdened by a large supply of information that cannot be assimilated, leading to breakdown: feelings of stress, confusion, pressure and anxiety when in an information overload state [6]; [9]; [10]; [14]; [20]; [21].

The **objective** definition is adopted in this paper. The terms 'requirements' and 'capacities' in the above definition can be measured in terms of the available time. The requirements refer to a given amount of information that has to be processed within a certain time period (Information needed to complete a task). The capacities refer to a given amount of information that has to be used within a certain time period (The quantity of information one can integrate into the decision making process).

2.2 Recent Research On Information Overload

2.2.1 Causes Of Information Overload.

As far as the corporate context is concerned, the main reasons for information overload can be related to five issues. These inductively generated categories of major overload causes are the **person** receiving, processing or communicating information; the **information** itself (its quantity, frequency or intensity, and quality or general characteristics); the **tasks or processes** which need to be completed by a person, team or organization; the **organizational design** (i.e., the formal and informal work structures), and the **information technology** that is used (and how it is used) in a company. The acronym **PITODIT** can be used to refer to these five constructs. Usually, information overload emerges not because of one of these factors, but because of a mix of all five causes.

Literature has indicated that to mitigate or to cope with information overload, all the other factors which interrelate with each other to cause the overload should also be considered in mitigating or coping with it [7]. From the literature analysis on information overload, future research directions have emerged. Literature has shown that the analysis of information overload should no longer be studied using models of linear cause and effect, but should rather be represented with cyclical structures and a focus on interdependencies. This is important since the complexity of the phenomenon is mainly given by the interconnectedness of its various variables. [7].

2.2.2 Symptoms Of Information

In **personal** situations, the symptoms exhibited are demotivation [1], satisfaction negatively affected [16]; [17], stress, confusion and cognitive strain [17]; [19]; [22]. The executive lacks learning anything since too little time is at his disposition [27]. There is greater tolerance of error in jobs performed [27] and lack of perspective [22]. The executive exhibits sense of loss of control which leads to a breakdown in communication [23]. With regards to **information**, sub-optimal decisions are made when information overload manifests itself. When information overload manifests itself, decision accuracy, quality [19]; [16]; [15], decision effectiveness are lowered [24]. Inefficient work [2], potential paralysis and delay of decisions [2]; [22] are some of the inevitable symptoms.

The complexity of **the task** is often discussed as a contributing factor when studying information overload. The work of Groff, Baron, and Moore [12] has taken the research a step further by analyzing the effects of distractions on the level of information anxiety. Their research shows that, in relation to task complexity, the level of information overload in complex tasks increases as the quantity and force of distractions increases.

In the organizational context, the presence of information overload is felt when there is overlapping and inconsistent information categories [8]. Executives ignore information and become highly selective thereby omitting significant information for organizational management [2]; [6]; [13]; [14]; [27]. Executives' loss control over information in the organisation [2]; [33]. Mangers lack critical evaluation of issues and become too credulous and analysis are very superficial [26]; [22]; [25].

In relation to **information technology**, limited information search and retrieval strategies are present. Search strategies through information sets become less systematic (this is less true for more experienced searchers) [29]. There are limited search directions for information in the information systems provided by the enterprise and move from compensatory search patterns to non-compensatory search patterns exist [5].

2.3 Costs of Social Networking to Organisations.

According to Wilson [32], there are five principal worries that management has with regard to social networking: perceived loss in staff productivity, loss of revenue, data leakage from staff gossiping freely in an open environment, damage to a company's reputation, scams practiced by "cyber crooks," and the open access to company information because of outdated passwords. There are many uses for the big four social networking sites. It is a concern to management and corporate executive officers that employees spend time on these websites while at work. One possible use of the networks that is a source of concern to management is the possible damage to the company's reputation that can be brought about by posts online. If an employee were to be angry, or have had a bad day, they might be inclined to take their anger out online.

This behaviour could damage a company's reputation. "Employers have the right to hold employees responsible for such conduct if the postings are used to 'attack the company' or 'harass co-workers'" [4]. Another concern, from a management standpoint, is the belief that the use of social networking websites is detrimental to the productivity of the employees who spend increasing amounts of time on these websites. Employees are given access to company equipment mainly computers and internet, in order to complete their jobs effectively and efficiently. According to Lange [18] "employers worry that staff are wasting time on websites during the day, weakening productivity and increasing security risks to the company by sharing data externally."

A company can face lawsuits, bad publicity, and decreased employee morale because of employees use of social networks. According to Greenwald [11], 55% of employees visit a social networking site at least once a week. Possible areas of company liability include sexual harassment, bullying, and threats of workplace violence, all potentially occurring during these visits. Sexual harassment occurs when one employee or supervisor makes continued, unwelcome sexual advances, requests for sexual favours, and other verbal or physical conduct of a sexual nature, to another employee or supervisor, against his or her wishes. These behaviours can, and do, happen online. The increased risk of liability can decrease productivity and cost the company a large amount of money, from dollars spent in defending against lawsuits and possible revenue lost due to damage to the reputation of the company.

3. METHODOLOGY

Information Overload related issues due to information communication technology referred to as social media such as: Twitter, Facebook, Google+, Foursquare, Pinterest etc. are used by employees in their work places. There are also other items such as: unnecessary interruptions, unproductive meetings, sending e-mail, reading/replying text messages, deleting spams, dealing with instant messages, etc. which hampers employees' productivity. In all, thirty eight (38) of these unproductive items were listed in a questionnaire for respondents (125 senior staff and 75 junior staff) in twenty (20) private Universities in Ghana to tick the appropriate time used on the items. The range provided in the questionnaire is [1] 1 – 10 minutes, [2] 11 – 20 minutes,....., [7] 1 – 4 hours, [8] 5 – 8 hours and [9] No idea. The median of these ranges are taken as the time used by an employee on the 38 items listed. [9] No idea is given no time i.e. "0" minutes. Salaries per month of all the senior and junior staff were collected from the accountants of the various institutions. The average salary per month was calculated for both categories of staff. The average was then used to calculate average salary per minute of each staff based on 20 working days per month. The average cost per minute, per hour, per day, per month and per year to the organisation when using these information overload related items were found. (See Appendixes 1 & 2, Tables 1 & 2).

4. RESULTS

This section presents the results of the research conducted on the senior and junior staff of the twenty (20) private Universities in Ghana. The research sets out to calculate the average cost to the organisations per staff per minute when staff (junior and senior) uses one of the information overload items between the following ranges of time: 1 – 10 minutes, 11 – 20 minutes, 21 – 30 minutes, 31 – 40 minutes, 41 – 50 minutes, 51 – 60 minutes, 1- 4 hours, 5 – 8 hours.

Table 1: Average cost to organization per junior staff using information overload related issues during working hours.

Average Costs	1 – 10mins	11 – 20min
Average cost to org. per staff per min per item (GH¢)	0.005	0.007
Average cost to org. per staff per hour per item (GH¢)	0.28	0.45
Average cost to org. per staff per day per item (GH¢)	2.22	3.57
Average cost to org. per staff per month per item (GH¢)	44.37	71.45
Average cost to org. per staff per year per item (GH¢)	532.39	857.35

Average Costs	21 – 30mins	31 – 40min
Average cost to org. per staff per min per item (GH¢)	0.006	0.010
Average cost to org. per staff per hour per item (GH¢)	0.37	0.58
Average cost to org. per staff per day per item (GH¢)	2.94	4.60
Average cost to org. per staff per month per item (GH¢)	58.77	92.04
Average cost to org. per staff per year per item (GH¢)	705.24	1,104.53

Average Costs	41 – 50mins	51 – 60min
Average cost to org. per staff per min per item (GH¢)	0.00	0.002
Average cost to org. per staff per hour per item (GH¢)	0.00	0.100
Average cost to org. per staff per day per item (GH¢)	0.00	0.80
Average cost to org. per staff per month per item (GH¢)	0.00	15.99
Average cost to org. per staff per year per item (GH¢)	0.00	191.87

Average Costs	61 – 240mins	241 – 480min
Average cost to org. per staff per min per item (GH¢)	0.009	0.00
Average cost to org. per staff per hour per item (GH¢)	0.54	0.00
Average cost to org. per staff per day per item (GH¢)	4.32	0.00
Average cost to org. per staff per month per item (GH¢)	86.43	0.00
Average cost to org. per staff per year per item (GH¢)	1,037.12	0.00

Source: Author

Table 2: Average cost to organization per senior staff using information overload related issues during working hours.

Average Costs	1 – 10mins	11 – 20min
Average cost to org. per staff per min per item (GH¢)	0.00210	0.00084
Average cost to org. per staff per hour per item (GH¢)	0.126	0.050
Average cost to org. per staff per day per item (GH¢)	1.007	0.403
Average cost to org. per staff per month per item (GH¢)	20.15	8.06
Average cost to org. per staff per year per item (GH¢)	241.79	96.72

Average Costs	21 – 30mins	31 – 40min
Average cost to org. per staff per min per item (GH¢)	0.00033	0.00012
Average cost to org. per staff per hour per item (GH¢)	0.01979	0.00720
Average cost to org. per staff per day per item (GH¢)	0.1583	0.0576
Average cost to org. per staff per month per item (GH¢)	3.166	1.151
Average cost to org. per staff per year per item (GH¢)	37.996	13.817

Average Costs	41 – 50mins	51 – 60min
Average cost to org. per staff per min per item (GH¢)	0.00009	0.00009
Average cost to org. per staff per hour per item (GH¢)	0.00540	0.00540
Average cost to org. per staff per day per item (GH¢)	0.04318	0.04318
Average cost to org. per staff per month per item (GH¢)	0.864	0.864
Average cost to org. per staff per year per item (GH¢)	10.36	10.36

Average Costs	61 – 240mins	241 – 480min
Average cost to org. per staff per min per item (GH¢)	0.00003	0.00
Average cost to org. per staff per hour per item (GH¢)	0.00180	0.00
Average cost to org. per staff per day per item (GH¢)	0.01439	0.00
Average cost to org. per staff per month per item (GH¢)	0.28785	0.00
Average cost to org. per staff per year per item (GH¢)	3.45	0.00

Source: Author

Appendices 1 and 2 show the full calculations of the average cost of information overload related issues.

5. DISCUSSION

Many senior staff spends between 1 to 10 minutes each day on reading text messages, replying to text messages and dealing with instant messages. They also spend appreciable time between 1 to 10 minutes replying e-mails, sending text messages, deleting unnecessary text messages, answering calls, making calls, deleting spams from e-mail boxes and unsubscribing from non-work related newsletters. Many again spend the organisations' time between 11 to 20 minutes daily using facebook and putting certain types of e-mail in certain types of folders. Much time between 31 minutes

8 hours is not spent on the 38 information overload related items. Majority of the senior staff are undecided on the time they spend daily on the usage of the items. (Appendix 2, Tables 4A and 4B).

The junior staff spend between 1 to 10 minutes daily on sending text messages, replying text messages, deleting unnecessary text messages, using twitter. They also spend between 1 to 10 minutes each day thinking on their work and deleting information. Much time between 31 minutes 8 hours is not spent on the 38 information overload related items. Majority of the junior staff are undecided on the time they spend daily on the usage of the items. (Appendix 2, Tables 4A and 4B). When both junior and senior staff use between 1 to 10 minutes daily, it costs the institutions GH¢532.39 and GH¢241.79 respectively per staff per item annually. When they use between 11 to 20 minutes daily, it costs them GH¢857.35 and GH¢96.72 respectively per staff per item annually.

For using between 1 to 10 minutes daily, 160.42 man-hours are lost daily by senior staff and 64 man-hours by junior staff. Using between 11 to 20 minutes daily by staff make their institutions loose GH¢103.33 and GH¢180.83 by junior and senior staff respectively.Reading and processing just 100 e-mail messages by staff between 1 to 10 minutes daily can cost the institution GH¢53,239.00 in respect of junior workers and GH¢24,179 in respect of senior workers annually.

Reading and processing just 100 e-mail messages between 11 to 20 minutes daily can cost the institution GH¢85,735 in respect of junior workers and GH¢9,671.79 in respect of senior workers annually.
A junior staff taking between 1 to 10 minutes filing, deleting, and sorting information can cost the institution GH¢1,596.00 annually, whilst a senior staff performing the same functions costs the institution GH¢725.37 annually. Again using between 11 to 20 minutes filing, deleting, and sorting by staff could cost the institution GH¢2,572.05 and GH¢290.16 annually by junior and senior staff respectively.

Twenty five (25) junior staff using between 1 to 10 minutes searching for relevant information which they cannot find costs the institution GH¢13,309.75 whilst the same number of senior staff could cost the institution GH¢6,044.75 annually. When they use between 11 to 20 minutes, it will cost the institution GH¢21,433.75 by junior staff and GH¢2,418.00 by senior staff annually.

When ten (10) junior staff takes between 1 to 10 minutes to recover when unnecessarily interrupted, the institution loses GH¢5,323.90 annually and ten (10) senior workers will cause the institution to lose GH¢2,417.95 annually. The same number of junior and senior staff taking 11 to 20 minutes to recover after interruptions will cause the institution to lose GH¢8,573.50 by junior workers and GH¢967.18 by senior staff annually.

Senior staff spend more time using social media such as facebook than junior staff in their various institutions. Pinterest is not common among senior staff whilst foursquare and pinterest are not common among junior staff. But if social media such as facebook, twitter, google+, foursquare and printest are used by a junior staff and a senior staff between 1 to 10 minutes, the institution will lose GH¢2,661.95 on the part of junior staff and GH¢1,208.95 on the part of senior staff annually. Between 11 to 20 minutes, the institution will lose GH¢4,286.75 on the part of junior staff and GH¢480.00 on the part of senior staff annually.

6. SUMMARY AND CONCLUSIONS

6.1 Summary

It's important to note that the issues covered in this paper are critical to institutions but staff in these institutions reading this article will probably not make it to the end, at least not without multiple interruptions and distractions. While reading it, staff will receive at least one or two phone calls, at least 10 to 20 e-mails between 1 to 10 minutes or between 11 to 20 minutes or more (whether they read them or not is another issue), a few instant messages, and a colleague may pop in, their mobile phone may ring or they may get a few text messages. They might get distracted and visit a Web site. When staff in these institutions work, they spend several minutes, even hours per day sending and receiving e-mail. Then they try to go onto other tasks, with varying degrees of success. The chances are that they will use a computer for a good portion of what they do, regardless of the job title. In all these, staff is unproductive when caught up in this information overload related issues which in turn are a loss to these institutions.

The interruptions, distractions, receiving e-mails, reading e-mails, receiving calls, making calls, seeing to a few instant messages, sending a few text messages, receiving text messages, visiting a Web site are costed for junior and senior staff of some institutions using the time they spent on these information related issues and their salaries. The result shows the average cost to the institution per staff per information overload related item used. Many staff using these items in big institutions costs them huge Ghana cedis thereby affecting their bottom line – profits per annum.

7. CONCLUSIONS

Today, for institutions with thousands of staff, information overload has become a major problem, costing them perhaps billions of Ghana cedis in lower productivity and hampered innovation. It may also be harmful to staff in a variety of ways, including lowering comprehension levels and skewing the work-life balance.

The problem of information overload is multi-faceted. It covers e-mail overload, interruptions, new technologies that compete for our attention, social networking and improved and ubiquitous connectivity, just to name a few issues.

Defining the problem from the foregoing isn't that simple either. It is not just a case of too much e-mail, too many interruptions, too many projects, or too many instant message sessions. It is how these items all mesh together - sometimes like an orchestra without a conductor.

In other words, whether sitting at a desk in the office, in a conference room, in one's home office, or at a client's, the likelihood of being able to complete a task without interruption is nearly nil.

8. RECOMMENDATIONS

In looking at an interruption, it is important to determine whether something is important, urgent, or both. Many staff simply do not differentiate, or see everything as both important and urgent. Something that is important may not require an immediate interruption, whereas something that is urgent would certainly be more likely to merit, and surely call for, an interruption. Importance can also vary, based on the particular needs of the group or organization.

Even more important to note is the fact that each staff has different priorities, different tasks, and a different idea of what is urgent or important. What is urgent and/or important to one staff at a given moment might not be as urgent and/or important to another. Staff may be constantly busy, but that doesn't make them either productive or efficient. It also doesn't mean that what they are doing is aligned with the strategic goals of the institution. Sometimes a staff might feel like a ping pong ball, bouncing around from task to task. The unending barrage of work - be it e-mail, meetings, or teleconferences, deleting unnecessary e-mails, sorting information, etc., never stops.

It appears that the more information staff has, the more they seem to generate and the less control they have over how they obtain it. Indeed it appears as if the role of computer- based communications is being obscured, moving from an effective communications medium to a problem that needs to be managed. In addition, rather than evolving in lockstep, the technologies staff use to communicate are coming at them at a faster pace than the corporate culture that must accept these tools into its midst is changing. This puts staff practically on the defensive from technology; just because an e-mail arrives shouldn't mean that staff should drop everything to respond to it. Just because staff can send an e-mail to all 10,000 employees in a company doesn't mean they should.

E-mail and social networking tools are far from being the only cause. Staff need to learn to recognize the trap of the tyranny of the convenient. And they need to understand what's important versus what's urgent and place their work in perspective with what others might be doing. What staff can expect in the near future is more information overload, not less. Information overload is, in part, a by-product of the lack of maturity of the information age. There is so much to learn as staff struggle to make technology work for the institutions they work. And information overload is not a new thing.

Staff has managed to increase the information production yet they have few tactics that seem to reduce the problem.

E-mail and social networking tools represent the lifeblood of how institutions conduct their business. They have to be managed. As institutions add different computer-mediated modes of communication, e-mail and social networking tools may become an archaic medium; staff may cling to them and be considered quaint, but for now they are the corporate lingua franca.

Staff should not get impatient when there's no immediate response to their message. They should keep their presence awareness state up-to-date and visible to others so they know whether they are busy or away. Staff should recognise that the intended recipient of their communications is not a mind reader and supply details in their messages accordingly.

Staff can help reduce information overload by choosing a communications medium wisely. Under which circumstances is instant messaging "better" than old-fashioned telephony? And under which circumstances might instant messaging be more appropriate than e-mail? Choosing the right modality for staff's needs will also help ensure that fewer interruptions occur.

Instant messaging is better than telephone when there are many people participating and all need to talk and be active, at least one participant is in an environment where people could listen in, and privacy or confidentiality is an issue and there are a number of many-to-many conversations taking place.

Telephone is better than instant messaging when there are many people participating passively and one person is speaking, a more personal touch is required and the nuances of voice matter. E-mail is better than instant messaging when the text needs to be memorialized and it contains an announcement to be sent to many people. Instant messaging is better than e-mail when an issue demands an immediate response, i.e. it is both urgent and important and the issue is relatively trivial, such as lunch plans.

Given that 45% of staff in any institution receives 50 or more e-mail messages per day, use social networking tools, deals with instant messages, etc. institutions still have a lot of work to do in managing the staff's attention for greater productivity.

Computer social networking has become part of the social activities which engage employees so much that they become addicted in such a way that it affects their progress, productivity and time which costs their organizations. Research need to be done to find the costs of information overload which are caused by each of the new social networks which are springing up on daily basis.

REFERENCES

[1] Baldacchino, C., C. Armistead and D. Parker, 2002. Information overload: It's time to face the problem. Management Services, 46, 18-19.

[2] Bawden, D, 2001. Information overload: Library & Information Briefings, 92, Online. Retrieved September 14, 2004 from http://litc.sbu.ac.uk/publications/lframe.html.

[3] Boyd, B., 2005. The (Staggering) Cost of Information Overload. Communication World Bulletin. September 2005.

[4] Breslin, J., A. Passant and S. Decker, 2009. The Social Semantic Web. New York: Springer Publishing.

[5] Cook, G. J., 1993. An empirical investigation of information search strategies with implications for decision support system design. Decision Sciences, 24, 683-699.

[6] Edmunds, A. and A. Morris, 2000. The problem of information overload in business organizations: A review on the literature. International Journal of Information Management, 20, 17-28.

[7] Eppler, M. J., and J. Mengis, 2004. The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. The Information Society, 20, 325-344.

[8] Eppler, M., 1998. Informative Action: An Analysis of Management and the Information Overload, (PhD thesis, HEC Management Studies). Geneva: University of Geneva.

[9] Farhoomand, A. F., and D.H. Drury, 2002. Managerial information overload. Communications of the ACM, 45, 27-131.

[10] Feather, J., 1998. The information society: A study of continuity and change. London: Library Association.

[11] Greenwald, J., 2009. Employees' social networking raises employers' liability risk, Business Insurance, Retrieved from http://www.businessinsurance.com/article/20090719/ISSUE01/307199966, July 2012.

[12] Groff, B. D., R.S. Baron and D.L. Moore, 1983. Distraction, attentional conflict, and dislike behavior. Journal of Experimental Social Psychology, 19, 359-38.

[13] Herbig, P. A. and H. Kramer, 1994. The effect of information overload on the innovation choice process. Journal of Consumer Marketing, 11, 45-54.

[14] Hiltz, S.R. and M. Turrof, 1985. Structuring computer-mediated communication systems to avoid information overload. Communications of the ACM 28(7), 680-689.

[15] Hwang, M.I. and J.W. Lin, 1999. Information Dimension, Information Overload and Decision Quality. Journal of Information Science, 25 (3), 213-218.

[16] Jacoby, J. 1984. Perspectives on information overload. Journal of consumer Research, 10, 432-436.

[17] Jones, Q., 1997. Virtual-communities, virtual-settlements & cyber-archaeology: A theoretical outline. Journal of Computer Mediated Communication, 3, 3. Retrieved August 2010 from http://www.ascusc.org/jcmc/vol3/issue3/jones.html

[18] Lange, P. G., 2008. Publicly Private and Privately Public: Social Networking on YouTube. Journal of Computer - Mediated Communication 13(1): 361.

[19] Malhotra, N. K., 1982. Information load and consumer decision making. The Journal of Consumer Research, 8, 419-431.

[20] O'Reilly, C. A., 1980. Individuals and information overload in organizations: Is more necessarily better? Academy of Management Journal, 23, 684-696.

[21] Rogers, E.M. and R. Agarwala-Rogers, 1975. Organizational communication. In: Hanneman, G.L., & McEwen, W.J. (eds.) Communication Behavior. Addision Wesley, Reading, Massachusetts, 218-236.

[22] Schick, A. G., L.A. Gorden and S. Haka, 1990. Information overload: A temporal approach. Accounting Organizations and Society, 15, 199-220.

[23] Schneider, S. C., 1987. Information overload: Causes and consequences. Human Systems Management, 7, 143-153.

[24] Schroder, H. M., M.J. Driver and S. Streufert, 1967. Human information processing – Individuals and groups functioning in complex social situations. New York: Holt, Rinehart, & Winston.

[25] Schultze, U. and B. Vandenbosch, 1998. Information overload in a groupware environment: now you see it, now you don't. Journal of organizational computing and electronic commerce 8(2), 127-148.

[26] Shenk, D. 1997. Data Smog: Surviving the Information Glut. San Francisco: Harper.

[27] Sparrow, P. R. 1999. Strategy and cognition: Understanding the role of management knowledge structures, organizational memory and information overload. Creativity and Innovation Management, 8, 140-149.

[28] Spira, J. 2011. Overload! How Too Much Information Is Bad For Your Organization. New York: John Wiley & Sons.

[29] Swain, M. R. And S.F. Haka, 2000. Effects of information load on capital budgeting decisions. Behavioural Research in Accounting, 12, 171-199.

[30] Tangri, R. 2003. Stress Costs, Stress Cures. Victoria: Trafford Publishing

[31] Tushman, M.L. and D.A. Nadler, 1978. Information processing as an integrating concept in organizational design. Academy of Management Review, Vol. 3 No.3, pp.613-24.

[32] Wilson, J. 2009. Social networking: the business case. ***Engineering & Technology***, 4, 54-56. Retrieved April 20, 2010, from Business Source Premier Database.

[33] Wurman, R. S. 1990. Information anxiety. What to do when information doesn't tell you what you need to know. New York: Bantam Books.

APPENDIX 1. Calculation of cost of Information overload.

Table 3A. Cost of Information Overload for Junior Staff

JUNIOR STAFF							
		MINUTES					
	TIME SPENT PER DAY ON:	**5.5**	**15.5**	**25.5**	**35.5**	**45.5**	**55.5**
		NO. OF JUNIOR STAFF					
1	Dealing with unnecessary interruptions.	25					25
2	Recovering in time to get back on track after interruptions.	50					
3	Doing unproductive work.			50			
4	Participating in unproductive meetings.		25				
5	Searching for relevant information you cannot find.			25			
6	Recreating information that already exists.				25		
7	Duplicating information.		25	25			
8	Thinking on your work.		50	25			
9	Reflecting on your work.	25	25		25		
10	Sending e-mails.	25	25				
11	Reading e-mails.	25			25		
12	Replying e-mails.	25	25				
13	Deleting unnecessary e-mail.	50					
14	Sending text messages.	75					
15	Reading text messages.	25	50				
16	Replying text messages.	75					
17	Deleting unnecessary text messages.	50	25				
18	Using Facebook.	25					
19	Using Twitter.	50					
20	Using Google+.	25					
21	Using Foursquare.						
22	Using Pinterest						
23	Answering calls.	25	25				
24	Making calls.	25	25				
25	Filing information.				25		

26	Deleting information.			50				
27	Sorting information.					25		
28	Changing from one task to another.					25		
29	Reading unnecessary information offline.			25				
30	Reading unnecessary information online.				25			
31	Deleting spams from e-mail box.		25	25				
32	Reading non-work related newsletters.					25		
33	Unsubscribing from non-work related Newsletters.					25		
34	Deleting non-work related Newsletters.				25			
35	Putting certain types of email in certain types of folders.				25			
36	Filing messages you might need someday.		25			25		
37	To mentally switch from one topic to another topic and back again.		25					
38	Dealing with instant messages.		25					
	TOTAL NO. OF STAFF		700	400	200	225	0	25
	TOTAL TIME (MINS) SPENT/DAY		3,850	6,200.00	5,100.00	7,987.50	0.00	1,387.50
	AVERAGE SALARY PER MONTH (75 JUNIOR WORKERS) (GH¢)	919.58						
	AVERAGE SALARY PER DAY (20 WORKING DAYS/MONTH) (GH¢)	45.98						
	AVERAGE SALARY PER HR (24 HOURS/DAY) (GH¢)	1.92						
	AVERAGE SALARY PER MIN (60 MINS/HOUR) (GH¢)	0.03						
	AVERAGE COST TO ORG. PER MIN (60		122.93	197.96	162.84	255.04	0.00	44.30

	MINS/HOUR) (GH¢)						
	AVERAGE COST TO ORG. PER HR (60 MINS/HOUR) (GH¢)	7,375.77	11,877.86	9,770.50	15,302.32	0.00	2,658.15
	AVERAGE COST TO ORG. PER DAY (8 HOURS PER DAY) (GH¢)	59,006.14	95,022.88	78,163.98	122,418.59	0.00	21,265.20
	AVERAGE COST TO ORG. PER MONTH (20 DAYS/MONTH) (GH¢)	1,180,122.89	1,900,457.65	1,563,279.68	2,448,371.85	0.00	425,304.03
	AVERAGE COST TO ORG. PER YEAR (12 MONTHS/YEAR) (GH¢)	14,161,474.71	22,805,491.74	18,759,356.11	29,380,462.15	0.00	5,103,648.35
	AVERAGE COST TO ORG.PER STAFF PER MIN (GH¢)	0.18	0.28	0.23	0.36	0.00	0.06
	AVERAGE COST TO ORG. PER STAFF PER HOUR (GH¢)	10.54	16.97	13.96	21.86	0.00	3.80
	AVERAGE COST TO ORG. PER STAFF PER DAY (GH¢)	84.29	135.75	111.66	174.88	0.00	30.38
	AVERAGE COST TO ORG. PER STAFF PER MONTH (GH¢)	1,685.89	2,714.94	2,233.26	3,497.67	0.00	607.58
	AVERAGE COST TO ORG. PER STAFF PER YEAR (GH¢)	20,230.68	32,579.27	26,799.08	41,972.09	0.00	7,290.93
	AVERAGE COST TO ORG.PER STAFF PER MIN PER ITEM (GH¢)	0.005	0.007	0.006	0.010	0.00	0.002
	AVERAGE COST TO ORG. PER STAFF PER HR PER ITEM (GH¢)	0.28	0.45	0.37	0.58	0.00	0.10
	AVERAGE COST TO ORG. PER STAFF PER DAY PER ITEM (GH¢)	2.22	3.57	2.94	4.60	0.00	0.80
	AVERAGE COST TO ORG. PER STAFF PER MONTH PER ITEM (GH¢)	44.37	71.45	58.77	92.04	0.00	15.99
	AVERAGE COST TO ORG. PER STAFF PER YEAR PER ITEM (GH¢)	532.39	857.35	705.24	1,104.53	0.00	191.87

Source: Author

Table 3B. Cost of Information Overload for Junior Staff

JUNIOR STAFF					
		MINUTES			
	TIME SPENT PER DAY ON:	**150**	**390**	**0**	**TOTAL**
		NO. OF JUNIOR STAFF			
1	Dealing with unnecessary interruptions.			25	75
2	Recovering in time to get back on track after interruptions.			25	75
3	Doing unproductive work.			25	75
4	Participating in unproductive meetings.			50	75
5	Searching for relevant information you cannot find.			50	75
6	Recreating information that already exists.			50	75
7	Duplicating information.			25	75
8	Thinking on your work.				75
9	Reflecting on your work.				75
10	Sending e-mails.			25	75
11	Reading e-mails.			25	75
12	Replying e-mails.			25	75
13	Deleting unnecessary e-mail.			25	75
14	Sending text messages.				75
15	Reading text messages.				75
16	Replying text messages.				75
17	Deleting unnecessary text messages.				75
18	Using Facebook.			50	75
19	Using Twitter.			25	75
20	Using Google+.			50	75
21	Using Foursquare.			75	75
22	Using Pinterest			75	75
23	Answering calls.	25			75
24	Making calls.	25			75
25	Filing information.			50	75
26	Deleting information.			25	75
27	Sorting information.			50	75
28	Changing from one task to another.			50	75
29	Reading unnecessary information offline.			50	75
30	Reading unnecessary information online.			50	75
31	Deleting spams from e-mail box.			25	75
32	Reading non-work related newsletters.			50	75
33	Unsubscribing from non-work related Newsletters.			50	75
34	Deleting non-work related Newsletters.			50	75
35	Putting certain types of email in certain types of folders.			50	75
36	Filing messages you might need someday.			25	75
37	To mentally switch from one topic to another topic and back again.			50	75
38	Dealing with instant messages.			50	75
	TOTAL NO. OF STAFF	50	0	1,250	

	TOTAL TIME (MINS) SPENT/DAY		7,500.00	0.00	0.00	
	AVERAGE SALARY PER MONTH (75 JUNIOR WORKERS) (GH¢)	919.58				
	AVERAGE SALARY PER DAY (20 WORKING DAYS/MONTH) (GH¢)	45.98				
	AVERAGE SALARY PER HR (24 HOURS/DAY) (GH¢)	1.92				
	AVERAGE SALARY PER MIN (60 MINS/HOUR) (GH¢)	0.03				
	AVERAGE COST TO ORG. PER MIN (60 MINS/HOUR) (GH¢)		239.47	0.00		1,022.55
	AVERAGE COST TO ORG. PER HR (60 MINS/HOUR) (GH¢)		14,368.38	0.00		61,352.98
	AVERAGE COST TO ORG. PER DAY (8 HOURS PER DAY) (GH¢)		114,947.04	0.00		490,823.84
	AVERAGE COST TO ORG. PER MONTH (20 DAYS/MONTH) (GH¢)		2,298,940.70	0.00		9,816,476.79
	AVERAGE COST TO ORG. PER YEAR (12 MONTHS/YEAR) (GH¢)		27,587,288.40	0.00		117,797,721.47
	AVERAGE COST TO ORG.PER STAFF PER MIN (GH¢)		0.34	0.00		1.46
	AVERAGE COST TO ORG. PER STAFF PER HOUR (GH¢)		20.53	0.00		87.65
	AVERAGE COST TO ORG. PER STAFF PER DAY (GH¢)		164.21	0.00		701.18
	AVERAGE COST TO ORG. PER STAFF PER MONTH (GH¢)		3,284.20	0.00		14,023.54
	AVERAGE COST TO ORG. PER STAFF PER YEAR (GH¢)		39,410.41	0.00		168,282.46
	AVERAGE COST TO ORG.PER STAFF PER MIN PER ITEM (GH¢)		0.009	0.00		0.04
	AVERAGE COST TO ORG. PER STAFF PER HR PER ITEM (GH¢)		0.54	0.00		2.31
	AVERAGE COST TO ORG. PER STAFF PER DAY PER ITEM (GH¢)		4.32	0.00		18.45
	AVERAGE COST TO ORG. PER STAFF PER MONTH PER ITEM (GH¢)		86.43	0.00		369.04
	AVERAGE COST TO ORG. PER STAFF PER YEAR PER ITEM (GH¢)		1,037.12	0.00		4,428.49

Source: Author

APPENDIX 2. Calculation of cost of Information overload.

Table 4A. Cost of Information Overload for Senior Staff

		SENIOR STAFF					
		MINUTES					
	TIME SPENT PER DAY ON:	**5.5**	**15.5**	**25.5**	**35.5**	**45.5**	**55.5**
		NO. OF SENIOR STAFF					
1	Dealing with unnecessary interruptions.	50			25		25
2	Recovering in time to get back on track after interruptions.	50	50		25		
3	Doing unproductive work.	50				25	
4	Participating in unproductive meetings.	25					
5	Searching for relevant information you cannot find.		50	25			
6	Recreating information that already exists.	25					25
7	Duplicating information.	25					
8	Thinking on your work.			25	25		
9	Reflecting on your work.			25	25		
10	Sending e-mails.	50	25	25			
11	Reading e-mails.	50	25	25			
12	Replying e-mails.	75		25			
13	Deleting unnecessary e-mail.	25	25				
14	Sending text messages.	75	50				
15	Reading text messages.	100	25				
16	Replying text messages.	100	25				
17	Deleting unnecessary text messages.	75	25				
18	Using Facebook.		75				
19	Using Twitter.	50					
20	Using Google+.	50	25	25			
21	Using Foursquare.	25					
22	Using Pinterest						
23	Answering calls.	75	25	25			
24	Making calls.	75		25			
25	Filing information.	50	25	25			
2	Deleting information.	50					

6								
27	Sorting information.		50	50				
28	Changing from one task to another.		50	50				
29	Reading unnecessary information offline.		25					
30	Reading unnecessary information online.		25					
31	Deleting spams from e-mail box.		75					
32	Reading non-work related newsletters.		25	25				
33	Unsubscribing from non-work related Newsletters.		75					
34	Deleting non-work related Newsletters.		50	25				
35	Putting certain types of email in certain types of folders.		50	75				
36	Filing messages you might need someday.		25		25		25	
37	To mentally switch from one topic to another topic and back again.		50	25			25	
38	Dealing with instant messages.		100					
	TOTAL NO. OF STAFF		1,750	700	275	100	75	75
	TOTAL TIME (MINS) SPENT/DAY		9,625.00	10,850.00	7,012.50	3,550.00	3,412.50	4,162.50
	AVERAGE SALARY PER MONTH (75 SENIOR WORKERS) (GH¢)	919.58						
	AVERAGE SALARY PER DAY (20 WORKING DAYS/MONTH) (GH¢)	45.98						
	AVERAGE SALARY PER HR (24 HOURS/DAY) (GH¢)	1.92						
	AVERAGE SALARY PER MIN (60 MINS/HOUR) (GH¢)	0.03						
	AVERAGE COST TO ORG. PER MIN (60 MINS/HOUR) (GH¢)		139.58	55.83	21.93	7.98	5.98	5.98

	AVERAGE COST TO ORG. PER HR (60 MINS/HOUR) (GH¢)	8,374.66	3,349.86	1,316.02	478.55	358.91	358.91
	AVERAGE COST TO ORG. PER DAY (8 HOURS PER DAY) (GH¢)	66,997.29	26,798.92	10,528.15	3,828.42	2,871.31	2,871.31
	AVERAGE COST TO ORG. PER MONTH (20 DAYS/MONTH) (GH¢)	1,339,945.85	535,978.34	210,562.92	76,568.33	57,426.25	57,426.25
	AVERAGE COST TO ORG. PER YEAR (12 MONTHS/YEAR) (GH¢)	16,079,350.17	6,431,740.07	2,526,755.03	918,820.01	689,115.01	689,115.01
	AVERAGE COST TO ORG.PER STAFF PER MIN (GH¢)	0.0798	0.0319	0.0125	0.0046	0.0034	0.0034
	AVERAGE COST TO ORG. PER STAFF PER HOUR (GH¢)	4.79	1.91	0.75	0.27	0.21	0.21
	AVERAGE COST TO ORG. PER STAFF PER DAY (GH¢)	38.28	15.31	6.02	2.19	1.64	1.64
	AVERAGE COST TO ORG. PER STAFF PER MONTH (GH¢)	765.68	306.27	120.32	43.75	32.82	32.82
	AVERAGE COST TO ORG. PER STAFF PER YEAR (GH¢)	9,188.20	3,675.28	1,443.86	525.04	393.78	393.78
	AVERAGE COST TO ORG.PER STAFF PER MIN PER ITEM (GH¢)	0.00210	0.00084	0.00033	0.00012	0.00009	0.00009
	AVERAGE COST TO ORG. PER STAFF PER HR PER ITEM (GH¢)	0.12593	0.05037	0.01979	0.00720	0.00540	0.00540
	AVERAGE COST TO ORG. PER STAFF PER DAY PER ITEM (GH¢)	1.00748	0.40299	0.15832	0.05757	0.04318	0.04318
	AVERAGE COST TO ORG. PER STAFF PER MONTH PER ITEM (GH¢)	20.14956	8.05982	3.16636	1.15140	0.86355	0.86355
	AVERAGE COST TO ORG. PER STAFF PER YEAR PER ITEM (GH¢)	241.79474	96.71790	37.99632	13.81684	10.36263	10.36263

Source: Author

Table 4B. Cost of Information Overload for Senior Staff

SENIOR STAFF					
		MINUTES			
	TIME SPENT PER DAY ON:	**150**	**390**	**0**	**TOTAL**
		NO. OF SENIOR STAFF			
1	Dealing with unnecessary interruptions.			25	125
2	Recovering in time to get back on track after interruptions.				125
3	Doing unproductive work.			50	125
4	Participating in unproductive meetings.			100	125
5	Searching for relevant information you cannot find.			50	125
6	Recreating information that already exists.			75	125
7	Duplicating information.			100	125
8	Thinking on your work.	25		50	
9	Reflecting on your work.			50	125
10	Sending e-mails.			25	125
11	Reading e-mails.			25	125
12	Replying e-mails.			25	125
13	Deleting unnecessary e-mail.			75	125
14	Sending text messages.				125
15	Reading text messages.				125
16	Replying text messages.				125
17	Deleting unnecessary text messages.			25	125
18	Using Facebook.			50	125
19	Using Twitter.			75	125
20	Using Google+.			25	125
21	Using Foursquare.			100	125
22	Using Pinterest			125	125
23	Answering calls.				125
24	Making calls.			25	125
25	Filing information.			25	125
26	Deleting information.			75	125
27	Sorting information.			25	125
28	Changing from one task to another.			25	125
29	Reading unnecessary information offline.			100	125
30	Reading unnecessary information online.			100	125
31	Deleting spams from e-mail box.			50	125
32	Reading non-work related newsletters.			75	125
33	Unsubscribing from non-work related Newsletters.			50	125
34	Deleting non-work related Newsletters.			50	125
35	Putting certain types of email in certain types of folders.				125
36	Filing messages you might need someday.			50	125
37	To mentally switch from one topic to another topic and back again.			25	125
38	Dealing with instant messages.			25	125
	TOTAL NO. OF STAFF	25	0	1,750	
	TOTAL TIME (MINS) SPENT/DAY	3,750.00	0.00	0.00	
	AVERAGE SALARY PER MONTH (75 SENIOR WORKERS) (GH¢) 919.58				
	AVERAGE SALARY PER DAY (20 WORKING 45.98				

	DAYS/MONTH) (GH¢)					
	AVERAGE SALARY PER HR (24 HOURS/DAY) (GH¢)	1.92				
	AVERAGE SALARY PER MIN (60 MINS/HOUR) (GH¢)	0.03				
	AVERAGE COST TO ORG. PER MIN (60 MINS/HOUR) (GH¢)		1.99	0.00		239.28
	AVERAGE COST TO ORG. PER HR (60 MINS/HOUR) (GH¢)		119.64	0.00		14,356.56
	AVERAGE COST TO ORG. PER DAY (8 HOURS PER DAY) (GH¢)		957.10	0.00		114,852.50
	AVERAGE COST TO ORG. PER MONTH (20 DAYS/MONTH) (GH¢)		19,142.08	0.00		2,297,050.02
	AVERAGE COST TO ORG. PER YEAR (12 MONTHS/YEAR) (GH¢)		229,705.00	0.00		27,564,600.29
	AVERAGE COST TO ORG.PER STAFF PER MIN (GH¢)		0.0011	0.00		0.14
	AVERAGE COST TO ORG. PER STAFF PER HOUR (GH¢)		0.07	0.00		8.20
	AVERAGE COST TO ORG. PER STAFF PER DAY (GH¢)		0.55	0.00		65.63
	AVERAGE COST TO ORG. PER STAFF PER MONTH (GH¢)		10.94	0.00		1,312.60
	AVERAGE COST TO ORG. PER STAFF PER YEAR (GH¢)		131.26	0.00		15,751.20
				0.00		
	AVERAGE COST TO ORG.PER STAFF PER MIN PER ITEM (GH¢)		0.00003	0.00		0.0036
	AVERAGE COST TO ORG. PER STAFF PER HR PER ITEM (GH¢)		0.00180	0.00		0.2159
	AVERAGE COST TO ORG. PER STAFF PER DAY PER ITEM (GH¢)		0.01439	0.00		1.7271
	AVERAGE COST TO ORG. PER STAFF PER MONTH PER ITEM (GH¢)		0.28785	0.00		34.5421
	AVERAGE COST TO ORG. PER STAFF PER YEAR PER ITEM (GH¢)		3.45421	0.00		414.5053

Source: Author

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Socio-Economic Implications of Wireless Sensor Networks with Special Reference to its Application in Agriculture

Pankaj Sharma
Department of Computer Sciences
Abhilashi Educational Society
Ner Chowk, Mandi, H.P.
pankajgrooves1977@gmail.com

ABSTRACT

Sensor Networks have been found to be one of the most important technologies for 21st century. Wireless sensor network are gaining greater attention from the research community. Sensor node which is the basic element of wireless node network is composed of sensing computation and wireless communication unit. These sensor nodes are hence capable of observing physical phenomenon, process the observed and received information and communicate the observed or processed information to the nearby sensor nodes to form a network of sensor nodes called Wireless Sensor Networks. The wireless networking capability of the sensor enabled nodes have resulted in various interesting applications ranging from surveillance, smart homes, precision agriculture, disaster detection, and traffic control to vehicular and supply chain management applications.

Keywords: Wireless Sensor Network, Agriculture, Monitoring, Application, ZigBee

African Journal of Computing & ICT Reference Format:
Pankaj Sharma (2013). Socio-Economic Implications of Wireless Sensor Networks with Special Reference to its Application in Agriculture. Afr J. of Comp & ICTs. Vol 6, No. 2. pp 31-40.

1. INTRODUCTION

A sensor network is a computer network Composed of a large number of sensor nodes. The evolution of wireless sensor network technology has enabled us to develop advanced systems for real time monitoring. In the present scenario wireless sensor networks are increasingly being used for precision agriculture. The advantages of using wireless sensor networks in agriculture are distributed data collection and monitoring, monitor and control of climate, irrigation and nutrient supply. Hence decreasing the cost of production and increasing the efficiency of production [1]. A wireless sensor network is a system comprised of radio frequency (RF) transceivers, sensors, microcontrollers and power sources.

Wireless sensor networks with self-organizing, self-configuring, self-diagnosing and self-healing capabilities have been developed to solve problems or to enable applications that traditional technologies could not address. Once available, these technologies would allow us to find many new applications that could not have been considered possible before [2]. The use of wireless sensor network for the large area is now becoming popular in green house technology of precision agriculture. The parameters of greenhouse to be controlled are increasing day by day so that it may cause the data traffic and congestion in the future [3].

An obvious advantage of wireless transmission is a significant reduction and simplification in wiring and harness. Wireless sensors allow otherwise impossible sensor applications, such as monitoring dangerous, hazardous, unwired or remote areas and locations. This technology provides nearly unlimited installation flexibility for sensors and increased network robustness. Furthermore, wireless technology reduces maintenance complexity and costs [4, 5]. Wireless Sensors Network (WSN) is the ubiquitous field showing wide spectrum of applications in various sectors. Wireless Sensor Networks (WSN) is an important and exciting new technology with great potential for improving many current applications in medicine, transportation, agriculture, industrial process control, and the military as well as creating new revolutionary systems in areas such as global-scale environmental monitoring, precision agriculture, home and assisted living medical care, smart buildings and cities, and numerous future military applications [6].

1.1 Mechanism of WSN

In ad hoc networks, wireless nodes self-organize into an infrastructure less network with a dynamic topology. Sensor networks share these traits, but also have several distinguishing features. The number of nodes in a typical sensor network is much higher than in a typical ad hoc network, and dense deployments are often desired to ensure coverage and connectivity; for these reasons, sensor network hardware must be cheap.

Nodes typically have stringent energy limitations, which make them more failure-prone. They are generally assumed to be stationary, but their relatively frequent breakdowns and the volatile nature of the wireless channel none the less result in a variable network topology. Ideally, sensor network hardware should be power-efficient, small, inexpensive, and reliable in order to maximize network lifetime, add flexibility, facilitate data collection and minimize the need for maintenance. In a generic sensor node, we can identify a power module, a communication block, a processing unit with internal and/or external memory, and a module for sensing and actuation.

Using stored energy or harvesting energy from the outside world are the two options for the power module. Energy storage may be achieved with the use of batteries or alternative devices such as fuel cells or miniaturized heat engines, whereas energy-scavenging opportunities are provided by solar power, vibrations, acoustic noise, and piezoelectric effects. The vast majority of the existing commercial and research platforms relies on batteries, which dominate the node size. Most sensor networks use radio communication, even if alternative solutions are offered by laser and infrared. Nearly all radio-based platforms use COTS (Commercial Off-The-Shelf) components. Popular choices include theTR1000 from RFM (used in the MICA motes) and the CC1000 from Chipcon (chosen for the MICA2 platform).More recent solutions use industry standards like IEEE802.15.4 (MICAz and Telosmotes with CC2420 fromChipcon) or pseudo-standards like Bluetooth. Typically, the transmit power ranges between −25 dBm. and 10 dBm while the receiver sensitivity can be as good as −110 dBm. Although low-power FPGAs might become a viable option in the near future, micro controllers (MCUs)are now the primary choice for processing in sensor nodes.

The key metric in the selection of an MCU is power consumption. Sleep mode deserve special attention, as in many applications low duty cycles are essential for lifetime extension. Just as in the case of the radio module, a fast wake-up time is important. The high sampling rates of modern digital sensors are usually not needed in sensor networks. The power efficiency of sensors and their turn-on and turn-off time are much more important. Additional issues are the physical size of the sensing hardware, fabrication, and assembly compatibility with other components of the system. Packaging requirements come into play, for instance, with chemical sensors which require contact with the environment [7].

1.2 Zigbee

ZigBee is an established set of specifications for wireless personal area networking (WPAN), i.e. digital radio connections between computers and related devices. ZigBee provides specifications for devices that have low data rates, consume very low power and are thus characterized by long battery life. ZigBee makes possible completely networked homes where all devices are able to communicate and be controlled by a single unit. ZigBee is a low-cost, low-power, wireless mesh network standard. The low cost allows the technology to be widely deployed in wireless control and monitoring applications. Low power-usage allows longer life with smaller batteries. Mesh networking provides high reliability and more extensive range.

ZigBee chip vendors typically sell integrated radios and microcontrollers with between 60 KB and 256 KB flash memory. ZigBee operates in the industrial, scientific and medical (ISM) radio bands; 868 MHz in Europe, 915 MHz in the USA and Australia, and 2.4 GHz in most jurisdictions worldwide. Data transmission rates vary from 20 to 900 kilobits/second. The ZigBee network layer natively supports both star and tree typical networks, and generic mesh networks. Every network must have one coordinator device, tasked with its creation, the control of its parameters and basic maintenance. Within star networks, the coordinator must be the central node. The ZigBee based drip irrigation system proves to be a real time feedback control system which monitors and controls all the activities of drip irrigation system efficiently. This will modernize the agriculture field and irrigation control by using ZigBee is one of the good technologies for controlling irrigation over large agricultural sector areas for growing of crops [8].

1.3 Types of wireless sensors

Current WSNs are deployed on land, underground, and underwater. Depending on the environment, a sensor network may be of following types [7]:

- Terrestrial WSN,
- Underground WSN,
- Underwater WSN,
- Multi-media WSN, and
- Mobile WSN

Terrestrial WSNs typically consist of hundreds to thousands of inexpensive wireless sensor nodes deployed in a given area, either in an ad hoc or in a pre-planned manner. In ad hoc deployment, sensor nodes can be dropped from a plane and randomly placed into the target area. In pre-planned deployment, there is grid placement, optimal placement, 2-d and 3-d placement and models. Underground WSNs and consist of a number of sensor nodes buried underground or in a cave or mine used to monitor underground conditions. Additional sink nodes are located above ground to relay information from the sensor nodes to the base station.

Underwater WSNs consist of a number of sensor nodes and vehicles deployed underwater. As opposite to terrestrial WSNs, underwater sensor nodes are more expensive and fewer sensor nodes are deployed. Autonomous underwater vehicles are used for exploration or gathering data from sensor nodes [9]. Multi-Media WSN have been proposed to enable monitoring and tracking of events in the form of multi-media such as video, audio, and imaging. Multi-media WSNs consist of a number of low cost sensor nodes equipped with cameras and microphones. These sensor nodes inter connect with each other over a wireless connection for data retrieval, process, correlation, and compression. Mobile WSNs consist of a collection of sensor nodes that can move on their own and interact with the physical environment. Mobile nodes have the ability sense, compute, and communicate like static nodes. A key difference is mobile nodes have the ability to reposition and organize itself in the network [10].

1.4 Challenges and Constraints

A sensor network faces different challenges and constraints [10, 11]:

Hardware Cost: The current cost of each individual sensor unit is still very high. Commercially available platforms cost in the order of Rs. 5000 per unit with temperature, humidity and light sensors when bought in large quantities. Capable sensors able to track human mobility inside buildings are costing around Rs.15000 per unit.

System Architecture: There is no unified system and networking architecture that is stable and mature enough to build different applications on top. Most of the applications and research prototypes are vertically integrated in order to maximize performance.

Wireless Connectivity: Wireless communication in indoor environments is still quite unpredictable using low-power consumption RF transceivers, in particular in clutter environments common inside buildings, with many interfering electromagnetic fields, such as the one produced by elevators, machinery and computers, among others.

Programmability: Some form of network re-programmability is desirable; doing so in energy and communication conservative form remains a challenge.

Security: The security challenges are at many levels.

• From the system point of view, it is critical that the information provided by the nodes be authenticated and the integrity verified, since this information provides the feedback loop to expensive equipment controlling power consumption in the building.

• From the user's point of view, it is also critical that this information cannot be easily spoofed and remains protected in the back end processor, since it may affect the privacy of users.

Programming Abstractions*:* A key to the growth of WSN is raising the level of abstraction for programmers. Currently, programmers deal with too many low levels details regarding sensing and node to node communication.

Algorithms*:* WSNs are composed of a large number of sensor nodes, therefore, an algorithm for a WSN is implicitly a distributed algorithm. A distributed algorithm is an algorithm designed to run on computer hardware constructed from inter connected processors. Distributed algorithms are used in many varied application areas of distributed computing, such as telecommunications, scientific computing, distributed information processing, and real-time process control.

1.5 Applications of WSN in Agriculture:

Drip Irrigation:

Drip irrigation, also known as trickle irrigation or micro irrigation or localized irrigation, is an irrigation method which saves water and fertilizer by allowing water to drip slowly to the roots of plants, either onto the soil surface or directly onto the root zone, through a network of valves, pipes, tubing, and emitter (Fig. 1). It is done with the help of narrow tubes which delivers water directly to the base of the plant. Modern drip irrigation has arguably become the world's most valued innovation in agriculture since the invention of the pact sprinkler in the 1930s, which offered the first practical alternative to surface irrigation. Drip irrigation may also use devices called micro-spray heads, which spray water in a small area, instead of dripping emitters. These are generally used on tree and vine crops with wider root zones.

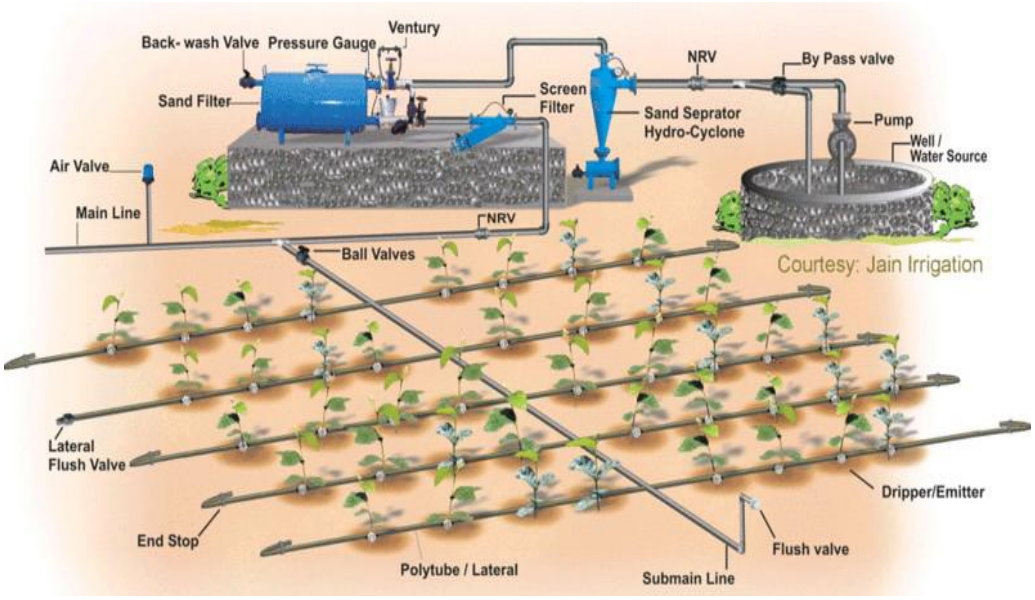


Fig. 1.: Layout showing working of drip irrigation [8]

In drip irrigation system, soil moisture sensing network is used to monitor the moisture contained in soil. Soil moisture is more generally considered within the context of hydrology, where it represents the immediate store of infiltrating rainfall, before it contributes to groundwater recharge. Three different sensors are used to monitor three layers of soil. Accordingly, further action is taken by microcontroller as the output of network is given to the microcontroller. Indicator indicates whether the soil is dry or wet. Microcontroller is the heart of the system; it controls the overall irrigation system. It takes the input from moisture sensor 1, 2, 3 etc. & according to the written program it turns ON or OFF the motor pump. It also indicates the condition of soil. Also it provides the data to the PC through ZigBee module. When soil is dry motor is on and when soil is wet motor is off. Thus microcontroller controls the operation of motor. ZigBee module is used generally to have wireless link between PC & the main irrigation system so that data can be logged into PC. AC or DC motor can be used for whole system. On the basis of soil moisture detection, motor ON/OFF working will be done. Provision of water and considering the need of water to the crop is done by controlling motor [8].

2. SENSORS FOR MONITORING WATER LEVEL

Normally the water level in a field will be uniform throughout the field. Water wells can be made as per the need and the water level in each well can be monitored. Water level sensors can be used for sensing the levels. These sensors can monitor the level according to the user needs. Normally there can be three levels, normal, high and low. These sensors are electro-mechanical devices. Even though electronic sensors are available due to the environmental conditions electro-mechanical devices are more applicable. The mechanical part in the devices will float on water and the electrical part will produce the signals based on the portion of the floating device. These values or signals generated by the sensor needs to be transmitted to the farmers. A farmer may have hectors of area as paddy field and need hundreds of sensors for monitoring.

Hence wired devices are not practical. Use of commonly used wireless communication technology is also not advisable due to complexity of the communication system and the high power needs of such systems. Proving power supply or having a solar panel is also not economical. High battery power may be needed by such devices and regular charging or replacement of the batteries is highly costly. Alternative solution is to provide power from a solar panel. The solar panels are of high cost and hence are after prove to theft. Low cost communication devices which needs low power and less maintenance, which can operate on a wireless architecture is the solution. The new generation wireless sensor networks can be considered for the situation.

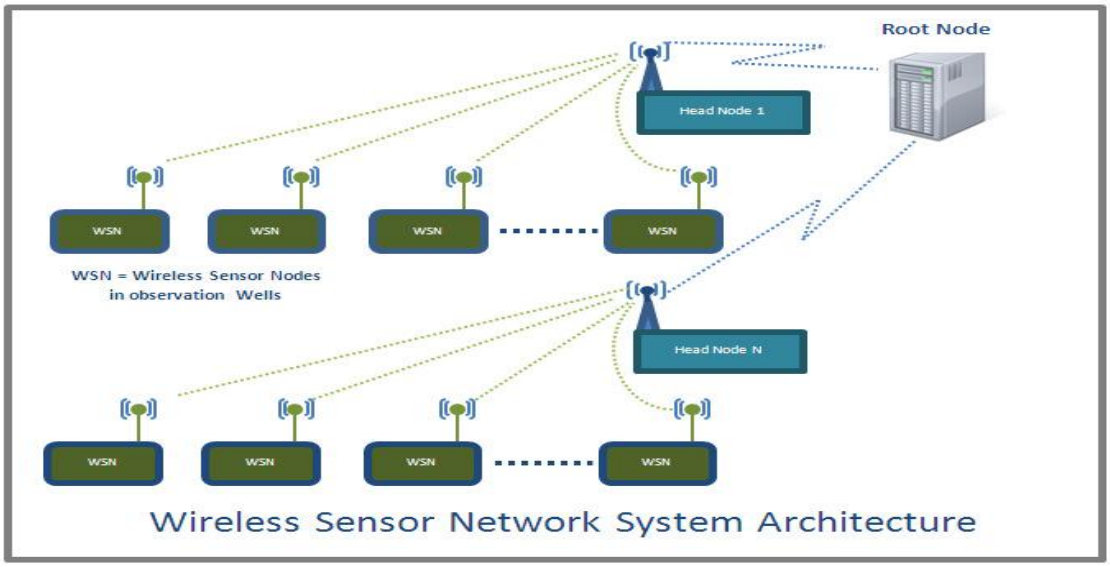


Fig. 2: WSN System architecture [1]

2.1 Data Transport Mechanism:
A tree like structure is followed in this system. The leaf level of the tree consists of sensors for obtaining the water level in the wells. There may be many such sensors so the power consumption of these devices is critical to the overall performance of the system. Therefore attempts are made to design a sensor node board with low power micro controllers, low power semiconductor based on MEMS sensors, ZigBee IEEE 802.15.4 wireless transceiver. These sensors will collect the data and communicate to the next higher level, the head nodes. These head nodes receive the data through the ZigBee module and process it over there. They are also equipped with GPRS Bridge modules for establishing communication with the root server node. By using any GSM network the message will be send to the root server if needed [1].Initial system architecture for data transport mechanism is as shown in the Fig 2.

3. WSN CONTROLLED UNMANNED AERIAL VEHICLES (UAVS) SPRAY IN AGRICULTURAL FIELDS

Unmanned Aerial Vehicles (UAV) is used to spray chemicals on an agricultural field. However, the neighbouring field, which may belong to another owner, must not be sprayed. Moreover, the UAV must respect their lane of operation (boundary). If the UAV used for spraying comes too close to the neighbouring field, or if there is a sudden change in the direction of the wind, the chemicals might fall on the neighbouring field and this must be avoided.
To be able to adjust the trajectory, we propose that the UAV gets information from the WSN deployed in the crop field (Fig. 3). Whether a sensor detects an excessive concentration of chemicals, the spraying UAV will be directed away from the border. The process of applying the chemicals is controlled by means of the feedback from the wireless sensors network deployed at ground level on the crop field. It has been divided into two modules: (i) the Behavioural Module and (ii) the Chemical Dispersion Module. The two modules run simultaneously, in an integrated way with socket-based communication. The Behavioural Module sends the current position of the UAV (x, y, z) to the Dispersion Module along with the UAV orientation and velocity (v).

Furthermore, in the Behavioural Module, the wind modelling is performed, which emulates changes in wind direction and velocity, and provides information to the Dispersion Module about changes in the environment. The Dispersion Module performs the calculation of the fall of the chemicals, by obtaining the position and fall time of each drop. The WSN, in turn, identifies the amount and position of the chemicals and returns this information to the Behavioral Module. Periodically, the UAV sends a broadcast message to the ground sensor nodes, requesting concentration in its area. The ground sensor nodes that receive this message, connect to the Dispersion Module and request its concentration using their positions (x, y, z) as parameters. In this way, they can respond by giving the concentration in this area to the UAV. By means of these response messages from the ground nodes, the UAV can call a decision manager instance to compute its decision and then change its route if necessary [12].

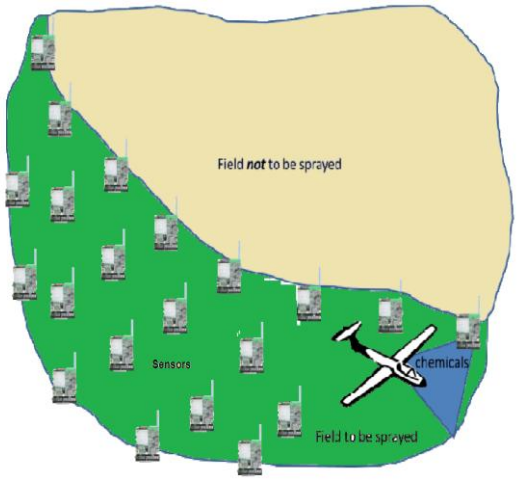


Fig. 3: Unmanned Aerial Vehicles [12]

4. GREENHOUSES

The automation and efficiency on greenhouse environment monitoring and control are crucial. In order to control and monitor the environmental factors, sensors and actuators are essential. Greenhouse crops can benefit a lot using WST, because inside the greenhouse the crop conditions such as climate and soil do not depend on natural agents. Thus, the implementations are easier than in outdoor applications. The first application of WSN in a greenhouse was reported in the year 2003. It was a monitoring and control system developed by means of Bluetooth [13]. A WSN in a greenhouse, that integrates a variety of sensors which can measure substrate water, temperature, electrical conductivity, daily photosynthetic radiation and leaf wetness in real-time.

Benefits came from an improved plant growth, more efficient water and fertilizer applications, together with a reduction in disease problems related to over-watering [14]. With the use of greenhouse concept, the farmer can produce different crops in different climates and various seasons. However, the farmer can easily keep the desired crop's environment conditions. To fulfil this requirement we need the environmental parameter sensors, such as temperature sensor, humidity sensor, CO_2 sensor etc. All these sensors can be connected to server or sink node without wire. This network can help to monitor and control all the environmental parameter of Precision Agriculture [15].

4.1 Other Applications

1. Cold Chain Monitoring

Many kinds of products have to be handled under controlled environmental quantities, such as temperature, humidity, vibrations, and exposure to light. Among these parameters, the temperature is usually a major concern due to its huge variety of effects. For example, if the temperature of some chilled foods exceeds specific limits, a great decrease in quality, along with an increase in the risk of food poisoning, can occur. The limits can be quite strict for chilled products with a storage temperature near 0˚C, where a rise in temperature of just a few degrees could cause microbial growth. The situation is more serious in the case of pharmaceutical products since an uncontrolled change in the temperature, even for short time intervals, could cancel the effectiveness of the product or even make it dangerous.

The chain that brings the temperature-sensitive products from the factory to the consumer through an uninterrupted series of steps under a controlled temperature is usually called the "cold chain." [16]. Perishable food products such as vegetables, fruit, meat or fish require refrigerated transports. Therefore, temperature is the most important factor when prolonging the practical shelf life of perishable food products. Studying and analyzing temperature gradients inside refrigeration rooms, containers, and trucks is a primary concern of the industry. The supply chain management of fresh foods requires fast decisions because goods are forwarded within hours after arrival at the distribution centre. Appropriate planning calls for more information than that which could be provided by standard RFID tracking and tracing.

Quality problems should be detected as quickly as possible, and alarms should be triggered when temperature gradients cross a threshold. Even if direct access to the means of transport is not possible, online notifications offer new opportunities for improved transport planning. The use of wireless sensors in refrigerated vehicles was proposed by Qingshan *et al*. The vehicles can host a variety of sensors to detect, identify, log, and communicate what happens during the journey, monitoring the status of perishable products in transport [17].

4.2. Health Care

As the world's population ages, those suffering from diseases of the elderly will increase. In-home and nursing- home pervasive networks may assist residents and their caregivers by providing continuous medical monitoring, memory enhancement, control of home appliances, medical data access, and emergency communication [18]. Advances in wireless sensor networking have opened up new opportunities in healthcare systems. In some modern hospital sensor networks are constructed to monitor patient physiological data, to control the drug administration track and monitor patients and doctors and inside a hospital. Pressure sensors monitor a patient's condition by providing accurate and reliable diagnostics in a broad range of conditions. Free scale devices are specifically designed for applications where high quality and reliability are especially important and are constructed using materials with a proven history in the health care industry [19].

Becoming mature enough to be used for improving the quality of life, wireless sensor network technologies are considered as one of the key research areas in computer science and healthcare application industries [20]. We are developing a network architecture for smart healthcare that will open up new opportunities for continuous monitoring of assisted and independent-living residents. High costs of installation and retrofit are avoided by using ad hoc, self-managing networks. Based on the fundamental elements of future medical applications (integration with existing medical practice and technology, real-time and long term monitoring, wearable sensors and assistance to chronic patients, elders or handicapped people), our wireless system will extend healthcare from the traditional clinical hospital setting to nursing and retirement homes, enabling telecare without the prohibitive costs of retrofitting existing structures [21].

4.3. Environmental Climate Monitoring Applications

Sensor networks are dense wireless networks of small, low-cost sensors, which collect and disseminate environmental data. Wireless sensor networks facilitate monitoring and controlling of physical environments from remote locations with better accuracy [22]. Large number of environmental applications makes use of WSNs. Sensor nodes are deployed in the environment for monitoring biodiversity. Recent developments in wireless network technology and miniaturization mean that for the first time, the natural environment can be realistically monitored. These systems can potentially provide new data for environmental science (e.g. climate models) as well as vital hazard warnings. This is particularly important in remote or dangerous environments where many fundamental processes have rarely been studied due to their inaccessibly [23]. Monitoring the behaviour of ice caps and glaciers is an important part of our understanding of the Earth's climate.

4.4. Disaster Management Applications

Ongoing research on deploying small-scale, battery-powered and wirelessly connected UAVs carrying cameras for disaster management applications. In this ''aerial sensor network'' several UAVs fly in formations and cooperate to achieve a certain mission. The ultimate goal is to have an aerial imaging system in which UAVs build a flight formation, fly over a disaster area such as wood fire or a large traffic accident, and deliver high-quality sensor data such as images or videos. These images and videos are communicated to the ground, fused, analyzed in real-time, and finally delivered to the user. In disaster management application, we introduce our aerial sensor network and its application in disaster situations. We discuss challenges of such aerial sensor networks and focus on the optimal placement of sensors. We formulate the coverage problem as integer linear program (ILP) and present first evaluation results.

The important task for which UAVs are used is to provide a bird's eye view and thus allow assessing the current situation. Usually, in disaster situations the first responders cannot rely on a fixed infrastructure and the available information (maps, etc.) may no longer be valid. The overall goal of collaborative microdrones (cDrones) project, hence, is to provide the first responders a quick and accurate overview of the affected area, typically spanning hundreds of thousands of square meters, and augment the overview image with additional information such as detected objects or the trajectory of moving objects. Covering such a large area with a single image from a UAV flying at low altitude (up to 100 m) is not possible. Moreover, a set of images is taken and stitched together for a large overview image. Due to the limitations of a single UAV and stringent time constraints we use multiple UAVs which form an airborne wireless sensor network.

The UAVs coordinate themselves during flight requiring as little control by a human operator as possible. UAVs

have to exchange flight data on a regular basis in order to coordinate themselves. This flight data includes the current position, speed, direction, etc. For individual UAVs this data can be exchanged every few seconds. But if two or more UAVs fly in a formation, the UAVs need to know each other's position more accurately. Thus, the position update interval is in the range of several milliseconds. Hence, communication links with low latency and a communication range of several hundred meters are required [24].

4.5. Speed Sensor in wind Turbine

Wind energy is presently the fastest growing renewable energy source in the world. However, the industry still experiences premature turbine component failures, which lead to increased operation and maintenance (O&M) costs and subsequently, increased cost of energy (COE). To make wind power more competitive, it is necessary to reduce turbine downtime and increase reliability. Condition monitoring may help by reducing the chances of catastrophic failures, enabling cost-effective operation and maintenance practices, and providing inputs to improve turbine operation, control strategy, and component design [25].

Wind power has become a leading source of renewable energy and is undergoing a remarkable expansion. Wind velocity is one of the key parameters determining the performance of a wind turbine. Beside the turbine control application, wind speed is widely used in wind-farm siting, power scheduling, and grid management. The fast-growing wind energy industry strongly depends on the accuracy and robustness of wind-speed measurements. Wind speed is usually measured by a mechanical anemometer installed at the nacelle top of a wind turbine. Prior research has considered wind-speed measurement, sensor calibration, prediction of wind speed, and control optimization on the basis of the estimated wind speed. The wind-speed sensor models were used as the reference wind-speed sensors for monitoring the performance of anemometers [26].

6. Fiber-Optic Sensors for the Exploration of Oil and Gas

A large variety of sensors are used in the oil service industry for the exploration and development of oil and gas. These sensors should operate under severe conditions of high pressure, up to 20,000 psi [1 atmospheric pressure = 14.7 psi], and high temperature, up to 175^0C. Owing to exploration activity in increasingly deeper wells, sensors that can function in higher-temperature and other hostile environments are required. Fiber-optic technology is a key enabler with its higher temperature capacity, multiplexing and distributed sensing capability, and small space requirements to meet these demanding applications [27].

Understanding crude oil composition early in a development process helps optimize resource exploitation. Such information is now available from a fluid analyzer using data from a formation dynamic tester tool to determine formation properties in open holes. Fluid analysis provides real-time information to optimize fluid sampling based on the measured in-situ composition. An absorption spectrometer and a fluorescence and gas detector are used. Down hole fluid is introduced into the flow line through the formation probe, and optical sensors analyze the fluid in the flow line. A fiber-optic sensing system offers many advantages for monitoring fluid movement in the larger reservoir-scale level through its capabilities in distributed sensing, multiplexing, and high-temperature operations. A fiber optic distributed temperature sensor (DTS) is the most popular fiber-optic sensor for down hole applications [28].

8. Using Remote Sensing in Drought and Flood:

Remote sensing technology along with geographic information system (GIS) has become the key tool for flood monitoring in recent years. Development in this field has evolved from optical to radar remote sensing, which has provided all weather capability compared to the optical sensors for the purpose of flood mapping. The central focus in this field revolves around delineation of flood zones and preparation of flood hazard maps for the vulnerable areas. In this exercise flood depth is considered crucial for flood hazard mapping and a digital elevation model (DEM) is considered to be the most effective means to estimate flood depth from remotely sensed or hydrological data [29].

Droughts and floods are water-related natural disasters which affect a wide range of environmental factors and activities related to agriculture, vegetation, human, wild life and local economies. Drought is the single most important weather-related natural disaster often aggravated by human action, since it affects very large areas for months and years and thus has a serious impact on regional food production, life expectancy for entire populations and economic performance of large regions or several countries. Remote sensing techniques make it possible to obtain and distribute information rapidly over large areas by means of sensors operating in several spectral bands, mounted on aircraft or satellites. A satellite, which orbits the Earth, is able to explore the whole surface in a few days and repeat the survey of the same area at regular intervals, whilst an aircraft can give a more detailed analysis of a smaller area, if a specific need occurs.

The spectral bands used by these sensors cover the whole range between visible and microwaves. Rapid developments in computer technology and the Geographical Information Systems (GIS) help to process Remote Sensing (RS) observations from satellites in a spatial format of maps - both individually and along with tabular data and "crunch" them together to provide a new perception - the spatial visualization of information of natural resources. The integration of information derived from RS techniques with other datasets - both in spatial and non-spatial formats provides tremendous potential for identification, monitoring and assessment of droughts and floods [30].

REFERENCES

[1] Simon, S. and Jacob, K. P. 2012. Wireless Sensor Networks for Paddy Field Crop Monitoring Application in Kuttanad. *International Journal of Modern Engineering Research* 2(4): 2017-2020.

[2] Shinghal, K. and Noor, A. 2011. Low Power pH Sensor for- Wireless Sensor Network Node Agricultural Application. *International Journal of Advances in Engineering & Technology*: pp.-2231-1963.

[3] Chaudhary, D., Nayse, S. and Waghmare, L.M. 2011.Application of Wireless Sensor Networks for Greenhouse Parameter Control in Precision Agriculture.*International Journal of Wireless & Mobile Networks*:Vol. 3.

[4] IEEE 1451.2 Standard, A Smart Transducer Interface for Sensors and Actuators. Transducer to Microprocessor Communication Protocols and Transducer Electronic Data Sheet (TEDS) Formats, Piscataway, NJ: IEEE Standards Department, 1998.

[5] Sensor Device Data/Handbook, DL200/D Rev. 4.1998. Motorola Semiconductor Products Sector, Austin, TX.Najafi, K., .Smart Sensors. *J. Micromechanics and Micro engineering*, Vol. 1, 1991, pp. 86.102.

[6] Stankovic, J.2008.When sensor and actuator networks cover the world.*ETRI Journal* 30(5), 627–633.

[7] Puccinelli, D. and Haenggi, M. 2005. Wireless Sensor Networks: Applications and Challenges

[8] of Ubiquitous Sensing. *IEEE Circuits and Systems Magazine*: 19-29.

[9] Awati, J.S. and Patil, V.S.2012. Automatic Irrigation Control by using wireless sensor networks. *Journal of Exclusive Management Science.*1 (6): 2277–5684.

[10] Schwiebert, L. Gupta, K. S. and Weinmann,J.2001."Research challenges inwireless networks of biomedical sensors". *Mobile Computing and Networking*: 151–165.

[11] Srivastava, N. 2010. Challenges of next-generation wireless sensor networks and its impact on society. *Journal of Telecommunications*, 1(1): 128-133.

[12] Kaler, E. B. S. and Kaler, E. M. K. 2010."Challenges in wireless sensor networks," In Proc. of ISCET. Available at: www.rimtengg. com/iscet/proceedings/pdfs/misc/176.pdf.

[13] Costa, F.G., Ueyama, J., Braun, T.,Pessin, G. 2012. The Use Of Unmanned Aerial Vehicles And Wireless Sensor Network Inagricultural Applications. Geoscience and Remote Sensing Symposium (IGARSS), IEEE International: 5045 – 5048.

[14] Liu, G. and Ying, Y. 2003. Application of Bluetooth technology in greenhouse environment, monitor and control. *J. Zhejiang Univ., Agric Life Sci*. 29: 329-334.

[15] Lea-Cox, J.D. Kantor, G. Anhalt, J. Ristvey, A. and Ross, D.S. 2007. A wireless sensor network for the nursery and greenhouse industry.In Southern Nursery Association Research Conference, Vol. 52.

[16] Burrell,J.Brooke, T. andBeckwith, R.2004. Vineyard computing: sensor networks in agricultural production. *IEEE Pervasive Computing*. 3(1): 38–45.

[17] Carullo, A.,Corbellini, S., Parvis, M. and Vallan, A. 2009. A Wireless Sensor Network for Cold-Chain Monitoring IEEE Transactions on Instrumentation and Measurement(58)5: 1405.

[18] Qingshan, S., Ying, L., Gareth, D. and Brown, D. 2004. Wireless intelligent sensor networks for refrigerated vehicle. In IEEE 6th Symposium on Emerging Technologies Mobile and Wireless Communication, Shanghai, China.

[19] Virone, G. Wood, A. Selavo, L. Cao, Q. Fang, L. Doan, T. He, Z. Stoleru, R. Lin, S.andStankovic, J.A.2006."An Advanced Wireless Sensor Network for Health Monitoring". Proceedings of the Transdisciplinary Conference on Distributed Diagnosis and Home Health Care.

[20] http://www.freescale.com/files/sensors/doc/fact_sheet/MEDPRESSENSFS.pdf

[21] Alemdar, H. andErsoy, C.2010. Wireless sensor networks for healthcare: A survey.*Computer Networks* 54: 2688–2710

[22] 21 Stankovic, J. A. Cao, Q. Doan, T. Fang, L. He, Z. Kiran, R. Lin, S. Son, S. Stoleru, R. and Wood, A.2005. "Wireless Sensor Networks for In-Home Healthcare: Potential and Challenges," in High Confidence Medical Device Software and Systems (HCMDSS) Workshop, Philadelphia

[23] Karl, H. and Willig, A.2005. Protocols and Architectures for Wireless Sensor Networks, John Wiley and Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, England.

[24] Martinez, K., Hart, K.J. and Ong, R. 2004.Environmental Sensor Networks.*IEEE Computer*, pp 1-6.

[25] Quaritsch, M., Kruggl, K., Wischounig-Strucl, D., Bhattacharya, S., Shah, M. and Rinner, B. 2010. Networked UAVs as aerial sensor network for disaster management applications *Elektro technik & Information stechnik* 127(3): 56–63.

[26] Sheng. S, Veers, P. 2011. Wind Turbine Drivetrain condition Monitoring – An Overview, Mechanical Failures Prevention Group: Applied Systems Health Management Conference, Virginia.

[27] Kusiak, A., Zheng H. and Zhang, Z. 2011. Virtual Wind Speed Sensor for Wind Turbines. *Journal of Energy Engineering*. 137(2): 59–69.

[28] Yamate, T. 2005. Fiber-Optic Sensors for the Exploration of Oil and Gas. Review of Laser Engineering 33(9): 582-586.

[29] Curtis, C., Kopper, R., Decoster, E., Guzmán-Garcia, A., Huggins, C., Knauer, L., Minner, M., Kupsch, N., Linares, L., Rough H. and Waite, M. 2002. "Heavy-Oil Reservoirs". *Oil field Review*: 30-51.

[30] Sanya, J. and Lu.X.X. 2004. Application of Remote Sensing in Flood Management with Special Reference to Monsoon Asia: A Review. *Natural Hazards* **33:** 283–301.

[31] Jeyaseelan, A.T. 1999. Droughts & Floods Assessment and Monitoring Using Remote Sensing and GIS. *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology*: 291-313.

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Study of Some E-Government Activities in South Africa

Surendra Thakur
Enterprise Development Unit
Durban University of Technology
Durban, South Africa
thakur@dut.ac.za

Shawren Singh
School of Computing
University of South Africa
Florida, South Africa
singhs@unisa.ac.za

ABSTRACT

This paper examines the nature and extent of e-government activities in South Africa. E-Government refers to the practice of leveraging ICT to deliver services to employees (G2E), to citizens (G2C), to business (G2B) and to other government departments (G2G). The research method used was through direct personal engagement through semi-structured interviews with 10 stakeholders at local, provincial and national government level. The sample comprised two senior provincial management, two senior municipal managers, two academic government researchers, one national government official and three businessmen connected with government. A literature review was also conducted with data sourced from journals, books, websites and popular media for evidence of e-Government not known to the respondent or a provincial (state) or national initiative. It finds although e-Government activities is being provided by, inter alia, government, citizens and business, the service is few, far between and expensive. Although government has committed to e-services, there are, with few exceptions, such services available. On the other hand citizens are rising in anger as they demand improved service levels. It determines some typically government-centric services have already perhaps out of necessity been usurped by citizen led initiatives. Business is also providing innovative e-services. This comes however with skewed costs that may agitate the digital divide. There is an increasingly pervasive broadband infrastructure from both public and private enterprise that represents an opportunity to roll e-services. The ubiquity of mobile devices and ICT may make-services possible. This is an enticing opportunity, particularly given the increasing and sometimes violent calls for service delivery. To achieve traction the authors argue a stronger advocacy role is needed to promote e-Government activities by all stakeholders namely government, business or civil society. This paper uses the Prossler-Krimmer Model to analyze e-government activities in South Africa including KwaZulu-Natal. It examines the technological, political, legislative and societal(sectors) respectively in South Africa and concludes e-Government is a major opportunities for all stakeholders through a combination or research, cross training, advocacy and a re-contextualization of e-government.

Keywords- Component; e-Government, e-Services, Civil Society, e-Readiness.

African Journal of Computing & ICT Reference Format:
S. Thakur & S. Singh (2013). Study of Some E-Government Activities in South Africa.
Afr J. of Comp & ICTs. Vol 6, No. 2. pp 41-.54

1. INTRODUCTION

The objective of e-Government is to improve service delivery to citizen and in so doing improve the efficiency of the governments activities. The application of Information and Communications Technology (ICT) facilitated by the Internet and web technology is the primary tool to enable the objectives of e-Government. Attention has been given to e-Government since the 1990s when it was realized that the Internet and web technology was being used to transform business processes and practices.

In this paper we look at some aspect of e-Government, the paper will outline: business driven e-Government, government driven e-Government and civil society driven e-Government.

1.1 Background

In recent years there has been growing pressure on government to deliver services in a more efficient accessible manner. These applications have been referred to as e-Government. It is difficult to define e-Government [1, 2] as it encompasses such a broad spectrum of activities. Currently, we have adopted the following working definition of e-Government as suggest by Grant and Chau [3]

"A broad-based transformation initiative, enabled by leveraging the capabilities [of] information and communication technology; to develop and deliver high quality, seamless, and integrated public services; to enable effective constituent relationship management; and to support the economic and social development goals of citizens, businesses, and civil society at local, state, national and international levels "

There appears to be a relatively low rate of awareness and successful use of computer applications or artifacts in the support of government services to the public. An example of improper usage of an e-Government application by stakeholders is the case of the Natural Resource Ministry's Scientific Information System where all decisions made by senior officials and senior committees did not take into? account the information provided by the e-Government system [4].

Customer-orientation has been proposed as a way in which governments have attempted to improve the quality of services offered to business and citizens [5, 6]. There is a debate in the literature as to which is the most appropriate term to use in describing the individuals who obtain the direct benefits from ICT systems in the e-Government environment and the terms customer or citizen are used [7, 8]. The term customer suggests a number of issues, one of which has to do with the organization satisfying a particular need and service. The term customer also suggests the possibility of the person receiving the service having some degree of freedom to choose the service provider. In the majority of e-Government systems the person who utilizes the service does not have any choice.

The post-apartheid South Africa government using a customer-orientation philosophy coined the term Batho Pele. Batho Pele is a Zulu word which means 'people first'. This philosophy requires a transformation in delivery modality that is just, transparent, economic, realistic, reasonable, and transformative [9]

Batho Pele is supported by four pillars, namely:

- Re-engineering and improving the back-office operations of government;
- Re-engineering and improving the front-office operations of government;
- Internal communication; and
- External communication [10].

Further, in light of scarce resources and the global financial crisis, government departments are experiencing increasing fiscal, public and political pressure to increase return of investment from their e-Government applications [11-13]. This could be one reason why citizens and indeed business are setting up their own service related websites and portals.

Complexity in e-Government applications pose a particular challenge [14] to designers and developers of e-Government applications. For example, Lips [15] reports on the Kafkabrigade project in the Netherlands. This projects attempts to tackle, what she refers to as, the 'wicked' problems or excessive administrative burdens from a citizen-centric point of view. The Netherlands national government has launched a website (www.kafkabrigade.nl) where Dutch citizens can post their problems concerning disjointed government departments. Relationships within e-Government structures are complex. Examples here are, government-to-citizen (G2C), citizen-to-government (C2G), government-to-government (G2G), government-to-business (G2B), business-to-government (B2G), government-to-employee (G2E), and employee-to-government (E2G)[16]. Government adopting other countries' e-Government systems poses a problem in implementation. Heeks [17] points out an e-Government system should be country specific and not an off-the-shelf system from another country.

Leadership and key stakeholders involvement affect the success of e-Government applications [18]. E-Government is about transforming government to be more citizen-centered. To achieve e-Government success requires active partnerships between government, citizens and the private sector [19]. There exists a complex set of dynamics with citizens and how they use e-Government services. Heeks [20] reports that citizen contact with government is relatively rare. Millard [21] found in Europe that e-Government citizens use government services on an average of 3.1 times a year compared with non-e-Government citizens who only tend to use government services 1.5 times a year. In Africa there appears to be no reliable statistics on the frequency of usage of e-Government systems by citizens.

Maumbe [22] argue e-Government applications can benefit from a comprehensive internal research program to address specific challenges to that application. Technological complexities, complex social processes and independent human agents according to Moodley [13] are not seriously factored into e-Government applications. Guanghua [23] in a similar light, contends e-Government applications are largely constrained within their social context. From a different perspective, the UK National Audit Office point out that "the ICT profession across both private and public sectors is immature in comparison to traditional professions such as medicine, law or accountancy. There is no core set of recognised qualifications and a very wide variety of entry points into the profession. This has made it harder for those in the senior civil service without ICT experience to understand the full value that the profession can deliver." [24]. The National Health System in the United Kingdom [25] illustrate the impact of the above mentioned problem on the success and failure of an e-Government system.

From the perspective of security, data privacy and data security is a concern with intragovernment and intergovernment data-sharing and data-mining [26, 27]. The recent hacking of several Malaysian e-Government websites [28] highlight the need for data privacy and data security in e-Government systems. In government departments there appear to be internal inefficiencies in terms of ICT and human resources [19]. According to Abrahams [29] e-Government applications face further challenges because of the fragmented nature of government administration and its communications processes. Emery [30] and Maumbe [27] points out e-Government models are primarily developed within a western context. There appears to be a need for a more context-centric e-Government model.

Another challenge that e-Government applications face is that of adoption, getting the new e-Government system implemented and used by the target audience [31]. Further, e-Government solutions are not a 'One Size fits All' solution [32]. Each application must take into account the variable factors influencing that application. Abrahams [29] asserts a significant number of e-Government initiatives are information based and only a few include interactive content. This is confirmed by Naidoo [33] that, for example, the South African government is predominantly concerned with efficiency improvement through information dissemination, rather than adopting a more comprehensive e-Government system. e-Government applications are also affected by digital divide issues [5, 34, 35].

Yet another challenge faced by e-Government applications is the use of inappropriate project management techniques [36]. Kaisara and Pather [37] contend even a well designed and developed e-Government systems has a latent defect. If policy-makers do not incorporate access and adoption strategies which take into account the environment that the e-Government system is designed in, the project will be faced with challenges. Further, a private sector solution cannot be imposed as an e-Government solution [17]. The National Data Bank (NDB) project was planned to provide a broad range of data and information support to users inside and outside Bangladesh [38]. In attempting to satisfy multiple levels of stakeholders in different sectors the project did not meet user expectations.

2. THE SOUTH AFRICAN E-GOVERMENT CHALLENGE

From an African perspective, Mutula [39] points out that providing an accurate status of e-Government applications in sub-Saharan Africa is not easy because major studies on e-Government do not fully cover all countries in the region. Specifically focusing on South Africa, the Internet user base is growing year by year at a very slow rate [40].

From this slow growth in the Internet user base we can conclude that there are several challenges facing the growth of the Internet in South Africa, despite South Africa being ranked fourth on the continent for Internet usage [41].

South African e-Government applications face various other challenges. Moodley [42] claims the exclusive emphasis being placed on ICT projects, "at the expense of careful analysis and consideration of the broader economic, social, and political elements that interact to improve the lives of individuals." Moodley [42] and Singh [43] further claim there are significant gaps in policy, understandings of ICT, understanding social process, bureaucratic incompetence and understanding "independent human agency". Put another way, if there is a problem in a system, the introduction of ICT will automate the problem.

Following on Agarwal's [9] evaluation of international e-Government applications, Moodley reiterates [42] South African ICT government programs are adopted and implemented in a top-down approach. This approach tends not to focus on the needs of the citizens. Singh [43] argues that ICT is a tool to achieve a particular desired objective, it is not the objective. This misunderstanding affects e-Government applications and may create further understanding gaps.

South African e-Government applications suffer further challenges. Some of these deal with the different dimensions of leadership, fragmentation of e-Government projects, the perceived value of e-Government projects, stakeholder involvement and project conflicts [44]. van Jaarsveldt [45] argues the next wave of South African e-Government projects should move away from pure information dissemination and move towards interactive service delivery projects. South Africa has to become more proactive rather than reactive to the development of e-Government applications. She further argues the South Africa government needs to consider the cost that is involved in further ICT developments; the sustainability of the application; the efficient utilisation of ICT goods and services especially in rural areas.

We have analyzed several United Nations e-Government surveys [46-50] with special focus on South Africa. In Table 1, we present South Africa's e-Government development ranking. This ranking encapsulates South Africa's performance in the field of e-Government.

TABLE 1.
South Africa's e-Government developments

Year	Ranked	Trend
2003	45	
2004	55	↓
2005	58	↓
2008	61	↓
2010	97	↓

The overall trend is that South Africa is falling lower in the ranking. The development ranking is composed of three subcomponents:

- Online service - this is the number of e-services available on a government website,
- Infrastructure - this component is a summary of the ICT infrastructure of the country, and
- Human capital - this is an index of the literacy of the citizens in the country.

In Table 2 there is a closer analysis of these subcomponents.

TABLE 2
South Africa's e-Government developments

Years	2003	2004	2005	2008	2010
Rank	45	55	58	61	97
Online Services	0.539	↓ 0.515	↑ 0.5692	↓ 0.5518	↓ 0.3079
Infrastructure	0.126	↓ 0.125	↓ 0.1234	↑ 0.1752	↓ 0.0476
Human Capital	0.88	↓ 0.83	↔ 0.83	↓ 0.8061	↑ 0.8432

Both online services and infrastructure are on a downward trend and human capital seems to be marginally increasing. Zuboff [51] indicated information systems could be used to automate, informate or tranformate processes and these concepts has influenced thinking in government [52]. From empirical evidence Zuboff's first notion of automate is well established in government, where computers have been performing routine tasks, such as updating records starting with the Bureau of the Census in the United States [53].

The idea of ICT for the purpose of providing information for managers has also been established in government for some time and applications such as those employed in the South African Revenue Service have been in use for the past 20 years. Data mining is used in the South African Revenue Service to identify anomalies in tax patterns and thereby identify individuals that are not fully complying with tax legislation in South Africa.

There is frequent yearly public government commitment by successive South African presidents to provide e-Government services in their state of the nation addresses [54]. This with few exceptions has simply not been adequately/met [55]. The notable exceptions are the national Department of Home Affairs projects which handle, inter alia, Identification and Passport application notification processes; and the successful, more substantive national South African Revenue Service (SARS) e-filing service**.**

It could be that government promises are propelled by citizens who appear to engage in violent, fatal protest demanding improved service delivery [56]. Examples here are Durban (July 2012); Umlazi (2012);Cape Town (2011); and Fiksberg (April 2011), during which civil activist Andries Tatene tragically died [56].
This paper, however, does not suggest that any e-Government initiative will be the panacea to these service delivery demands. It does appear that certainly more e-government applications are possible.

3. RESEARCH METHODOLOGY

The purpose of this exploratory research is to understand the nature and extent of e-Government activities in South Africa from the perspectives of business driven e-Government initiatives, government driven e-Government and civil society driven e-Government programs.

When undertaking research, it is essential to make use of a structured research methodology to ensure the research has integrity (i.e. that it is reliable, valid and can be "reproduced") [57]. According to Leedy and Ormrod [58], a research methodology is an operational framework in which the facts are placed so that their meaning may be seen more clearly. The region of interest for this study was South Africa. The parameter of interest is nature, perceived extent and perceived usage of e-Government artifacts[1].

Some data was collected using direct personal engagement through semi-structured interviews with 10 stakeholders at local, provincial and national government level. The respondents comprised two senior provincial

[1]The words artefact is deliberately used over applications due to the convergent nature of e-government.

management, two senior municipal managers, two senior government academic researchers, One national researcher and three (businessmen who are involved in e-Government activities. A search was undertaken of the municipal websites for evidence of e-Government activates. A literature review was also conducted with data sourced from journals, books, websites and trade journals for evidence of e-Government activates at provincial (state) or national initiative.

3.1 Limitations of the study

This research is exploratory and therefore it is claimed the findings are merely indicative .It is not the intention of exploratory research of this nature to be generalized.

4. BUSINESS AND E-SERVICES

The most publicized impact of G2B has been the implementation of the SARS tax revenue management system. The system has high adoption from business. This is a vote of confidence as it demonstrates perceived value. In other areas business has been creating and simultaneously leveraging e-infrastructure to introduce e-services such as e-commerce and online interactive services. For example the KZN branch of Business against Crime, has leveraged mesh-network technologies in local geographical areas such as central town blocks to alerts participants of imminent danger or challenges [59]. There is also an active and significant private armed response unit, that responds to connected alarm services and or silent activation [60].

South Africa has a well-developed banking infrastructure, with 2,698 bank branches, 8,785 automatic teller machines (ATMs) and no fewer than 109,454 point-of-sale (POS) devices [61]. South Africa has a third of all the ATMs in Africa [61]. South Africa has the most penetrative ICT infrastructure in Africa with a third of the continent installed ATM's. However, its e-services are amongst the most expensive in the world with a survey placing it most costly out of 27 countries surveyed. The fact that 18 of the 27 countries surveyed offer this as a free service implies this should perhaps also be considered in South Africa [62].

For e-commerce to grow there must be very advanced physical, secure, always-on encrypted e-infrastructure. The irony is this always-on service is provided at a premium [61]. This penalizes the poor who typically make very frequent low value withdrawals [61]. This penalty is both geographic and economic as rural folk travel distances to access an ATM. This infrastructure represents an opportunity to increase access and create e-services.

On the other hand, South Africa's communication telephone call rates mobile rates rank the country 32nd most-expensive out of 42 African countries surveyed [63]. This has not discouraged usage by the poor and may be viewed as further economic disempowerment. In other spheres business has (possibly opportunistically) been offering typical government-centric services such as private health, private education, private tolled roads and security. This service comes at a price and mitigates this class divide where the rich can afford "premium services". The adoption and growth of these offerings demonstrates that business will engage where there is a business case for a particular activity; e-Government will certainly be no exception.

This discussion implies there is sufficient infrastructure *in-place* for increased B2C, G2G, G2C and G2B activities. This opportunity may be leveraged to drive creation of artifacts (supply) and increase usage (demand). This is an important opportunity which will require engagement and analysis. This section shows that the e-services exist and that there is an opportunity to leverage the existing physical infrastructure and skills, created by business to produce e-services which will drive adoption.

4.1 Some e-government initiatives in at provincial level

KwaZulu-Natal launched an e-Government initiative in 2011 called KznOnLine. It is perceived to be largely static. The website is not interactive only providing digital content at present. Nevertheless this is a necessary, though not necessarily sufficient step towards an interactive Gov 2.0 interface. In general, there is little evidence of Gov 2.0 engagement, on this or any other site. The site displayed a hotline number and two Social Media links namely:a FaceBook link (www.FaceBook.com/KZNONLINE) as of July 2012and a Twitter link @KznGov as of October 2011. There were 30 Facebook posts, 33 tweets with 20 registered followers as at July 31 2012. The Online portal does not display a traffic counter, which makes number-of-hits difficult.

The Provincial Department of Economic Development also launched a website to facilitate job creation http://www.kznunemployedgrads.gov.za/. This site had two socialmedia links and a Facebook link:http://www.facebook.com/pages/KZN-Unemployed-Graduates-Database/187978511243616 and a Twitter link @kznugrads. The FaceBook link is unfortunate and difficult to remember although it did have four posts and 204 likes on 5 August 2012 while the Twitter account had seven Tweets and 33 followers.

This low activity on both websites probably does not have to do with the lack of awareness as no public advertisement have been read by the authors in the media in the last year advertising the social media links nor has any government official informed the authors of the site. One of 10 respondents is aware of the website but did not know its location or its content. Current municipal website may be characterized as informate systems. There are ongoing efforts to create municipal systems in KZNwith mixed success such as the eThekwini Revenue Management System (RMS). RMS also known as LogoSoft project (an acronym for Local Government Software) was touted to be the best municipal program. eThekwini created processes to market RMS to developing countries *before* it was even developed several years ago. The project started in 2003with a projected cost of R250m [64]. The system has also missed several deadlines. The current costs are R474mandcity manager S'bu Sithole has stated the municipality will enlist the services of an independent risk assurer to test the controversial system as he cannot give a guarantee that the new billing system is watertight [65].

The Johannesburg City Council is experiencing similar challenges with both business and citizens engaging in legal action over perceived poor services such as inconsistent bills and disrupted services [65]. In a landmark hearing, government was given 14 days to deliver textbooks to 6000 schools in Limpopo province six months after the academic year started. They did not meet the deadline [92]. It is not unreasonable to assume that part of this failure relates to poor Government systems. This type of failures is not a South African aberration as Dada (2006) informs: "Specifically dealing with e-Government systems, there is an emerging body of research dealing with information systems failure." This is vindication of Bannister who in Jenner [66] asserted "Alas, the history of government computing is dotted with what, in the worst case can only be described as computer disasters, and in the best as poor investments". A conjecture, strongly supported by the authors is a lack of strong technically competent staff with strategic vision being appointed to run ICT or drive such e-Government projects.

Newcastle in KZN purchased 32 iPads for its councilors. The intent is to move towards 'paperless' council meetings. The council calculated the investment will repay itself in 24 months with the savings from photocopying and reprinting expenses. The councilors are involved in ongoing training [93]. The mayor also reflected that digital data will make information exaction and prevent simple (and sometimes) costly errors arising from illegible duplicate copies and or transcribing errors. This suggests an emerging green or carbon friendly value to e-Government solutions. Ladysmith also purchased 53 Notebooks for city councilors.

The aim here was to reduce the cost of paper for meetings calculated at R2m per annum [55]. Although Moodley [40] may view these as top-down projects, the benefits of inculcating an e-culture, the resultant economic cost and paper saving should make this project popular. The metro of eThekwini has embraced digital communication. This has resulted in the creation of a call center that handles over 300,000 calls per month with a range of municipal services such as citizens reporting water and light faults, handling queries for rates and other municipal services, waste removal and tree-felling. The success of this innovation has resulted in this solution being sold to other cities in South Africa by the company Analog [67].

The State President also created a call center or hotline for citizens to log challenges. The launch led to a flood of calls, stimulating a denial-of-service attack. This represents empirical evidence of the scope and scale of frustrations experienced at a national level in South Africa. At a strategic level the province of KZN, proactively commissioned a research entitled "KwaZulu-Natal e-Government Roadmap." This is being circulated to a limited list for comment, before general release.
This section suggests that e-Government can assist with e-transparency reduce errors.

4.3 Civil society's initiatives to govern

The ubiquity of mobile telephony has displaced several government functions with citizen-led initiatives leveraging the web and social media to self-manage civic societal functions. At an international level, social media tools have been used for perceived positive crowd sourcing and perceived negative crowd sourcing. The Arab Spring in North Africa is a current example of a perceived positive case where ICT convergence and social media was used for information sharing to effect regime change [68].On the other hand, the 2011riots in the UK, was an example where riot sourcing was used for destructive criminal use.

A further case is the Ushahidi[2] website phenomenon in Kenya [69]. Here the website was used as a crowd sourcing source needed to assist during the carnage that took place in the 2007 elections [69]. A legacy of this innovative intervention is that the Kenyans recognise the strength of Ushahidi and now (re)use it to monitor government projects for success, failure and evidence of collusion or corruption. This introduced a form of e-transparency into the societal framework [69].

[2]Ushahidi means testimony

A negative consequence of social media communication services was the insidious use of Facebook and Twitter for *riot-sourcing* or the sourcing of fellow like-minded criminals in London in 2011 [70]. In order to avoid early police intervention, the location of the gathering of criminals was covertly planned by use of BlackBerry's Private Peer-to-Peer Messenger Service quaintly called BBM. eBlock is a South African community project striving to fight crime by providing the community with information about crime and to empower its members by using Geo-referencing [71]. The eBlockWatch incident map shows all crime related incidents reported by our members. The incidents are organised by geographical location used to see what is happening in your neighborhood [71].

In one South African incident, a carjacking hostage, stuck in the boot of his car, used Twitter to inform his girlfriend of his predicament. She used Twitter and posted the message "Be on the look for DSS041GP my boyufriend (sic) has just been hijacked and is in the boot please RT." The RT standing for ReTweet. This message went viral and the hostage was saved two hours later[72, 73]. Pigspotter, demonstrated the power of positive crowd-sourcing when it also joined and coordinated the hunt to find a carjacked hostage using its twitter account: http://twitter.com/pigspotter[72, 73].

There is anecdotal evidence from senior municipal officers to support the view that local rural folk in Ladysmith use SMS to alert fellow countrymen of closed roads, flooded rivers, veld fires and other impending disasters [55]. The authors suggest an initiative where emergency tweets or SMSs are broadcast free. This will expedite adoption and consequent usage, saving lives and time and return government to the moral high ground in a pragmatic fashion.

SMS and social media may also be used for less ethical reasons where for example the selfsame Pigspotter also claims that it is 'not about defeating the ends of justice' informs folk of where police are setting roadblock [72, 73], because 'police are corrupt as well'. On the other hand Social Media provides rich hunting grounds for perverse folk to prey on naïve youth. This has already happened when a bright young 14 year girl was persuaded to 'elope' by an older manipulator, who pretended to be younger. He carefully nurtured the young lady on a social network over an extended period of time to persuade the youngster to leave home. When saved she was relieved [74].

This civic-inventiveness has little corresponding match at local indeed any government level, even though there is increasing evidence of efforts to start this process. The above shows the power of communication (perceived positive and perceived negative) and shows how with creativity convergence can be positively leveraged to provide services in emergencies. The next section points an important counter-example.

5. TOWARDS AN INCREASED AND IMPROVED E-GOVERNMENT PARADIGM

It is a reasonable conjecture e-Government usage and availability particularly in developing countries is not at desired level. South Africa is one of the most unequal societies in the world, with the gap between rich and poor increasing. e-Government represents an opportunity to leverage the infrastructure to develop artifacts that assist the marginalized by providing more accessible local services that save citizens transport time and costs. Here a simple entry level e-Government option will be to automate, highly interactive, time consuming, routine, non-contested services such as vehicle license queries and renewals; rate queries and so on. The benefit is compounded as the freed resources may be used for other local government services.

The ubiquity of mobiles provides a new opportunity. Many users experience the internet in a very different manner through this device. Cambell and Im[75]point e-Government was conceptualized when 'mobiles were not in existence.' The lack of fitness-for-purpose of mobiles for e-Government mitigates the creation and adoption of e-Government services, while paradoxically its ubiquity militate adoption. It may be that we need a new m-government strategy to leverage this opportunity. Intriguingly Cambell and Im[75] predict the majority of users will experience Internet for the first time on a mobile device. Consequently researchers leapfrog the e-Government argument and now support the Web 2.0 technology to the extent that we now have an established field known as Gov. 2.0. O' Reilly [76] refer to Web 2.0 as 'a platform that extends to all connected devices.' He further asserts that this is a mechanism for social cohesion and cooperation.

The authors have found little direct literature or documentation alluding to direct spend of ICT for e-Government projects in South Africa. On the other hand, an analysis of local government Integrated Development Plans (IDPs) suggest that many municipalities are increasingly embarking on projects that leverage ICT for the provision of services. This represents a considerable opportunity; given this papers conjecture that e-Government is not well understood. The e-readiness of South Africa, in general, and KZN in particular to deliver e-Government is considered in the next section.

5.1 e-Readiness of South African local government

Prosser and Krimmer [77] proposed a model to evaluate a nation's readiness to implement electronic voting (e-voting) (Fig 1.), which is a special form of participation to enable democracy. Thakur [78] used this model to evaluate e-democracy as a possibility for South Africa. The dimensions of this model are: technological, legislative, societal and political. The model affirms that many criteria must simultaneously be satisfied in order to implement e-voting.

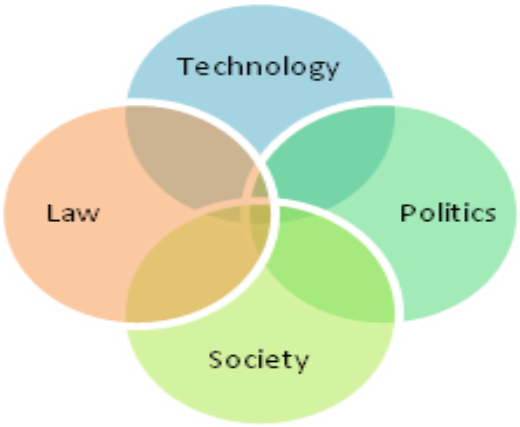


Figure 1: The Prosser-Krimmer e-Democracy Model [77]

This model is now extended to evaluate the e-readiness of South Africa for e-government. This model is analytically used to briefly consider whether and how e-Government may be driven.

5.1.1 Society

One reason why services are not being developed or perhaps adopted could be a general lack of knowledge and awareness of e-Government possibilities and activities. The authors have been involved in the creation of an e-Government course to extend the knowledge of and influence the usage of e-government. They have been surprised that just one of the 11 respondents all involved in government was aware of the KZN Online Portal. The fact that we have competing and sometimes conflicting terms such as Gov 2.0, e-Governance, e-Government, m-Gov, Web 2.0 and Municipality in the Cloud appear to further obscure the issue[3].

To realise the potential of e-Government certain requisite technical and programming skills are needed to develop software artifacts. Skilled graduates are available given the high number of graduates, many of whom are unemployed and could be cross trained in ICT to produce artifacts. This represents an opportunity to create employment. Further, 'a build it and they may use it' principal is advocated to create the services which may drive demand. When one considers the SARS and Home Affairs examples it is clear that creating perceived useful artifacts will drive adoption and that strong advocacy through sustained advertising will sustain and may grow adoption.

This section suggests that South Africa needs a strategy to create awareness at all levels and that KZN has the people or? warmware to create artifacts with some intervention such as cross training.

5.1.2 Technological role

A strong supply of relevant artifacts is necessary though not sufficient to create demand. Once again one looks at *eBlockWatch*, SARS e-filing and PigSpotter as projects that garner support through high availability, up-to-date-relevance and advertisement. An e-Government ecosystem requires appropriate ICT infrastructure to provide location independent access to artifacts that provide services. It is already noted that South Africa has the best ICT infrastructure in Africa. There is empirical evidence the digital divide gap is diminishing as government is steadily increasing the number of access points in rural South Africa. This is reflected in Table 3. In KZN for example the Department of Rural Development installed 64 access points or i-centres between 2010/11 and 2011/12 and intends to deploy a further 70 i-centre in KZN. The Department of Economic Development installed 15 telecentres or inforcentres and wants to install a further 300. The Universal Services Agency (USAASA) installed 30 telecentres and intends to install 300 more [Table 3]. The table does not include public private partnerships (PPP) that resulted in access for rural or the unemployed which will probably grow the number of access points. From the data in Table 3it is reasonable to assume the increase in government investment in ICT infrastructure would contract the digital divide.

TABLE 3. Government Commitment to ICT Infrastructure

Year	Department of Rural Development [79]	Department of Economic Development [80]	USAASA [81]
2010/11	37	5	100
2011/12	27	10	44
2012/13	35	150	200
2013/14	35	150	150

The National Department of Communication has through a grant established an e-Skills Hub in KZN to advocate e-government through interventions such as skills upliftment, research and development, which an author is part of. [3]Some of these terms are deliberately stated for completeness but not discussed in this paper.

At a national level, there is robust activity in rolling broadband across the country. The national coverage is at about 60% and growing with a steady decline in access cost. In addition the ubiquity of mobile devices provides instant mobile access and connectivity. This implies that the issue of access is becoming less of a challenge.

5.1.3 Legislative role

It is possible the legislative front may be used to create an e-Government act like the USA (2002) or the Austria (2005) acts respectively. This will set the terms of reference for future government projects to include an e-Government component. This is a good intervention that will reap medium term benefits.

There are several legislative acts that stakeholders who engage with citizens using ICT should be aware of. These are the Electronic Communication Act (ECA), the Promotion of Access to Information Act (PAIA)and the Independent Communications Authority of South Africa Act (ICASA) [82]. These are listed for completeness but not discussed further.

5.1.4 Political Role

The role of national, regional and local government is central to the promotion, creation, adoption and sustained usage of e-government. KZN has launched a portal, created social media links and developed a roadmap which shows government commitment to the process. The advocacy roll has been mooted. It is reasonable that the creation of an e-Government center will nurture an e-Government awareness and culture. However any innovation or 'disruptive change' needs strong strategic and political support by an appropriately influential champion [83]. In this case it is fitting the e-Government portal is being driven by the Provincial Premiers office. However advocacy without appropriate technology is doomed. This shows the intertwined multi-faceted nature of the model.

However current ICT initiatives are not geared towards *e-Government* – they are geared towards *access*. For this access points to deliver e-government, we argue that a national policy and champions to drive the advocacy needed. A mistaken belief by officials is that e-Government is the availability of speeches, policy and legislation on accessible media such as website. This is important for information dissemination and e-transparency and may even represent the first step towards an e-Government culture. This creates a static website that is not interactive. There is no white paper or policy for e-Government in South Africa at a national level. KZN is a provincial leader to drive and address this gap. At a national level, however, South Africa is evaluating e-democracy with an international study of electronic voting to neutrally evaluate the possibility of e-participation [78].

We leveraged the Prossler-Krimmer framework, to show there is ICT infrastructure (hardware) and IT savvy graduates (warmware) who can be cross trained to produce e-Government artifacts (software).

This section proposes that we need strategic and operational awareness, functionally technically competent people and relevant artifacts to drive demand.

6. OPPORTUNITIES AND FURTHER WORK

At an artifact level there is a developmental opportunity that may well assist in job creation through the cross training of underemployed or unemployed artisans. The benefit is compounded as freed resources may be used for other local government services. The opportunity to automate a service is an opportunity to evaluate how a service is delivered and what functionality may be added/removed to optimize such a service. Adding a computer to a problem with automate the problem not fix the problem.

The perceptions of local senior municipality officials understanding of e-Government has not really been tested in any rigorous method to the knowledge of the authors. Olugbara [84] is undertaking a large scale analysis in South Africa. If there was such knowledge, we argue there will be considerably better e-services artifacts or directed efforts to produce such software. An analysis of local government Integrated Development Plans (IDPs) suggest many ICT projects may even be e-Government projects. This represents considerable scope for further systemic research and development.

At a cursory artifact level the following services are suggested. This is just a small sample. One, geo-referencing or tagging of potholes for speed address– these may even be sent to navigation software to assist drivers and prevent accidents. Two, the broadcast of particular local context sensitive information for example seasonal health alerts in particular geographic areas with malaria or TB. Three, a more just-in-time feature may be about a rabid dog loose in a particular area. The latter is a kind of highly localised 'breaking news' or 'news alert' feature.

7. CONCLUSION

This preliminary assessment of the e-Government activities in South Africa indicates while there is a digital divide, the ubiquity of the mobile represents an opportunity to be leveraged within a so called m-government framework to deliver e-government. This form of engagement dovetails with the availability of such devices through the youth increasing dependence and usage of selfsame.

The development of two portals, while not currently heavily used is a welcome nonverbal commitment towards e-Government. On the other hand, low uptake of these sites, validates some of the assertions that advocacy and perceived usefulness will drive usage. The establishment of an e-skills hub in KZN with a focus on e-Government will assist in both the advocacy pull and the technology and artifact push. Further the commissioning of the e-Government KZN roadmap is evidence of government taking a pragmatic route towards this genre of service delivery.

The high number of static websites presents both an opportunity and a platform to create interactive services. Web 2.0 represents an opportunity for the developing municipalities to leapfrog e-Government into Gov 2.0.
Both Newcastle and Ladysmith are examples of good interventions, although top-down. This cultural shift may yet lead to good practice. However, the efficiency and effectiveness of these technological interventions should be methodologically assessed for their impact for translation to other environments. Municipalities should, in addition, develop bottom-up simple entry level projects to automate low level and non-contested civic activities. It is further recommended that the advocacy component be practiced inside-out to get staff geared towards e-government by educating them about business-led and citizen-led e-government activities. Can such outside-in innovations be leveraged in a developmentally?

An overarching national policy that informs strategy, promoted by champions, legislated by a legal framework demand ICT-enabled services that leverages human capital is the e-Government's quintessential quadruple.
There appears to exist a need for a managed approach to the development of e-Government artifacts between business driven initiatives, government initiatives and civil societies initiatives. This may be in the form of production of technology such as Application Programmer Interface (API) for artifacts, facilitated engagement through advocacy, and research and development colloquiums. Finally this paper gravely points to the increasingly violent service delivery unrest. For South Africa to succeed and grow peacefully, e-Government potential must translate to credible, repeatedly reusable, accessible artifacts.

REFERENCES

[1] S. Kim and D. Kim, "South Korean Public Officials' Perception of Values, Failure, and Consequences of Failure in E-Government Leadership," Public Performance & Management Review, vol. 26, pp. 360-375, 2003.

[2] L. A. Joia, "Building Government-to-Government Enterprise," in Electronic Government: Concepts, Methodologies, Tools, and Applications, A. Anttiroiko, Ed. Hershey: Information Science Reference, 2008, pp. 1761-1768

[3] G. Grant and D. Chau, "Developing a Generic Framework for E-Government," Journal of Global Information Management, vol. 13, pp. 1-30, 2005.

[4] R. Heeks,2002,Success/Failure Case Study No.3: Problems for a Natural Resource Ministry's Scientific Information System,http://www.egov4dev.org/success/case/fishsis.shtml,2011,25 July

[5] N. Helbig, J. R. Gil-García, and E. Ferro, "Understanding the complexity of electronic government: Implications from the digital divide literature," Government Information Quarterly, vol. 26, pp. 89-97, 2009.

[6] F. Bannister, "Serving the Citizen: A proposed Model for IT value in Public Administration," Southern African Business Review, vol. 4, pp. 33-44, 2000.

[7] H. Mintzberg, "Managing Government, Governing Management," Harvard Business Review, 1996.

[8] B. Brewer, "Citizen or customer? Complaints handling in the public sector," International Review of Administrative, vol. 73, pp. 549-556, 2007.

[9] M. A. Mofolo and W. Smith, "Making use of 'Batho Pele' principles to improve service-delivery in municipalities," Journal of Contemporary Management, pp. 430 - 440, 2009.

[10] Anon, "TheBatho Pele Handbook.A Service Delivery Improvement Guide," Department of Public Service and Administration, 2003

[11] M. P. Gupta, J. Bhattacharya, and A. Agarwal, "Evaluating e-government," in e-Governance: Case Studies, A. Agarwal, Ed. Hyderabad: Universities Press, 2007, pp. 1-56

[12] R. Naidoo and W. Palk, "Are e-Government Investments Delivering Against Expected Payoffs? Evidence from the United Kingdom and South Africa," presented at IST-Africa 2010 Conference, Durban, 2010

[13] S. Moodley, "Deconstructing the South African Government's Information and Communication Technologies for Development Discourse," Africa Insight, vol. 35, pp. 3-12, 2005.

[14] S. Purao and K. Desouza, "Looking for Clues to Failures in Large-Scale, Public Sector Projects: A Case Study," presented at 44th Hawaii International Conference on System Sciences, Hawaii, 2011

[15] M. Lips, "Before, After or During the Reforms? Towards Information-Age Government in New Zealand," Policy Quarterly, vol. 4, pp. 21-26, 2008.

[16] A.-V. Annttiroiko, "A Brief Introduction to the Field of E-Government," in Electronic Government:Concepts, Methodologies, Tools, and Applications, vol. 1, M. Khosrow-Pour, Ed. Hershey: Information Science Reference, 2008, pp. xii-lxxv

[17] R. Heeks,2003,Most eGovernment-for-Development Projects Fail: How Can Risks be Reduced?,http://idpm.man.ac.uk/publications/wp/igov/index.shtml,2 May,2011

[18] S. C. Y. Luk, "The impact of leadership and stakeholders on the success/failure of e-government service: Using the case study of e-stamping service in Hong Kong," Government Information Quarterly, vol. 26, pp. 594-604, 2009.

[19] M. Farelo and C. Morris, "The Status of E-government in South Africa," presented at IST Africa Conference, Pretoria, South Africa, 2006.http://hdl.handle.net/10204/966

[20] R. Heeks,2006,Benchmarking eGovernment: Improving the National and International Measurement, Evaluation and Comparison of eGovernment,http://www.sed.manchester.ac.uk/idpm/publications/wp/igov/index.htm,2 May,2011

[21] J. Millard, "eGovernance and eParticipation: lessons from Europe in promoting inclusion and empowerment," presented at United Nations (UNDESA) multi-stakeholder advisory group workshop "eParticipation and eGovernment: understanding the present and creating the future", part of the "International eParticipation and Local Democracy Symposium: promoting social inclusion via eParticipation", Budapest, Hungary, 2006

[22] B. M. Maumbe, W. J. Taylor, H. Wesso, and G. Erwin, "E-Value Creation in a Government Web Portal in South Africa," in Encyclopaedia of Portal Technologies and Applications, A. Tatnall, Ed. Hershey: IGI Global, 2007, pp. 384-390

[23] L. Guanghua, "E-Government, People And Social Change: A Case Study In China," The Electronic Journal on Information Systems in Developing Countries, vol. 38, pp. 1-23, 2009.

[24] National Audit Office, "Cross-government: Information and Communications Technology in government Landscape Review," Comptroller and auditor General, London 14 February 2011

[25] R. Brooks, "System failure!," in Private Eye, 2007, pp. 17-24

[26] J. Fedorowicz, J. L. Gogan, and M. J. Culnan, "Barriers to Interorganizational Information Sharing in e-Government: A Stakeholder Analysis," The Information Society, vol. 26, pp. 315-329, 2010.

[27] B. M. Maumbe, V. Owei, and H. Alexander, "Questioning the pace and pathway of e-government development in Africa: A case study of South Africa's Cape Gateway project," Government Information Quarterly, vol. 25, pp. 757-777, 2008.

[28] J. Ribeiro,2011,Many Malaysian Government Sites Hacked,http://www.pcworld.com/businesscenter/article/230421/many_malaysian_government_sites_hacked.html,2011,24 July

[29] L. Abrahams, "E-Governance Policy 1999-2009: Paths And Limitations To Progress," Journal of Public Administration, vol. 44, pp. 1011-1026, 2009.

[30] L. Emery, A. Gordon, G. Krieshok, and A. Pompe,2009,Information Policy Country Report: South Africa,http://hdl.handle.net/2027.42/64841,22 July,2011

[31] R. Heeks and R. Santos,2009,Understanding Adoption of e-Government: Principals, Agents and Institutional Dualism,http://www.sed.manchester.ac.uk/idpm/research/publications/wp/igovernment/index.htm,2 May,2011

[32] A. Y. Al-Eryani, "E-Government Services In Yemen: Success and Failure Factors," in Faculty of Science Bulletin, vol. 22: Sana'a University, 2009, pp. 61-74

[33] G. Naidoo, S. Singh, and N. Levine, "An Overview of Internet Developments and their Impact on E-Government in South Africa," in Handbook of Research on E-Services in the Public Sector: E-Government Strategies and Advancements, A. T. A. Ajeeli and Y. A. L. Al-Bastaki, Eds. Hershey: IGI Gobal, 2011, pp. 63-77

[34] G. B. Guomundsdottir, "Approaching the Digital Divide in South Africa," presented at NETREED Conference, Beitostolen,Norway, 2005

[35] C. G. Hossan, M. W. Habib, and I. Kushchu,2006,Success and Failure Factors for e-Government projects implementation in developing countries: A study on the perception of government officials of Bangladesh,http://www.mgovernment.org/resurces/euromgvo2006/index.html,10 June,2011

[36] D. Sarantis, D. Askounis, and S. Smithson, "Critical Appraisal on Project Management Approaches in e-Government," presented at 7th International Conference on ICT and Knowledge Engineering, Bangkok, 2009

[37] G. Kaisara and S. Pather, "e-Government in South Africa: e-service quality access and adoption factors," presented at 11th Annual Conference On World Wide Web Applications, Port Elizabeth, 2009.(http://www.zaw3.co.za

[38] R. Heeks,2002,Success/Failure Case Study No.24: The National Data Bank Project: An Expensive Lesson for Bangladesh,http://www.egov4dev.org/success/case/ndb.shtml,25 July,2011

[39] S. M. Mutula, "Comparison of sub-Saharan Africa's e-government status with developed and transitional nations," Information Management & Computer Security, vol. 16, pp. 235-250, 2008.

[40] S. Singh, "Towards Understanding e-Government Application Development," in E-Governance Policies & Practices, N. Bhatt and A. Aggarwal, Eds. New Delhi: Excel India Publishers, 2011

[41] Internet World Stats,2011,Africa Top Internet Countries,http://www.internetworldstats.com/stats1.htm,23 July,2011

[42] S. Moodley, "The Promise of E-Development? A Critical Assessment of the State ICT for Poverty Reduction Discourse in South Africa," Perspectives on Global Development and Technology, vol. 4, pp. 1-26, 2005.

[43] S. Singh, "The South African 'Information Society', 1994–2008: Problems with Policy, Legislation, Rhetoric and Implementation," Journal of Southern African Studies, vol. 36, pp. 209-227, 2010.

[44] R. Matavire, W. Chigona, D. Roode, E. Sewchurran, Z. Davids, A. Mukudu, and C. Boamah-Abu, "Challenges of eGovernment Project Inplementation in a South African Context," The Electronic Journal of Information Systems Evaluation, vol. 13, pp. 153-164, 2010.

[45] L. C. van Jaarsveldt, "Advances in ICT beneficial to government and education in a post-apartheid South Africa, Adopting E-Governance," in Adopting E-governance, G. P. Sahu, Ed. New Delhi, India: GIFT publishers, 2008.http://www.iceg.net/2007/books/book3.html

[46] R. Kerby, R. S. Kim, M. Mimicopoulos, J. Rabinovitch, P. Spearing , and K. Yao, "United Nations E-Government Survey 2010: Leveraging e-government at a time of financial and economic crisis," United Nations, New York ST/ESA/PAD/SER.E/131, 2010.http://www.unpan.org/egovkb/global_reports/08report.htm

[47] S. Hafeez, R. Kerby, and J. Roy, "United Nations e-Government Survey 2008: From e-Government to Connected Governance," United Nations, New York ST/ESA/PAD/SER.E/112, 2008.http://www.unpan.org/egovkb/global_reports/08report.htm

[48] S. Hafeez, "UN Global E-government Readiness Report 2005: From E-government to E-inclusion," United Nations, New York UNPAN/2005/14, 2005.http://www.unpan.org/egovkb/global_reports/08report.htm

[49] S. Hafeez, "UN Global E-government Survey 2003," United Nations, New York 2003.http://www.unpan.org/egovkb/global_reports/08report.htm

[50] S. Hafeez, "Un Global E-Government Readiness Report 2004: Towards Access For Opportunity," United Nations, New York UNPAN/2004/11, 2004.http://www.unpan.org/egovkb/global_reports/08report.htm

[51] S. Zuboff, In the age of the smart machine. New York: Basic Books, 1988

[52] K. G. Nygren, ""Monotonized administrators" and "personalized bureaucrats" in the everyday practice of e-government Ideal-typical occupations and processes of closure and stabilization in a Swedish municipality," Transforming Government: People, Process and Policy, vol. 4, pp. 322-337, 2010.

[53] J. Agar, The Government Machine: A Revolutionary History of the Computer. Cambridge: MIT Press, 2003

[54] State of the Nation addresses 2010 2011 2012. www.dpsa.gov.za
[55] G. Reddy,2012,Interview with Ladysmith Municipal Manager,
[56] P. Alexander, "Rebellion of the poor: South Africa's service delivery protests - a preliminary analysis," Review of African Political Economy, vol. 37, pp. 25-40, 2010.
[57] D. Remenyi and B. Williams, "Some Aspects of Methodology for Research in Information Systems," Journal of Information Technology, vol. 10, pp. 191-201, 1995.
[58] P. D. Leedy and J. E. Ormrod, Practical Research: Planning and Design, 9th ed. Boston: Pearson, 2010
[59] Business Against Crime, " Interview with KZN CEO," 2012
[60] J. Nair,2012,Business against crime Interview and discussion of e-business and KZN e-government,
[61] W5,2010,Update on Regulation of Branchless Banking in South Africa,http://www.cgap.org/gm/document-1.9.42404/Updated_Notes_On_Regulating_Branchless_Banking_South_Africa.pdf,5 Agust,2012
[62] E. Pickworth,2012,SA's ATM fees 'highest of 27 countries',http://www.businessday.co.za/articles/Content.aspx?id=174253,15 June,2012
[63] E. Calandro, A. Gillwald, and C. Stork,2012,Africa Prepaid Mobile Price Index 2012: South Africa,www.researchictafrica.net/docs/SA_Mobile_Prepaid_policy_brief.pdf,2012,5 August
[64] O. M. Suthcliff,2005,The value of the RIMS project,
[65] W4,2012,Doubt plagues new Durban billing system,http://www.iolproperty.co.za/roller/news/entry/doubt_plagues_new_durban_billing,5 August,2012
[66] S. Jenner, Realising benefits from Government ICT investment - a fool's errand? Reading. UK.: Academic Books Publishing International Limited, 2009
[67] L. Naidoo,2012, Interview with e-Government company: Analog,
[68] L. Anderson,2011,Demystifying the Arab Spring parsing the differences between Tunisia, Egypt, and Libya,http://www.ssrresourcecentre.org/wp-content/uploads/2011/06/Anderson-Demystifying-the-Arab-Spring.pdf,1 June,2012
[69] Ushahidi Website,2012,Ushahidi: Kenya monitoring function,http://ushahidi.com
[70] S. Makri, E. G. Toms, L. McCay-Peet, and A. Blandford, "Encouraging Serendipity in Interactive Systems: An Introduction." UCL Interaction Centre: University College London, Gower Street, London, 2012
[71] W5,2012,eBlock website,www.eblockwatch.co.za,5 August,2012
[72] PigSpotterwebsite,2012,www.pigspotter.co.za,
[73] PigSpotterTweets,http://twitter.com/pigspotter,
[74] W6,2012,Missing teenager Kaylene Lewis has been found,http://www.jbaynews.com/missing-teenager-kaylene-lewis-has-been-found/,
[75] J. Cambell and T. Im, "Beyond ubiquity: Mobile government theory and practice," presented at 2012 KAPA conference, Soul National University., 2012
[76] http://ssrn.com/abstract=1008839T. Oreilly, "What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," Communications & Strategies, vol. 1, pp. 17, 2005.
[77] A. Prosser and R. Krimmer, "The Dimensions of Electronic Voting Technology, Law, Politics and Society," Electronic Voting in Europe Technology, Law, Politics and Society, 2004.
[78] S. Thakur, "Commissioned research. E-voting: A cross border experience. Unpublished at time of writing," IEC. South Africa 2012
[79] "Comprehensive Rural Development Programme: ICT," Department of rural development and land reform, 2012
[80] L. Nadasen,2012,KZN ICT Summary. Department of Economic Development and Tourism,
[81] K. Mahlangu,2012,Universal Service and Access Agency of South Africa (USAASA). Business Development Officer,
[82] Independent Communications Authority of South Africa Act (ICASA),www.icasa.org.za,
[83] E. Roger, Diffusion of Innovations, 4th ed. New York: The Free Press, 1995
[84] O. O. Olugbara, "An analysis of the extent of e-government services in South Africa - unpublished research," KZN e-government e-skills hub, South Africa 2012

About the Authors

Colin Thakur is currently the Director of Enterprise Development Unit of Durban University of Technology, South Africa. He is also the Coordinator of the KwaZuluNatal e-Skills Hub. He is a seasoned ICT practitioner. He is researching Electronic Voting with a developing world context. He is an International elections observer and has delivered two keynote addresses on this topic.

Shawren Singh is a senior lecturer in the School of Computing at the University of South Africa and a PhD candidate at the University of the Witwatersrand. He is currently conducting research in the e-Government space, with special focus on the design and development of e-Government applications.

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Recursive Shortest Route Algorithm Using Abstract Data Type, Graph

O.E. Oguike
Department of Computer Science
University of Nigeria
Nsukka, Nigeria
osondu.oguike@unn.edu.ng

ABSTRACT

The shortest route algorithm is an important algorithm in the mathematical sciences because it determines the most efficient route between two nodes in a network. Among the variants of the shortest route algorithm that are available, different data structures can be used to define the algorithm. The abstract data type, graph is one of the highly structured data structures or abstract data type that has important applications to mathematical and computer sciences. Mathematically, it can be defined as a collection of two sets, the set of nodes and the set or table/relation of edges. Recursive algorithm is very efficient when implemented, especially when the recursive case has one recursive call, it is also very structured when implemented. This paper presents a recursive design of the Dijkstra's shortest route algorithm, using the operations of abstract data type, graph.

Keywords- Graph, abstract data type, shortest route algorithm, network, dynamic data structure, recursive algorithm.

African Journal of Computing & ICT Reference Format:
O.E. Oguike (2013). Recursive Shortest Route Algorithm Using Abstract Data Type, Graph. Afr J. of Comp & ICTs. Vol 6, No. 2. pp 55-66..

1. INTRODUCTION

The shortest route algorithm is an important algorithm in mathematical sciences and computer science in particular. This is because it determines the most efficient route in communication network, water pipeline network, oil pipeline network, road network etc. Variants of the shortest route algorithm exist, which are, Dijkstra's shortest route algorithm, Floyd's shortest route algorithm etc. Data structures, like array, list and graph can be used to design the shortest route algorithm. Using the graph data structure to design the shortest route algorithm, graph can be considered as adjacency matrix or adjacency list [1]. In this paper, graph will be considered, based on its mathematical definition. Mathematically, graph is defined as a collection of two sets, the sets of nodes and the table or relation of edges [2].

Based on this mathematical definition of graph, the abstract data types, set and table/relation will be used to form the abstract data type, graph. The underlying ADT that will be used for the set is list, while the underlying ADTs that will be used for the table are set and list. This paper will apply the shortest route algorithm to GSM telecommunication network, with satellite distance between the various nodes of the network [6]. Suppose we have a GSM telecommunication network, as shown in figure 1, with appropriate satellite distance between the nodes of the network.

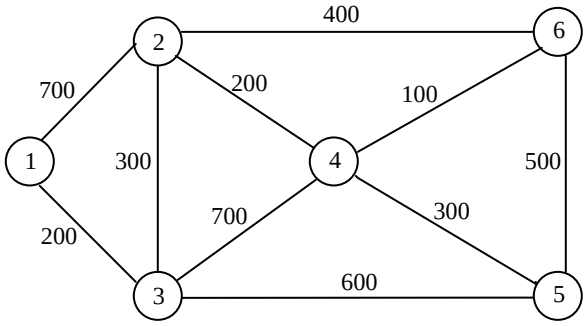


Figure 1. GSM Network with Satellite Distance

The network diagram shown above can be represented as two set, which are the set of nodes, represented as {1, 2, 3, 4, 5, 6}, and the table of edges/triple, which is shown in the table below:

Table 1: Table of Triple (Edges with Distance Weight)

SNode	DNode	Distance	SNode	DNode	Distance
1	2	700	2	6	400
2	1	700	6	2	400
1	3	200	3	5	600
3	1	200	5	3	600
2	3	300	4	6	100
3	2	300	6	4	100
3	4	700	4	5	300
4	3	700	5	4	300
2	4	200	6	5	500
4	2	200	5	6	500

The abstract data type set will be used to represent the set of nodes of the network. The algorithms for the operations of the ADT, set are emptyset, isemptyset, insertset, removeset and memberset. They will be designed using the algorithms for the operations of the underlying ADT, list, which include, emptylist, isemptylist, conslist, headlist and taillist. Furthermore, the abstract data type, table of edges will have the following attributes, source node, destination node and satellite distance. The primary keys of the table/relation will be the source node and the destination node. The algorithms for the operations of the abstract data type, table of edges/triple are emptytabletriple, isemptytabletriple, membertabletriple, locatedistance, removetabletriple and insertabletriple. They will be designed using the following algorithms for the operations of set of triple, emptysetriple, isemptysetriple, insertsetriple, removesetriple and membersetriple.

In a similar manner, the algorithms for the operations of set of triple will be designed using the following algorithms for the operations of the ADT, list of triple, emptylistriple, isemptylistriple, conslistriple, sourcehead, destinationhead, distancehead and tailistriple. The algorithms for the operations of graph are emptygraph, addnode, addedge, isemptygraph, contains, isadjacent, deletenode, deletedge. They will be designed using the algorithms for the operations of the set of nodes and table of edges/triple. Finally, relevant algorithms that will be used in the recursive Dijkstra's shortest route algorithm will be designed using the algorithms for the operations of graph. The hierarchy of the underlying ADTs used to represent graph can be shown in figure 2.

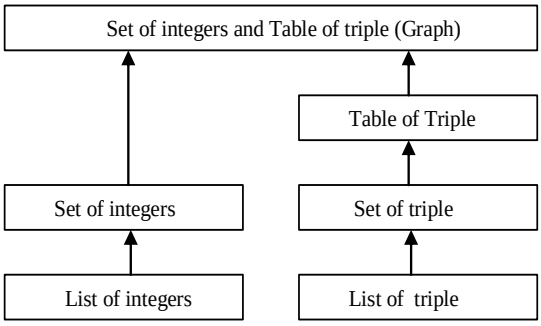


Figure 2. Hierarchy of the various abstract data types

2. SURVEY OF RELATED LITERATURE

According to [4], Abstract data type is "a kind of data abstraction where a type's internal form is hidden behind a set of access functions; values of the type are created and inspected only by calls to the access functions". Applications of graph algorithm have been the focus of some literature, for different configurations of network. In [2], the authors reviewed the use of fuzzy logic in optimal path selection for an ad hoc network The authors in [3] outlined the various applications of Graph Theory to Computer Science. According to them, "the major role of graph theory in computer science is in the development of graph algorithms", which can be used to solve problems that have been modeled using graph, such problem, according to them includes: shortest path algorithm in a network, finding minimal spanning tree, finding graph planarity etc. They also identified some of the languages that can be used to support graph theory concept.

The author in [5] identified other operations of graph, e.g. difference graph, union graph, and intersection graph. These operations, according to him can be used for other applications. Furthermore, the author in [8] presented a version of Dijkstra's shortest route algorithm, which was based on temporary and permanent labeling of the nodes, until, if possible all the nodes of the graph have been labeled permanently. In [1], [7], presentations of some of the abstract data types, like, list, set, relations and graph were made, with relevant applications. The authors in [9] traced the history of graph theory, according to them; Graph theory, as an area of study within mathematics began with a paper published by Leonhard Euler in 1736. According to them, the study of graph algorithm as an area with computing is relatively new. They also presented two ways of presenting graph, which are adjacency matrix and adjacency list. The authors in [10] suggested a modification of the Dijkstra's shortest route algorithm, they suggested a multi parameter Dijkstra's shortest route algorithm that will consider other factors like distance, time and congestion of the edge during the computation of the shortest route.

Furthermore, in [11], the authors used priority queue and C programming language to implement and evaluate the performance of Dijkstra's shortest route algorithm. Comparative analysis of all the algorithms that solve the shortest route problem were made by the authors in [12], while in [13], the authors surveyed the various applications of shortest path routing, which include, transportation GIS, network analysis, operations research, artificial intelligence and robotics.

3. DESIGN OF ALGORITHMS FOR THE OPERATIONS OF ABSTRACT DATA TYPE, SET OF NODES

The algorithms for the various operations of the abstract data type, set of nodes can be designed, using the various algorithms for the operations of the underlying abstract data type, list. First, the design of the algorithms for the operations of the abstract data type, list will be considered. The structure of the abstract data type list can be defined in a java class called mylist as follows:

Int data
Mylist link

The algorithms for the operations of the abstract data type, list can be defined as follows.

Mylist emptylist()
1. Determine emptylist
1.1 emptylist = null
2 Display emptylist

The algorithm, emptylist returns the null list and it takes nothing as parameter.

Boolean isemptylist(mylist l)
1. Read l
2 Determine isemptylist
2.1 IF l = null THEN
2.1.1 Isemptylist = true
ELSE
2.1.2 isemptylist = false
3 Display isemptylist

The algorithm, isemptylist takes a list as parameter and it returns true if the list is empty, otherwise it returns false.

int headlist(mylist l)
1 Read l
2 Determine headlist
2.1 IF isemptylist(l) THEN
2.1.1 Headlist=("No head for empty list")
ELSE
2.1.2 headlist = l.data
3 Display headlist

The algorithm, headlist takes a list of integers (nodes) as paramete. If the list is not empty, it returns the first integer (node) on the left; otherwise, it returns an appropriate error message.

Mylist taillist(mylist l)
1, Read l
2. Determine taillist
2.1 IF isemptylist(l) THEN
2.1.1 taillist = l
ELSE
2.1.2 taillist = l.link
3. Display taillist

The algorithm, taillist takes a list of integers (nodes) as parameter. If the list is not empty, it removes the integer number on the left hand side of the list, and returns the remaining list; otherwise, it returns the given list.

Mylist conslist(int x, mylist l)
1 Read data
1.1 Read x
1.2 Read l
2 Determine conslist
2.1 Declare and allocate temp of the type mylist
2.2 Temp.data = x
2.3 Temp.link = l
2.4 Conslist = temp
3 Display conslist

The algorithm, conslist takes two parameters, an integer and a list of integers. The algorithm makes the integer to be part of list of integers. Having designed the algorithms for the operations of the list abstract data type, these algorithms will be used to design the algorithms for the operations of the abstract data type, set. The following are the algorithms for the operations of the abstract data type, set.

Myset emptyset()
1 emptyset = emptylist()
2 Display emptyset

The algorithm, emptyset takes nothing as parameter and it returns an emptyset.

Boolean isemptyset(myset s)
1 Read s
2 Determine isemptyset
2.1 isemptyset = isemptylist(s)
3 Display isemptyset

The algorithm, isemptyset takes a set as parameter, and it returns true if the set is empty, otherwise it returns false.

Boolean memberset(integer x, myset s)
1 Read data
1.1 Read x
1.2 Read s
2 Determine memberset
2.1 IF isemptyset(s) THEN
2.1.1 Memberset = false
ELSE
2.1.2 IF (x = headlist(s)) THEN
2.1.2.1 memberset = true
ELSE
2.1.2.2 memberset = memberset(x, taillist(s))
3 Display memberset

The algorithm, memberset takes an integer and a set of integers as parameters and it returns true if the integer is a member of the set of integers, otherwise, it returns false.

```
Myset insertset(int x, myset s)
1    Read data
     1.1    Read x
     1.2    Read s
2    Determine insertset
     2.1    IF memberset(x, s) THEN
            2.1.1    insertset = s
            ELSE
            2.1.2    insertset = conslist(x,  s)
3    Display insertset
```

The algorithm, insertset takes an integer and a set as parameters and it inserts the integer into the set provided that the integer is not a member of the set.

```
Myset removeset(int x, myset s)
1    Read data
     1.1    Read x
     1.2    Read s
2    Determine removeset
     2.1    IF memberset(x, s) THEN
            2.1.1    IF (x =  headlist(s)) THEN
                     2.1.1.1    removeset =
                                taillist(s)
                     ELSE
                     2.1.1.2    removeset =
                                insertset(headlist
                                (s),  removeset(x,
                                taillist(s)))
            ELSE
            2.1.2    removeset = s
3    Display removeset
```

The algorithm, removeset takes an integer and a set of integers as parameters and it removes the integer from the set, if it is a member, otherwise, nothing will be removed.

4. DESIGN OF ALGORITHM FOR THE OPERATIONS OF ABSTRACT DATA TYPE, TABLE OF EDGES

The algorithms for the operations of the table of edges can be designed using the algorithms for operations of the underlying abstract data type, set of triples. While the algorithms for the operations of the abstract data type, set of triple can be designed using algorithms for the operations of the abstract data type, list of triple. We begin by designing the algorithms for the operations of list of triples. The structure of the list of triples can be defined in a java class called listriple as follows:

```
int source
int destination
double distance
listriple link
```

The following are the design of algorithms for the operations of the list of triple.

```
listriple emptylistriple()
1    Determine emptylistriple
     1.1    emptylistriple = null
2.   Display emptylistriple
```

The algorithm takes nothing as parameter and it returns an empty list of triple.

```
boolean isemptylistriple(listriple l)
1    Read l
2    Determine isemptylistriple
     2.1    IF l = null THEN
            2.1.1    Isemptytriple = TRUE
            ELSE
            2.1.2    isemptytriple = FALSE
3    Display isemptylistriple
```

The algorithm, isemptylistriple takes a list of triple as parameter and it returns true if the list of triple is empty, otherwise, it returns false.

```
int sourcehead(listriple l)
1    Read l
2    Determine sourcehead
     2.1    IF isemptylistriple(l) THEN
            2.1.1    sourcehead = "Empty list does
                     not have head"
            ELSE
            2.1.2    sourcehead = l.source
3    Display sourcehead
```

The algorithm, sourcehead takes a list of triple as parameter and it returns an appropriate error message if the list of triple is empty, otherwise, it returns the source attribute of the list of triple.

```
int destinationhead(listriple l)
1    Read l
2    Determine destinationhead
     2.1    IF isemptylistriple(l) THEN
            2.1.1    desinationhead = "Empty list
                     does not have head"
            ELSE
            2.1.2    desinationhead = l.destination
3    Display destinationhead
```

The algorithm, destinationhead takes a list of triple as parameter and it returns an appropriate error message, if the list of triple is empty, otherwise, it returns the destination attribute of the list of triple.

```
Double  distancehead(listriple l)
1.   Read l
2    Determine distancehead
     2.1    IF isemptylistriple(l) THEN
            2.1.1    distancehead = "Empty list
                     does not have head"
            ELSE
            2.1.2    distancehead = l.distance
3    Display distancehead
```

The algorithm, distancehead takes a list of triple as parameter and it returns an appropriate error message, if the list of triple is empty, otherwise, it returns the distance attribute of the list of triple.

```
Listriple tailistriple(listriple l)
1   Read l
2   Determine tailistriple
    2.1   IF isemptylistriple(l) THEN
          2.1.1   tailistriple = l
          ELSE
          2.1.2   tailistriple = l.link
3   Display tailistriple
```

The algorithm, tailistriple takes a list of triple, l as parameter and it return l, if it is empty, otherwise, it returns a list of triple after taking away the head.

```
listriple conslistriple(int x, int y; double z;  listriple l)
1   Read data
    1.1   Read x
    1.2   Read y
    1.3   Read z
    1.4   Read l
2.  Determine conslistriple
    2.1   Declare and allocate, temp of the
          type listriple
    2.2   Temp.source= x
    2.3   Temp.destination = y
    2.4   Temp.distance = z
    2.5   Temp.link = l
    2.6   Conslistriple = temp
3.  Display conslistriple
```

The algorithm, conslistriple takes four parameters, which represent the source, destination, distance and the list of triple. The algorithm forms the record of the head and makes it part of the list of triple. These algorithms for the operations of list of triple will be used to design algorithms for the operations of set of triple. The following are the design of algorithms for the operations of set of triple:

```
setriple emptysetriple()
1   Determine emptysetriple
    1.1   emptysetriple = emptylistriple()
2   Display emptysetriple
```

The algorithm, emptysetriple takes nothing as parameter, and it uses the algorithm, emptylistriple to return an empty list of triple as an empty set of triple.

```
Boolean isemptysetriple(setriple l)
1   Read l
2   Determine isemptysetriple
    2.1   isemptysetriple = isemptylistriple(l)
3   Display isemptysetriple
```

The algorithm, isemptysetriple takes a set of triple l, and it uses the algorithm, isemptylistriple to return true if the set of triple is empty, otherwise, it return false.

```
Boolean membersetriple(int x. y; setriple l)
1   Read data
    1.1   Read x
    1.2   Read y
    1.3   Read l
2   Determine membersetriple
    2.1   IF isemptysetriple(l) THEN
          membersetriple = false
    ELSE
    2.2   IF (sourcehead(l) = x) AND
          (destinationhead(l) = y) THEN
          2.2.1   Membersetriple = true
          ELSE
          2.2.2   membersetriple
                  membersetriple(x,y,
                  tailistriple(l))
3   Display membersetriple
```

The algorithm, membersetriple takes three parameters, which represent source node, destination node and the set of triple. The algorithm returns true if the pair of the source node and the destination node is a member of the set of triple, otherwise, it returns false.

```
Setriple insertsetriple(int x; int  y; double z;setriple l)
1   Read data
    1.1   Read x
    1.2   Read y
    1.3   Read z
    1.4   Read l
2   Determine insertsetriple
    2.1   IF membersetriple(x, y, l) THEN
          2.1.1   insertsetriple = l
          ELSE
          2.1.2   insertsetriple =
                  conslistriple(x,y,z,l)
3   Display insertsetriple
```

The algorithm, insertsetriple takes four parameters, which represent the following, source node, destination node, distance and the set of triple. The algorithm uses some of the algorithms for the operations of listriple to determine if the pair of source node and destination nodes is a member of the set of triple. If the pair is a member, it does not insert it into the set of triple, otherwise, it inserts the source node, destination node and the distance into the set of triple.

```
Setriple removesetriple(int x, y; setriple l)
1   Read data
    1.1   Read x
    1.2   Read y
    1.3   Read l
2   Determine removesetriple
    2.1   IF membersetriple(x,y l) THEN
          2.1.1   IF x = sourcehead(l) AND
                  y = destinationhead(l)
                  THEN
                  2.1.1.1   Removesetriple =
                            tailistriple(l)
                  ELSE
                  2.1.1.2   removesetriple =
                            conslistriple(sour
                            cehead(l),
                            destinationhead(l
                            ),
                            distancehead(l),
                            removesetriple(x,
                            y, tailistriple(l)))
                  ELSE
    2.1.2   removesetriple = l
3   Display removesetriple
```

The algorithm, removesetriple takes three parameters, which represent the source node, destination node and the set of the triple. The algorithm uses the algorithms for the operations of list of triple and set of triple to determine if the pair source node and destination node is a member of the set of triple, if it is, it looks for the pair of source node and destination node and removes it from the set of triple, otherwise, nothing will be removed. The algorithms for the operations of the abstract data type, tabletriple can be defined using the algorithms for the operations of the underlying abstract data type, setriple. The primary keys of the abstract data type, tabletriple are source and destination fields. The following are the design of algorithms for the operations of the abstract data type, table of triple.

```
tabletriple emptytabletriple()
1   Determine emptytabletriple
    1.1   emptytabletriple = emptysetriple()
2   Display emptytabletriple
```

The algorithm, emptytabletriple takes nothing as parameter and it uses the algorithm for the operation of the abstract data type, setriple to return an emptytable of triple.

```
Boolean isemptytabletriple(tabletriple l)
1   Read l
2   Determine isemptytabletriple
    2.1   isemptytabletriple = isemptysetriple(l)
3   Display isemptytabletriple
```

The algorithm, isemptytabletriple takes one parameter, which is a table of triple. The algorithm uses the algorithm for the operation of setriple to return true if the table is empty, otherwise, it returns false.

```
Boolean membertabletriple(int x;  int  y; tabletriple l)
1   Read data
    1.1   Read x
    1.2   Read y
    1.3   Read l
2   Determine membertabletriple
    2.1   membertabletriple =
          membersetriple(x, y, l)
3   Display membertabletriple
```

The algorithm, membertabletriple takes three parameters, which represent the source node, destination node and a table of triple. The algorithm uses the algorithm for the operation of setriple to return true if the pair of the source and destination nodes is in the table of triple, otherwise, it returns false.

```
Double locatedistance(int x; int  y; tabletriple l)
1   Read data
    1.1   Read x
    1.2   Read y
    1.3   Read l
2   Determine locatedistance
    2.1   IF membertabletriple(x, y, l) THEN
          2.1.1   IF x = sourcehead(l) AND
                  y = destinationhead(l)
                  THEN
                  2.1.1.1   locatedistance =
                            distancehead(l)
                  ELSE
                  2.1.1.2   locatedistance =
                            locatedistance(x,
                            y, tailistriple(l))
          ELSE
          2.1.2   locatedistance = "The pair
                  of primary key does not
                  exist"
3   Display locatedistance
```

The algorithm, locatedistance takes three parameters, which represent the source, destination nodes, and a table of triple. The algorithm uses the algorithms for the operations of tabletriple to determine if the pair that forms the primary key is a member of the primary key of the table of triple. If it is, it looks for the primary key and returns the corresponding distance attribute; otherwise, it returns an appropriate error message.

Tabletriple insertabletriple(int x; int y; double z; tabletriple l)
1 Read data
1.1 Read x
1.2 Read y
1.3 Read z
1.4 Read l
2 Determine insertabletriple
2.1 insertabletriple = insertsetriple(x, y, z, l)
3 Display insertabletriple

The algorithm, insertabletriple takes four parameters, which represent the source and destination nodes, the distance attribute and a table of triple. The algorithm uses the algorithm for the operation of setriple to insert the source, destination and distance attributes into the table of triple.

Tabletriple removetabletriple(int x;int y; tabletriple l)
1 Read data
1.1 Read x
1.2 Read y
1.3 Read l
2 Determine removetabletriple
2.1 removetabletriple = removesetriple(x, y, l)
3 Display removesetriple

The algorithm, removetabletriple takes three parameters, which represent the source and destination nodes, and a table of triple. The algorithm uses the algorithm for the operation of setriple to remove the record that has the source and destination nodes as primary key, from the table of triple. The following algorithms will be used to sort the abstract data ***type, tabletriple in ascending order of distance.***

Tabletriple sortabletriple(tabletriple t)
1 Read t
2 Determine sortabletriple
2.1 IF isemptytabletriple(t) THEN
2..1.1 sortabletriple = t
ELSE
2.1.2 sortabletriple = insertordertable(sourcehead(t), destinationhead(t), distancehead(t), sortabletriple(tailistriple(t)))
3 Display sortabletriple
Tabletriple insertordertable(int x; int y; double z; tabletriple t)
1 Read data
1.1 Read x
1.2 Read y
1.3 Read z
1.4 Read t
2 Determine insertordertable
2.1 IF isemptytabletriple(t) THEN
2.1.1 Insertordertable = insertabletriple(x, y, z, t)
ELSE
2.1.2 IF z <= distancehead(t) THEN
2.1.2.1 Insertordertable = insertabletriple(x, y, z, t)
ELSE
2.1.2.2 insertordertable = insertabletriple(sourcehead(t), destinationhead(t), distancehead(t), insertordertable(x, y, z, tailistriple(t)))

3. Display insertordertable

5. DESIGN OF ALGORITHMS FOR THE OPERATIONS OF ABSTRACT DATA TYPE, GRAPH

The algorithms for the operations of the abstract data type, graph can be designed using the algorithms for the operations of the abstract data types, set of nodes and table of edges/triple. Since some of the algorithms return a particular abstract data type, therefore some of the algorithms for the operations of the abstract data type, graph will be considered as two different algorithms. The following are the design of algorithms for the operations of abstract data type, graph.

myset emptygraph1()
1 Determine emptygraph1
1.1 emptygraph1 = emptyset()
2 Display emptygraph1
Tabletriple emptygraph2()
1 Determine emptygraph2
1.1 emptygraph2 = emptytabletriple()
2. Display emptygraph2

The algorithm, emptygraph has been designed as two algorithms, emptygraph1 and emptygraph2. The algorithm, emptygraph1 takes nothing as parameter and it uses the algorithm for the operation of the abstract data type, set of nodes to return an empty set of node. Similarly, the algorithm, emptygraph2 takes nothing as parameter, but it uses the algorithm for the operation of the abstract data type, table of edges to return an empty table of edges.

myset addnode1(int x; myset s; tabletriple t)
1 Read data
1.1 Read x
1.2 Read s
1.3 Read t
2 Determine addnode1
2.1 addnode1 = insertset(x,s)
3 Display addnode1
tabletriple addnode2(int x; myset s; tabletriple t)
1 Read data
1.1 Read x
1.2 Read s
1.3 Read t
2 Determine addnode2
2.1 addnode2 = t
3 Display addnode2

The algorithm, addnode has been designed as two algorithms, addnode1 and addnode2. Both of them take three parameters, which are the node, x to be added into the set of nodes, s, and the table of edges/triple, t. Addnode1 uses the algorithm for the operation of the abstract data type, set of nodes to add the node into the set of nodes, and it returns a set of nodes. On the other hand, addnode2 returns the table of edges/triple that was passed to it as parameter.

myset addedge1(int x; int y; int z; myset s; tabletriple t)
1 Read data
1.1 Read x
1.2 Read y
1.3 Read z
1.4 Read s
1.5 Read t
2 Determine addedge1
2.1 Addedge1 = s
3 Display addedge1
Tabletriple addedge2(int x; int y; int z; myset s; tabletriple t)
1 Read data
1.2 Read x
1.3 Read y
1.3 Read z
1.4 Read s
1.5 Read t
2 Determine addedge2
2.1 IF memberset(x,s) AND memberset(y,s) THEN
2.1.1 addedge2 = insertabletriple(x,y,z,t)
ELSE
2.1.2 addedge2 = t
3. Display addedge2

The algorithm, addedge has been designed as two algorithms, addedge1 and addedge2. Both of them take five parameters, which are two integer numbers, x and y that form the new edge and the distance z between them, which will be added into the table of edges, t and a set of node, s. Addedge1 returns the set of nodes, s that was passed to it as parameter, while addedge2 uses the algorithm for the operations of set of nodes to confirm if each of the two nodes that form the edges is a member of the set of nodes. If each of them is a member, it uses the algorithm for the operation of table of edges to insert the triple data, x,y and z into the table of triple t, otherwise ,they will not be inserted.

Boolean isemptygraph(myset s, tabletriple t)
1 Read data
1.1 Read s
1.2 Read t
2 Determine isemptygraph
2.1 IF isemptyset(s) AND isemptytabletriple(t) THEN
2.1.1 isemptygraph = true;
ELSE
2.1.2 isemptygraph = false;
3 Display isemptygraph

The algorithm, isemptygraph takes two parameters, a set of nodes and a table of triple. The algorithm uses the algorithms for the operations of set of nodes and table of edges to determine if the graph is empty. The graph can only be empty if the set of nodes and the table of edges are empty; otherwise, the graph is not empty.

Boolean contains(int x; myset l; tabletriple t)
1 Read data
1.1 Read x
1.2 Read l
1.3 Read t
2 Determine contains
2.1 IF memberset(x,l) THEN
2.1.1 contains = true
ELSE
2.1.2 contains = false
3 Display contains

The algorithm, which is called, contains takes a graph and a node as parameters. The graph consists of a set of nodes and a table of edges. The algorithm uses the algorithm for the operations of set of nodes to determine if the node x is a member of the set of nodes, l.

```
Boolean isadjacent(int x; int y; myset s; tabletriple t)
1   Read data
    1.1     Read x
    1.2     Read y
    1.3     Read s
    1.4     Read t
2   Determine isadjacent
    2.1     IF membertabletriple(x,y,t) THEN
            2.1.1   isadjacent = true
    ELSE
            2.1.2   isadjacent = false
3   Display isadjacent
```

The algorithm, iaadjacent takes four parameters, which are two nodes, x and y, a set of nodes, s and a table of edges, t. The algorithm uses algorithm for the operation of table of edges to determine if the two nodes x and y are adjacent nodes of the graph.

```
Myset deletenode1(int x, myset s, tabletriple t)
1   Read data
    1.1     Read x
    1.2     Read s
    1.3     Read t
2   Determine deletenode1
    2.1     IF membertable(x,t) THEN
            2.1.1   deletenode1 = s
            ELSE
            2.1.2   deletenode1 =
                    removeset(x,s)
3   Display deletenode1
Tabletriple deletenode2(int x, myset s, tabletriple t)
1   Read data
    1.1     Read x
    1.2     Read s
    1.3     Read t
2   Determine deletenode2
    2.1     deletenode2 = t
3   Display deletenode2
```

The algorithm, deletenode has been designed as two algorithms, deletenode1 and deletenode2. Each of them takes three parameters, which are the node x to be deleted from the set of nodes, s and the table of edges, t. The algorithm, deletenode1 first ensures that the node to be deleted is not used to form an edge. Once it confirms this, it removes the node from the set of nodes, otherwise, it does not delete the node from the set of nodes, afterwards, it returns a set of nodes. The algorithm, membertable, which is designed below is used to confirm if the node is not used to form an edge before deleting. The algorithm, deletenode2 returns a table of edges.

```
Boolean membertable(int x, tabletriple t)
1   Read data
    1.1     Read x
    1.2     Read t
2   Determine membertable
    2.1     IF isemptytabletriple(t) THEN
            2.1.1   membertable = false
            ELSE
            2.1.2   IF x = sourcehead(t) OR x
                    =   destinationhead(t)
                    THEN
            2.1.2.1 Membertable = true
            ELSE
            2.1.2.2 membertable =
                    membertable(x,tailistriple(t))
3   Display membertable
```

The algorithm, membertable, takes two parameters, a node , x, which is an integer and tabletriple t, which is a table of triple. The algorithm returns true if the node x is used to form an edge, otherwise, it returns false.

```
Myset deletedge1(int x, int y, myset s, tabletriple t)
1   Read data
    1.1     Read x
    1.2     Read y
    1.3     Read s
    1.4     Read t
2   Determine deletedge1
    2.1     deletedge1 = s
3   Display deletedge1
Tabletriple deletedge2(int x, int y, myset s, tabletriple
t)
1   Read data
    1.1     Read x
    1.2     Read y
    1.3     Read s
    1.4     Read t
2   Determine deletenode2
    2.1     deletenode2 =
            removetabletriple(x.y,t)
3   Display deletedge2
```

The algorithm, deletedge has been designed as two algorithms, deletedge1 and deletedge2. Both of them take four parameters, which are, two integer numbers that form the edge to be deleted, a set of node and a table of edges. The algorithm, deletedge1 returns the set of nodes, while the algorithm, deletedge2 uses the algorithm for the operation of abstract data type, tabletriple to remove the two integer numbers that form an edge from the table of edges, afterwards, it returns a table of edges.

6 DESIGN OF RECURSIVE DIJKSTRA'S SHORTEST ROUTE ALGORITHM

The algorithms for the operations of the abstract data type, graph will be used as the underlying abstract data type to design the recursive Dijkstra's shortest route algorithm.

The following algorithms will be used to design the recursive Dijkstra's shortest route algorithm.

```
tabletriple neighbourtable(int x, myset s, tabletriple t, tabletriple p)
1   Read data
    1.1   Read x
    1.2   Read s
    1.3   Read t
    1.4   Read p
2   Determine neighbourtable
    2.1   IF isemptytabletriple(t) THEN
          2.1.1   neighbourtable= emptytabletriple()
          ELSE
          2.1.2   IF (x = sourcehead(t)) AND (NOT memberperm(destinationhead(t), p))) THEN
                  2.1.2.1   Neighbourtable = insertabletriple(sourcehead(s), destinationhead(s), distancehead(t), neighbourtable(x, s, tailistriple(t), p))
                  ELSE
                  2.1.2.1   neighbourtable = neighbourtable(x, s, tailistriple(t), p)
3   Display neighbourtable
```

The algorithm, neighbourtable determines all the neighbours of a node and put them in a table. It uses an algorithm called, memberperm, which helps to ensure that the neighbours of the node has not been labeled permanently.

```
tabletriple templabels(double x, tabletriple t, tabletriple p)
1.  Read Data
    1.1   Read x
    1.2   Read t
    1.3   Read p
2.  Determine templabels
    2.1   IF (isemptytabletriple(t)) THEN
          2.1.1   templabels = emptytabletriple()
          ELSE
          2.1.2
          IF(destinationmember(destinationhead(t), p) THEN
                  2.1.2.1   templabels = templabels(x, tailistriple(t), p)
                  ELSE
                  2.1.2.2   templabels = insertabletriple(sourcehead(t), destinationhead(t), distancehead(t)+ x, templabels(x, tailistriple(t), p))
3   Display templabels
```

The algorithm, templabels labels as temporary labels all the neighbours of the current source node that have not been lebelled as permanent labels. The above algorithm uses the algorithm, destinationmember, which is defined below:

```
Boolean destinationmember(int x, listriple t)
1.  Read Data
    1.1   Read x
    1.2   Read t
2.  Determine destinationmember
    2.1   IF (isemptytabletriple(t)) THEN
          2.1.1   destinationmember = FALSE
          ELSE
          2.1.2   IF (x = destinationhead(t)) THEN
                  2.1.2.1   destinationmember = TRUE
                  ELSE
                  2.1.2.2   destinationmember = destinationmember(x,tailistriple(t))
3.  Display destinationmember
```

The recursive Dijkstra's shortest route algorithm will use another algorithm, called tabletripleunion, this algorithm takes two tables of triple and joins them together, and the algorithm is defined below:

```
Tabletriple tabletripleunion(tabletriple t1, tabletriple t2)
1   Read Data
    1.1   Read t1
    1.2   Read t2
2   Determine tabletripleunion
    2.1   IF (isemptytabletriple(t1)) THEN
          2.1.1   tabletripleunion = t2
          ELSE
          2.1.2   tabletripleunion = insertabletriple(sourcehead(t1), destinationhead(t1), distancehead(t1), tabletripleunion(tailistriple(t1), t2));
3.  Display tabletripleunion
```

These algorithms can be used to design the recursive Dijkstra's shortest route algorithm, which is designed below.

tabletriple shortestroute(int sourcenode, tabletriple temp, tabletriple perm, myset s, tabletriple t)

1. ***Read Data***
   - ***1.1 Read sourcenode***
   - ***1.2 Read temp***
   - ***1.3 Read perm***
   - ***1.4 Read s***
   - ***1.5 Read t***
2. ***Determine shortestroute***
   - ***2.1 IF (isemptytabletriple(temp))***
     - ***2.1.1 shortestroute = perm;***

     ELSE
     - ***2.1.2 newsourcenode = destinationhead(sortabletriple(temp))***
     - ***2.1.3 pivotdistance1= distancehead(sortabletriple(temp));***
     - ***2.1.4 sampletemp = removetabletriple(sourcehead(sortabletriple(temp)), destinationhead(sortabletriple(temp)), temp);***
     - ***2.1.5 sampleneighbours = null;***
     - ***2.1.6 sampleneighbours= neighbourtable(newsourcenode, s, t);***
     - ***2.1.7 sampletemp1 =templabels(pivotdistance1 ,sampleneighbours, perm);***
     - ***2.1.8 sampletemp =tabletripleunion(sampletemp1, sampletemp);***
     - ***2.1.9 sampleperm =insertabletriple(sourcehead(sortabletriple(sampletemp)), destinationhead(sortabletriple(sampletemp)), distancehead(sortabletriple(sampletemp)), perm);***
     - ***2.10 sampleperm = shortestroute(n, sampletemp, sampleperm, s, t);***
3. ***Displaysampleperm***

7. IMPLEMENTATION OF THE ALGORITHMS

Java programming language was used to implement all the algorithms for the operations of the following abstract data types, mylist, myset, listriple, setriple, tabletriple and graph. Each of these abstract data types was implemented in a Java class, therefore, the class hierarchy is similar to the hierarchy of the abstract data type, which was shown in figure 2. Furthermore, all the algorithms that the recursive Dijkstra's shortest route algorithm used, including the recursive Dijkstra's shortest route algorithm were implemented, and a test program was written to test the implemented algorithms. The test program used the implemented algorithms to construct a network as a set of nodes and a table of triple, afterwards the test program used the implemented shortest route algorithm and other implemented algorithms to determine the shortest route between a source node and every other nodes of the network, and the corresponding shortest distance between the source node and any other node of the network.

8 RESULT OF THE IMPLEMENTED ALGORITHM

The test program has been tested and for any source node of the network, it produces the shortest route between the source node and any other node of the network with the corresponding distance. The result is presented as a table of triple, (A, B, C), where letter B denotes the destination node, while letter A denotes the sequence node, and letter C denotes the shortest distance from the source node to node B. Therefore, the result of the shortest route from node 1 to any other node of the network diagram shown in figure 1 can be presented as a table shown below in table 2.

Table 2: Result of the Implemented Algorithm

Sequence Node	Destination Node	Shortest Distance
3	5	800
4	6	800
2	4	700
3	2	500
1	3	200
0	1	0

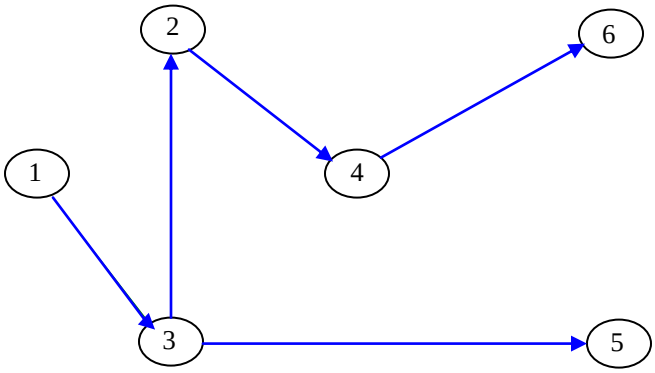


Figure 3: Shortest route from node 1 to any other node

The result, which is shown in table 2 can be represented in the network diagram in figure 3. The colored route shows the shortest route from node 1 to any of the destination node.

9. CONCLUSION

This paper has been able to use a novel approach to represent graph as a collection of two sets, the set of nodes and the table/relation of triple (edges). The paper has designed novel algorithms for each of the operations of the abstract data types that the graph ADT uses. Furthermore, the recursive Dijkstra's shortest route algorithm has been designed and implemented using the algorithms for the operations of the ADT, graph, and other ADTs, mylist, myset, listriple, setriple and tabletriple. All the algorithms have been implemented using Java programming language, and a test program has been written to test the implemented algorithms. The result of the test program shows that for a given graph, the shortest route from a given source node to any other node in the network will be determined.

REFERENCES

[1] R. Harrison, "Abstract Data Types in MODULA-2, John Wiley & Sons, 1989.

[2] Bernard Kolman and Robert C. Busby, "Discrete Mathematical Structures f or Computer Science", Prentice-Hall, 1984.

[3] S.G. Shrinivas, S.Vetrivel and Dr. N.M.Elango. "Applications of Graph Theory In Computer Science An Overview", International Journal of Engineering Science and Technology. Vol. 2(9), 2010

[4] Denis Howe. Free on-line dictionary of computing. foldoc.org, 1993.

[5] Graph Theory, Keijo Ruohonen , 2008, http://math.tut.fi/~ruohonen/GT_English.pdf

[6] O.E. Oguike and M.N. Agu (Mrs), "GSM Network Algorithm for the Most Efficient Route Between any two Geographic Area in the Network, In Proc. of the 2002 National Conference of Computer Association of Nigeria, now Nigeria Computer Society, Abuja, Nigeria.

[7] Gregory L. Heileman, "Data Structures, Algorithms, and Object Oriented Programming", McGraw-Hill International Editions, 1996.

[8] Hamdy A. Taha, "OperationResearch An Introduction", Prentice-Hall of India, 1999.

[9] Allen B. Tucker, Bradley W. James, Charles E. Keleman, "Fundamentals of Computing II: Abstraction, Data Structure and Large Software Systems, Mc-Graw Hill.

[10] Nishtha Kesswani and Dinesh Gopalani, "Design and Implementation of Multi-Parameter Dijkstra's (MPD) Algorithm: A Shortest Path Algorithm for Real Road Network", Int. Journal of Advances in Engineering Research, Vol. 2, Issue III, September 2011.

[11] P. Biswas, P. K. Mishra and N. C. Mahanti, *Computational Efficiency of Optimized Shortest Path Algorithm", International Journal of Computer Science & Applications, 2005, *Vol. II, No. II, pp. 22 – 37*

[12] Shweta Srivastava, "Comparative Analysis of Algorithms for Single Source Shortest Path Problem", International Journal of Computer Science and Security (IJCSS), Volume (6) : Issue (4) : 2012

[13] S.Meena Kumari & Dr.N.Geethanjali, "A Survey on Shortest Path Routing Algorithms for Public Transport Travel", Global Journal of Computer Science and Technology, Vol. 9 Issue 5 (Ver 2.0), 2010.

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

An Effective Measurement of Data Security in a Cloud Computing Environment

A.A. Elusoji
Computer Technology Department
Yaba College of Technology, Yaba
Lagos State, Nigeria
elusoji872@yahoo.com

L.N. Onyejegbu
Computer Science Department
University of Port-Harcourt, Rivers –State
nneka2k@yahoo.com

O.S. Ayodele
Computer Science Department
Kogi State Polytechnic Lokoja
Kemtemmy2009@yahoo.com

ABSTRACT
The increasingly sophisticated network infrastructure and increased bandwidth developed in recent years has dramatically enhanced the stability of various application services available to users through the Internet, thus marking the beginning of cloud computing network services. The security for Cloud Computing is emerging field because of its performance, high availability, and least cost and since there is a critical need to securely store, manage, share and analyze massive amounts of complex web applications, it is important that clouds be secure. Service providers must have a viable way to protect their clients' data, especially to prevent the data from disclosure by unauthorized insiders. This paper has been written to focus on the problem of data security. It describes the approach to securing cloud computing based on analysis of Cloud Security treats and Technical Components of Cloud Computing.

Keywords: Cloud Computing, Security threats, Service provider Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS)

African Journal of Computing & ICT Reference Format:
A.A. Elusoji, L.N. Onyejegbu & O.S. Ayodele (2013). An Effective Measurement of Data Security in a Cloud Computing Environment. Afr J. of Comp & ICTs. Vol 6, No. 2. pp 67-76

1. INTRODUCTION
There are numerous security issues for cloud computing as it encompasses many technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Therefore, security issues for many of these systems and technologies are applicable to cloud computing. For example, the network that interconnects the systems in a cloud has to be secure. Furthermore, virtualization paradigm in cloud computing results in several security concerns. For example, mapping the virtual machines to the physical machines has to be carried out securely. Data security involves encrypting the data as well as ensuring that appropriate policies are enforced for data sharing. In addition, resource allocation and memory management algorithms have to be secure.

Each type of cloud computing model—public, private or hybrid—faces different levels of IT risk. In the private cloud delivery model, the cloud owner does not share resources with any other company. Private clouds are owned and operated by a single organization, delivering IT services within the constraints of their own network perimeter. In the public cloud computing model, IT activities and functions are provided as a service that can be billed on a pay-per-use or subscription basis via the Internet from external suppliers, using resources not owned by the consumer. The sharing of IT resources in a public, multitenant environment can help improve utilization rates and can reduce costs significantly, while maintaining access to high quality technology. In a public cloud, an organization rents IT resources instead of having to invest in their own physical IT infrastructure or maintain under-utilized equipment to service peak loads.

Instead, they can scale usage up or down, according to need, with costs directly proportional to need. Many organizations embrace both public and private cloud computing by integrating the two models into hybrid clouds.

2. OVERVIEW OF CLOUD COMPUTING

Cloud computing services such as Amazon EC2 and Windows Azure are becoming More and more popular but it seems many people are still unclear as to what exactly the buzzword "Cloud computing" actually means. In its simplest form, the principle of Cloud computing is the provision of computing resources via a network. It shifts the responsibility of configuring, deploying and maintaining computing infrastructure from clients to Cloud providers. Providers generally expose an interface for clients to interact with their resources as if they were their own standalone resource; however often a number of resources may be aggregated on the same computer or cluster of computers. The user does not necessarily know the details of the location, equipment or configuration of their resources, rather they are provided with a "virtualised" computer resource hosted in "the Cloud".

There are distinctions among the most common cloud service models as shown in Figure 1. Available to anyone with Internet access, cloud service models include:-

Software as a service (SaaS) cloud model
SaaS clients rent usage of applications running within the Cloud‟s provider infrastructure. It Enables software to be delivered from a host source over a network as opposed to installations or implementations

Platform as a Service (PaaS) cloud model
PaaS Cloud providers offer an application platform as a service, for example Google App Engine. This enables clients to deploy custom software using the tools and programming languages offered by the provider. Clients have control over the deployed applications and environment-related setting. It Enables operating systems and middleware services to be delivered from a managed source over a network.

Infrastructure as a Service (IaaS) cloud model
IaaS delivers hardware resources such as CPU, disk space or network components as a service. These resources are usually delivered as a virtualization platform by the Cloud provider and can be accessed across the Internet by the client. It Enables the entire infrastructure to be delivered as a service over a network, including storage, routers, virtual systems, hardware and servers

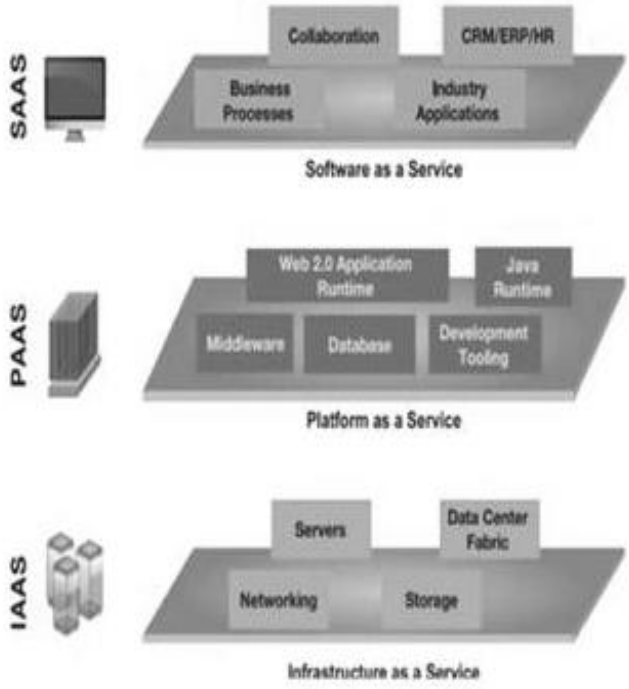


Figure 1. Cloud computing models

3. TECHNICAL COMPONENTS OF CLOUD COMPUTING

As shown in the Figure 2, key functions of a cloud management system is divided into four layers, respectively the Resources & Network Layer, Services Layer, Access Layer, and User Layer. Each layer includes a set of functions:

- The Resources & Network Layer manages the physical and virtual resources.
- The Services Layer includes the main categories of cloud services, namely, NaaS, IaaS, PaaS, SaaS/CaaS, the service orchestration function and the cloud operational function.
- The Access Layer includes API termination function, and Inter-Cloud peering and federation function.
- The User Layer includes End-user function, Partner function and Administration function.

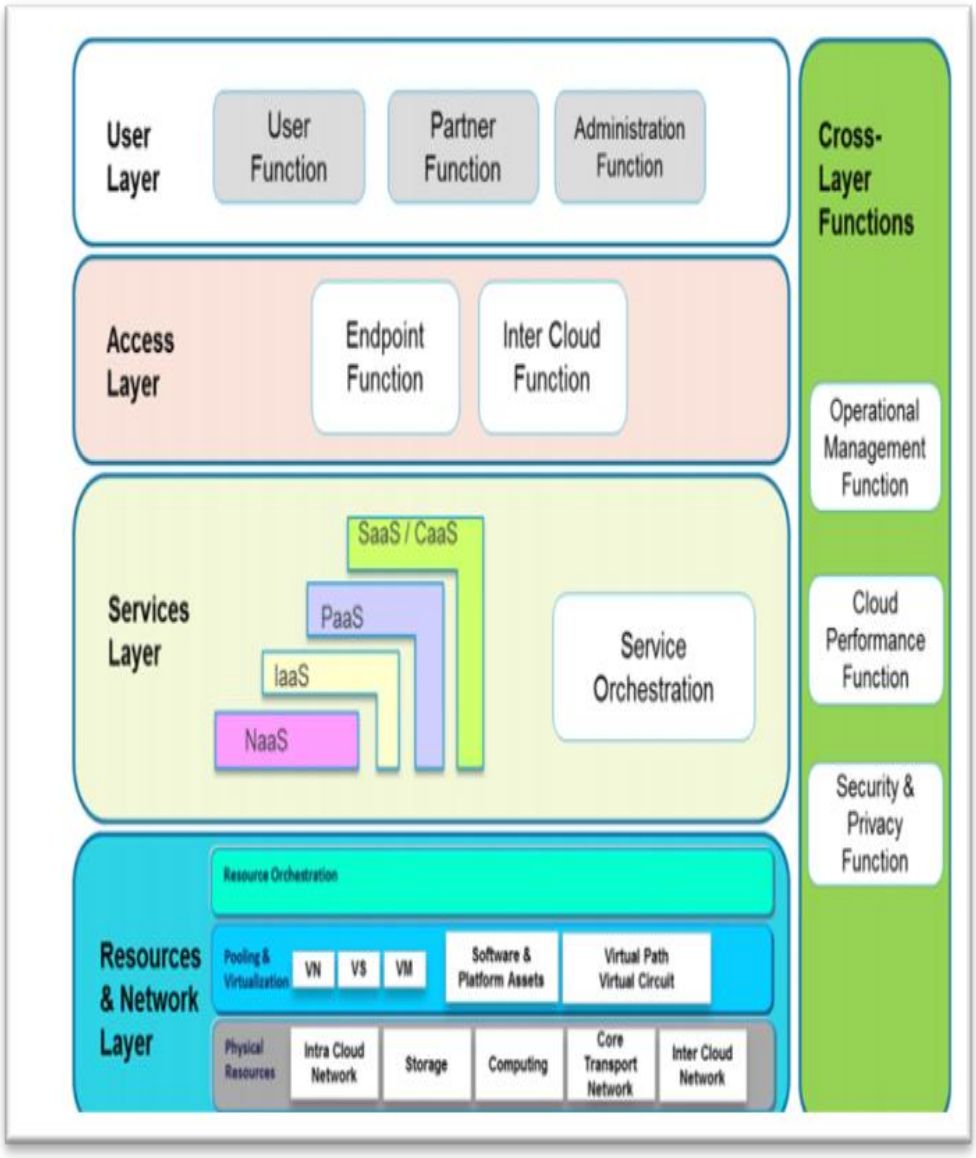


Figure 2. The cloud computing components

Other functions like Management, Security & Privacy, etc. are considered as cross layer functions that covers all the layers. The main principle of this architecture is that all these layers are supposed to be optional. This means that a cloud provider who wants to use the reference architecture may select and implement only a subset of these layers. However, from the security perspective, the principal of separation requires each layer to take charge of certain responsibilities. In event the security controls of one layer are by passed (e.g. access layer), other security functions could compensate and thus should be implemented either in other layers or as cross-layer functions.

4. SECURING DATA IN THE CLOUDS

A common approach to protect user data is that user data is encrypted before it is stored. In a cloud computing environment, a user's data can also be stored following additional encryption, but if the storage and encryption of a given user's data is performed by the same service provider, the service provider's internal staff (e.g., system administrators and authorized staff) can use their decryption keys and internal access privileges to access user data. From the user's perspective, this could put his stored data at risk of unauthorized disclosure. In which if a user (either employee or anonymous) want to access the data if it belongs to protection then user have to register itself (if he is already registered need not require further registration.

Now suppose the user registered itself for accessing data, Organization will provide username and password for authentication. At the same time organization sends the username to cloud provider.

1. Request for access data
2. Send the signal to redirect person
3. Redirects

Now when user sends request along with username to access the data to cloud provider, the cloud provider first check in which ring requested data belong. If authentication is required, it first checks the username in its own directory for existence, if the username does not exist it ask the user to register itself.

If the username matches it redirect the request to company for authentication.

(1) Send password for authentication
(2) Redirect to access resource
(3) Request redirected

Now the user sends password for authentication, and after authentication it redirect the request to cloud provider to access resource .If user-name and password doesn't match then user is not allow to access their account. And also in some case if hacker wants to hack the account of a particular user then in that case hacker gets only the fake database of the account is there to access the account by hitting the user-name and password, if limit become cross then hacker get's the fake database.

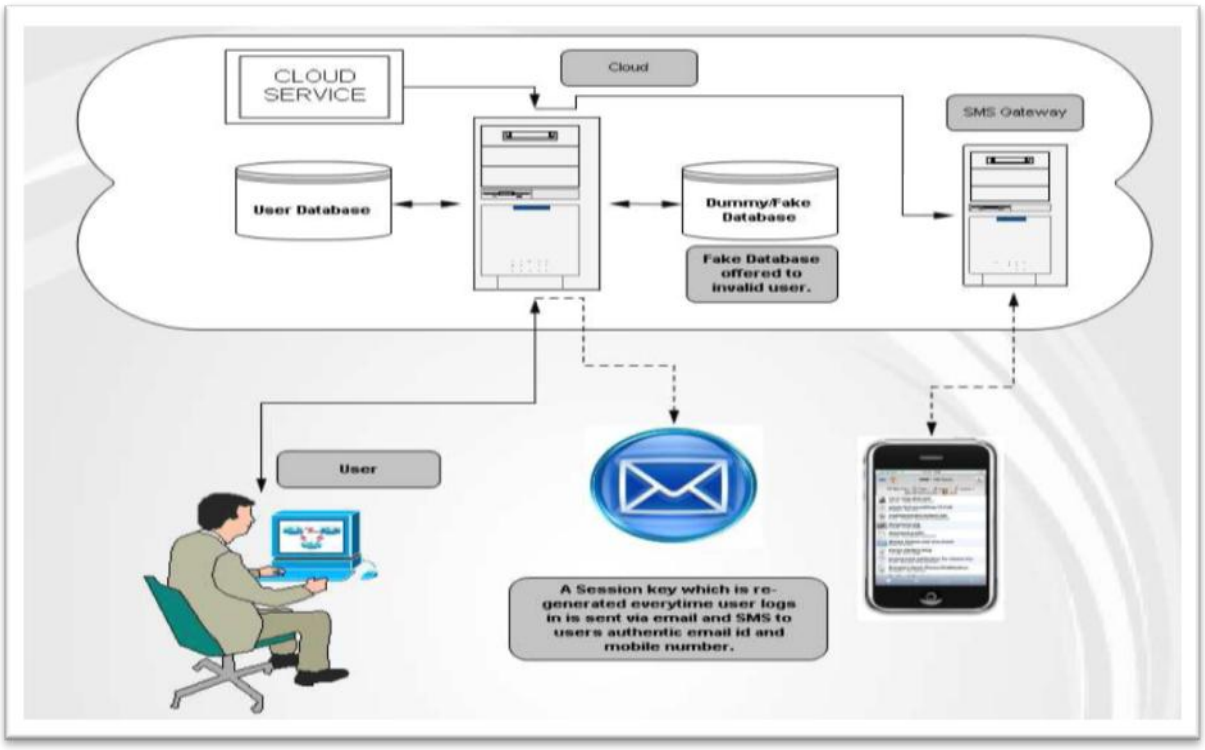


Figure 3: Authentication cycle

5. THREATS TO SECURITY IN THE CLOUD

There are several significant threats that should be considered before adopting the paradigm of cloud computing in e-learning. These threats are described as follows:-

1) Abuse and Nefarious use of cloud: Cloud services providers often targeted for their weak registration system and limited fraud detection capabilities. This paves way to the spammers, malicious code authors and other cybercriminals can misuse the various types of services including unlimited bandwidth and storage facilities offered by the cloud providers. Misuse includes creating spam, decoding and cracking of passwords, executing malicious codes to access rich information such as question papers, learning materials, assessments etc.

2) Insecure Software Access: Various software interfaces and APIs are used by the cloud users in e-learning to access and manage the cloud services. These APIs play an integral part during provisioning, management, orchestration and monitoring of the processes running in a cloud environment. Hence these APIs needs to be secured and should include features of authentication, access control, encryption and activity monitoring. Many security issues will be raised if cloud service providers believe on weak set of APIs.

3) Malicious Insider: Malicious employees who are working in the provider‟s or user site can be able to perform insider attacks. This insider can steal the confidential data of cloud users in e-learning. Malicious insider can easily get the cloud users confidential data such as password, cryptographic keys and files. It will affect the standards and trust of cloud users in e-learning. As a result, it can cause damage on both financial grounds as well as organisation reputation

4) Data Separation: Virtual Machine (VMs) are virtualized based on the physical hardware of cloud providers and stores the e-learning user‟s applications supplied by the cloud providers due to the cloud virtualization. These VMs are isolated from each other by cloud providers in order to maintain the security of users. These VMs are managed by hypervisor who are the main source of managing the virtualized cloud platform so as to provide virtual memory as well as CPU scheduling policies to VMs. Hypervisors are mainly targeted by the hackers since they are residing between VMs and hardware. Strong isolation is needed to ensure that VMs are not able to access the activities of other VMs under the same cloud computing providers. Even though several vendors offers strong security mechanism to protect the cloud supervisors, however sometimes security of VMs is compromised

5) Data Loss or Leakage: Operational failures, unreliable data storage and inconsistent use of encryption keys will lead to a data loss. Operational failure includes deletion, incomplete deletion or alteration without any backup of the source e-learning content. It may be either intentionally or unintentionally. Unreliable data storage means storing a data on unreliable media which cannot be recoverable if the data is lost. Inconsistent use of encryption keys will lead to unauthorized access and data loss such as destruction of sensitive and confidential information. It will definitely affect the reputation of the company.

6) Hijacking: Controlling the users account through the unauthorised access by the hackers is referred as account or service hijacking. It includes phishing, fraud and exploitation of software vulnerabilities.It is not enough to secure the sensitive and confidential information through the common way of authentication and authorization process e-learning.

7) Unknown Risk: It is essential for the every e-learning user to know the software versions, security practices, software code updates and intrusion attempts. Cloud service providers usually advertised these futures and functionality with the necessary details such as internal security procedure, configuration hardening, patching, auditing and logging. E-learning users must be aware and clarified how their data and related files are stored. On the other hand, e-learning user may unaware of the unknown risk profile which may include serious threats.

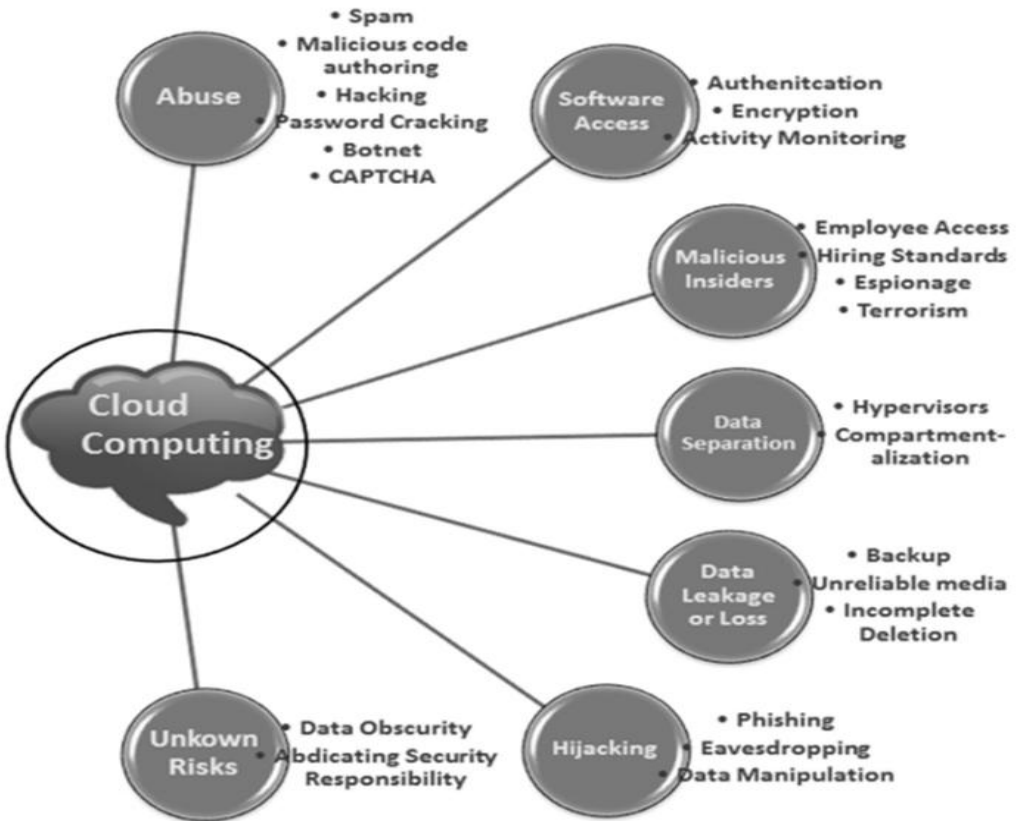


Fig.4: Seven threats to security in cloud

6. GUIDANCE FOR SECURITY CONCERN IN CLOUD BASED E-LEARNING

There are various steps given by the cloud service providers to ensure the security concern in the cloud computing which could be applied to cloud based e-learning. Few guidelineshave been given by the organizations such as Cloud Standards Customer Council, Intel, Microsoft etc to build the security in the cloud are as follows:

A. **Steps of Security for Cloud Computing**[13][14]
This is designed to help public cloud consumers evaluate and compare security provided in key areas from different cloud providers. This steps has to be followed by the institutions to ensure the security for cloud computing before going for cloud services in their e-learning systems.

1) **Step 1: Ensure effective governance, risk and compliance processes**: Security controls available in cloud computing are very much similar to traditional IT environments. But educational institution should understand their own level of risk tolerance and focus on mitigating the risks that institutions cannot afford to neglect.

2) **Step 2: Audit operational and business processes**: Audits should be carried out by the educational institution appropriately by assigning skilled staff and set of controls should be established to meet the institutions security requirements.

3) **Step 3: Manage people, roles and identities**: Educational institutions needs to control users‟ roles and privileges as it manage thousands of users such as students and staff who access cloud applications and services, each with different roles and rights.
4) **Step 4: Ensure proper protection of data and information**: Educational institution should ensure the proper protection of data and information. Additional focus on data security is needed because of the distributed nature of the cloud computing infrastructure and the shared responsibilities that cloud computing involves.

5) **Step 5: Enforce privacy policies**: Educational institutions are responsible for defining policies to address any privacy concerns and in increasing awareness of data protection within the institutions. In addition to that they should ensure the adherence of cloud providers to the defined privacy policies.

6) **Step 6: Assess the security provision for cloud applications**: Educational institutions must apply the same diligence to application security for both physical security and infrastructure security. Applications should not be compromised at any cost to avoid any additional risk.

7) **Step 7: Ensure cloud networks and connections are secure**: Educational institution should check for certain external network perimeter safety measures from cloud providers to ensure the secured connections and network.
8) **Step 8: Evaluate security controls on physical infrastructure and facilities:** Educational institutions should concern about the physical infrastructure and facilities provided by the cloud providers and it are an important consideration for security of any IT system especially in a cloud based e-learning.

9) **Step 9** : **Manage security terms in the cloud Service Legal Agreement(SLA):** As cloud services involves more than one organisation, responsibilities of user and the service provider must be made clear in SLA for better understanding. Educational institution should double check the terms in the SLA.

10) **Step 10**: **Understand the security requirements of the exit process**: Exit process must allow the educational institution to retrieve their data in a suitable secured form. It includes clarity on backup retention and deletion.

B. Seven Steps for building security in the cloud
This is a helpful guide designed for IT managers in ensuring best practices to follow in the cloud in order to help in building the security in the cloud.

Fig. 5: Seven Steps of security for Cloud Computing

1) Step 1: Start security planning early: The best way to approach cloud security is to integrate it with overall cloud planning early in the process. By this way, educational institution can use a threat based approach to planning for deployments of their specific workloads, security requirements and specific cloud delivery model and architecture.

2) Step 2: Identify vulnerabilities for selected services: It is the responsibility of the educational institution to identify the vulnerabilities for the selected services in the cloud computing. It is also important to understand while a fill-the-gap approach may seem to work on a particular vulnerability, but it may expose the unknown vulnerabilities in other areas. Best approach is to review the specific service architecture and then layer technologies to develop a strong security net that protects data, applications and platform and network at all levels irrespective of chosen cloud model

3) Step 3: Four things to mitigate security vulnerability: Four things an IT manager can do to mitigate security vulnerabilities in cloud based e-learning. Intel recommends prioritizing security investment through a risk assessment to determine the order and timing for building this level of trust and compliance into cloud ecosystem in four areas as follows:-

- Encrypt to protect data that rests or moves in the cloud especially in public clouds.
- Establish a trusted foundation to secure educational institution data center platform and infrastructure and protect clients.
- Build higher assurance into compliance to streamline auditing.
- Establish and verify identities before educational institution federate by controlling access from trusted cloud users in e-learning and trusted systems.

4) Step 4: Protect data: Protect data in motion, in process and at rest. Encrypt the data wherever it is in the cloud: at rest, in process, or in motion. Since, data doesn‟t stay in one place on any network and this is especially true in case of data in the cloud based e-learning.

5) Step 5: Secured platform: Securing both client and server platforms in cloud based e-learning are very important as there is increase trend in malware threats. It will facilitate the additional enforcement point which builds trust between servers and between servers and clients. The best to start is with a trusted hardware based root and extend the chain through the critical controlling software layers.

6) Step 6: Extend trust across federated clouds: Additional layer of complexity to the current security equation should be added in cloud based e-learning as it evolves the vision of federated cloud relationships across several cloud infrastructure. Managing identities and access-management policies including standards-based single sign-in (SSO), strong authentication, account provisioning, API security and audit capability can built the trusted access to the cloud and across clouds. Simple usernames and passwords are not adequate for cloud security since it can be easily compromised. In federated cloud environment, strong second-factor authentication is essential for secure SSO.

7) Step 7: Choose the right cloud service provider: Choosing the right cloud service provider is tedious process as it involves many levels from the cloud delivery model and architecture to specific applications. In addition to that the countless interdependencies and relationships are exists among the vendors both in terms of technological and business related. Cloud users in e-learning needs to know about the data and platform protections for the services they offered.

C. Checklist for Cloud Security[17]
Microsoft has given the checklist for IT managers to ensure the security in cloud based e-learning as follows:-

1) Integration : Integration points needs to be checked with the security and identity management technologies currently available in the educational institution such as active directory, controls for role-based access and entity-level applications.

2) Privacy: Educational institution should make sure that cloud service includes data encryption, effective data anonymization and mobile location privacy.

3) Access: Educational institutions should aware of the means of preventing inadvertent access when the resources are placed in a shared cloud infrastructure. Cloud Provider‟s policy on accidental release of protected data must be carefully read by the educational institutions.

4) Jurisdiction: The location of a cloud provider‟s operation can affect the privacy laws that apply to the data it hosts. Educational institutions need to check the data whether it is to be reside within their legal jurisdiction.

REFERENCES

[1] D.Kasi Viswanath, S.Kusuma and Saroj Kumar Gupta,[July 2012] “Cloud Computing Issues and Benefits Modern Education”, Global Journal of Computer Science and Technology Cloud & Distributed., Vol. 12 Issue 10 Version 1.0 pp.15-19.

[2] Md. Anwar Hossain Masud, Xiaodi Huang[2012], “An E-learning System Architecture based on Cloud Computing” ,World Academy of Science, Engineering and Technology

[3] Paul POCATILU [2010], “Cloud Computing Benefits for E-learning Solutions”, O economics of Knowledge, Vol. 2, Issue 1, 1Q.

[4] A. Fern´andez, D. Peralta, F. Herrera2, and J.M. Ben´ıtez, [2012] “An Overview of E-Learning in Cloud Computing”,Available:http://sci2s.ugr.es/publications/ficheros

[5] Cloud Computing,[Online] [2012] Available:http://en.wikipedia.org/wiki/Cloud_computing accessed on November 2012.

[6] Ajith Singh. N, M. Hemalatha, [2012] “Cloud Computing for Academic Environment”,International Journal of Information and Communication Technology Research,Vol. 2 No. 2, Feb.

[7] Bhruthari G. Pund*,Prajakta P. Deshmukh, “Appliance of Cloud Computing on E-Learning” International Journal of Computer Science and Management Research, Vol. 1 Issue 2 Sep.2012.

[8] Ivan I Ivanov, [“Cloud Computing in Education: The Intersection of Challenges and Opportunities”

[9] Ahmed E. Youssef, [2012] “Exploring Cloud Computing Services and Applications”,Journal of Emerging Trends in Computing and Information Sciences,Vol. 3, No. 6, July 2012.

[10] Mervat Adib Bamiah & Sarfraz Nawaz Brohi,[2011] “Seven Deadly Threats and Vulnerabilities in Cloud Computing”, International Journal Of Advanced Engineering Sciences And Technologies, Vol No. 9, Issue No. 1, pp.087 – 090

[11] Kangchan Lee Electronics and Telecommunications Research Institute chan@etr.re.kr

[12] Joshi Ashay Mukundrao (2012) Enhancing Security in Cloud Computing. D.Y. Patil College Of Engineering, Akurdi, Pune University of Pune, Maharashtra, India

[13] S.Hameetha Begum* T.Sheeba S.N.Nisha Rani (2013). Security in Cloud based E-Learning Computing. Volume 3, Issue 1, January 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com &Muscat College Computing &Muscat College ECE &Fatima Michael Engg. College Oman Oman Tamilnadu, India

[14] Kevin Hamlen, Murat Kantarcioglu,Latifur Khan,Bhavani Thuraisingham [2010]. Security Issues for Cloud Computing International Journal of Information Security and Privacy, 4(2), 39-51, April-June 2010

[15] IBM Global Technology Services Technical White Paper June 2011 Security and high availability in cloud
computing environments

[16] BM Global Technology Services Technical White Paper June 2011 Security and high availability in cloud computing environment

[17] Assessing Cloud Node Security Context Information Security whitepapers@contextis.com March 2011

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Soft Computing Techniques for Weather and Climate Change Studies

A.B. Adeyemo
Department of Computer Science
University of Ibadan
Ibadan, Nigeria
sesanadeyemo@gmail.com

ABSTRACT

The Weather is a continuous, data-intensive, multidimensional, dynamic process that makes weather forecasting a formidable challenge. Weather forecasting involves predicting how the present state of the atmosphere will change. Climate is the long-term effect of the sun's radiation on the rotating earth's varied surface and atmosphere. The Day-by-day variations in a given area constitute the weather, whereas climate is the long-term synthesis of such variations. A simple, long-term summary of weather changes, however, is still not a true picture of climate. To obtain this requires the analysis of daily, monthly, and yearly patterns. This paper presents the use of Self Organizing Maps (SOM) and Co-Active Neuro Fuzzy Inference System (CANFIS) soft computing techniques for weather and climate change studies using historical data collected from the city of Ibadan, Nigeria between 1951 and 2009. The results show that the soft computing techniques can be used for knowledge discovery in weather prediction and climate change studies.

Keywords: Weather Forecasting, Climate Change, Self Organizing Maps (SOM), Co-Active Neuro-Fuzzy Inference System (CANFIS)

African Journal of Computing & ICT Reference Format:
A.B. Adeyemo (2013).Soft Computing for Weather and Climate Change Studies.
Afr J. of Comp & ICTs. Vol 6, No. 2. pp 77-.90.

1. INTRODUTION

Weather can be described as the state of the atmosphere at a given time and place. Most weather activities takes place in the troposphere which is the lowest layer of the atmosphere. Weather is measured and described in a variety of ways by meteorologists who are scientists that study and predict weather. Weather forecasting entails predicting how the present state of the atmosphere will change. It has been one of the most scientifically and technologically challenging problems around the world in the last century. This is due mainly to two factors: first, it's used for many human activities and secondly, due to the opportunism created by the various technological advances that are directly related to this concrete research field, like the evolution of computation and the improvement in measurement systems (Casas et al., 2009).

To make an accurate prediction is one of the major challenges facing meteorologist all over the world. Since ancient times, weather prediction has been one of the most interesting and fascinating domain. Scientists have tried to forecast meteorological characteristics using a number of methods, some of these methods being more accurate than others (Elia, 2009).

Some methods of forecasting are:

- Persistence Method: This method assumes that the conditions at the time of the forecast will not change. For example, if it is sunny and 67degrees today, the persistence method predicts that it will be sunny and 67 degrees the next day. The method works well when weather patterns change very little and features on the weather maps move very slowly. However, if weather conditions change significantly from day to day, the persistence method usually breaks down and is not the best forecasting method to use although it's the simplest method. While it is assumed that it only works for shorter-term forecasts (e.g. a forecast for a day or two), actually one of the most useful roles of the persistence forecast is predicting long range weather conditions or making climate forecasts
- Climatology Method: This method involves averaging weather statistics accumulated over many years to make the forecast. For example to predict the weather for a city on a particular day involves assembling all the weather data that has been recorded for that day over the years and taking its average. This method works well only when the weather pattern is similar to that expected for the chosen time of

year. If the pattern is unusual for the given time of year, the climatology method will often fail

- Analog Method: This method is a slightly more complicated method of producing a forecast. It involves examining the days forecast scenario and remembering a day in the past when the weather scenario looked similar (an analog). The forecaster would predict that the weather in this forecast will behave the same as it did in the past. The method is difficult to use because it is virtually impossible to find a perfect analog.
- Numerical Weather Prediction: This method uses the power of computers to make a forecast. Complex computer programs, also known as forecast models that run on supercomputers provide predictions on many atmospheric variables such as temperature, pressure, wind, and rainfall. A forecaster examines how the features predicted by the computer will interact to produce the day's weather. Numerical Weather Prediction models are complex mathematical models and require a lot of computational power to solve these equations. Also the time taken to produce results limits the methods ability to provide very short-term forecasts. Even though the method may be costly and time consuming, it provides more accurate results for forecasts of both short and longer time steps ranging from one hour and beyond.

Modern weather forecasts are made by obtaining present weather conditions by ground observations, observations from ships and aircraft, radiosondes, doppler radar, and satellites. This information is sent to meteorological centers where the data is collected, analyzed, and a variety of charts, maps, and graphs are plotted using high-speed computers which are then used to develop surface and upper-air maps with the help of meteorologists who correct the maps for errors. The computer is also used to predict the future state of the maps. Climate is the long-term effect of the sun's radiation on the rotating earth's varied surface and atmosphere.

The Day-by-day variations in a given area constitute the weather, whereas climate is the long-term synthesis of such variations. Weather is measured by thermometers, rain gauges, barometers, and other instruments, but the study of climate relies on statistics which can be handled efficiently by computers. A simple, long-term summary of weather changes, however, is still not a true picture of climate. To obtain this requires the analysis of daily, monthly, and yearly patterns (Fairbridge, 2007).

Climate change is a significant and lasting change in the statistical distribution of weather patterns over periods ranging from decades to millions of years. It may be a change in average weather conditions or the distribution of events around that average (e.g., more or fewer extreme weather events). The term is sometimes used to refer specifically to climate change caused by human activity, as opposed to changes in climate that may have resulted as part of Earth's natural processes. Climate change today is synonymous with anthropogenic global warming. While the term global warming has been used to refer to surface temperature increases, climate change includes global warming and everything else that increasing greenhouse gas amounts will affect (Wikipedia, 2011).

Weather prediction and Climate change/prediction problems differ with respect to the time scale being considered. While weather predictions are on a much shorter time scale such as next few hours, days or weeks, Climatic forecasts are on a longer time scale such as hundreds or thousands of years. Therefore a variation of 5^oC change in temperature from one day to the next day is not as significant as a 5^oC in global climatic temperature. The mathematical models for explaining and predicting future weather and climatic conditions are complex non-linear dynamical systems which are currently being processed with the aid of powerful super computer systems running massively parallel algorithms. However, a lot of historical data whose origins coincides with the advent of modern weather forecasting, which started with the invention of the barometer in 1644, has been collected over the years. With this database of weather/climatic data available, data mining techniques which have proven to be efficient at solving complex non-linear problems can be applied to both weather prediction and climate change/forecasting problems.

In this work rainfall and weather data for the city of Ibadan were analyzed using Self Organizing Maps (SOM) and Co-Active Neuro Fuzzy Inference System (CANFIS) soft computing techniques for knowledge discovery and prediction of weather parameters such as wind-speed, sunshine irradiation, minimum and maximum temperature, and rainfall intensity using historical data.

2. SELF-ORGANIZING MAPS

Clustering groups data into sets in such a way that the intra-cluster similarity is maximized while the inter-cluster similarity is minimized (Zengyou et al, 2003). Self-Organizing Maps (SOM) are competitive networks that provide a "topological" mapping from the input space to the clusters (Kohonen, 1999). The SOM working algorithm is a variant of multidimensional vectors clustering of which the Kmeans clustering algorithm is an example of this type of algorithm (Statsoft, 2002).

Competitive learning is an adaptive process in which the neurons in a neural network gradually become sensitive to different input categories, sets of samples in a specific domain of the input space. A kind of a division of labor emerges in the network when different neurons specialize to represent different types of inputs. The specialization is enforced by competition among the neurons: when an input $\mathbf{x}$ arrives, the neuron that is best able to represent it wins the competition and is allowed to learn it even better, as described in Kohonen (1999).

If there exists an ordering between the neurons, i.e., the neurons are located on a discrete lattice, the self-organizing map, the competitive learning algorithm can be generalized: if not only the winning neuron but also its **neighbors** on the lattice are allowed to learn, neighboring neurons will gradually specialize to represent similar inputs, and the representations will become **ordered** on the map lattice. This is the essence of the SOM algorithm (Kaski, 1997). The neurons represent the inputs with reference vectors $\mathbf{m}_i$, the components of which correspond to synaptic weights. One reference vector is associated with each neuron called **unit** in a more abstract setting. The unit, indexed with *c*, whose reference vector is nearest to the input $\mathbf{x}$ is the winner of the competition:

$$c = c(\mathbf{x}) = \arg\min_i \{\|\mathbf{x} - \mathbf{m}_i\|^2\} \quad \ldots\text{Eq 1}$$

Usually Euclidean metric is used, although other choices are possible as well. The winning unit and its neighbors adapt to represent the input even better by modifying their reference vectors towards the current input. The amount the units learn will be governed by a neighborhood kernel *h*, which is a decreasing function of the distance of the units from the winning unit on the map lattice.

If the locations of units *i* and *j* on the map grid are denoted by the two-dimensional vectors $\mathbf{r}_i$ and $\mathbf{r}_j$, respectively,

then $h_{ij}(t) = h(\|\mathbf{r}_i - \mathbf{r}_j\|; t)$,

where *t* denotes time.

During the learning process at time *t* the reference vectors are changed iteratively according to the following adaptation rule, where $\mathbf{x}(t)$ is the input at time *t* and $c = c(\mathbf{x}(t))$ is the index of the winning unit:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)] \quad \ldots\text{Eq 2}$$

In practice the neighborhood kernel is chosen to be wide in the beginning of the learning process to guarantee global ordering of the map, and both its width and height decrease slowly during learning. The learning process consisting of winner selection by Equation 1 and adaptation of the synaptic weights by Equation 2, can be modeled with a neural network structure, in which the neurons are coupled by inhibitory connections (Kaski, 1997)

A problem with the clustering methods is that the interpretation of the clusters may be difficult. Most clustering algorithms prefer certain cluster shapes, and the algorithms will always assign the data to clusters of such shapes even if there were no clusters in the data. Therefore, if the goal is not just to compress the data set but also to make inferences about its cluster structure, it is essential to analyze whether the data set exhibits a clustering tendency. The results of the cluster analysis need to be validated, as well (Kaski, 1997). Another potential problem is that the choice of the number of clusters may be critical: for example in Kmeans clustering different kinds of clusters may emerge when *K* is changed. Good initialization of the cluster centroids may also be crucial; some clusters may even be left empty if their centroids lie initially far from the distribution of data. Clustering can be used to reduce the amount of data and to induce a categorization. In exploratory data analysis, however, the categories have only limited value as such. The clusters should be illustrated somehow to aid in understanding of what they are like (Kaski, 1997). The SOM is a special case in that it can be used at the same time both to reduce the amount of data by clustering, and for projecting the data nonlinearly onto a lower-dimensional display.

Generally, standard SOMs are applied to feature values of numeric type which usually uses an Euclidean function to calculate the distances between input vectors and reference vectors. During the learning, the update of reference vectors is performed by incremental or arithmetic operations. Unfortunately, these calculations are not practical on categorical values. Although categorical data has been discussed in some clustering algorithms, it is not directly addressed in SOMs due to the limitation of learning laws. A traditional approach is to translate categories to numeric numbers in data preprocess and then perform standard SOMs on the transformed data (Chen and Marques, 2005).

2.1 Co-Active Neuro Fuzzy Inference System (CANFIS)

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a huge number of highly interconnected artificial neurons working in unison to solve specific problems.

The artificial neuron is an information processing unit that is fundamental to the operation of a neural network. ANNs, like people, learn by example. An ANN is configured for a particular application, such as pattern recognition or data classification, through a learning process. Fuzzy logic is a form of logic used in systems where variables can have degrees of truthfulness or false hood represented by a range of values between 1 (true) and 0 (false). With fuzzy logic, the outcome of an operation can be expressed imprecisely rather than as a certainty. For example, instead of being true or false, an outcome might have such meaning as probably true, possibly true, and probably false.

A hybrid neuro-fuzzy system is a fuzzy system that uses a learning algorithm based on gradients or inspired by the neural networks theory (heuristical learning strategies) to determine its parameters (fuzzy sets and fuzzy rules) through the patterns processing (input and output). A neuro-fuzzy system can be interpreted as a set of fuzzy rules. This system can be totally created from input output data or initialized with the *à priori* knowledge (fuzzy rules). The resultant system by fusing fuzzy systems and neural networks has as advantages of learning through patterns and the easy interpretation of its functionality. There are several different ways to develop hybrid neuro-fuzzy systems; therefore, being a recent research subject, each researcher has defined its own particular models.

These models are similar in its essence, but they present basic differences. Examples of these systems include: Fuzzy Adaptive learning Control Network (FALCON), Adaptive Neuro Fuzzy Inference System (ANFIS), Generalized Approximate Reasoning based Intelligent Control (GARIC), Fuzzy Inference and Neural Network in Fuzzy Inference Software (FINEST), FUzzy Net (FUN), Evolving Fuzzy Neural Network (EFuNN), Self Constructing Neural Fuzzy Inference Network (SONFIN) and Neuro-Fuzzy Control (NEFCON), (Abraham and Baikunth, 2000).
ANFIS is perhaps the first integrated hybrid neuro-fuzzy model and belongs to the class of rule-extracting systems using a decompositional strategy, where rules are extracted at the level of individual nodes within the neural network. After extraction, rules are aggregated to form global behavior descriptions (Zanchettin et al, 2010). The ANFIS architecture (figure 1) consists of a five-layer structure where in the first layer, the node output is the degree to which the given input satisfies the linguistic label associated to the membership functions.

The parameters in the first layer are referred to as premise parameters. In the second layer, each node function computes the firing strength of the associated rule. In general, any T-norm operators that perform fuzzy AND can be used as the node function in this layer. Each node i in third layer calculates the ratio of the ith rule firing strength for the sum of firing strength of all rules. The fourth layer is the product of the normalized firing level and the individual rule output of the corresponding rule. Parameters in this layer are referred to as consequent parameters. The single node function of the fifth layer computes the overall system output as the sum of all incoming signals. Only Layer 1 and Layer 4 contain modifiable parameters. (Zanchettin et al, 2010)

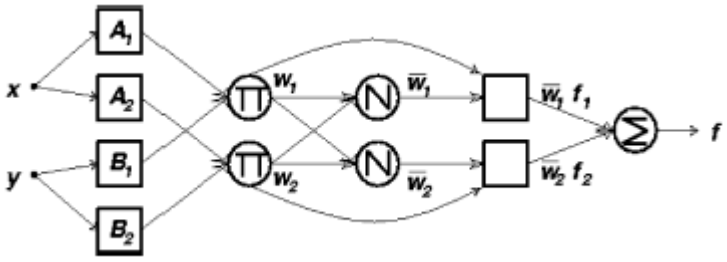


Figure 1: ANFIS system (implementing Tsukamoto fuzzy inference system

The architecture of ANFIS is a one-output fuzzy inference system based on an adaptive network. CANFIS is a generalized form of ANFIS. CANFIS allows more than one output with the advantage of non-linear rule formations. The CANFIS model (figure 2) integrates fuzzy inputs with a modular neural network to quickly solve poorly defined problems (Heydari and Talaee, 2011).

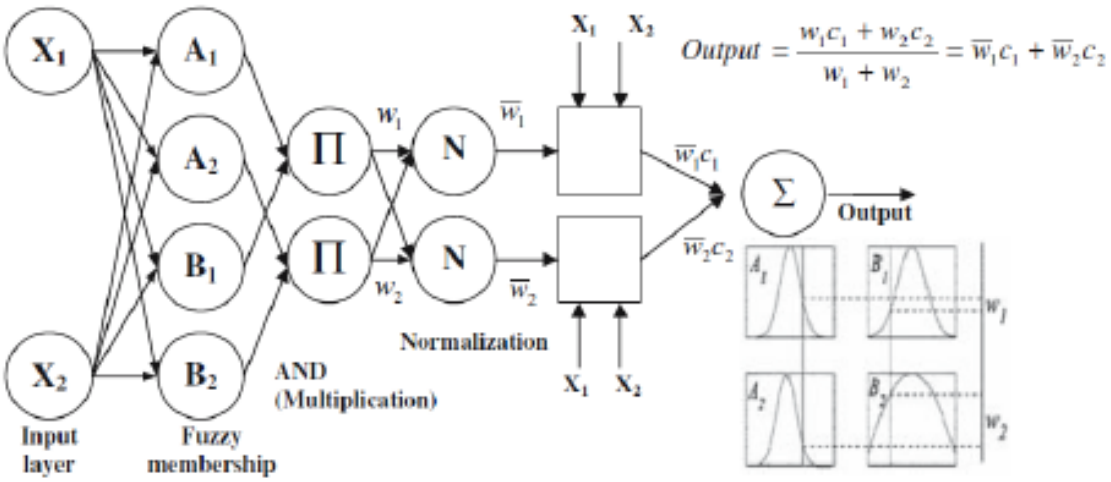


Figure 2: The CANFIS system

The fundamental concepts of CANFIS can be found in Heydari and Talaee (2011), and in J.S.R. Jang, C.T. Sun, and E. Mizutani in Neuro-Fuzzy and Soft Computing, published by Prentice Hall in 1997 where it was originally proposed. CANFIS also uses a five layer modular network structure that uses two fuzzy structures: the Tsukamoto model and the Takagi–Sugeno–Kang (TSK) model and uses the Bell and Gaussian Membership Functions. The CANFIS model is more computationally intensive than most other models.

3. MATERIALS AND METHODS

For knowledge discovery from the rainfall data using SOM clustering, the rainfall data from Ibadan, Oyo State, Nigeria, covering the period between 1951 and 2008 was used. The rainfall data was clustered using NeuroXL Clusterize software. The attributes of the dataset are: year, month and volume of rainfall in millimeters. This is presented in table 1.

Table 1: Attributes of Ibadan Rainfall Dataset

Attribute	Type	Data Coding	Description
Year	Numeric	Numeric (Ordinal)	Year considered e.g. 1951, 1952, 1953 etc
Month	Categorical	Numeric (Ordinal)	Month considered e.g. Jan, Feb, Mar etc
Rainfall	Numeric	Numeric (Continous)	Monthly rainfall (mm)

Clusters were generated and the arithmetic mean and standard deviation of each cluster group was computed. As a measure of central tendency the mean describes the central location of data (that is the center of gravity of the data), it is usually used with other statistical measures such as the standard deviation because it can be affected by extreme values in the data set and therefore be biased. The standard deviation describes the spread of the data and is a popular measure of dispersion. It measures the average distance between a single observation and its mean. For weather prediction, the weather data was collected from Ibadan, Oyo State, Nigeria from Ibadan Synoptic Airport through the Nigerian Meteorological Agency, Oyo State office which covered a period of 120 months from January 2000 to December 2009 (Folorunsho and Adeyemo, 2012) was used. The data attributes of the weather dataset are presented in table 2.

Table 2: Attributes of Meteorological Dataset

Attribute	Type	Description
Year	Numeric	Year considered
Month	Numeric	Month considered
Wind speed	Numeric	Wind run in km
Radiation	Numeric	The amount of solar irradiation
MinTemp	Numeric	The monthly Minimum Temperature
MaxTemp	Numeric	Maximum Temperature
Rainfall	Numeric	Total monthly rainfall

The CANFIS program was used for the weather parameters prediction program because CANFIS can support multiple output variables, while ANFIS supports one output variable. The NeuroSolutions implementation of the CANFIS program was used. NeuroSolutions incorporates a number of standard parameters which can be used to evaluate the performance of a neural network model. Those relevant to the study as presented in NeuroSolutions are:

a. Mean Squared Error: The mean squared error (MSE) and the normalized mean squared error (NMSE) can be used to determine how well the network output fits the desired output. It however does not necessarily reflect whether the two sets of data move in the same direction. The mean squared error is simply two times the average cost, For instance, by simply scaling the network output, the MSE can be changed without changing the directionality of the data.

b. The correlation coefficient (r): The value of the correlation coefficient ranges between [-1, 1]. When r =1 there is a perfect positive correlation. When r = -1, there is perfectly linear negative correlation. When r =0 there is no correlation between the two quantities investigated. Intermediate values describe partial correlations.

c. Learning curve: The learning curve that shows how the mean square error evolves with the training iteration is a quantity that can be used to check the progress of learning. The difficulty of the task and how to control the learning parameters can be determined form the learning curve. When the learning curve is flat, the step size is increased to speed up leaning, when the learning curve oscillates up and down the step size is deceased.

In the extreme, the error curve increases steadily upwards, showing that learning is unstable, at this point the network is reset.

4. RESULTS AND DISCUSSION

4.1 Rainfall data clustering using SOM

Three SOM software used were used and the one the gave the best result was selected. These are: NeuroXL Clusterizer, a proprietary Artificial Neural Network clustering software implemented in Microsoft Excel; NNClust, a SOM software implemented in Microsoft Excel and Pittnet Neural Network Educational Software, an open source SOM software coded using C++.

The NClust SOM clustering software was trained using a starting learning rate of 0.9 and was trained over 100 epochs. The software was programmed to normalize (scale) the data automatically between 1 and -1, and only works with numeric values. Non numeric values are treated as missing values which are replaced by the column mean. The software was set to use a square SOM grid of 10x10, which also determines the maximum number of clusters to be generated. The software generated only nine clusters from the dataset and increasing the training cycle did not improve the results.

The Pitnett Neural Network software was trained using a starting learning rate of 0.9 and was set to train over 100 epochs, although the software stops training as soon as the maximum number of clusters have been generated. The software was programed to normalize (scale) the data automatically between 0 and 1. The network requires the number of clusters expected to be specified apriori. This number is used in conjunction with the number of input signals to determine the SOM grid size. This number was set to twenty. The software generated 18 clusters, and the mean and standard deviation of these clusters were computed. Table 4 presents the analyses of the 18 clusters, while figure 5 presents the chart of the clusters and rainfall average.

Table 4: Analysis of the 18 clusters from Pitnett

SNo	Cluster	Year	Month	Min Rainfall	Max Rainfall	Rainfall Average	Rainfall SD
1	1	1971 - 1978	3,4	199.8	350.1	270.6333313	75.47989326
2	2	1975 - 1994	8,9,10,11,12	58.5	235.8	172.5448282	53.81027918
3	4	1974 - 1996	4,5,6,7,8	156.6	312.4	240.9055549	44.76238685
4	5	1989 - 2008	6,7,8,9,10	146.1	377	232.3911102	54.96919143
5	6	1951 - 1972	4,5,6,7	237	399.8	310.8666662	60.64680031
6	7	1977 - 2008	8,10,11,12	0	71.5	18.40169495	23.75895108
7	8	1951 - 1976	8,9,10,11,12	0	152.4	42.46027352	44.08310212
8	9	1965 - 1998	8,9,10,11,12	0	195.8	91.58064508	58.98524609
9	10	1963 - 1999	8,9,10	255.5	415	314.3277808	50.33098414
10	11	1964 - 1960	9,10	266.4	324.8	298.0666707	29.36363096
11	12	1982 - 2008	1,2,3,4,5	7	194	46.51124946	54.05845832
12	14	1997, 2003	4,6	242.6	343.5	293.0999985	71.27636566
13	15	1951 - 1975	1,2,3,4,5,6,7	0	129.8	30.90303033	34.458354
14	16	2000 - 2008	4,5	129.9	206.3	173.5333337	24.30319025
15	17	1951 – 1972	7,8,9,10	51.1	256.7	167.2795464	65.28443524
16	18	1992 – 2008	5,6,7,8,9,10	9	215.6	111.080001	50.7121805
17	19	1952 – 1972	5,6,7,8	118.1	289.1	209.7083346	52.38667157
18	20	1951 – 1974	3,4,5,6,7	28.4	209.8	151.3382366	41.06354325

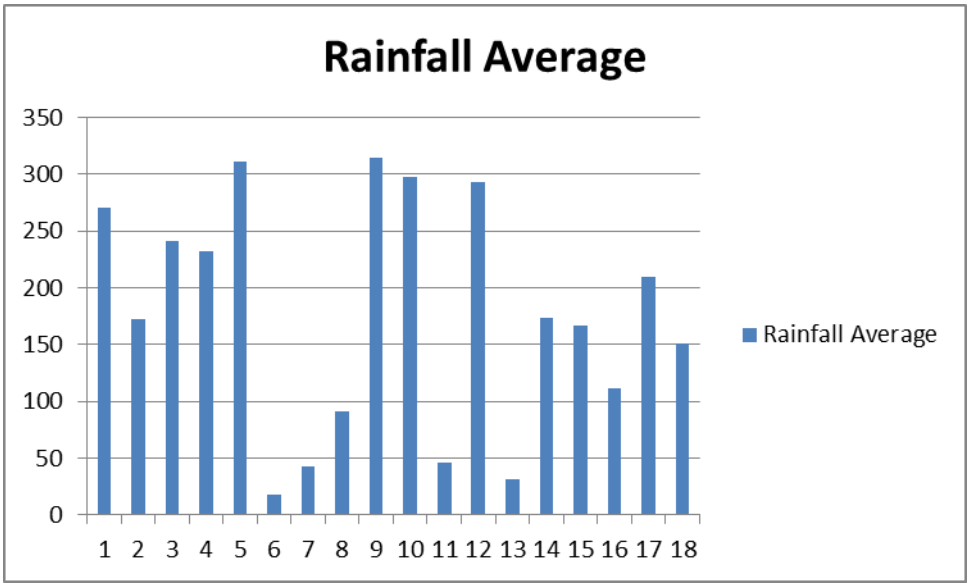


Figure 5: Chart of clusters and rainfall average from Pitnett

The NeuroXL Clusterizer network was trained using a starting learning rate of 0.9 and was trained over 100 epochs. The software can be set to either scale (normalize) the data or not. The number of clusters to be retrieved from the data can also be set and different values were used. The mean and standard deviation of these clusters were computed. From the results the clustering run which gave 20 clusters was selected, since it gave clusters which had the minimum standard deviation values. The analyses of the twenty clusters are presented in table 5 while figure 6 shows a chart of the clusters and rainfall average.

Table 5: Analysis of the twenty clusters from NeuroXL Clusterizer

Clusters	Years	Months	Min Rainfall	Max Rainfall	Rainfall Av	Rainfall SD
1	1970 -1994	1,2,3,8,9.11.12	13.6	34.8	21.91111111	5.623599541
2	1994 – 2008	1,2,3	0	15.1	3.02962963	4.993288943
3	1994 – 2008	11,12	0	9	1.077777778	2.664042827
4	1976 – 1993	1,2,4,11,12	0	10.2	1.523809524	2.809288559
5	1951 1974	1,2,3,6,7,8,9,11,12	19.3	55.6	34.57692308	10.59335595
6	1951 1979	2,3,4,5,7,8,11	56.4	102.5	76.54565217	14.24529094
7	1951 -1976	1,2,3,11,12	0	18.3	3.980645161	5.451119263
8	2003 -2008	1,3,4,11,12	18.7	36.6	28.47777778	5.92574421
9	1975 – 2006	1,2,3,4,7,8,911,12	37.5	61.1	49.15483871	6.728686782
10	1981 – 2008	2,3,4,5,7,8,11	60.2	84.1	69.84615385	6.213870341
11	1951 – 1982	3,4,5,7,8,9,10	103.1	150.6	127.79375	13.05282893
12	1967 – 2004	2,3,4,5,7,11	83.6	105.9	95.35769231	6.886053911
13	1977 – 2008	3,4,6,7,8,9,10,12	101.6	127.6	112.1684211	7.512881725
14	1982 -2008	4,5,6,8,9,10	124.7	159.5	142.35	10.57428408
15	1951 – 2008	2,3,4,5,6,7,8,9,10	147.8	186.8	168.3053333	10.61280325
16	1952 – 2006	2,4,5,6,7,8,9,10,11	183.3	223.8	203.1868852	10.76423825
17	1951 – 2008	4,5,6,7,8,9,10	224.8	266.8	246.3188679	11.926555
18	1957 – 2008	4,5,6,7,8,9,10	269.3	312.4	288.1685714	12.17958203
19	1960 – 2003	4,5,6,8,9,10	320.2	364.3	340.9583333	14.0279888
20	1951 – 1999	6,7,8,10	363.5	415	385.7625	19.3775966

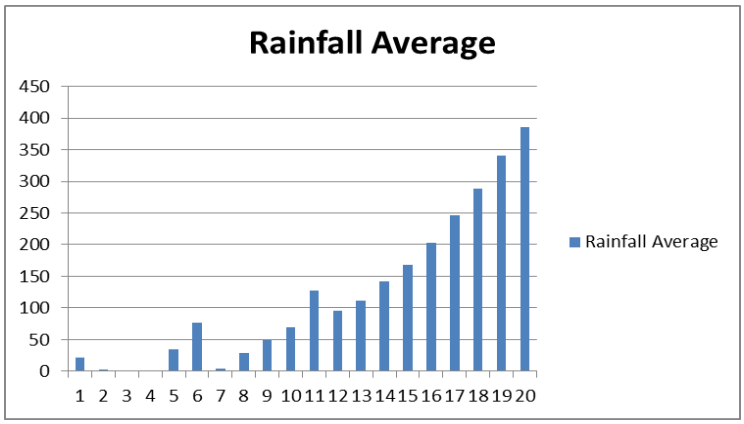


Figure 6: Chart of clusters and rainfall average from NeuroXL Clusterizer

Some known facts about the climatic condition of Ibadan (Wikipedia, 2013) are: Ibadan has tropical wet and dry climate (by the Koppen Climate Classification), it has a lengthy wet season (between March to October) with a break in August, which divides the wet season into two. There are two peaks for rainfall in the wet season which is in June and September. The dry season is from November to February, while the driest month is December (Wikipedia, 2013). The weather condition over the country had been stable relatively between 1971 and 2000, when departures from the normal conditions could be noticed (Nigeria Climate Review, 2010). Some significant events that took place due to excessive rainfall levels were the floods disasters recorded in Ibadan in 1951, 1955, 1960, 1963, 1969, 1978, 1980, 2011 and 2012. Table 3 presents some of the known average weather data (temperature, rainfall, humidity and sunshine) for Ibadan, while figures 3 present a chart of the known average rainfall and precipitation for Ibadan.

Table 3: Climate data for Ibadan (Source: Wikipedia, 2013)

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Record high °C (°F)	37 (99)	39 (102)	38 (100)	38 (100)	35 (95)	33 (91)	31 (88)	31 (88)	36 (97)	33 (91)	34 (93)	35 (95)	39 (102)
Average high °C (°F)	33 (91)	34 (93)	34 (93)	33 (91)	32 (90)	29 (84)	28 (82)	27 (81)	29 (84)	30 (86)	32 (90)	33 (91)	31 (88)
Average low °C (°F)	21 (70)	22 (72)	23 (73)	23 (73)	22 (72)	22 (72)	21 (70)	21 (70)	22 (72)	22 (72)	22 (72)	21 (70)	22 (72)
Record low °C (°F)	10 (50)	12 (54)	18 (64)	18 (64)	18 (64)	18 (64)	16 (61)	16 (61)	17 (63)	18 (64)	14 (57)	14 (57)	10 (50)
Rainfall mm (inches)	8 (0.31)	23 (0.91)	76 (2.99)	125 (4.92)	145 (5.71)	163 (6.42)	132 (5.2)	74 (2.91)	170 (6.69)	152 (5.98)	43 (1.69)	10 (0.39)	1,121 (44.13)
Avg. precipitation days	1	2	5	9	11	12	12	10	15	12	4	1	94
% humidity	76	71	75	78	82	86	88	88	86	84	80	76	81
Mean monthly sunshine hours	170	198	170	170	170	141	85	57	85	141	198	198	1,783

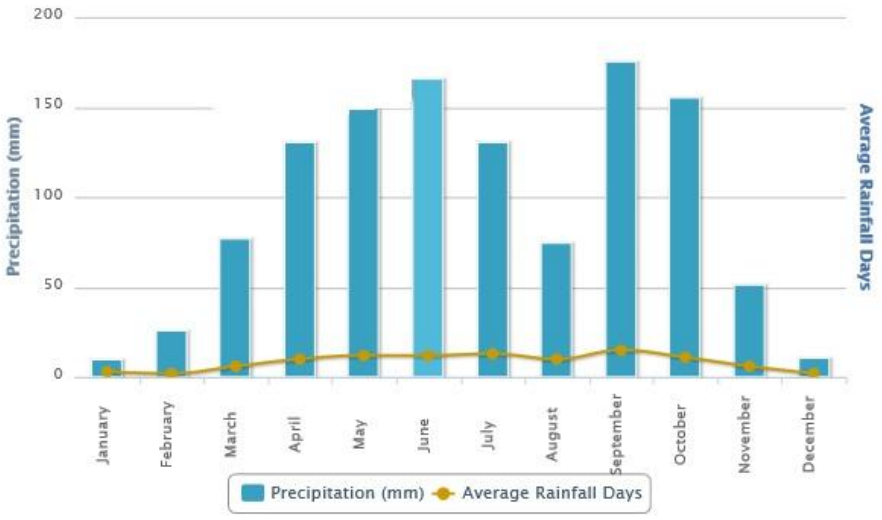


Figure 3: Average Rainfall and Precipitation for Ibadan (Source: http://www.worldweatheronline.com/Ibadan-weather-averages/Oyo/NG.aspx)

The clusters generated by the NeuroXL Clusterizer software gave the best performance because they had the lowest standard deviation values. It was also observed that the clusters represented rainfall intensity bands (groups) which corresponded to some known climatic and weather events in Ibadan between 1951 and 2008. For example the data records in cluster #20 (table 4) contains periods of peak rainfalls recorded in Ibadan, which also resulted in flooding disasters in the town. None of the clusters generated by the other software were able to drill down to this level.

Table 4: Cluster #20 of NeuroXL Clusterizer

Year	Month	Rainfall_mm
1951	7	369.1
1960	6	385.3
1963	8	363.5
1965	6	399.8
1965	7	369.3
1980	8	415
1993	7	377
1999	10	407.1

Also the data in cluster #2, #3, and #7 corresponds to the dry season period. Table 5 shows the data records in cluster #3. Cluster #2 represents dry season months which occurred in January, February and March between 1994 and 2008, while cluster #3 represents dry season months which occurred in November and December between 1994 and 2008. Cluster #7 contains data which represents the dry season months which occurred between November and March, 1951 to 1976.

4.2 Weather prediction using CANFIS

The Neurosolutions CANFIS network used for the weather parameters prediction program had seven input neurons in its input layer and five output neurons in its output layer. The inputs were the year, month, wind-speed, radiation, minimum temperature, maximum temperature and rainfall while the output was the predicted wind-speed, minimum temperature, maximum temperature and rainfall intensity. The network was trained for 1000 epochs (cycles) using Levenberg Marquardt learning. Batch training was used and the training stopping criteria was on increasing MSE on the cross validation samples. The best network result selected used the Takagi–Sugeno–Kang (TSK) model with five (5) Bell shaped membership functions.

Table 5: Cluster #3 of NeuroXL Clusterizer

Year	Month	Rainfall_mm
1994	12	0
1995	12	3.4
1996	11	0
1996	12	0
1999	12	0
2000	11	0
2000	12	0
2001	12	0
2002	12	0
2003	12	0
2004	11	0
2004	12	0
2005	11	0
2005	12	0
2006	12	0
2007	12	7
2008	11	0
2005	8	9

Table 6 presents the network training report, Figure 4 presents the learning curve, table 7 presents the network testing report, figure 5 presents the sensitivity analysis result which shows the input variables contribution to the output result, and table 8 presents the predicted network output for rainfall in year 2010, while figure 6 presents this information in graphical form. Figure 7 shows the column chart (extracted from figure 6) for the predicted rainfall for 2010. It is noted that the rainfall pattern follows the monthly profile shown in figure 3. Figure 8 presents a chart of the historical weather parameter (windspeed, minimum and maximum temperature, and rainfall averages) for Ibadan for 2010 downloaded from www.Tutiempo.net/en (Weather station: 652080; Latitude: 7.43; Longitude: 3.9; Altitude: 228).

Table 6: Network training report

Best Networks	*Training*	*Cross Validation*
Epoch #	1000	46
Minimum MSE	0.014810273	0.046169199
Final MSE	0.014810273	0.078883972

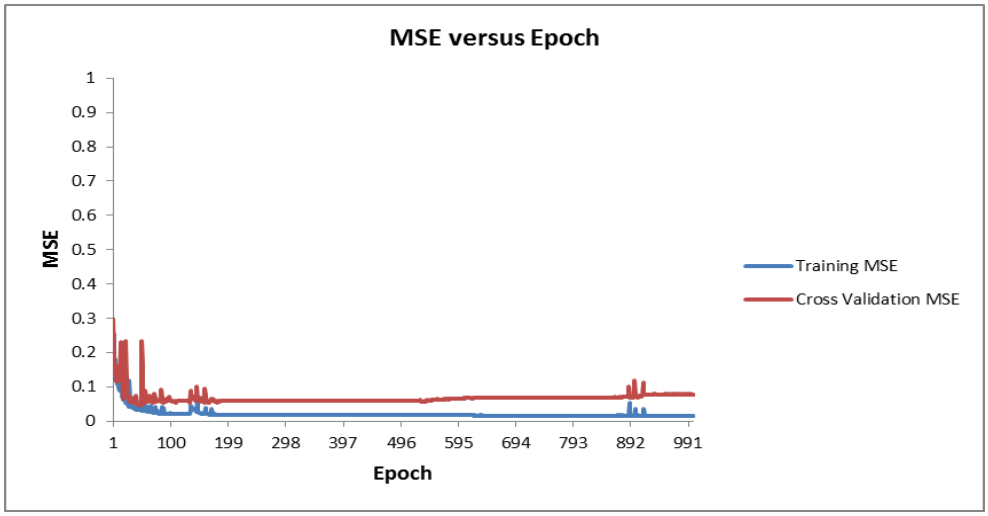


Figure 4: Network learning curve

Table 7: Network testing report

Performance	*Wind*	*Radiation*	*MinTemp*	*MaxTemp*	*Rainfall*
MSE	619.5845617	5.120205621	12.22381097	3.838489961	11595.28926
NMSE	1.972268028	1.37819717	20.5391322	1.108895982	1.07732946
R	0.750344634	0.519604583	-0.299922647	0.543774968	0.712360315

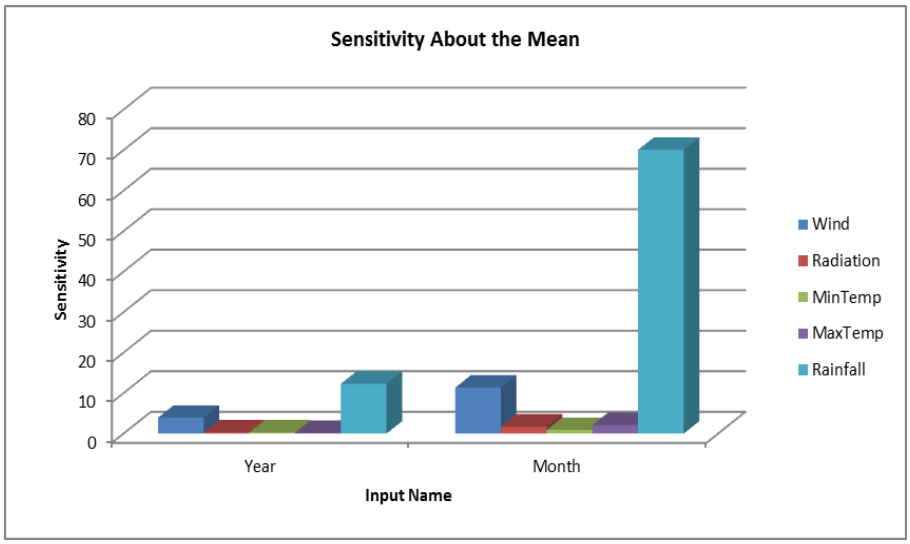


Figure 5: Sensitivity analysis result

Table 8: Predicted network output

Year	Month	Wind	Radiation	MinTemp	MaxTemp	Rainfall
2010	1	149.4943	9.768377	24.98293	33.76771	18.19179
2010	2	161.5787	10.2203	25.55385	33.27231	38.88475
2010	3	206.1897	10.55984	26.82792	32.44958	67.3745
2010	4	173.7567	10.53395	26.93608	31.41716	82.03233
2010	5	156.6988	10.58878	28.79529	29.56042	110.219
2010	6	154.0258	10.36948	30.9407	27.44971	134.1238
2010	7	165.9713	4.785612	30.88352	25.95928	15.42722
2010	8	148.2829	10.60675	25.2678	29.87004	57.98322
2010	9	145.3576	10.31745	21.5661	30.43777	210.909
2010	10	138.3491	8.763394	21.98761	30.14349	174.9529
2010	11	128.6779	13.77576	24.64287	29.83381	-10.7456
2010	12	122.3363	15.84211	25.05434	30.11157	-40.6949

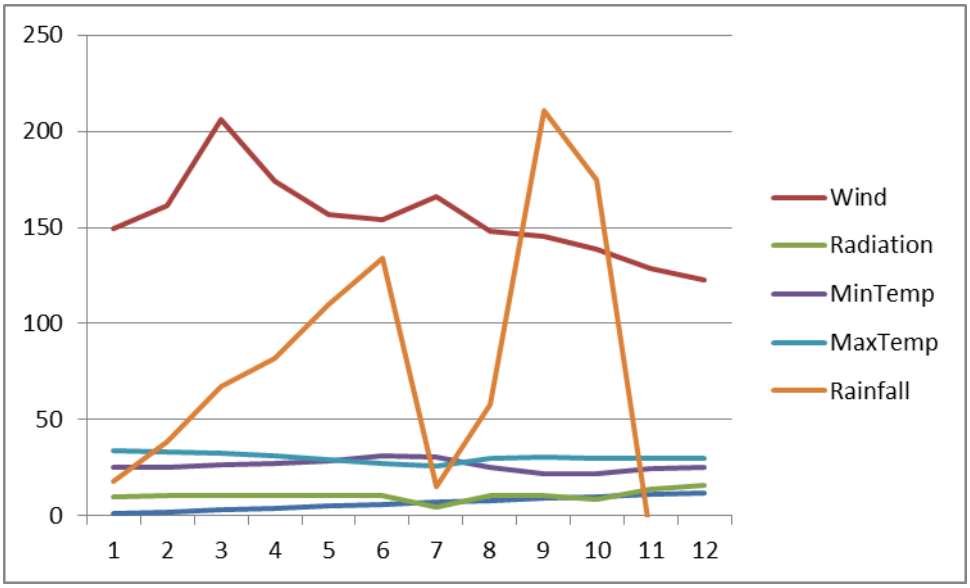


Figure 6: Predicted network output

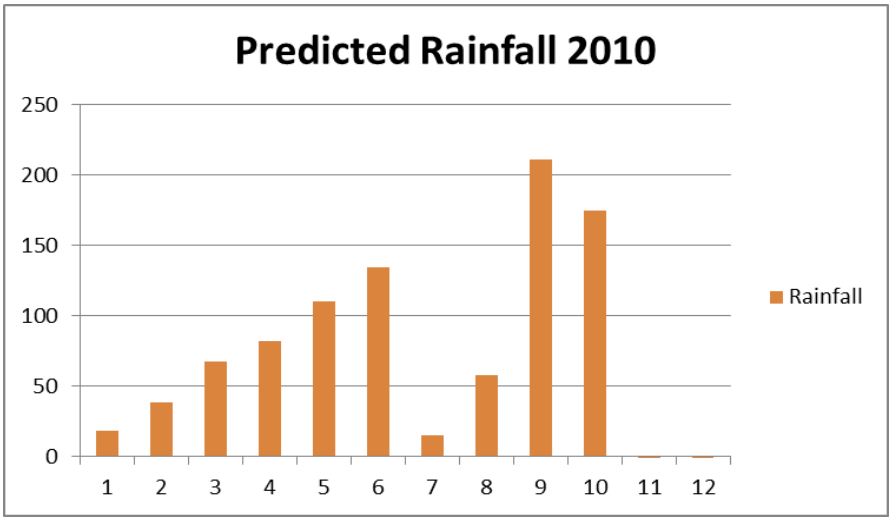


Figure 7: Predicted Rainfall for 2010

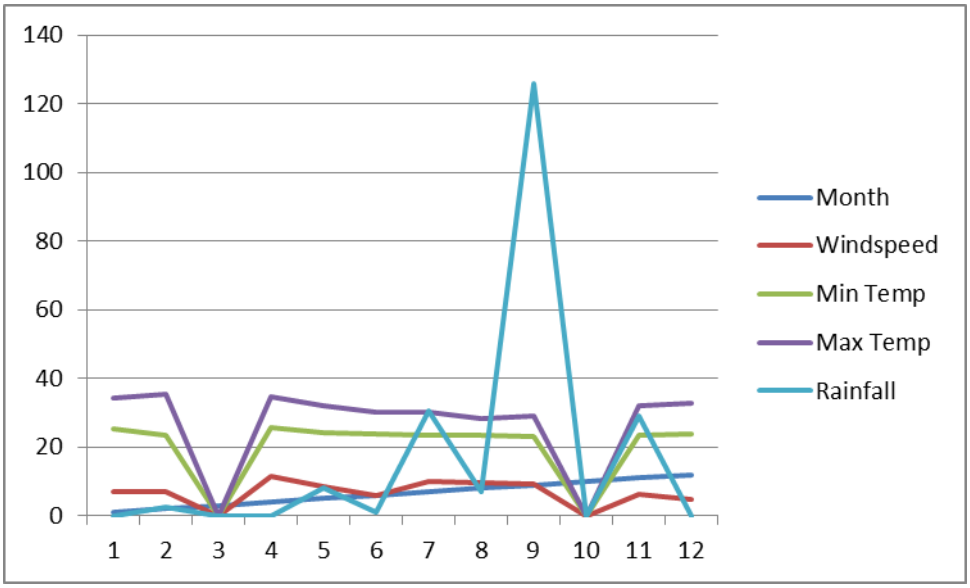


Figure 8: Ibadan average rainfall 2010 (www.Tutiempo.net/en)

5. CONCLUSION

This work presents the use of soft computing techniques (SOM and CANFIS) for knowledge discovery and prediction of rainfall and weather parameters. Clusters were generated from the rainfall data which are representative of some of the known climatic events in the town. The prediction of weather parameters carried out with the CANFIS technique shows that it can be used for long range weather forecasts (in this study of 12 months ahead), on a time scale within the current weather pattern (of within 20 to 30 years interval) of the data set used. A shift in the weather pattern forecasted using historical data may be indicative of climate change activities.

ACKNOWLEDGEMENT

The author acknowledges the contributions of Mrs A. K. Ojo and Mr Olaiya who collected the data used for this work.

REFERENCES

[1] Abraham A. and Baikunth N., (2000), "Designing Optimal Neuro-Fuzzy Systems for Intelligent Contro"l, In Proceedings of The Sixth International Conference on Control, Automation, Robotics and Vision, (ICARCV 2000 - Singapore),(CD ROM Proceeding), Wang J.L. (Editor)

[2] Abraham A., (2004), Adaptation of Fuzzy Inference System Using Neural Learning, Computer Science Department, Oklahoma State University, USA, ajith.abraham@ieee.org, http://ajith.softcomputing.net

[3] Casas D. M, Gonzalez A.T, Rodrígue J. E. A., Pet J. V., 2009, "Using Data-Mining for Short-Term Rainfall Forecasting", *Notes in Computer Science*, Volume 5518, 487-490

[4] Chang C., Ding Z., (2004), "Categorical data visualization and clustering using subjective factors", Data & Knowledge Engineering, Published by Elsevier B.V.

[5] Chen N. and Marques N. C., (2005), "An Extension of Self-Organizing Maps to Categorical Data", Proceedings of the 12th Portuguese conference on progress in Artificial Intelligence , pp 304 - 313, (Sringer-Verlag Berlin, Heidelberg ©2005)

[6] Elia G. P., 2009, "A Decision Tree for Weather Prediction", Universitatea Petrol-Gaze din Ploiesti, Bd. Bucuresti 39, Ploiesti, Catedra de Informatică, Vol. LXI, No. 1

[7] Fairbridge R. W., 2007, "Climate" *Microsoft® Student 2008 [DVD]*, Redmond, WA: Microsoft Corporation, 2007.

[8] Folorunsho O., Adeyemo A. B., (2012), "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", I.J. Information Engineering and Electronic Business, 2012, 1, 51-59, Published Online February 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijieeb.2012.01.07, Copyright © 2012 MECS

[9] Heydari M and Talaee P. H., (2011), "Prediction of flow through rockfill dams using a neuro-fuzzy computing technique", The Journal of Mathematics and Computer Science (TJMCS), Vol.2, No.3 (2011), pp 515-528, Available online at http://www.TJMCS.com

[10] Kaski S., (1997), "Data exploration using self-organizing maps", Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82, Espoo 1997, 57 pp. Published by the Finnish Academy of Technology. ISBN 952-5148-13-0. ISSN 1238-9803. UDC 681.327.12:159.95:519.2

[11] Kohonen T, (1999), "The Self-Organizing Map (SOM)", Helsinki University of Technology, Laboratory of Computer and Information Science, Neural Networks Research Centre, Quinquennial Report (1994-1998), (Downloaded from http://www.cis.hut.fi/research/reports/quinquennial/ January 2006).

[12] Nigeria Climate Review, 2010, Nigerian Meteorological Agency, www.nimetng.org

[13] Pelczer I. J. and Cisneros H. L., (2008), "Identification of rainfall patterns over the Valley of Mexico", 11th International Conference on Urban Drainage, Edinburgh, Scotland, UK, 2008

[14] Statsoft Electronic Statistics Textbook, (2002), Copyright, 1984-2003, (http://www.statsoftinc.com/txtbook/glosd.html #Data Mining), Downloaded June 2002.

[15] Wikipedia, 2011, "Climate change" *From Wikipedia - the free encyclopedia*, retrieved from http://en.wikipedia.org/wiki/Climate_change in August 2011

[16] Wikipedia, (2013), "Rain" *From Wikipedia - the free encyclopedia*, retrieved from http://en.wikipedia.org/wiki/Rain in February 2013

[17] Zanchettin C., Minku L. L., Ludermir T. B., (2010), "Design of Experiments in Neuro-Fuzzy Systems", International Journal of Computational Intelligence and Applications, Vol. 9, No. 2 (2010) 137–152

[18] Zengyou He, Xiaofe I Fe, Shengchun Deng, (2003), "Clustering Mixed Categorical and Numeric Data", Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, P. R. China

Vol 6. No. 2, June 2013
African Journal of Computing & ICT

www.ajocict.net

Option Pricing in a GARCH Framework

[1]D. Allenotor
University of Benin
Benin City, Nigeria
dallen@uniben.edu.ng
+2347056009584

B. Ola
Technobeacon Consulting Ltd
IT Security Director
London, UK
olawest@yahoo.com

R.K. Thulasiram
Dept. of Computer Science
University of Manitoba, R3T 2N2
Winnipeg, Canada
tulsi@cs.umanitoba.ca

[1]***Correspondence Author***

ABSTRACT
There is a compelling need to accurately and efficiently compute option values. Existing literature shows that models based on constant stock volatilities have been widely used in option valuation. However, stock volatilities change constantly in real life situations. The introduction of the Auto Regressive Conditional Heteroskedasticity (ARCH) model and subsequently, the Generalized Auto Regressive Conditional Heteroskedasticity (GARCH) model provides a framework for valuing options using time-varying volatilities. In this paper, we explore the pricing of European styled call options using an analytical approximation of the GARCH option pricing model that is developed in existing literatures.

Keyword – option pricing, financial option, volatility, GARCH.

African Journal of Computing & ICT Reference Format:
D. Allenotor, B. Ola & R.K. Thulasiram (2013). Option Pricing in a GARCH Framework
Afr J. of Comp & ICTs. Vol 6, No. 2. pp 91-104

1. INTRODUCTION

In recent times, option pricing has received a considerable attention in finance literature. Options provide a veritable means, with which investors derive optimum yield from investments. Options can be used to reduce risk of losses (hedging), take risks (speculation), and benefit from arbitrage opportunities [8]. The need to accurately and efficiently compute option values cannot be underscored because inaccurate estimation of option values could result in catastrophic and disastrous consequences.

1.1 Financial Option
A financial option is a contract that gives the right to its holder to exercise the option. That is, a financial option (see, for example [8]) gives the right to buy or sell an asset (for example, a stock) under certain future terms for a given period.

Since the holder has right but without obligations, the option has value. An option is a derivative because its value depends on the value of another asset which is called the underlying asset. Examples of underlying assets are stocks, foreign currencies, stock indices, debt instruments, futures contracts, and commodities.

Two types of options are call options (calls) and put options (puts). A call option gives the option holder the right to buy an underlying asset by a certain date for a certain price. A put option gives the option holder the right to sell an underlying asset by certain date for a certain price. The option holder may decide to use or not to use the option. If the holder decides to buy/sell the underlying asset using the option, it is said that the option is exercised.

The price at which the asset can be bought or sold is called the strike price or exercise price of the option. The date when the contract expires is known as the maturity or expiration date. These two types of options are available in many different styles. A European option can be exercised only on the maturity date, while an American option can be exercised any time up to the maturity date. The writer of the option gets the price of the option or the option premium, when the contract is agreed upon and the writer accepts potential liabilities in the future.

If the asset price S is less than the strike price K of the option, the holder of a call option may not want to exercise the option because the same asset can be bought from the market at a lower price. However, whenever S is greater than K, the call option can be exercised and the holder of the option can make some profit equivalent to $(S-K)$ since the holder can buy the underlying asset for a price at K and sell immediately at price S. The difference between the asset price and strike price at maturity is generally referred to as the pay-off. At any time until maturity, the difference between asset and strike price is called the intrinsic value of a call option. This would be the pay-off if the option was exercised instantly. If C denotes call option and P denotes a put option, then the general formula for the value of a call option C and a put option P at maturity [8] is given as:

$$\left.\begin{aligned} C &= \max(0; S-K) \text{ or} \\ p &= \max(0; K-S) \end{aligned}\right\} \quad (1)$$

Equation (1) implies that higher asset prices give higher values for a call options. For put options, it is just the opposite. The length of time to maturity affects the option value. For example, for an American option, a longer time to maturity increases the option value because the holder of the option has all the exercise opportunities open through the life of the option. For a European option, the effect of time to maturity is ambiguous. However, a longer time to maturity increases the value of a European call option, where there are no dividends during the life of the option.

Models such as the classical Black-Scholes model [1], Binomial trees [17], Trinomial trees [8], Finite difference methods [21], Finite element methods, Multithreaded Algorithms [19], and the L_0 stable algorithm [21] are widely used in computing option values. These methods are premised on the assumption that stocks have constant volatilities. Option prices are very sensitive to changes in volatilities. The use of varying or stochastic volatility based models to compute option prices has started gaining prominence in literature.

In this paper, an algorithm based on the analytical approximation of the GARCH option pricing model [5], for computing European styled option values is presented. The algorithm is implemented in a high level language, results and observations are reported. The contributions of this paper are:

(a) Formulation of a computational algorithm for an analytical approximation of the GARCH option pricing model presented in [5].
(b) Provision of a robust, reliable, and reusable implementation in a high level language for the analytical approximation of the GARCH option pricing model in [5].
(c) This implementation is expected to outperform implementations in Mathematical software such as GAUSS, S-Plus, Maple, Matlab, and Mathematica in execution speed, efficiency, and memory usage.

The rest of this paper is organized as follows. Section 2 provides the related work and background materials for GARCH understanding. In Section 3, we provide an overview of the GARCH model and integrate it with financial option pricing. Section 3 also describes GARCH based computational algorithm. Our experiments are provided in Section 4. The results of our experiments and the discussion of the results are provided in Section 5. Section 6 concludes the paper.

2. BACKGROUND AND RELATED WORK

Volatility is a critical parameter in option valuation. Many financial data series have been observed to exhibit time varying volatilities, which follow a GARCH process. The need to use an accurate estimate of volatility for computing option values cannot be underscored, especially in a market with rapidly and constantly changing underlying asset price system.

The introduction of ARCH [6] model and subsequently GARCH [2] model has provided a veritable framework for estimating conditional variance of time-varying volatilities. Various models of GARCH have appeared subsequently in literature. Some of these models include the Exponential Generalized Auto Regressive Conditional Heteroskedasticity (EGARCH), Linear Generalized Auto Regressive Conditional Heteroskedasticity (LGARCH), Multiplicative Generalized Auto Regressive Conditional Heteroskedasticity (MGARCH), and Nonlinear Asymmetric Generalized Auto Regressive Conditional Heteroskedasticity (NGARCH). A partial taxonomy of these GARCH models and their variance specifications is available in [10].

GARCH models were not applied to option pricing until the first approach in [4]. Duan [4] and most other GARCH option pricing models that appeared subsequently in literature were continuous time based models. However, the continuous time models do not have closed form solutions. Ritchken and Hsieh [11] state that continuous time volatility models are difficult to implement, this is attributable to difficulty in observing and identifying volatilities of continuous-time models. Option values cannot be estimated from continuous time models, since volatility levels cannot be captured [11].

Heston and Nandi [7] introduced a closed form solution of GARCH option pricing model for pricing European styled options. Ritcthken and Trevor [10] developed a closed form lattice based algorithm for pricing discrete time options in the GARCH framework, Cakici and Topyan [3] modified the algorithm introduced by [10]. The lattice based algorithm developed by Ritchken and Trevor [10] is based on some assumptions that deviate sharply from real life situations. For instance, the values of the risk free rate and risk unit premium were assumed to be zero. Moreover, the lattice based algorithm is not amenable and congenial to implementation in a high level language; the approach used in estimating the jump parameter is more intuitive than descriptive. Hence, estimating this jump parameter is intricate and complicated.

Duan et al. [5] developed an analytical approximation of the GARCH option pricing model for pricing European styled options. The NGARCH specification was adopted in the analytical approximation of the GARCH option pricing model. This paper is based on the analytical approximation developed in [5]. The reasons for using [5] to illustrate option pricing in a GARCH framework include: the mathematical expressions derived in the analytical approximation is amenable to implementation in a high level language and the assumptions made in the paper are close to real life situations.

3. OVERVIEW OF THE GARCH OPTION PRICING MODEL

Duan [4] presented a discrete time model for option pricing. Let S_t represent the asset price at time t, r represents the continuously compounded one-period risk free rate of return, λ represents unit risk premium, the stock price dynamics in locally risk-neutralized probability measure Q can be written as:

$$\ln \frac{S_{t+1}}{S_t} = r - \tfrac{1}{2} h_{t+1} + \sqrt{h_{t+1}} \varepsilon_{t+1} \quad (1)$$

$$h_{t+1} = \beta_0 + \beta_1 h_t + \beta_2 h_t (\varepsilon_t - \theta - \lambda)^2 \quad (2)$$

$$\varepsilon_t \Big| \sim N(0,1) \text{ under measure } Q \quad (3)$$

where h_t is the initial conditional variance and h_{t+1} is the conditional variance at time t_{t+1}, θ is the leverage effect parameter, $\beta_0, \beta_1,$ and β_2 are GARCH parameters respectively, $\varepsilon_t = \varepsilon_{t+\lambda}$ is a standard normal variable with mean 0 and variance 1 under the locally risk-neutralized probability measure Q.

The specification of the model above is non-linear asymmetric (NGARCH), with a change in the leverage parameter from θ to $\theta + \lambda.$ The relevant underlying parameters required for computing option prices are: $\beta_0, \beta_1, \beta_2,$ and $\theta + \lambda.$

Whenever the value of $\beta_1 + \beta_2[1 + (\theta + \lambda)^2] < 1,$ the process represented by Equation (2) is said to have a first order weak stationarity and follows the GARCH process. The initial conditional variance of stock return h_1 is given by:

$$h_1 = \frac{\beta_0}{(1 - \beta_1 - \beta_2[1 + (\theta + \lambda)^2])} \quad (4)$$

The risk-neutralized system in Equation (4) provides a framework for pricing options.

3.1 Analytical Approximations of European Call Option Prices

Duan [5] obtained analytical approximations of European call option prices using Edgeworth expansion. Readers interested in Edgeworth expansion should consult [9]. The analytical approximation of the European call based on the first two moments (mean and variance) is given by:

$$C_2 = S_0 e^{\delta \sigma \rho T} N(\tilde{d}) - K e^{-rT} N(\tilde{d} - \sigma_{\rho T}) \quad (5)$$

where $S_0, K,$ and T, represent stock price, strike price, and maturity period respectively. $N(\cdot)$ is an expression for the cumulative normal distribution function, $\sigma_{\rho T}$ represents standard deviation - which is the square root of the second moment, the expression for $\delta, d,$ and Δ are given by:

$$\delta = \left(\mu_{\rho T} - rT + \frac{1}{2}(\sigma_{\rho T})^2\right)\left(\sigma_{\rho T}\right)^{-1} \quad (5a)$$

$$d = \frac{\ln\left(\frac{S_0}{K}\right) - rT + \frac{1}{2}(\sigma_{\rho T})^2}{\sigma_{\rho T}} \quad (5b)$$

$$\Delta = d + \delta \quad (5c)$$

The analytical approximations of the European call based on the first three moments: mean, variance, and skewness is given by:

$$C_2 = C_3 + \kappa_3 A_3 \quad (6)$$

where C_2 is given in Equation (5) as $S_0 e^{\delta\rho T} N(\tilde{d}) - Ke^{-rT} N(\tilde{d} - \sigma_{\rho T})$, κ_3 is the value of the third moment (skewness), and A_3 is given by:

$$A_3 = \tfrac{1}{3!} S_0 e^{\delta\sigma_{\rho T}} \left[(2\sigma_{\rho T}) n(\tilde{d}) - (\sigma_{\rho T})^2 N(\tilde{d})\right] \quad (6a)$$

The analytical approximations of the European call based on the first four moments: mean, variance, skewness, and kurtosis is given by:

$$C_4 = C_2 + \kappa_3 A_3 + (\kappa_4 - 3) A_4 \quad (7)$$

while A_3 is as given in Equation (6a), κ_4 is the value of the fourth moment (kurtosis) and A_4 is given by :

$$A_4 = \tfrac{1}{4!} S_0 e^{\delta\sigma_{\rho T}} [(\tilde{d})^2 - 1 - 3\sigma_{\rho T}(\tilde{d} - \sigma_{\rho T}))n(\tilde{d}) + (\sigma_{\rho T})^3 N(\tilde{d})] \quad (7a)$$

3.2 Specific Problem Statement

With the emergence of various financial derivatives, option pricing techniques have assumed vital roles in financial markets. Stock volatility is a crucial metric in options valuation. Accurate estimation of stock volatilities, and subsequently option values in markets with rapidly changing underlying asset price systems is paramount to the finance world. Continuous time models were developed to handle this twin problem. However, observing and determining volatilities for continuous time models are difficult; hence, estimating corresponding option values is difficult using continuous time models.

Close form analytical and discrete time numerical techniques provide a very feasible framework for estimating conditional variance of time-varying volatilities and subsequently their corresponding option prices.

In this paper, an algorithm and a program for valuing European styled options, based on the analytical approximation of the GARCH option pricing model developed in [5] was developed. The analytical approximation of European call options obtained in this paper is based on the first two moments (mean and variance).

3.3 Solution Strategy and Implementation

The solution strategy adopted is presented in this section. The computational algorithm derived from the mathematical equations in [5] for analytically approximating the value of a European call option is presented below. A flowchart showing the steps in this algorithm is shown in Figure 1.

3.4 Computational Algorithm

The algorithm is presented in steps.

Algorithm 1: GARCH

1. *Initialization:* Initialize the values of GARCH parameters, $\beta_0, \beta_1, \beta_2, \lambda + \theta$ to their appropriate values. The values of the GARCH parameters should satisfy the following conditions:
   $\beta_0 \geq 0, \beta_1 > 0$ and $\beta_2 > 0$, and $\beta_1 + \beta_2 < 1$,
   $\lambda + \theta$ must be ≥ 0.5 and ≤ 1.0
2. *Test stochastic process for first order weak stationarity and GARCH process.*
   Compute **mvalue** as:
   $mvalue = \beta_1 + (\beta_2(1 + (\lambda + \theta)^2)$
   if mvalue < 1.0
   then
   print (stochastic process has 1st order weak stationarity and follows a GARCH process)
   else
   print (stochastic process has no 1st order weak stationarity and follows a GARCH process)
   go to last step
   Compute h_1 (the initial unconditional variance of asset return) given in Equation (4) as:
   $h_1 = \frac{\beta_0}{(1 - \beta_1 - \beta_2[1 + (\theta + \lambda)^2])}$

3. *Compute the moment conditions u_1 and u_2:*
Compute the following parameter values initially as:
   i. $lamda2 = (\lambda + \theta)^2$
   ii. $lamda4 = (\lambda + \theta)^4$
   iii. $lamda6 = (\lambda + \theta)^6$
   Compute the first and second moment conditions as follows:
   $u_1 = \beta_2(1.0 + lamda2) + \beta_2$
   $u_2 = (\beta_2)^2(3.0 + lamda2 + lamda4) + 2\beta_1\beta_2(1.0 + lamda2) + (\beta_1))$
   test whether GARCH parameters are within admissible region
   if $u_1 \le 0.98$ and $u_2 \le 0.98$ then
   print (GARCH) parameter values are out of range) print (results obtained may be unrealistic)

4. *Input European option parameters interactively or from a file.*
   input value for (maturity time t in days)
   input value for risk free rate (rdt), where rdt can be $0.05, 0.06, 0.12$
   input value for initial stock price S_t
   input value for strike price K

5. *Compute first moment1.*
   moment1 $= \left(rdt * \frac{t}{365}\right) - 0.5 * V_1$
   where $V_1 = \sum_{i=1}^{t} E_0^Q[h_i]$
   and
   $E_0^Q[h_i] = \beta_0(1-\mu_1^{i-1})(1-\mu_1)^{-1} + \mu_1^{i-1}h_1$

6. *Compute second moment2*
   moment2
   $= t^2r^2 - tr\sum_{i=1}^{t} E_0^Q[h_i] + \frac{1}{4}S_{D1} + S_{D2} - S_{D3}$
   where $S_{D1} = E_0^Q \sum_{i=1}^{t}\sum_{j=1}^{t-1} h_i h_j$,
   $S_{D2} = \sum_{i=1}^{t} E_0^Q[h_i]$, and
   $S_{D3} = E_0^Q \sum_{i=1}^{t}\sum_{j=1}^{t} h_i\sqrt{h_j\varepsilon_j}$

7. *Compute variance and standard deviation*
   *variance = moment2-moment1*moment1*
   stadev $= \sqrt{\text{variance}}$

8. *Compute the cumulative normal distribution of necessary parameters*

9. *Compute the following parameters*
   $d = \frac{\ln(\frac{S_0}{K}) - rT + \frac{1}{2}(\sigma_{\rho T})^2}{\sigma_{\rho T}}$
   $\delta = (\mu_{\rho T} - rT + \frac{1}{2}(\sigma_{\rho T})^2)(\sigma_{\rho T})^{-1}$
   $\Delta = d + \delta$

10. *Compute European call option value analytical approximation based on the first two moments*
   callvalue
   $= S_0 e^{\delta\sigma_{\rho t}} N(\tilde{d}) - Ke^{-rt} N(\tilde{d} - \sigma_{\rho t})$

11. *Repeat Steps 4 – 9 to compute more option values for the European options or go to Step 13.*

12. *Print European call option value.*

13. *STOP.*

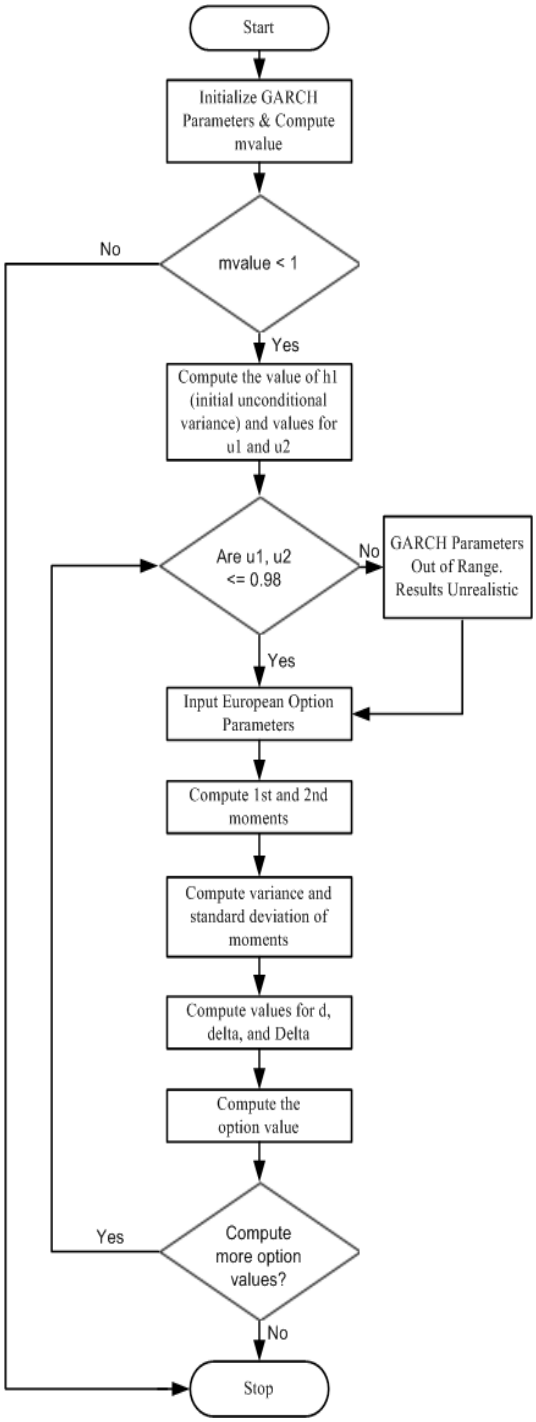


Figure 1: Flowchart Showing Steps in Analytical Approximation of European Call Option.

4. EXPERIMENTAL FRAMEWORK

These experiments were carried out on a system with the following configuration: Intel Core Duo CPU 4.2GHz Processor, 4.0 GB of RAM, 1.00 TB hard Drive space, and Windows 7 Professional operating system. The codes were run on the machine *swan.cs.umanitoba.ca*, with UNIX operating system.

4.1 Experiment 1a

The experiments carried out in this paper are outlined below.
Experiment 1a aims at obtaining option values for at-the-money European call option, with interest rate at 5%. In Experiment 1a, the value of β_0 was set to 0.00001, β_1 was set to 0.7. Other parameters used are $\beta_2 = 0.1, \lambda + \theta = 0.5, h_1 = h_1 * 1.0,$ and interest rate = 0.05. These data were used to compute option value for an asset with stock price (St) 50 and Strike price (K) 50 for different maturity periods. Figure 2 shows the graph for this experiment, while Table 1 is a summary of the input and output parameters.

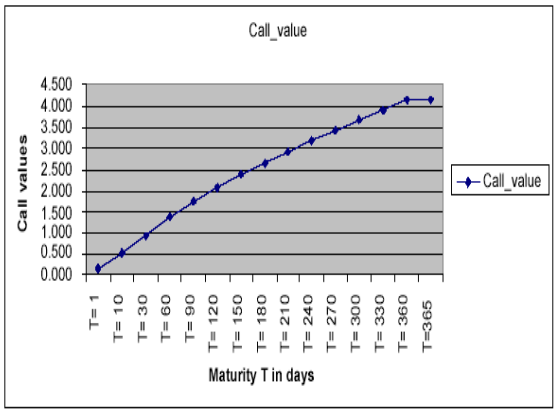


Figure 2: Graph of Table 1

Table 1: Input and Output Parameters for Experiment 1a

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	0.150	0.510	0.930	1.378	1.747
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	2.076	2.378	2.662	2.930	3.190
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	3.440	3.680	3.918	4.147	4.149

With the parameters stated above, we observed that the analytical scheme computed realistic at-the-money European call option values for all maturity periods considered.

4.2 Experiment 1b

Experiment 1b focus to obtain option values for at-the-money European call option, with interest rate of 12%. The parameters in Experiment 1a were used in this experiment, except the interest rate which was changed from 5% to 12%. Figure 3 shows the graph of the results obtained and a summary input (maturity (T) in days and the captured call value is provided in Table 2.

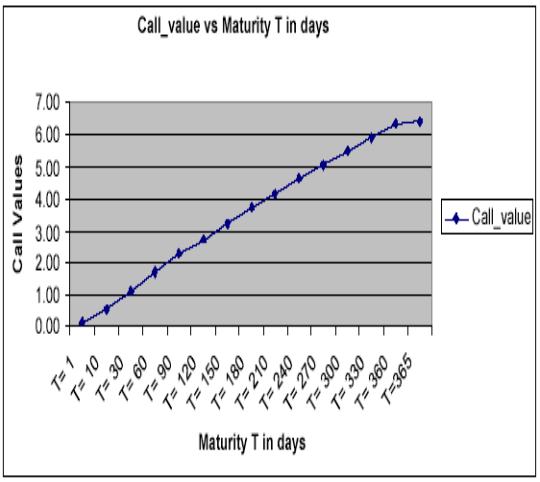


Figure 3: Call Value and Maturity in Days

Table 2: Input and Output Parameters for Experiment 1b

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	0.16	0.56	1.09	1.71	2.26
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	2.77	3.25	3.72	4.18	4.63
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	5.07	5.50	5.93	6.35	6.42

Using interest rate at 12% in Experiment 1b, it is observed that the analytical scheme computed realistic at-the-money European call option values for all maturity periods under review. This result confirms the analytical approximation of the GARCH option pricing model can handle at-the-money European call options with varied interest rates.

4.3 Experiment 2a

The aim of this experiment is to ascertain if the analytical approximation of the GARCH option pricing model can handle in-the-money European call options at a specific interest rate. Using the following parameters

$$\beta_0 = 0.00001, \beta_1 = 0.7, \beta_2 = 0.1, \lambda + \theta = 0.5, h_1 = h_1 * 1.0,$$

and interest rate of 5%, we compute option values for an asset with stock price (St) 55 and Strike price (K) 50 for various maturity periods. Figure 4 shows the results for the computation. Table 3 shows the summary of the input and output parameters for the experiment.

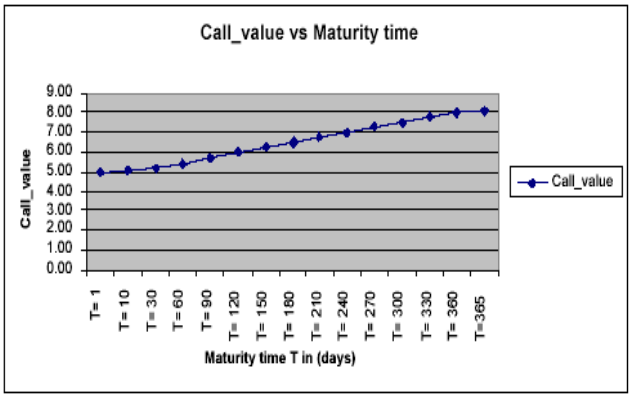


Figure 4: Graph for Table 3

Table 3: Input and Output Parameters for Experiment 2a

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	5.01	5.07	5.21	5.46	5.72
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	5.99	6.26	6.52	6.78	7.03
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	7.28	7.52	7.76	7.99	8.04

With the parameters above, the analytical scheme computed realistic in-the-money European call option value for all maturity periods used, with interest rate at 5%.

4.4 Experiment 2b

The aim of experiment 2b is to determine if the analytical approximation of the GARCH option pricing model can handle in-the-money European call options. The parameters used in this experiment where the same as Experiment 2a but with a change in the interest rate which was increased from 5% to 12%. The results obtained are depicted in Figure 5 and Table 4 shows the summary the input and output parameters.

With interest rate at 12%, it is observed that the analytical scheme computed realistic in-the-money European call option values for all maturity periods considered. This result substantiate the analytical approximation of the GARCH option pricing model can handle in-the-money European call options with varied interest rates.

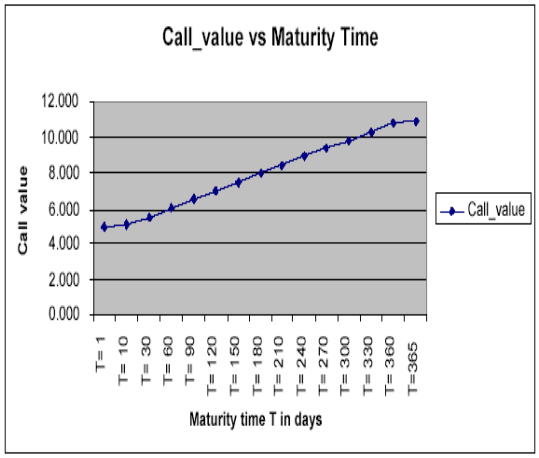


Figure 5: Graph for Table 3

Table 4: Input and Output Parameters for Experiment 2b

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	5.02	5.16	5.49	6.01	6.52
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	7.03	7.53	8.01	8.50	8.97
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	9.43	9.82	10.34	10.79	10.86

4.5 Experiment 3a and 3b

Experiment 3 aims at determining whether the analytical approximation of the GARCH option pricing model can compute realistic out-of-the money European styled call option values. With the following parameters $\beta_0 = 0.00001, \beta_1 = 0.7, \beta_2 = 0.1,$ $\lambda + \theta = 0.5,$ $h_1 = h_1 * 1.0,$ and interest rate 5%. These data were used to compute option values for an asset with stock price (St) 48 and Strike price (K) 50 for different maturity periods. The results of these computations are shown in the summary table of input and output parameters in Table 5. The Zeros entries for call value indicates outrageous option values generated in this experiment.

Table 5: Input and Output Parameters for Experiment 3a

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	0.00	0.00	0.00	0.00	0.00
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	0.00	0.00	0.00	0.00	0.00
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	0.00	0.00	0.00	0.00	0.00

With the parameters above, the analytical scheme was unable to compute realistic option values for out-of-the-money European call options. When the interest in this experiment was changed to 12%, the analytical scheme was still unable to compute realistic option values for out-of-the-money European call options. This result is shown in Table 6.

Table 6: Input and Output Parameters for Experiment 3b

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	0.00	0.00	0.00	0.00	0.00
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	0.00	0.00	0.00	0.00	0.00
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	0.00	0.00	0.00	0.00	0.00

4.6 Experiment 4

The aim of Experiment 4 is to determine the term structure of the first two moments for the cumulative returns under NGARCH. Figure 6 shows the graph of the first moment against the maturity period. Table 7 summarizes the input and output values. Figure 7 shows the graph of the second moment against the maturity period, while Table 8 shows the summary of the input and output parameters.

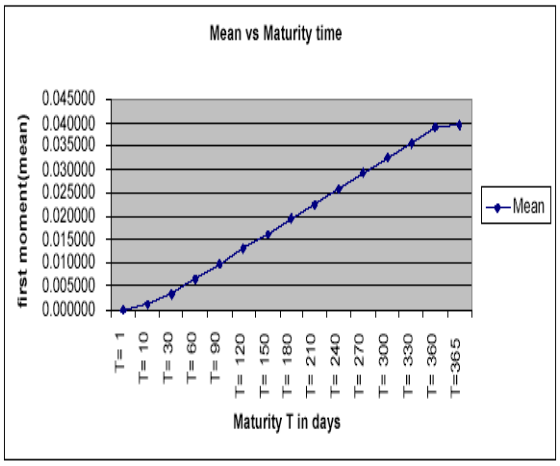


Figure 6: Graph for Table 7

Table 7: Input and Output Parameters for Experiment 4a

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	0.0001	0.0011	0.0033	0.0065	0.0098
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	0.0130	0.0163	0.0195	0.0228	0.0260
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	0.0293	0.0325	0.0358	0.0390	0.0396

In Figure 7, the variance is increasing with maturity time. This relationship exhibited is due to the conditional variance used being constant. In Figure 6, the first moment was increasing steadily and proportionately with maturity period, this is also attributable to the constant volatility persistence level.

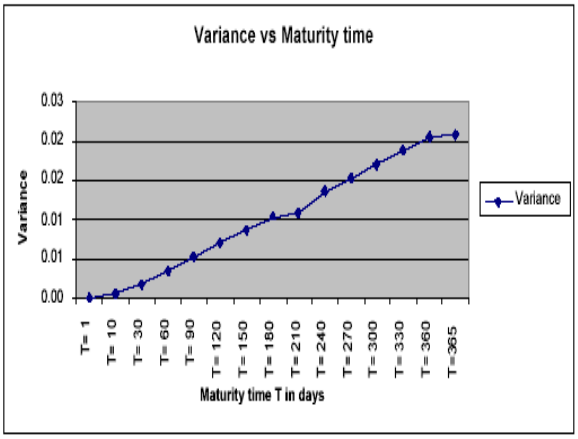


Figure 7: Graph for Table 8

Table 8: Input and Output Parameters for Experiment 4b

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	0.0001	0.0006	0.0017	0.0034	0.0051
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	0.0069	0.0086	0.0103	0.0109	0.0137
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	0.0154	0.0171	0.0188	0.0205	0.0207

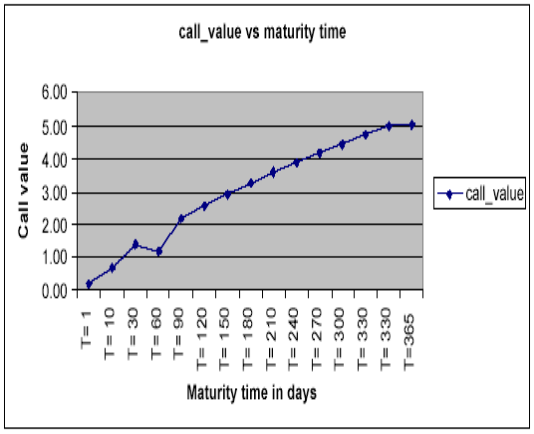


Figure 8: Graph for Table 9

Table 9: Input and Output Parameters for Experiment 5

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	0.20	0.67	1.38	1.20	2.20
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	2.59	2.95	3.29	3.60	3.90
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	4.19	4.47	4.73	4.99	5.04

4.7 Experiment 5

The aim of this experiment is to ascertain the effect of changing the risk unit premium on the European call option value. The following parameters where used in this

experiment $\beta_0 = 0.00001, \beta_1 = 0.7, \beta_2 = 0.1,$

$\lambda + \theta = 0.5,\ h_1 = h_1 * 1.0,$ and interest rate 5%. The data used to compute option value for the asset are; stock price (St) 50 and Strike price (K) 50, for different maturity days. Figure 8 shows the graph of the experiment and the summary of the input and output parameter is shown in Table 9.

With the value of the risk unit premium at 1.0, it is observed that the option values were computed realistic. Although, option values were slightly higher than when the risk unit premium was 0.5, this is attributable to the increase in volatility persistence as the value of μ_1 and μ_2 increased, with values close to 1. With the value of the risk unit premium set at 1.0, it is observed that realistic option values were computed. Although option values were slightly higher than when the risk unit premium was 0.5, this is attributable to the increase in volatility persistence as the value of μ_1 and μ_2 increased, with values close to 1. However, the option values of at-the-money European call options obtained in the maturity periods used where realistic.

4.8 Experiment 6

The aim of this experiment is to determine the effect of changing the GARCH parameters $\beta_0, \beta_1, \beta_2$ on the value of European call options. The following parameters where used in this experiment $\beta_0 = 0.000015,\ \beta_1 = 0.65, \beta_2 = 0.15,$

$\lambda + \theta = 0.5,\ h_1 = h_1 * 1.0,$ and interest rate 5%. The data used to compute option value for the asset are; stock price (St) 50 and Strike price (K) 50, for different maturity days. Figure 9 shows the graph for this experiment, while Table 10 shows a summary of the input and output data.

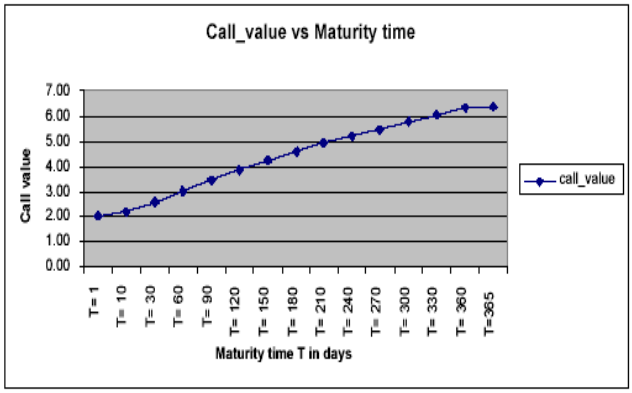


Figure 9: Graph of Table 10

Table 10: Input and Output Parameters for Experiment 6

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	2.01	2.14	2.54	3.06	3.50
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	3.89	4.26	4.60	4.92	5.23
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	5.52	5.80	6.08	6.35	6.39

With slight changes made to values of the GARCH parameters $\beta_0, \beta_1,$ and β_2; it is observed that the analytical scheme computed realistic at-the-money European call option values for all maturity periods considered.

4.9 Experiment 7a

This experiment is aimed at observing the effect of changing the conditional variance on the value of the European call option. The value of the conditional variance was increased by a ratio of 1.1, i.e., $h_1 = h*1.1$. The other parameters used in this experiment are $\beta_0 = 0.000015, \beta_1 = 0.65, \beta_2 = 0.15,$ $\lambda + \theta = 0.5, h_1 = h_1 * 1.0,$ and interest rate 5%. The data used to compute option value for the asset are; stock price (St) 50 and Strike price (K) 50, for different maturity days. Figure 10 shows the graph of the experiment Table 11 shows the summary of input and output data captured during the experiment.

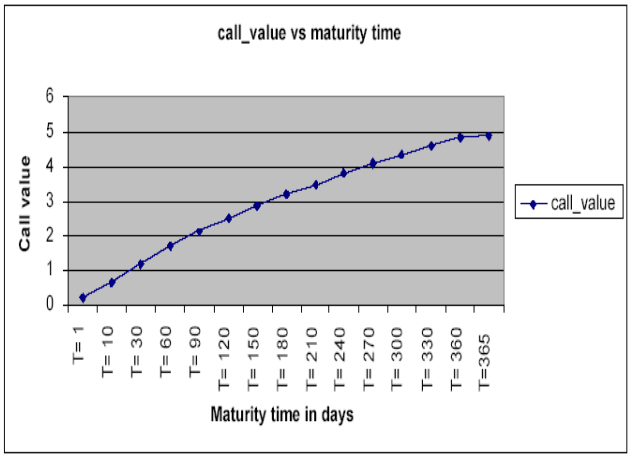


Figure 10: Graph of Table 11

Table 11: Input and Output Parameters for Experiment 7a

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	0.195	0.641	1.153	1.691	2.127
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	2.511	2.861	3.187	3.495	3.788
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	4.069	4.340	4.602	4.857	4.899

From the results, we observe that with an increase in the conditional variance h_1; by a ratio of 1.1, it is observed that the analytical approximation computed realistic at-the-money European call option values for all the maturity periods used.

4.10 Experiment 7b

This experiment also aims at observing the effect of reducing the conditional variance on the value of the European call option the value of the conditional variance was decreased by a ratio of 0.9, i.e., $h_1 = h * 0.9$. The other parameters and data used are the same as in Experiment 7a. Figure 11 shows the graph of the experiment carried out, while Table 12 shows a summary for input and output data.

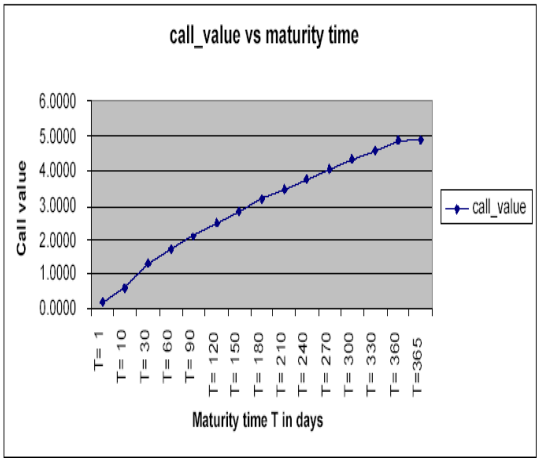


Figure 11: Graph of Table 12

Table 12: Input and Output Parameters for Experiment 7b

Maturity (T) in days	T=1	T=10	T=30	T=60	T=90
Call value	0.1852	0.6247	1.3100	1.6831	2.1210
Maturity (T) in days	T=120	T=150	T=180	T=210	T=240
Call value	2.5057	2.8564	3.1829	3.4909	3.7843
Maturity (T) in days	T=270	T=300	T=330	T=360	T=365
Call value	4.0656	4.3367	4.5991	4.8539	4.8957

With decrease in conditional variance h_1 by a ratio of 0.9, it is observed that the analytical approximation of the GARCH option pricing model, computed realistic at-the-money European call option values for all maturity periods considered.

6. RESULTS AND DISCUSSIONS

In the course of the experiments performed in the preceding section, changes were made to the GARCH parameters: β_0 β_1, β_2 risk unit premium $\lambda + \theta$, conditional variance h_1, the one-period free risk rate of return, the maturity time, stock price, and strike price.

Overall, from various experiments carried out, it is observed that the analytical approximation of the GARCH option pricing model is well suited for short maturity period at-the-money and in-the-money European call options. The analytical scheme performed poorly in computing option values for out-the-money European call options.

7. CONCLUSIONS AND FUTURE WORK

In this paper, a computational algorithm for analytical approximation of the GARCH option pricing model (based on the first two moments), and used for valuing European styled call options was presented. The algorithm was implemented in C; various experiments were carried out on the implementation.

The experiments conducted confirms the analytical approximation of the GARCH option pricing model (based on the first two moments), very suitable for at-the-money and in-the-money European styled call options with short maturity periods. In the future, this work could benefit from parallel programming. Also, the work can be extended to compute for higher moments.

ACKNOWLEDGMENT

The first and third authors acknowledge gratefully the University Research Grant Program (URGP) of the University of Manitoba and the Natural Sciences and Engineering Research Council (NSERC) of Canada for the partial financial support.

REFERENCES

[1] Black F. and Scholes M., "The Pricing of Options and Corporate Liabilities", *Journal of Political Economy*, Vol.81, pp. 637-659, 1973.

[2] Bollerslev T., "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, Vol. 31, pp. 307-327, 1986.

[3] Cakici N. and Topyan K., "GARCH versus Constant Volatility: A Lattice Approach in Option Pricing" *Proceedings of the 1999 Association of Employment Practices and Principles Conference*, 1999. New York, USA.

[4] Duan J. C., "The GARCH Option Pricing Model", *Journal of Mathematical Finance*, Vol. 5, No. 1, pp.13-32, 1995.

[5] Duan J. C., Gauthier G. and Simonato J. G., "An Analytical Approximation for the GARCH Option Pricing Model", *Journal of Computational Finance*, Vol. 2, pp. 75-116, 2001.

[6] Engle R., "Autoregressive Conditional Heteroskedasticity, with Estimates of the Variance of UK Inflation", *Journal of Econometrica*, Vol. 50, pp. 987-1108, 1982.

[7] Heston S. L. and Nandi S., "A Closed-Form GARCH Option Pricing Model", *Review of Financial Studies*, Fall 2000.

[8] J. C. Hull, Options, Futures, and Other Derivatives. Prentice Hall, 7th Edition, 2009.

[9] Jarrow R. and Rudd A., "Approximate Option Valuation for Arbitrary Stochastic Process", *Journal of Financial Economics*, pp.307-327, 1982.

[10] Ritcthken P. and Trevor R., "Pricing Options under Generalized and Stochastic Volatility Processes", *Journal of Finance*, Vol. 54, No. 1, pp. 377-402, 1999.

[11] Ritchken P. and Hsieh K. C., "An Emperical Comparison of GARCH Option Pricing Model", Unpublished Working Paper, September 2000.

[12] Xhao X, Scarrott C. J., and Marco R., "GARCH Dependence in Extreme Value Models with Bayesian Inference", Journal of *Mathematics and Computers in Simulation*, Vol. 81, No. 7, pp.1430-1440, 2011.

[13] Liu S. and Liu Y., "Threshold-GARCH Option Pricing: A Trinomial Tree Approach", *In Proceedings of the 2nd International Conference on Innovative Computing, Information and Control*, 2007, USA.

[14] Thavaneswaran A., Appadoo S. S., and Paseka A., "Weighted Possibilistic Moments of Fuzzy Numbers with Applications to GARCH Modeling and Option Pricing", *Mathematical and Computer Modeling: An International Journal*, Vol. 49, No. 1-2, pp. 352-368, 2009.

[15] Breton M. and Frutos J., "Option Pricing Under GARCH Processes Using PDE Methods", *Journal of Operations Research*, Vol. 58, No. 4-Part-2, pp. 1148-1157, 2010.

[16] Ben-Ameur H, Breton M., and Martinez J., "Dynamic Programming Approach for Valuing Options in the GARCH Model", *Journal of Management Science*, Vol. 55, No. 2, pp. 252-266, 2009.

[17] A. V. Gerbessiotis. Architecture independent parallel binomial tree option price valuations. *Parallel Computing*, 30(2), 2004.

[18]J. C. Cox, S. A. Ross, and M. Rubinstein, "Option pricing: A Simplified Approach," Journal of Financial Economics, vol. 3, no. 7, pp. 229–263, 1979.

[19] R. K. Thulasiram, L. Litov, H. Nojumi, C. T. Downing, and G. R. Gao., "Multithreaded Algorithms for Pricing a Class of Complex Options," in Proc. (CDROM) of the 15th IEEE/ACM Int'l. Parallel and Distributed Processing Symposium (IPDPS), Anchorage (Alaska) USA, April, 2001.

[20] R. K. Thulasiram and P. Thulasiraman, "Performance Evaluation of a Multithreaded Fast Fourier Transform Algorithm for Derivative Pricing," *The Journal of Supercomputing*, vol. 26, no. 1, pp. 43–58, 2003.

[21] R. K. Thulasiram, C. Zhen, A. Chabra, P. Thulasiraman, and A. Gumel, "A Second Order L_0 Stable Algorithm for Evaluating European Options," International Journal of High Performance Computing and Networking (IJHPCN), no. 5-6, pp. 311–320, 2006.

[22] D. Tavalla and C. Randall, Pricing Financial Instruments: The Finite Difference Method. John Wiley and Sons, New York, NY, 2000.

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Analysis of Network-Based Intrusion Detection and Prevention System in an Enterprise Network Using Snort Freeware

[1]O.B. Lawal
Computer Science Department,
Olabisi Onabanjo University Consult, Ibadan, Nigeria
lawal5@yahoo.com

A. Ibitola
Department of Computer and Information Science
Lead City University
Ibadan, Nigeria

O.B. Longe
Department of Computer Science
University of Ibadan
Ibadan, Nigeria
longeolumide@fulbrightmail.org

[1]Corresponding Author: lawal5@yahoo.com

ABSTRACT
Information Systems and Networks are subject to electronic attacks. Attempts to breach information security are rising every day, along with the availability of the Vulnerability Assessment tools that are widely available on the internet, for free, as well as for a commercial use. Tools such as SubSeven, BackOrifce, Nmaps, L0ftCrack, can all be used to scan, identify, probe, and penetrate systems on the network. Firewalls are put in place to prevent unauthorized access to the Enterprise Networks. Unfortunately, firewalls alone are not enough to protect our systems. This paper describes the characteristics of Network-Base IDPS technologies, outlines the necessity of the implementation of Intrusion Detection Systems in the enterprise environment and a brief evaluation of Snort® Freeware technology.

Keywords: Intrusion detection, Intrusion Prevention, IDPS, Network, Firewalls, Snort Freeware

African Journal of Computing & ICT Reference Format
O.B. Lawal, A. Ibitola & O.B. Longe (2013). Analysis of Network-Based Intrusion Detection and Prevention System in an Enterprise Network Using Snort Freeware. Afr J. of Comp & ICTs. Vol 6, No. 1. Pp 105-120

1. INTRODUCTION
Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices [1]. Intrusion prevention is the process of performing intrusion detection and attempting to stop detected possible incidents. Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, attempting to stop them, and reporting them to security administrators.

In addition, organizations use IDPSs for other purposes, such as identifying problems with security policies, documenting existing threats, and deterring individuals from violating security policies. IDPSs have become a necessary addition to the security infrastructure of nearly every organization [1]. IDPSs typically record information related to observed events, notify security administrators of important observed events, and produce reports. Many IDPSs can also respond to a detected threat by attempting to prevent it from succeeding[1]. They use several response techniques, which involve the IDPS stopping the attack itself, changing the security environment (e.g., reconfiguring a firewall), or changing the attack's content [2].

Snort® is an open source network intrusion prevention and detection system (IDS/IPS) developed by Sourcefire [27]. Combining the benefits of signature, protocol, and anomaly-based inspection, Snort is the most widely deployed IDS/IPS technology worldwide [27]. With millions of downloads and nearly 400,000 registered users, Snort has become the de facto standard for IPS [27].

2. INTRUSION DETECTION AND PREVENTION PRINCIPLES

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible *incidents*, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices. Incidents have many causes, such as malware (e.g., worms, spyware), attackers gaining unauthorized access to systems from the Internet, and authorized users of systems who misuse their privileges or attempt to gain additional privileges for which they are not authorized. Although many incidents are malicious in nature, many others are not; for example, a person might mistype the address of a computer and accidentally attempt to connect to a different system without authorization [5].

Endorf [8] defines an *intrusion detection system* (IDS) has software that automates the intrusion detection process. An *intrusion prevention system* (IPS) is software that has all the capabilities of an intrusion detection system and can also attempt to stop possible incidents.

IDS and IPS technologies offer many of the same capabilities, and administrators can usually disable prevention features in IPS products, causing them to function as IDSs [1]. Accordingly, for brevity the term *intrusion detection and prevention systems* (*IDPS)* is used throughout the rest of this paper to refer to both IDS and IPS technologies.

IDS and IPS technologies offer many of the same capabilities, and administrators can usually disable prevention features in IPS products, causing them to function as IDSs [1]. Accordingly, for brevity the term *intrusion detection and prevention systems* (*IDPS)* is used throughout the rest of this guide to refer to both IDS and IPS technologies. Any exceptions are specifically noted.

2.1 Uses of IDPS Technologies

IDPSs are primarily focused on identifying possible incidents [5]. For example, an IDPS could detect when an attacker has successfully compromised a system by exploiting a vulnerability in the system. The IDPS could then report the incident to security administrators, who could quickly initiate incident response actions to minimize the damage caused by the incident.

The IDPS could also log information that could be used by the incident handlers. Many IDPSs can also be configured to recognize violations of security policies. For example, some IDPSs can be configured with firewall ruleset-like settings, allowing them to identify network traffic that violates the organization's security or acceptable use policies. Also, some IDPSs can monitor file transfers and identify ones that might be suspicious, such as copying a large database onto a user's laptop [4].

Many IDPSs can also identify reconnaissance activity, which may indicate that an attack is imminent [4]. For example, some attack tools and forms of malware, particularly worms, perform reconnaissance activities such as host and port scans to identify targets for subsequent attacks. An IDPS might be able to block reconnaissance and notify security administrators, who can take actions if needed to alter other security controls to prevent related incidents. Because reconnaissance activity is so frequent on the Internet, reconnaissance detection is often performed primarily on protected internal networks [2].

In addition to identifying incidents and supporting incident response efforts, organizations have found other uses for IDPSs, including the following:

- **Identifying security policy problems.** An IDPS can provide some degree of quality control for security policy implementation, such as duplicating firewall rulesets and alerting when it sees network traffic that should have been blocked by the firewall but was not because of a firewall configuration error.
- **Documenting the existing threat to an organization.** IDPSs log information about the threats that they detect. Understanding the frequency and characteristics of attacks against an organization's computing resources is helpful in identifying the appropriate security measures for protecting the resources. The information can also be used to educate management about the threats that the organization faces.
- **Deterring individuals from violating security policies.** If individuals are aware that their actions are being monitored by IDPS technologies for security policy violations, they may be less likely to commit such violations because of the risk of detection.

Because of the increasing dependence on information systems and the prevalence and potential impact of intrusions against those systems, IDPSs have become a necessary addition to the security infrastructure of nearly every organization [9].

2.2 Key Functions of IDPS Technologies

There are many types of IDPS technologies, which are differentiated primarily by the types of events that they can recognize and the methodologies that they use to identify incidents [7]. In addition to monitoring and analyzing events to identify undesirable activity, all types of IDPS technologies typically perform the following functions:

- **Recording information related to observed events.** Information is usually recorded locally, and might also be sent to separate systems such as centralized logging servers, security information and event management (SIEM) solutions, and enterprise management systems.
- **Notifying security administrators of important observed events.** This notification, known as an *alert*, occurs through any of several methods, including the following: e-mails, pages, messages on the IDPS user interface, Simple Network Management Protocol (SNMP) traps, syslog messages, and user-defined programs and scripts. A notification message typically includes only basic information regarding an event; administrators need to access the IDPS for additional information [1].
- **Producing reports.** Reports summarize the monitored events or provide details on particular events of interest.

Some IDPSs are also able to change their security profile when a new threat is detected [7]. For example, an IDPS might be able to collect more detailed information for a particular session after malicious activity is detected within that session. An IDPS might also alter the settings for when certain alerts are triggered or what priority should be assigned to subsequent alerts after a particular threat is detected [7]. IPS technologies are differentiated from IDS technologies by one characteristic: IPS technologies can respond to a detected threat by attempting to prevent it from succeeding. They use several response techniques, which can be divided into the following groups:

- **The IPS stops the attack itself.** Examples of how this could be done are as follows:
  - Terminate the network connection or user session that is being used for the attack
  - Block access to the target (or possibly other likely targets) from the offending user account, IP address, or other attacker attribute
  - Block all access to the targeted host, service, application, or other resource.
- **The IPS changes the security environment.** The IPS could change the configuration of other security controls to disrupt an attack. Common examples are reconfiguring a network device (e.g., firewall, router, switch) to block access from the attacker or to the target, and altering a host-based firewall on a target to block incoming attacks. Some IPSs can even cause patches to be applied to a host if the IPS detects that the host has vulnerabilities.
- **The IPS changes the attack's content.** Some IPS technologies can remove or replace malicious portions of an attack to make it benign. A simple example is an IPS removing an infected file attachment from an e-mail and then permitting the cleaned email to reach its recipient. A more complex example is an IPS that acts as a proxy and *normalizes* incoming requests, which means that the proxy repackages the payloads of the requests, discarding header information. This might cause certain attacks to be discarded as part of the normalization process.

Another common attribute of IDPS technologies is that they cannot provide completely accurate detection [1] When an IDPS incorrectly identifies benign activity as being malicious, a *false positive* has occurred. When an IDPS fails to identify malicious activity, a *false negative* has occurred. It is not possible to eliminate all false positives and negatives; in most cases, reducing the occurrences of one increases the occurrences of the other. Many organizations choose to decrease false negatives at the cost of increasing false positives, which means that more malicious events are detected but more analysis resources are needed to differentiate false positives from true malicious events. Altering the configuration of an IDPS to improve its detection accuracy is known as *tuning* [1].

Most IDPS technologies also offer features that compensate for the use of common evasion techniques [1]. *Evasion* is modifying the format or timing of malicious activity so that its appearance changes but its effect is the same. Attackers use evasion techniques to try to prevent IDPS technologies from detecting their attacks. For example, an attacker could encode text characters in a particular way, knowing that the target understands the encoding and hoping that any monitoring IDPSs do not. Most IDPS technologies can overcome common evasion techniques by duplicating special processing performed by the targets. If the IDPS can "see" the activity in the same way that the target would, then evasion techniques will generally be unsuccessful at hiding attacks [10].

2.3 Common Detection Methodologies

IDPS technologies use many methodologies to detect incidents. Sections 2.3.1 through 2.3.3 discuss the primary classes of detection methodologies: signature-based, anomaly-based, and stateful protocol analysis, respectively. Most IDPS technologies use multiple detection methodologies, either separately or integrated, to provide more broad and accurate detection [1].

2.3.1 Signature-Based Detection

According to [10], a *signature* is a pattern that corresponds to a known threat. *Signature-based detection* is the process of comparing signatures against observed events to identify possible incidents.5 Examples of signatures are as follows:

- A telnet attempt with a username of "root", which is a violation of an organization's security policy
- An e-mail with a subject of "Free pictures!" and an attachment filename of "freepics.exe", which are characteristics of a known form of malware
- An operating system log entry with a status code value of 645, which indicates that the host's auditing has been disabled.

Signature-based detection is very effective at detecting known threats but largely ineffective at detecting previously unknown threats, threats disguised by the use of evasion techniques, and many variants of known threats. For example, if an attacker modified the malware in the previous example to use a filename of "freepics2.exe", a signature looking for "freepics.exe" would not match it [10].

Signature-based detection is the simplest detection method because it just compares the current unit of activity, such as a packet or a log entry, to a list of signatures using string comparison operations. Signature-based detection technologies have little understanding of many network or application protocols and cannot track and understand the state of complex communications [10]. For example, they cannot pair a request with the corresponding response, such as knowing that a request to a Web server for a particular page generated a response status code of 403, meaning that the server refused to fill the request. They also lack the ability to remember previous requests when processing the current request. This limitation prevents signature-based detection methods from detecting attacks that comprise multiple events if none of the events contains a clear indication of an attack [10].

2.3.2 Anomaly-Based Detection

Anomaly-based detection is the process of comparing definitions of what activity is considered normal against observed events to identify significant deviations [1]. An IDPS using anomaly-based detection has *profiles* that represent the normal behavior of such things as users, hosts, network connections, or applications. The profiles are developed by monitoring the characteristics of typical activity over a period of time. For example, a profile for a network might show that Web activity comprises an average of 13% of network bandwidth at the Internet border during typical workday hours [1].

The IDPS then uses statistical methods to compare the characteristics of current activity to thresholds related to the profile, such as detecting when Web activity comprises significantly more bandwidth than expected and alerting an administrator of the anomaly. Profiles can be developed for many behavioral attributes, such as the number of e-mails sent by a user, the number of failed login attempts for a host, and the level of processor usage for a host in a given period of time [1]. The major benefit of anomaly-based detection methods is that they can be very effective at detecting previously unknown threats [1]. For example, suppose that a computer becomes infected with a new type of malware. The malware could consume the computer's processing resources, send large numbers of e-mails, initiate large numbers of network connections, and perform other behavior that would be significantly different from the established profiles for the computer [10].

An initial profile is generated over a period of time (typically days, sometimes weeks) sometimes called a *training period*. Profiles for anomaly-based detection can either be static or dynamic [1]. Once generated, a static profile is unchanged unless the IDPS is specifically directed to generate a new profile. A dynamic profile is adjusted constantly as additional events are observed. Because systems and networks change over time, the corresponding measures of normal behavior also change; a static profile will eventually become inaccurate, so it needs to be regenerated periodically. Dynamic profiles do not have this problem, but they are susceptible to evasion attempts from attackers [8]. For example, an attacker can perform small amounts of malicious activity occasionally, then slowly increase the frequency and quantity of activity. If the rate of change is sufficiently slow, the IDPS might think the malicious activity is normal behavior and include it in its profile. Malicious activity might also be observed by an IDPS while it builds its initial profiles.

Inadvertently including malicious activity as part of a profile is a common problem with anomaly-based IDPS products. (In some cases, administrators can modify the profile to exclude activity in the profile that is known to be malicious.) [8]. Another problem with building profiles is that it can be very challenging in some cases to make them accurate, because computing activity can be so complex. For example, if a particular maintenance activity that performs large file transfers occurs only once a month, it might not be observed during the training period; when the maintenance occurs, it is likely to be considered a significant deviation from the profile and trigger an alert. Anomaly-based IDPS products often produce many false positives because of benign activity that deviates significantly from profiles, especially in more diverse or dynamic environments [8].

Another noteworthy problem with the use of anomaly-based detection techniques is that it is often difficult for analysts to determine why a particular alert was generated and to validate that an alert is accurate and not a false positive, because of the complexity of events and number of events that may have caused the alert to be generated [1].

2.3.3 Stateful Protocol Analysis

Stateful protocol analysis is the process of comparing predetermined profiles of generally accepted definitions of benign protocol activity for each protocol state against observed events to identify deviations [1]. Unlike anomaly-based detection, which uses host or network-specific profiles, stateful protocol analysis relies on vendor-developed universal profiles that specify how particular protocols should and should not be used. The "stateful" in stateful protocol analysis means that the IDPS is capable of understanding and tracking the state of network, transport, and application protocols that have a notion of state. For example, when a user starts a File Transfer Protocol (FTP) session, the session is initially in the unauthenticated state.

Unauthenticated users should only perform a few commands in this state, such as viewing help information or providing usernames and passwords [8]. An important part of understanding state is pairing requests with responses, so when an FTP authentication attempt occurs, the IDPS can determine if it was successful by finding the status code in the corresponding response. Once the user has authenticated successfully, the session is in the authenticated state, and users are expected to perform any of several dozen commands. Performing most of these commands while in the unauthenticated state would be considered suspicious, but in the authenticated state performing most of them is considered benign [8].

2.4 Types of IDPS Technologies

There are many types of IDPS technologies [1]. For the purposes of this paper, they are divided into the following four groups based on the type of events that they monitor and the ways in which they are deployed:

i. Network-Based, which monitors network traffic for particular network segments or devices and analyzes the network and application protocol activity to identify suspicious activity [1]. It can identify many different types of events of interest. It is most commonly deployed at a boundary between networks, such as in proximity to border firewalls or routers, virtual private network (VPN) servers, remote access servers, and wireless networks.

ii. Wireless, which monitors wireless network traffic and analyzes its wireless networking protocols to identify suspicious activity involving the protocols themselves. It cannot identify suspicious activity in the application or higher-layer network protocols (e.g., TCP, UDP) that the wireless network traffic is transferring. It is most commonly deployed within range of an organization's wireless network to monitor it, but can also be deployed to locations where unauthorized wireless networking could be occurring [10].

iii. Network Behavior Analysis (NBA), which examines network traffic to identify threats that generate unusual traffic flows, such as distributed denial of service (DDoS) attacks, certain forms of malware (e.g., worms, backdoors), and policy violations (e.g., a client system providing network services to other systems) [1]. NBA systems are most often deployed to monitor flows on an organization's internal networks, and are also sometimes deployed where they can monitor flows between an organization's networks and external networks (e.g., the Internet, business partners' networks).

iv. Host-Based, which monitors the characteristics of a single host and the events occurring within that host for suspicious activity. Examples of the types of characteristics a host-based IDPS might monitor are network traffic (only for that host), system logs, running processes, application activity, file access and modification, and system and application configuration changes. Host-based IDPSs are most commonly deployed on critical hosts such as publicly accessible servers and servers containing sensitive information [8].

Some forms of IDPS are more mature than others because they have been in use much longer [10]. Network-based IDPS and some forms of host-based IDPS have been commercially available for over ten years (which form the base of this paper) [10]. Network behavior analysis software is a somewhat newer form of IDPS that evolved in part from products created primarily to detect DDoS attacks, and in part from products developed to monitor traffic flows on internal networks. Wireless technologies are a relatively new type of IDPS, developed in response to the popularity of wireless local area networks (WLAN) and the growing threats against WLANs and WLAN clients [6].

3. NETWORK-BASED IDPS

A network-based IDPS monitors network traffic for particular network segments or devices and analyzes network, transport, and application protocols to identify suspicious activity [9].

3.1 Networking Overview

TCP/IP is widely used throughout the world to provide network communications. TCP/IP communications are composed of four layers that work together [1] [3] [6]. When a user wants to transfer data across networks, the data is passed from the highest layer through intermediate layers to the lowest layer, with each layer adding more information. The lowest layer sends the accumulated data through the physical network; the data is then passed up through the layers to its destination.

Essentially, the data produced by a layer is encapsulated in a larger container by the layer below it. The four TCP/IP layers, from highest to lowest, are shown in Figure 3-1.

Application Layer. This layer sends and receives data for particular applications, such as Domain Name System (DNS), Hypertext Transfer Protocol (HTTP), and Simple Mail Transfer Protocol (SMTP).
Transport Layer. This layer provides connection-oriented or connectionless services for transporting application layer services between networks. The transport layer can optionally ensure the reliability of communications. Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) are commonly used transport layer protocols.
Internet Protocol (IP) Layer (also known as Network Layer). This layer routes packets across networks. IPv4 is the fundamental network layer protocol for TCP/IP. Other commonly used protocols at the network layer are IPv6, Internet Control Message Protocol (ICMP), and Internet Group Management Protocol (IGMP).
Hardware Layer (also known as Data Link Layer). This layer handles communications on the physical network components. The best known data link layer protocol is Ethernet.

Figure 3.1 - TCP/IP Layers [1]

The four TCP/IP layers work together to transfer data between hosts. Network-based IDPSs typically perform most of their analysis at the application layer. They also analyze activity at the transport and network layers both to identify attacks at those layers and to facilitate the analysis of the application layer activity (e.g., a TCP port number may indicate which application is being used). Some network-based IDPSs also perform limited analysis at the hardware layer [3].

3.1.1 Application Layer

The application layer enables applications to transfer data between an application server and client. An example of an application layer protocol is Hypertext Transfer Protocol (HTTP), which transfers data between a Web server and a Web browser. Other common application layer protocols include Domain Name System (DNS), File Transfer Protocol (FTP), Simple Mail Transfer Protocol (SMTP), and Simple Network Management Protocol (SNMP). There are hundreds of unique application layer protocols in common use, and many more that are not so common. Regardless of the protocol in use, application data is generated and then passed to the transport layer for further processing [3].

3.1.2 Transport Layer

The transport layer is responsible for packaging data so that it can be transmitted between hosts. Most applications that communicate over networks rely on the transport layer to ensure reliable delivery of data [1][6]. Generally, this is accomplished by using TCP. When loss of some application data is not a concern (e.g., streaming audio, video), or the application itself ensures reliable delivery of data, UDP is typically used. UDP is connectionless; one host simply sends data to another host without any preliminary negotiations. Each TCP or UDP packet has a source port number and a destination port number. One of the ports is associated with a server application on one system; the other port is associated with a corresponding client application on the other system. Client systems typically select any available port number for application use, whereas server systems usually have a static port number dedicated to each application. Although UDP and TCP ports are very similar, they are distinct from each other and are not interchangeable [1].

3.1.3 Network Layer

The network layer, also known as the IP layer, is responsible for handling the addressing and routing of data that it receives from the transport layer. After the network layer has encapsulated the transport layer data, the resulting logical units are referred to as *packets*. Each packet contains a *header*, which is composed of various *fields* that specify characteristics of the transport protocol in use; optionally, packets may also contain a *payload*, which holds the application data. The IP header contains a field called IP Version, which indicates which version of IP is in use. Typically this is set to 4 for IPv4; but the use of IPv6 is increasing, so this field may be set to 6 instead. Other significant IP header fields are as follows:

- **Source and Destination IP Addresses.** These are the "from" and "to" addresses that are intended tindicate the endpoints of the communication. Example of IP addresses are 10.3.1.70 (IPv4) and 1000::2F:8A:400:427:9BD1 (IPv6).
- **IP Protocol Number.** This indicates which network or transport layer protocol the IP payload contains. Commonly used IP numbers includes 1 (ICMP), 6(TCP), 17 (UDP) and 50 (Encapsulating Security Payload [ESP]).

The network layer is also responsible for providing error and status information involving the addressing and routing of data [7]; it does this with ICMP. ICMP is a connectionless protocol that makes no attempt to guarantee that its error and status messages are delivered. Because it is designed to transfer limited information, not application data, ICMP does not have ports; instead, it has message types, which indicate the purpose of each ICMP message. Some message types also have message codes, which can be thought of as subtypes [3].

3.1.4 Hardware Layer

As the name implies, the hardware layer, also called the data link layer, involves the physical components of the network, including cables, routers, switches, and network interface cards (NIC) [1][7]. The hardware layer also includes various hardware layer protocols, with Ethernet being the most widely used. Ethernet relies on the concept of a media access control (MAC) address, which is a unique six-byte value (such as 00-02-B4-DA-92-2C) that is permanently assigned to a particular NIC. Each *frame*, the logical unit at the hardware layer, contains two MAC addresses, which indicate the MAC address of the NIC that just routed the frame and the MAC address of the next NIC to which the frame is being sent. As a frame passes through networking equipment (such as routers and firewalls) on its way between the original source host and the final destination host, the MAC addresses are updated to refer to the local source and destination. Several separate hardware layer transmissions may be linked together within a single network layer transmission [3].

In addition to the MAC addresses, each frame also contains an EtherType value, which indicates the protocol that the frame's payload contains (typically IP or Address Resolution Protocol [ARP]). When IP is used, each IP address maps to a particular MAC address. (Because multiple IP addresses can map to a single MAC address, a MAC address does not necessarily uniquely identify an IP address) [1].

3.2.1 Typical Components

A typical network-based IDPS is composed of sensors, one or more management servers, multiple consoles, and optionally one or more database servers (if the network-based IDPS supports their use). All of these components are similar to other types of IDPS technologies, except for the sensors [1]. A network-based IDPS sensor monitors and analyzes network activity on one or more network segments. The network interface cards that will be performing monitoring are placed into *promiscuous mode*, which means that they will accept all incoming packets that they see, regardless of their intended destinations. Most IDPS deployments use multiple sensors, with large deployments having hundreds of sensors. Sensors are available in two formats:

Appliance. An appliance-based sensor is comprised of specialized hardware and sensor software. The hardware is typically optimized for sensor use, including specialized NICs and NIC drivers for efficient capture of packets, and specialized processors or other hardware components that assist in analysis. Parts or all of the IDPS software might reside in firmware for increased efficiency. Appliances often use a customized, hardened operating system (OS) that administrators are not intended to access directly [7].

Software Only. Some vendors sell sensor software without an appliance. Administrators can install the software onto hosts that meet certain specifications. The sensor software might include a customized OS, or it might be installed onto a standard OS just as any other application would [1] [3].

3.2.2 Network Architectures and Sensor Locations

Organizations should consider using management networks for their network-based IDPS deployments whenever feasible. If an IDPS is deployed without a separate management network, organizations should consider whether or not a VLAN is needed to protect the IDPS communications [11]. In addition to choosing the appropriate network for the components, administrators also need to decide where the IDPS sensors should be located. Sensors can be deployed in one of two modes [11]:

Inline. An *inline sensor* is deployed so that the network traffic it is monitoring must pass through it, much like the traffic flow associated with a firewall. In fact, some inline sensors are hybrid firewall/IDPS devices, while others are simply IDPSs [1]. The primary motivation for deploying IDPS sensors inline is to enable them to stop attacks by blocking network traffic. Inline sensors are typically placed where network firewalls and other network security devices would be placed—at the divisions between networks, such as connections with external networks and borders between different internal networks that should be segregated. Inline sensors that are not hybrid firewall/IDPS devices are often deployed on the more secure side of a network division so that they have less traffic to process. Figure 3-2 shows such a deployment. Sensors can also be placed on the less secure side of a network division to provide protection for and reduce the load on the dividing device, such as a firewall [1][7].

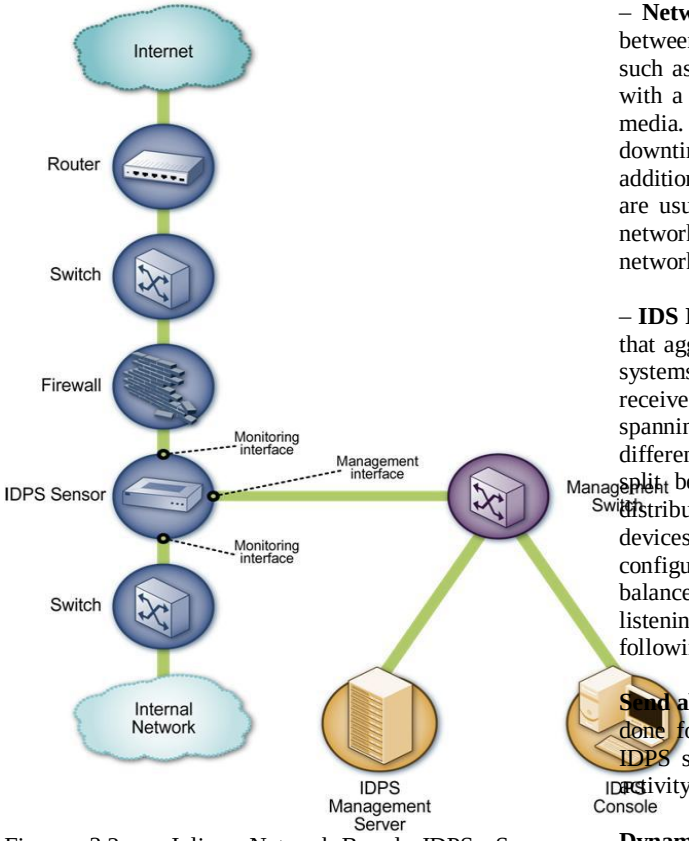


Figure 3.2 - Inline Network-Based IDPS Sensor Architecture Example [1]

Passive. A *passive sensor* is deployed so that it monitors a copy of the actual network traffic; no traffic actually passes through the sensor [1]. Passive sensors are typically deployed so that they can monitor key network locations, such as the divisions between networks, and key network segments, such as activity on a demilitarized zone (DMZ) subnet. Passive sensors can monitor traffic through various methods, including the following:

- **Spanning Port.** Many switches have a *spanning port*, which is a port that can see all network traffic going through the switch. Connecting a sensor to a spanning port can allow it to monitor traffic going to and from many hosts. Although this monitoring method is relatively easy and inexpensive, it can also be problematic. If a switch is configured or reconfigured incorrectly, the spanning port might not be able to see all the traffic [12].

– **Network Tap.** A *network tap* is a direct connection between a sensor and the physical network media itself, such as a fiber optic cable. The tap provides the sensor with a copy of all network traffic being carried by the media. Installing a tap generally involves some network downtime, and problems with a tap could cause additional downtime. Also, unlike spanning ports, which are usually already present throughout an organization, network taps need to be purchased as add-ons to the network.

– **IDS Load Balancer.** An *IDS load balancer* is a device that aggregates and directs network traffic to monitoring systems, including IDPS sensors. A load balancer can receive copies of network traffic from one or more spanning ports or network taps and aggregate traffic from different networks (e.g., reassemble a session that was split between two networks). The load balancer then distributes copies of the traffic to one or more listening devices, including IDPS sensors, based on a set of rules configured by an administrator. The rules tell the load balancer which types of traffic to provide to each listening device. Common configurations include the following:

Send all traffic to multiple IDPS sensors. This could be done for high availability or to have multiple types of IDPS sensors perform concurrent analysis of the same activity.

Dynamically split the traffic among multiple IDPS sensors based on volume. This is typically done to perform load balancing so that no sensor is overwhelmed with the amount of traffic and corresponding analysis [1][3].

Split the traffic among multiple IDPS sensors based on IP addresses, protocols, or other characteristics. This could be done for load balancing purposes, such as having one IDPS sensor dedicated to Web activity and another IDPS sensor monitoring all other activity. Splitting traffic could also be done to perform more detailed analysis of certain types of traffic (e.g., activity involving the most important hosts).

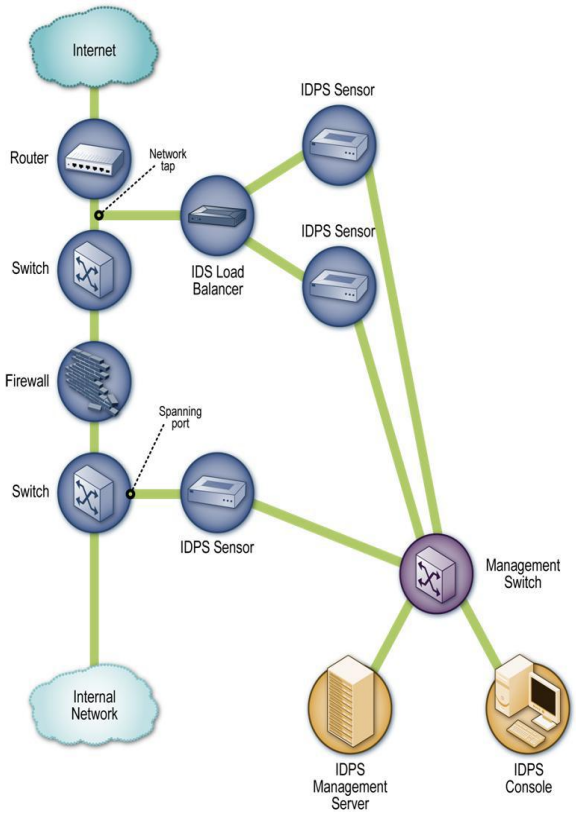


Figure 3.3 - Passive Network-Based IDPS Sensor Architecture Example

3.3 Security Capabilities

Network-based IDPS products provide a wide variety of security capabilities. These are described below [14].

3.3.1 Information Gathering Capabilities

Some network-based IDPSs offer limited information gathering capabilities, which means that they can collect information on hosts and the network activity involving those hosts. Examples of information gathering capabilities are as follows:

- **Identifying Hosts.** An IDPS sensor might be able to create a list of hosts on the organization's network arranged by IP address or MAC address. The list can be used as a profile to identify new hosts on the network [14].
- **Identifying Operating Systems.** An IDPS sensor might be able to identify the OSs and OS versions used by the organization's hosts through various techniques. For example, the sensor could track which ports are used on each host, which could indicate a particular OS or OS family (e.g., Windows, Unix). Another technique is to analyze packet headers for certain unusual characteristics or combinations of characteristics that are exhibited by particular OSs; this is known as *passive fingerprinting*. Some sensors can also identify application versions (as described below).
- **Identifying Applications.** For some applications, an IDPS sensor can identify the application versions in use by keeping track of which ports are used and monitoring certain characteristics of application communications. For example, when a client establishes a connection with a server, the server might tell the client what application server software version it is running, and vice versa. Information on application versions can be used to identify potentially vulnerable applications, as well as unauthorized use of some applications [12].
- **Identifying Network Characteristics.** Some IDPS sensors collect general information about network traffic related to the configuration of network devices and hosts, such as the number of hops between two devices. This information can be used to detect changes to the network configuration.

3.3.2 Logging Capabilities

Network-based IDPSs typically perform extensive logging of data related to detected events [1]. This data can be used to confirm the validity of alerts, to investigate incidents, and to correlate events between the IDPS and other logging sources. Data fields commonly logged by network-based IDPSs include the following:

- Timestamp (usually date and time)
- Connection or session ID
- Event or alert type
- Rating (e.g., priority, severity, impact, confidence)
- Network, transport, and application layer protocols
- Source and destination IP addresses
- Source and destination TCP or UDP ports, or ICMP types and codes
- Number of bytes transmitted over the connection
- Decoded payload data, such as application requests and responses
- State-related information (e.g., authenticated username)
- Prevention action performed (if any) .

Most network-based IDPSs can also perform packet captures. Typically this is done once an alert has occurred, either to record subsequent activity in the connection or to record the entire connection if the IDPS has been temporarily storing the previous packets [14].

3.3.3 Detection Capabilities

Network-based IDPSs typically offer extensive and broad detection capabilities [15]. Most products use a combination of signature-based detection, anomaly-based detection, and stateful protocol analysis techniques to perform in-depth analysis of common protocols; organizations should use network-based IDPS products that use such a combination of techniques. The detection methods are usually tightly interwoven; for example, a stateful protocol analysis engine might parse activity into requests and responses, each of which is examined for anomalies and compared to signatures of known bad activity. Some products also use the same techniques and provide the same functionality as network behavior analysis (NBA) software [14]. This section discusses the following aspects of detection capabilities: (1) Types of events detected, (2) Detection accuracy, (3) Tuning and customization and (4) Technology limitations.

3.3.3.1 Types of Events Detected

The types of events most commonly detected by network-based IDPS sensors includes:

Application layer reconnaissance and attacks (e.g., banner grabbing, buffer overflows, format string attacks, password guessing, malware transmission). Most network-based IDPSs analyze several dozen application protocols. Commonly analyzed ones include Dynamic Host Configuration Protocol (DHCP), DNS, Finger, FTP, HTTP, Internet Message Access Protocol (IMAP), Internet Relay Chat (IRC), Network File System (NFS), Post Office Protocol (POP), rlogin/rsh, Remote Procedure Call (RPC), Session Initiation Protocol (SIP), Server Message Block (SMB), SMTP, SNMP, Telnet, and Trivial File Transfer Protocol (TFTP), as well as database protocols, instant messaging applications, and peer-to-peer file sharing software [1].

Transport layer reconnaissance and attacks (e.g., port scanning, unusual packet fragmentation, SYN floods). The most frequently analyzed transport layer protocols are TCP and UDP [16].

Network layer reconnaissance and attacks (e.g., spoofed IP addresses, illegal IP header values). The most frequently analyzed network layer protocols are IPv4, ICMP, and IGMP. Many products are also adding support for IPv6 analysis. The level of IPv6 analysis that network-based IDPSs can perform varies considerably among products.

Unexpected application services (e.g., tunneled protocols, backdoors, hosts running unauthorized application services). These are usually detected through stateful protocol analysis methods, which can determine if the activity in a connection is consistent with the expected application protocol, or through anomaly detection methods, which can identify changes in network flows and open ports on hosts.

Policy violations (e.g., use of inappropriate Web sites, use of forbidden application protocols). Some types of security policy violations can be detected by IDPSs that allow administrators to specify the characteristics of activity that should not be permitted, such as TCP or UDP port numbers, IP addresses, Web site names, and other pieces of data that can be identified by examining network traffic [1].

Some IDPSs can also monitor the initial negotiation conducted when establishing encrypted communications to identify client or server software that has known vulnerabilities or is misconfigured. This can include application layer protocols such as secure shell (SSH) and Secure Sockets Layer (SSL), and network layer virtual private networking protocols such as IP Security (IPsec) [3].

3.3.3.2 Detection Accuracy

Historically, network-based IDPSs have been associated with high rates of false positives and false negatives [20]. Most of the early technologies relied primarily on signature-based detection, which by itself is accurate only for detecting relatively simple well-known threats. Newer technologies use a combination of detection methods to increase accuracy and the breadth of detection, and generally the rates of false positives and false negatives have declined. Another common problem with network-based IDPSs' accuracy is that they typically require considerable tuning and customization to take into account the characteristics of the monitored environment [17].

Ideally, network-based IDPSs would be able to interpret all network activity just as the endpoints do [21]. For example, different types of Web servers can interpret the same Web request in different ways. Stateful protocol analysis techniques often attempt to do this by replicating the processing performed by common types of clients and servers. This allows the sensors to improve their detection accuracy slightly. Many attackers employ client and server-specific processing characteristics, such as handling character encodings, in their attacks as evasion techniques. Organizations should use network-based IDPSs that can compensate for the use of common evasion techniques [14].

3.3.3.3 Tuning and Customization

As mentioned in Section above, network-based IDPSs usually require extensive tuning and customization to improve their detection accuracy. Examples of tuning and customization capabilities are thresholds for port scans and application authentication attempts, blacklists and whitelists for host IP addresses and usernames, and alert settings. Some products also provide code editing features, which is usually limited to signatures but in some cases may allow access to additional code, such as programs used to perform stateful protocol analysis [1].

Some network-based IDPSs can use information regarding the organization's hosts to improve detection accuracy [1][3]. For example, an IDPS might allow administrators to specify the IP addresses used by the organization's Web servers, mail servers, and other common types of hosts, and also specify the types of services provided by each. This allows the IDPS to better prioritize alerts; for example, an alert for an Apache attack directed at an Apache Web server would have a higher priority than the same attack directed at a different type of Web server. Some network-based IDPSs can also import the results of vulnerability scans and use them to determine which attacks would likely be successful if not blocked. This allows the IDPS to make better decisions on prevention actions and prioritize alerts more accurately [1].

3.3.3.4 Technology Limitations

Although network-based IDPSs offer extensive detection capabilities, they do have some significant limitations. Three of the most important are analyzing encrypted network traffic, handling high traffic loads, and withstanding attacks against the IDPSs themselves. These limitations are discussed below [28].

Network-based IDPSs cannot detect attacks within encrypted network traffic, including virtual private network (VPN) connections, HTTP over SSL (HTTPS), and SSH sessions. As previously mentioned, some network-based IDPSs can do some analysis of the setup of encrypted connections, which can identify that the client or server software has known vulnerabilities or is misconfigured [28].

Network-based IDPSs may be unable to perform full analysis under high loads. Passive IDPS sensors might drop some packets, which could cause some incidents to go undetected, especially if stateful protocol analysis methods are in use [1]. For inline IDPS sensors, dropping packets under high loads causes disruptions in network availability; also, delays in processing packets could cause unacceptably high latency. To avoid this, organizations using inline IDPS sensors should select ones that can recognize high load conditions and either pass certain types of network traffic through the sensor without performing full analysis (i.e., partial or no analysis) or drop low-priority traffic to reduce load [1].

IDPS sensors are susceptible to various types of attacks [2]. Attackers can generate unusually large volumes of traffic, such as distributed denial of service (DDoS) attacks, and anomalous activity (e.g., unusually fragmented packets) to attempt to exhaust a sensor's resources or cause it to crash. Another attack technique, known as *blinding*, generates network traffic that is likely to trigger many IDPS alerts in a short period of time; typically, the network traffic is specially crafted to take advantage of typical configurations of IDPS sensors.

In many cases, the blinding traffic is not intended to actually attack any targets. An attacker runs the "real" attack separately at the same time as the blinding traffic [2].

3.3.4 Prevention Capabilities

Network-based IDPS sensors offer various prevention capabilities [1], including the following:

Ending the Current TCP Session. A passive sensor can attempt to end an existing TCP session by sending TCP reset packets to both endpoints; this is sometimes called *session sniping*.

Performing Inline Firewalling. Most inline IDPS sensors offer firewall capabilities that can be used to drop or reject suspicious network activity.

Throttling Bandwidth Usage. If a particular protocol is being used inappropriately, such as for a DoS attack, malware distribution, or peer-to-peer file sharing, some inline IDPS sensors can limit the percentage of network bandwidth that the protocol can use. This prevents the activity from negatively impacting bandwidth usage for other resources.

Altering Malicious Content. As described earlier some inline IDPS sensors can sanitize part of a packet, which means that malicious content is replaced with benign content and the sanitized packet sent to its destination. A sensor that acts as a proxy might perform automatic normalization of all traffic, such as repackaging application payloads in new packets [1].

Reconfiguring Other Network Security Devices. Many IDPS sensors can instruct network security devices such as firewalls, routers, and switches to reconfigure themselves to block certain types of activity or route it elsewhere. This can be helpful in several situations, such as keeping an external attacker out of a network and quarantining an internal host that has been compromised

Running a Third-Party Program or Script. Some IDPS sensors can run an administrator-specified script or program when certain malicious activity is detected. This could trigger any prevention action desired by the administrator, such as reconfiguring other security devices to block the malicious activity. Third-party programs or scripts are most commonly used when the IDPS does not support the prevention actions that administrators want to have performed.

4. ADVANTAGES OF NETWORK BASED INTRUSION DETECTION SYSTEMS:

The followings are merits of NIDS [19].

1. Lower Cost of Ownership: Network based IDS can be deployed for each network segment. An IDS monitors network traffic destined for all the systems in a network segment. This nullifies the requirement of loading software at different hosts in the network segment. This reduces management overhead, as there is no need to maintain sensor software at the host level.

2. Easier to deploy: Network based IDS are easier to deploy as it does not affect existing systems or infrastructure. The network-based IDS systems are Operating system independent. A network based IDS sensor will listen for all the attacks on a network segment regardless of the type of the operating system the target host is running.

3. Detect network based attacks: Network based IDS sensors can detect attacks, which host-based sensors fail to detect. A network based IDS checks for all the packet headers for any malicious attack. Many IP-based denial of service attacks like TCP SYN attack, fragmented packet attack etc. can be identified only by looking at the packet headers as they travel across a network. A network based IDS sensor can quickly detect this type of attack by looking at the contents of the packets at the real time.

4. Retaining evidence: Network based IDS use live network traffic and does real time intrusion detection. Therefore, the attacker cannot remove evidence of attack. This data can be used for forensic analysis. On the other hand, a host-based sensor detects attacks by looking at the system log files. Lot of hackers are capable of making changes in the log files so as to remove any evidence of an attack.

5. Real Time detection and quick response: Network based IDS monitors traffic on a real time. So, network based IDS can detect malicious activity as they occur. Based on how the sensor is configured, such attack can be stopped even before they can get to a host and compromise the system. On the other hand, host based systems detect attacks by looking at changes made to system files. By this time critical systems may have already been compromised.

6. Detection of failed attacks: A network based IDS sensor deployed outside the firewall (as shown in picture1 above) can detect malicious attacks on resources behind the firewall, even though the firewall may be rejecting these attempts. This information can be very useful for forensic analysis. Host based sensors do not see rejected attacks that could never hit a host inside the firewall.

5. TYPES OF INTRUSION DETECTION MECHANISMS

Any enterprise or organisation transacting over the network should require intrusion detection and preventing mechanism. Below I provide a list of vendors that offer Intrusion Detection products and services. Products vary from freeware to commercially available [24].

Freeware:

- Snort - http://www.snort.org/
- Shadow

Commercially Available:

- RealSecure from ISS - http://www.iss.net/customer_care/resource_center/product_lit/
- NetProwler from Symantec- http://enterprisesecurity.symantec.com/products/products.cfm?ProductID=50&PID=5863267
- NFR - http://www.nfr.com/

6. SNORT OVERVIEW

Snort® is an open source network intrusion prevention and detection system (IDS/IPS) developed by Sourcefire [26]. Combining the benefits of signature, protocol, and anomaly-based inspection, Snort is the most widely deployed IDS/IPS technology worldwide. With millions of downloads and nearly 400,000 registered users, Snort has become the de facto standard for IPS [24].

The Snort® open source intrusion detection and prevention technology was created in 1998 by Martin Roesch [24], the founder of Sourcefire®. Snort uses a rule-driven language which combines the benefits of signature, protocol and anomaly-based inspection methods. With its dramatic speed, power and performance, Snort quickly gained momentum. With nearly 4 million downloads to date, Snort has become the single most widely deployed intrusion detection and prevention technology in the world [24].

The wide availability of open source Snort brings many advantages.
- Because the source code is open, development can occur at a markedly accelerated pace compared to proprietary models
- A vast community of security experts that continually reviews, tests, and proposes improvements to the code
- Security engineers and specialists the world over write Snort rules for new and evolving threats every hour of the day, often in record time.

As a result, the Snort open source community has a well-earned reputation for extraordinary organization and dedication.

6.1 Uses
Snort's open source network-based intrusion detection system (NIDS) has the ability to perform real-time traffic analysis and packet logging on Internet Protocol (IP) networks. Snort performs protocol analysis, content searching, and content matching. The program can also be used to detect probes or attacks, including, but not limited to, operating system fingerprinting attempts, common gateway interface, buffer overflows, server message block probes, and stealth port scans [25].

Snort can be configured in three main modes: sniffer, packet logger, and network intrusion detection. In sniffer mode, the program will read network packets and display them on the console. In packet logger mode, the program will log packets to the disk. In intrusion detection mode, the program will monitor network traffic and analyze it against a rule set defined by the user. The program will then perform a specific action based on what has been identified [25].

6.2 Required Software
Before installing Snort you need to verify that you have a number of software packages installed [27]. These are: Libpcap, PCRE, Libnet and Barnyard.

If you're downloading Snort binaries the only requirements are WinPcap and Barnyard for Windows users [27].

- **Libpcap** In the field of computer network administration, pcap (packet capture) consists of an application programming interface (API) for capturing network traffic. Unix-like systems implement pcap in the libpcap library; Windows uses a port of libpcap known as WinPcap [25].

Monitoring software may use libpcap and/or WinPcap to capture packets traveling over a network. libpcap and WinPcap also support saving captured packets to a file and reading files containing saved packets. Snort uses these files to read network traffic and analyze it [25].

- **Perl Compatible Regular Expressions** (PCRE) is a regular expression C library inspired by Perl's external interface, written by Philip Hazel. The PCRE library is incorporated into a number of prominent open-source programs such as the Apache HTTP Server, the PHP and R scripting languages, and Snort [25].

- **Libdnet** is a generic networking API that provides access to several protocols.

- **Barnyard** is an output system for Snort. Snort creates a special binary output format called unified. Barnyard2 reads this file, and then resends the data to a database back-end. Unlike the database output plugin, Barnyard2 manages the sending of events to the database and stores them when the database temporarily cannot accept connections [26].

- **DAQ** is the Data-Acquisition API that is necessary to use Snort version 2.9.0 and above [24].

6.3 Snort Rules
Once you've downloaded and installed Snort, you must download and maintain a ruleset in order for Snort to have the latest detection capabilities [27].

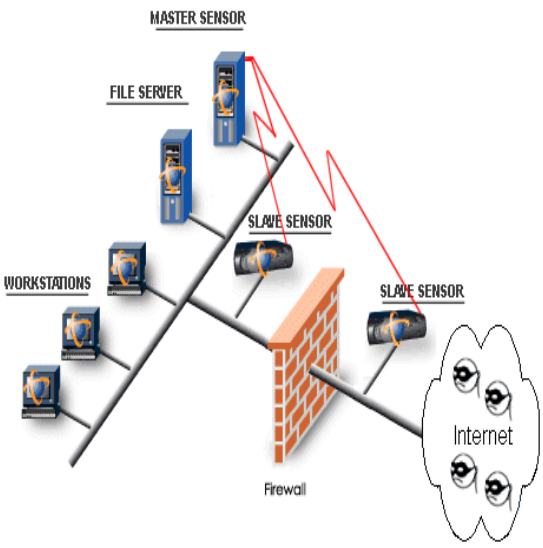
Sourcefire VRT Certified Rules
Sourcefire Vulnerability Research Team (VRT) Rules are the official rules of snort.org. Each rule is developed and tested using the same rigorous standards the VRT uses for Sourcefire customers. These rules are distributed under the VRT Certified Rules License Agreement. This license agreement allows you to study and modify VRT rules but restricts commercial redistribution [26].

There are two ways Snort users can obtain these rules:

- Subscribers: Real-time access to VRT Certified Rules Updates requires a paid subscription.
- Registered Users: Registered users of Snort.org are able to download and use VRT rules free of charge 30 days after their initial release date [24].

6.4 Keeping your Snort Rules Updated
Users may opt to manually download and updates rules files, however most Snort users automate the process using PulledPork, an open source perl script. If you plan on using PulledPork to manage VRT Rules updates you'll need to login to snort.org and generate an Oinkcode to properly configure PulledPork [24].

6.5 Installing Snort
32/64bit Windows Intrusion Detection System (WinIDS) Guided Install



Source - Michael E. Steele @ winsnort.com - [27]

6.5.1 Pre-installation Tasks

- Downloading and extracting the 'WinIDS - Core Software Pack', and the 'WinIDS - (32/64bit) Software Support Pack [27].
- Download The 'WinIDS - Core Software Pack' to a temporary location.

Depending on the processors architecture being used (32bit or 64bit)

- Open an Explorer window and navigate to the location of the 'winids-csdp-xx.xx.xx.zip' file, right-click the 'winids-csdp-xx.xx.xx.zip' file, highlight and left-click 'Extract all...', in the 'Files will be extracted to this folder:' dialog box type 'd:\temp' (less the outside quotes), left-click and uncheck the 'Show extracted files when complete' radio box, left-click extract, in 'Password:' dialog box type 'w1nsn03t.c0m' (less the outside quotes), and left-click 'OK'.

Right-click the 'winids-sdp-xnn-xx.xx.xx.zip' file, highlight and left-click 'Extract all...', in the 'Files will be extracted to this folder:' dialog box type 'd:\temp' (less the outside quotes), left-click and uncheck the 'Show extracted files when complete' radio box, left-click extract, in 'Password:' dialog box type 'w1nsn03t.c0m' (less the outside quotes), left-click 'OK', and eXit the explorer window [27].

6.5.2 Installing WinPcap

The full installation guide can be found at http://winsnort.com/index.php?module=Pages&func=display&pageid=49 [27].

7. RECOMMENDATIONS

Implementing the following recommendations should facilitate more efficient and effective intrusion detection and prevention system use for any enterprise or organisation [1],[3].

Organizations should ensure that all IDPS components are secured appropriately [1]**.**
Securing IDPS components is very important because IDPSs are often targeted by attackers who want to prevent the IDPSs from detecting attacks or want to gain access to sensitive information in the IDPSs, such as host configurations and known vulnerabilities. All components' operating systems and applications should be kept fully up-to-date, and all software-based IDPS components should be hardened against threats. Specific protective actions of particular importance include creating separate accounts for each IDPS user and administrator, restricting network access to IDPS components, and ensuring that IDPS management communications are protected appropriately, such as encrypting them or transmitting them over a physically or logically separate network.

Administrators should maintain the security of the IDPS components on an ongoing basis, including verifying that the components are functioning as desired, monitoring the components for security issues, performing regular vulnerability assessments, responding appropriately to vulnerabilities in the IDPS components, and testing and deploying IDPS updates. Administrators should also back up configuration settings periodically and before applying updates to ensure that existing settings are not inadvertently lost.

Organizations should consider using multiple types of IDPS technologies to achieve more comprehensive and accurate detection and prevention of malicious activity.
The four primary types of IDPS technologies— network-based, wireless, NBA, and host-based— each offer fundamentally different information gathering, logging, detection, and prevention capabilities. Each technology type offers benefits over the others, such as detecting some events that the others cannot and detecting some events with significantly greater accuracy than the other technologies. In many environments, a robust IDPS solution cannot be achieved without using multiple types of IDPS technologies. For most environments, a combination of network-based and host-based IDPS technologies is needed for an effective IDPS solution. Wireless IDPS technologies may also be needed if the organization determines that its wireless networks need additional monitoring or if the organization wants to ensure that rogue wireless networks are not in use in the organization's facilities.

NBA technologies can also be deployed if organizations desire additional detection capabilities for denial of service attacks, worms, and other threats that NBAs are particularly well-suited to detecting. Organizations should consider the different capabilities of each technology type along with other cost-benefit information when selecting IDPS technologies [3].

Organizations planning to use multiple types of IDPS technologies or multiple products of the same IDPS technology type should consider whether or not the IDPSs should be integrated.
Direct IDPS integration most often occurs when an organization uses multiple IDPS products from a single vendor, by having a single console that can be used to manage and monitor the multiple products. Some products can also mutually share data, which can speed the analysis process and help users to better prioritize threats [1]. A more limited form of direct IDPS integration is having one IDPS product provide data for another IDPS product (but no data sharing in the opposite direction). Indirect IDPS integration is usually performed with security information and event management (SIEM) software, which is designed to import information from various security-related logs and correlate events among them. SIEM software complements IDPS technologies in several ways, including correlating events logged by different technologies, displaying data from many event sources, and providing supporting information from other sources to help users verify the accuracy of IDPS alerts [1].

Before evaluating IDPS products, organizations should define the requirements that the products should meet.
Evaluators need to understand the characteristics of the organization's system and network environments, so that a compatible IDPS can be selected that can monitor the events of interest on the systems and/or networks. Evaluators should articulate the goals and objectives they wish to attain by using an IDPS, such as stopping common attacks, identifying misconfigured wireless network devices, and detecting misuse of the organization's system and network resources. Evaluators should also review their existing security policies, which serve as a specification for many of the features that the IDPS products need to provide. In addition, evaluators should understand whether or not the organization is subject to oversight or review by another organization. If so, they should determine if that oversight authority requires IDPSs or other specific system security resources. Resource constraints should also be taken into consideration by evaluators. Evaluators also need to define specialized sets of requirements for the following:

- Security capabilities, including information gathering, logging, detection, and prevention
- Performance, including maximum capacity and performance features
- Management, including design and implementation (e.g., reliability, interoperability, scalability, product security), operation and maintenance (including software updates), and training, documentation, and technical support
- Life cycle costs, both initial and maintenance costs.

When evaluating IDPS products, organizations should consider using a combination of several sources of data on the products' characteristics and capabilities.
Common product data sources include test lab or real-world product testing, vendor-provided information, third-party product reviews, and previous IDPS experience from individuals within the organization and trusted individuals at other organizations. When using data from other parties, organizations should consider the fidelity of the data because it is often presented without an explanation of how it was generated. There are several major challenges in performing in-depth hands-on IDPS testing, such as the considerable resources needed and the lack of a standard test methodology and test suites, which often make it infeasible. However, limited IDPS testing is helpful for evaluating security requirements, performance, and operation and maintenance capabilities [1], [3].

8. CONCLUSION

IDS are becoming the logical next step for many organizations after deploying firewall technology at the network perimeter. IDS can offer protection from external users and internal attackers, where traffic doesn't go past the firewall at all.

However, the following points are very important to always keep in mind. If all of these points are not adhered to, an IDS implementation along with a firewall alone can not make a highly secured infrastructure.

1. Strong identification and authentication:
An IDS uses very good signature analysis mechanisms to detect intrusions or potential misuse; however, organizations must still ensure that they have strong user identification and authentication mechanism in place.

2. Intrusion Detection Systems are not a solution to all security concerns:
IDS perform an excellent job of ensuring that intruder attempts are monitored and reported. In addition, companies must employ a process of employee education, system testing, and development of and adherence to a good security policy in order to minimize the risk of intrusions.

3. An IDS is not a substitute for a good security policy:
As with other security and monitoring products, an IDS functions as one element of a corporate security policy. Successful intrusion detection requires that a well-defined policy must be followed to ensure that intrusions and vulnerabilities, virus outbreaks, etc. are handled according to corporate security policy guidelines.

4. Human intervention is required:
The security administrator or network manager must investigate the attack once it is detected and reported, determine how it occurred, correct the problem and take necessary action to prevent the occurrence of the same attack in future.

Lastly, Tight integration between host and network based IDS is very much necessary. As shown in Fig. 3.2 and 3.3, it is advised to use network based IDS inside and outside the firewall or between each firewall in a multi-layered environment and host based IDS on all critical or key hosts. Also, as shown in Picture1, it is important although not always necessary to have an integrated deployment of host based and network based Intrusion Detection Systems.

As security continues to move to the center stage, managers and network administrators alike are beginning to focus their attention on intrusion-detection technology. While modern-day IDSes are far from bulletproof, they can add significant value to established information-security programs. With vendors working on eliminating the shortcomings of Intrusion Detection Systems, the future looks brighter for this technology.

REFERENCES

[1] Bace, Rebecca, *Intrusion Detection*, Macmillan Technical Publishing, 2000.
[2] Bejtlich, Richard, *The Tao of Network Security Monitoring: Beyond Intrusion Detection*, Addison-Wesley, 2004.
[3] Crothers, Tim, *Implementing Intrusion Detection Systems: A Hands-On Guide for Securing the Network*, 2002.
[4] Endorf, Carl et al, *Intrusion Detection and Prevention*, McGraw-Hill Osborne Media, 2003.
[5] Kruegel, Chris et al, *Intrusion Detection and Correlation: Challenges and Solutions*, Springer, 2004.
[6] Nazario, Jose, *Defense and Detection Strategies Against Internet Worms*, Artech House Publishers, 2003.
[7] Northcutt, Stephen and Novak, Judy, *Network Intrusion Detection: An Analyst's Handbook, Third Edition*, New Riders, 2003.
[8] Rash, Michael et al, *Intrusion Prevention and Active Response: Deployment Network and Host IPS*, Syngress, 2005.
[9] NIST SP 800-86, *Guide to Integrating Forensic Techniques into Incident Response*, http://csrc.nist.gov/publications/nistpubs/.
[10] NIST SP 800-94, *Guide to Intrusion Detection and Prevention Systems (IDPS)*: Recommendations of the National Institute of Standards and Technology by Karen Scarfone and Peter Mell.
[11] NIST SP 800-95, *Guide to Web Services Security (DRAFT)*, available at http://csrc.nist.gov/publications/drafts.html.
[12] NIST Special Publication (SP) 800-61, *Computer Security Incident Handling Guide*, available at http://csrc.nist.gov/publications/nistpubs/.
[13] Distributed Intrusion Detection System (DShield), http://dshield.org/indexd.html
[14] IETF Intrusion Detection Exchange Format (idwg) Working Group, http://www.ietf.org/html.charters/OLD/idwg-charter.html
[15] An Introduction to Intrusion Detection Systems, http://www.securityfocus.com/infocus/1520
[16] Evaluating Intrusion Prevention Systems, http://www.cioupdate.com/article.php/3563306
[17] Intrusion Detection System Overview, http://www.webopedia.com/TERM/I/intrusion_detection_system.html
[18] Recommendations for Deploying an Intrusion-Detection System, http://searchsecurity.techtarget.com/tip/1,289483,sid14_gci781471,00.html
[19] http://www.sans.org/reading_room/whitepapers/detection/understanding-intrusion-detection-systems_337
[20] http://www.iana.org/assignments/version-numbers
[21] http://www.iana.org/assignments/protocol-numbers.
[22] http://www.iana.org/assignments/icmp-parameters.
[23] http://www.iana.org/assignments/ethernet-numbers
[24] http://www.snort.org/start/rules
[25] http://en.wikipedia.org/wiki/Snort_(software)
[26] http://www.sourcefire.com/security-technologies/open-source/snort
[27] http://winsnort.com/index.php?module=Pages&func=display&pageid=49. Written by Michael E. Steele.
[28] http://www.juniper.net/solutions/literature/white_papers/200063.pdf

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Impacts of D-STATCOM on Voltage Stability

Saeed Mohammadi
Kermanshah University of Technology
sd.mohammadi1@gmail.com

ABSTRACT

power quality is a major issue in the distribution system. There will be problem occurs regarding reactive power transfer in distribution system due to large power angle even with substantial voltage magnitude gradient. Here a DSTATCOM is used as a FACT device which can compensate reactive power. D-STATCOM is three phase voltage source converter used to compensate voltage and make the system stable by absorbing and generating reactive power. Here Simulink model and control system is designed in MATLAB Simulink environment. Here modeling of the D-STATCOM is also given.

Keywords: D-STATCOM, Distribution System, Line Voltage, Voltage Stability.

African Journal of Computing & ICT Reference Format
Saeed Mohammadi (2013). Impacts of D-STATCOM on Voltage Stability.
Afr J. of Comp & ICTs. Vol 6, No. 1. Pp 121-128

1. INTRODUCTION

Power quality is set of electrical boundaries that allow the piece of equipment to function in its intended manner without significant loss of performance or life expectancy. The electrical device like electric motor, a transformer, a generator, a computer, a printer, communication equipment, or a house hold appliance. All of these devices and others react adversely to power quality issues, depending on the severity of problems. Reactive power cannot be transmitted across large power angle even with substantial voltage magnitude gradient. Reactive power should be generated close to the point of consumption. We can make several reason to minimize reactive power transfers.

- It is inefficient during high real power transfer and require substantial voltage magnitude gradient
- It causes high real and reactive power losses
- It can lead to damaging temporary overvoltage's following load rejections
- It requires larger equipment size for transformer and cables

Due to this here a D-STATCOM as shunt FACT devices is used. A distribution static synchronous compensator (D-STATCOM) is a fast response, solid-stat power controller that provides flexible voltage control at the point connection to the utility distribution feeder for power quality (PQ) improvements. It can regulate the bus voltage by absorbing or generating reactive power from system to the converter and converter to the system at the point of common coupling.

2. D-STATCOM

The D-STATCOM is three phase shunt connected power electronics based device. It is connected near the load at the distribution system. It is also a one type of the voltage-source converter, which converts a DC input voltage into AC output voltage in order to compensate the active and reactive power needed by the system. The DSTATCOM mainly consists of DC voltage source behind self-commutated inverters using IGBT and coupling transformer. A three phase IGBT based current controlled voltage source inverter with a dc bus capacitor is used as a DSTATCOM.

D-STATCOM improves supply power factor, provide load balancing & improve load terminal voltage. DSTATCOM limits the short circuit current, improves the system transient stability limit and increases the load ability of the system. D-STATCOM controller is highly effective in improving the power quality at the distribution level by making the voltage stable. A voltage Source converter is connected to bus via three phase transformer. A voltage source converter is a power electronics device, which can generate sinusoidal voltage with required magnitude frequency and phase angle. The VSC is used to either completely replace the voltage or to inject the missing voltage. Here missing voltage is the difference between the nominal voltage and the actual voltage. The solid state electronics in the converter is then switched to get the desired output. The filter bank at the inverter output to absorb harmonics.

A 10000 μF capacitor is used as a dc voltage source for the inverter. PCC is the point of common coupling at which the generation or absorption of reactive Power takes Place to and from the system and the device. At the distribution voltage level, the switching device is generally the IGBT due to its lower switching losses and reduced size. D-STATCOM has advantage than other devices like conventional SVC that it can operate at lower voltage and it does not produce lower order harmonics.

3. FUNCTION OF D-STATCOM

Fig 1 shows a simplified diagram of a STATCOM connected to a typical distribution network represented by an equivalent network.

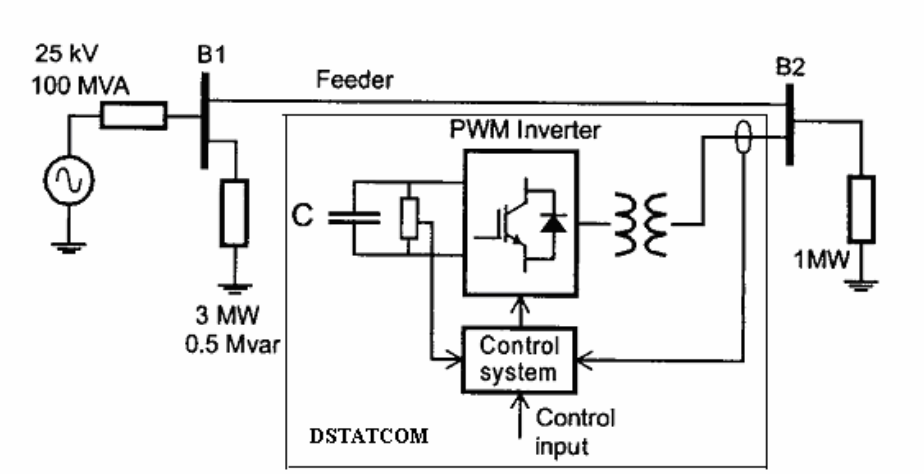


Fig 1 simplified diagram of a D-STATCOM connected to a distribution network

The D-STATCOM consists mainly of a PWM inverter connected to the network through a transformer. The dc link voltage is provided by the capacitor which is charged with the power taken from the network. The control system ensures the regulation of the bus voltage and the dc link voltage. The principle operation of the DSTATCOM depends upon reactive current generation, so (I) varies as:

$$I = \frac{V - V0}{X}$$

Where *V0*, *V*, *X* are the output voltage of the IGBT-based inverter, the system voltage, the total circuit reactance (transformer leakage reactance plus system short circuit reactance) respectively

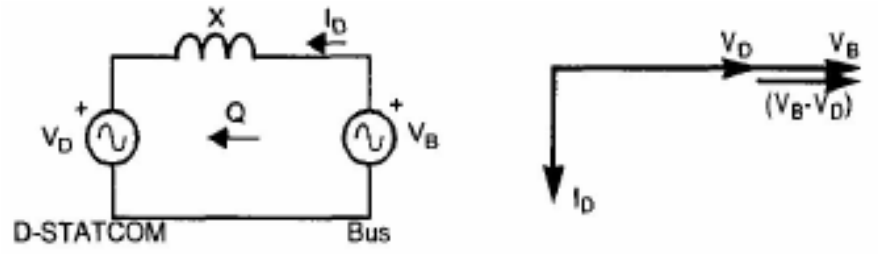


Figure 2 D-STATCOM operation
(a) Inductive operation, (b) capacitive operation

When the secondary voltage (VD) is lower than the bus voltage (VB), the D-STATCOM acts like an inductance absorbing reactive power from the bus. When the secondary voltage (VD) is higher than the bus voltage (VB), the DSTATCOM acts like a capacitor generating reactive power to the bus.

4. MODELLING APPROACH

There are three modelling approach, average modelling, detailed modelling and phase to phase modelling. Generally phase to phase modelling is not used for D-STATCOM.
In the average modeling the IGBT Voltage-Sourced Converters (VSC) are represented by equivalent voltage sources generating the AC voltage averaged over one cycle of the switching frequency. This model does not represent harmonics, but the dynamics resulting from control system and power system interaction is preserved. This model allows using much larger time steps (typically 50 microseconds), thus allowing simulations of several seconds.

The detailed model includes detailed representation of power electronic IGBT based converters. In order to achieve an acceptable accuracy with the 1680 Hz switching frequency used. the model must be discredited at a relatively small time step (5 microseconds). This model is well suited for observing harmonics and control system dynamic performance over relatively short periods of times (typically hundreds of milliseconds to one second). Here a detailed modeling is used for the D-STATCOM.

5. D-STATCOM CONTROLLER

The control diagram is shown in Fig 3 it consist of several subsystem: a phase locked loop (PLL), two measurement system, a current regulation loop, a voltage loop, and a DC link voltage regulator. The PLL is synchronized to the fundamental of the transformer primary voltage to provide the synchronous reference (sinωt and cosωt) required by the abc-qd transformation. The measurement block computes the d-axis and q-axis component of voltage and current by executing an abc-qd transformation in the synchronous reference determined by sinωt and cosωt provided by PLL An outer current regulation loop consisting of an AC voltage regulator and a DC voltage regulator. The output of the AC voltage regulator is the reference current iqref for the current regulator.

The output of the DC voltage regulator is the reference current idref for the current regulator. Iq is in quadrature with voltage which control reactive power flow. Id is in phase with voltage which control active power flow An inner current regulation loop consists of a current regulator. The current regulator controls the magnitude and phase of the voltage generated by the PWM converter from the Idref and Iqref reference currents produced respectively by the DC voltage regulator and AC voltage regulator The main function is to operate the converter power switches so as to generate a fundamental output voltage waveform with the demanded magnitude and phase angle in synchronism with the ac system.

6. MODEL DESCRIPTION AND SIMULATION RESULTS

Modeling the D-STATCOM includes the power network and its controller in simulink environment require electrical block from power system block set. The DSTATCOM of ±3 Mvar is connected to a 25-kV distribution network as shown in figure 4. The feeding network represented by bus B1 followed by 21- km feeder which is modeled by a pi-equivalent circuit connected to bus B2. The D-STATCOM output is coupled in parallel with the network through a step-up 1.25/25-kV delta-star transformer. At the output of D-STATCOM a filter bank is provided to absorb harmonics. The primary of this transformer is fed by a voltage-source PWM inverter consisting of two IGBT bridges.

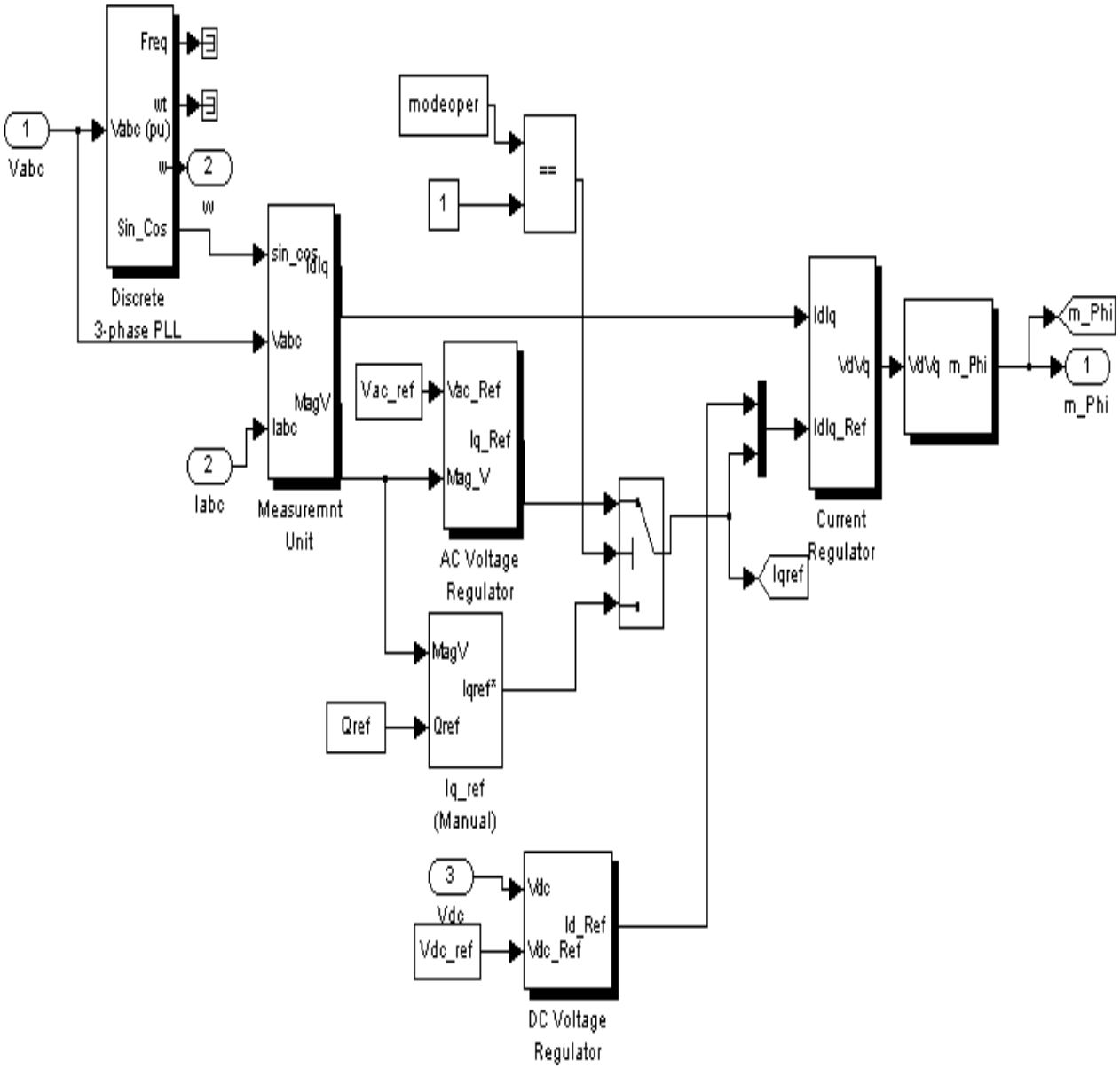


Fig 3 D-STATCOM control system

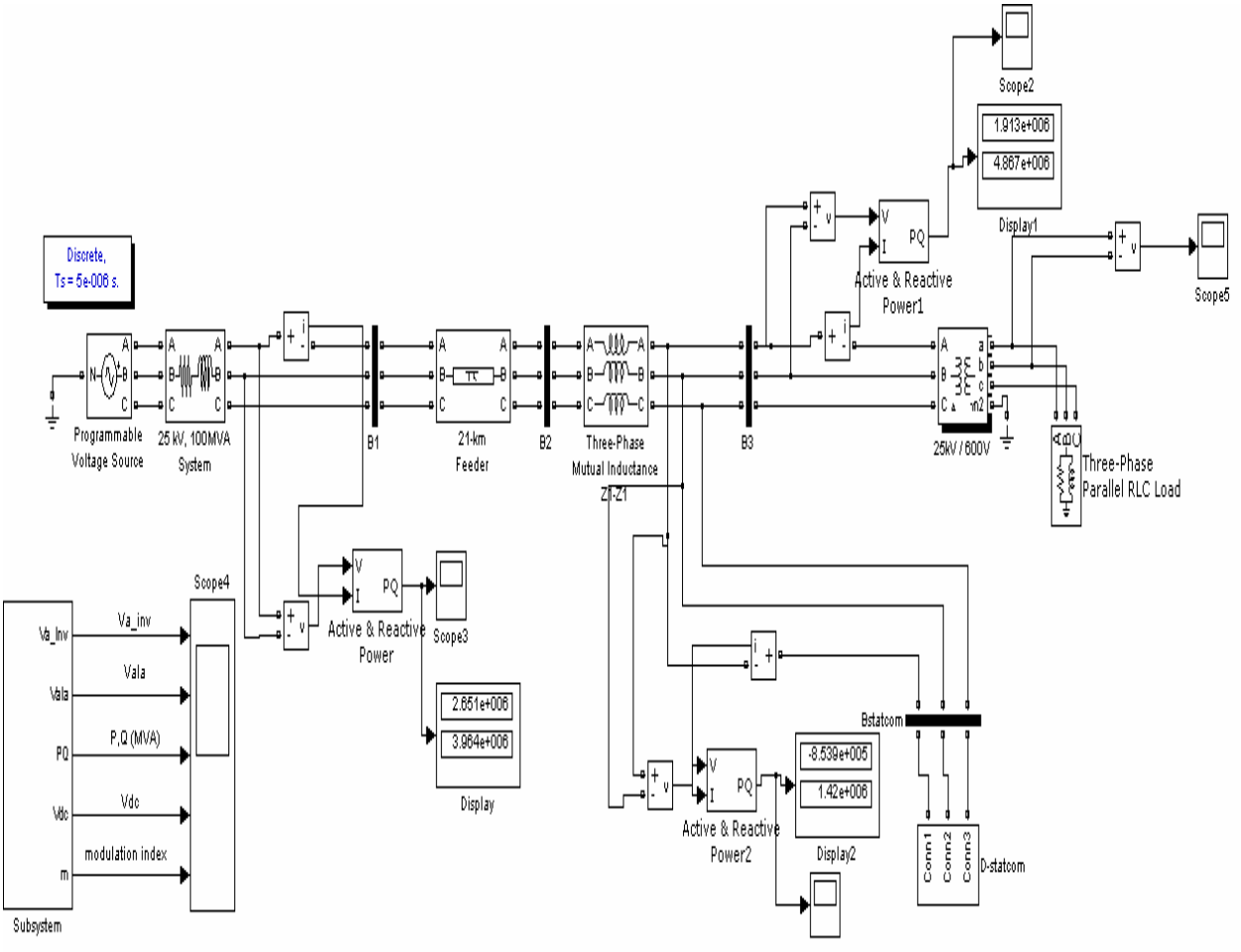


Fig 4. Simulink Model for D-STATCOM with Distribution System

10000 µF capacitor is used as a dc voltage source for the inverter. A Distribution Static Synchronous compensator (DSTATCOM) is used to regulate voltage on a 25-kV distribution network. The D-STATCOM regulates bus B3 voltage by absorbing or generating reactive power. This reactive power transfer is done through the leakage reactance of the coupling transformer by generating a secondary voltage in phase with the primary voltage (network side). This voltage is provided by a voltage-sourced PWM inverter. When the secondary Voltage is lower than the bus voltage, the DSTATCOM acts like an inductive absorbing reactive power. When the secondary voltage is higher than the bus voltage, The D-STATCOM acts like a capacitor generating reactive power. For the voltage compensation:

QL = QG + Qstatcom (1)

Whrer QL is reactive power at load side, QG is reactive power at generating side (sending end), Qstatcom is reactive power generated by D-STATCOM By comparing results we can show that the equation (1) can be approximately satisfied which shows that voltage compensation has been done with the help of D-STATCOM. Performance characteristics of D-STATCOM in distribution system is given below:

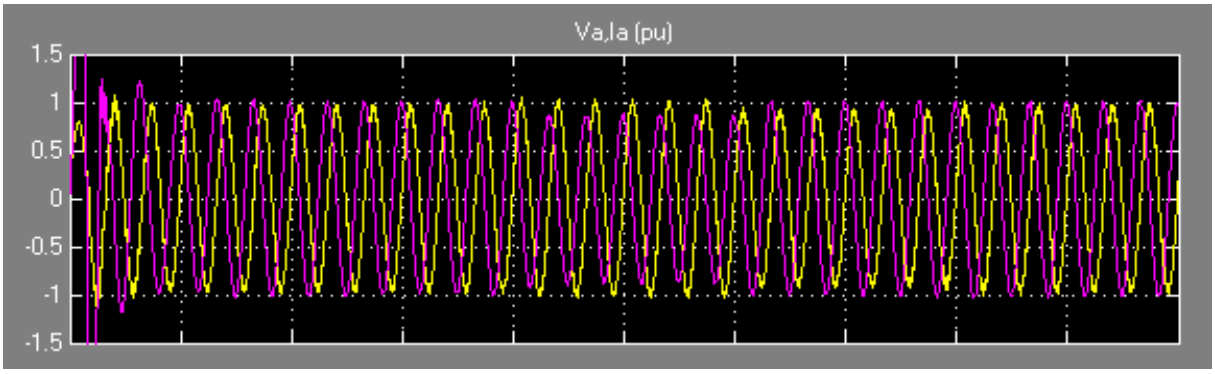


Fig 5 phase voltage and current

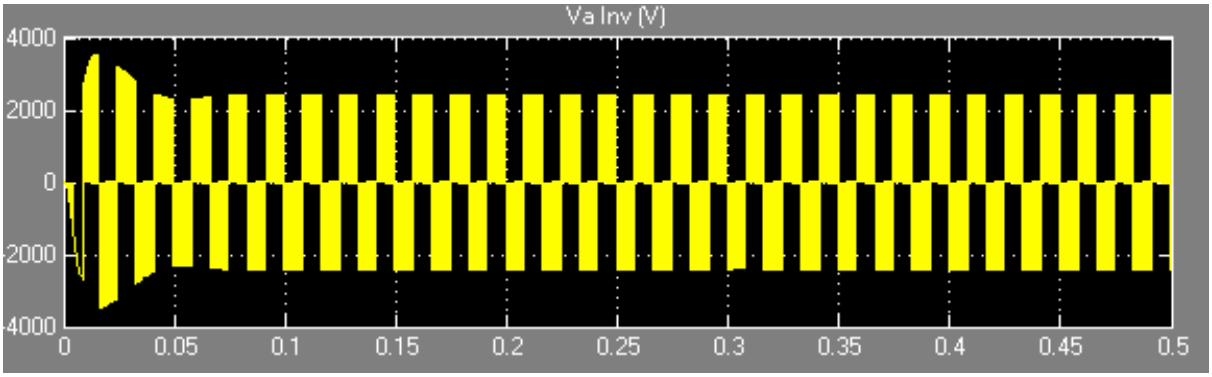


Fig 6 inverter waveform

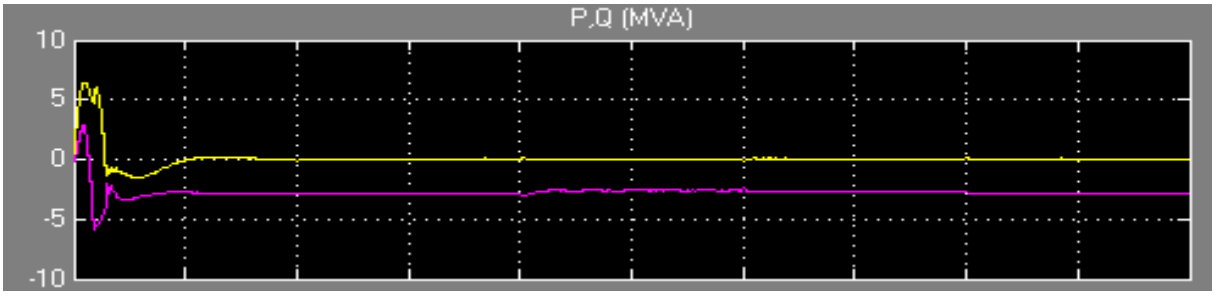


Fig 7 real and reactive power

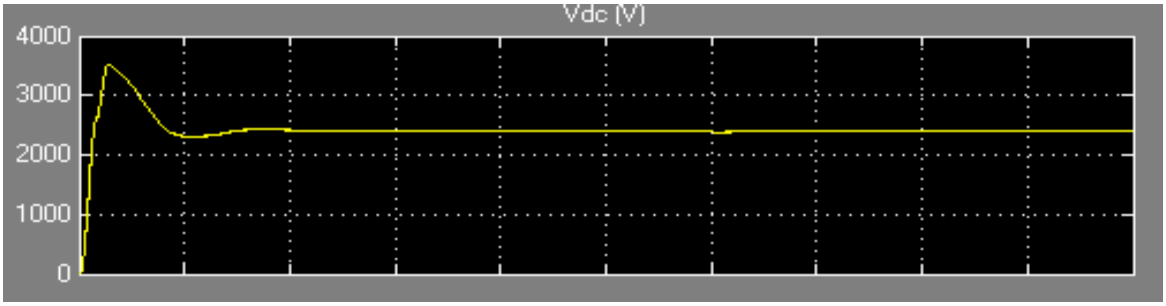


Fig 8 dc link voltage

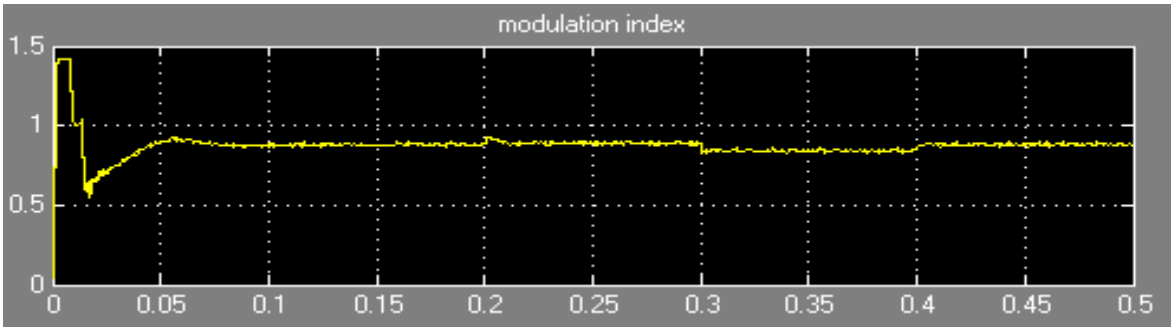


Fig 9 modulation index

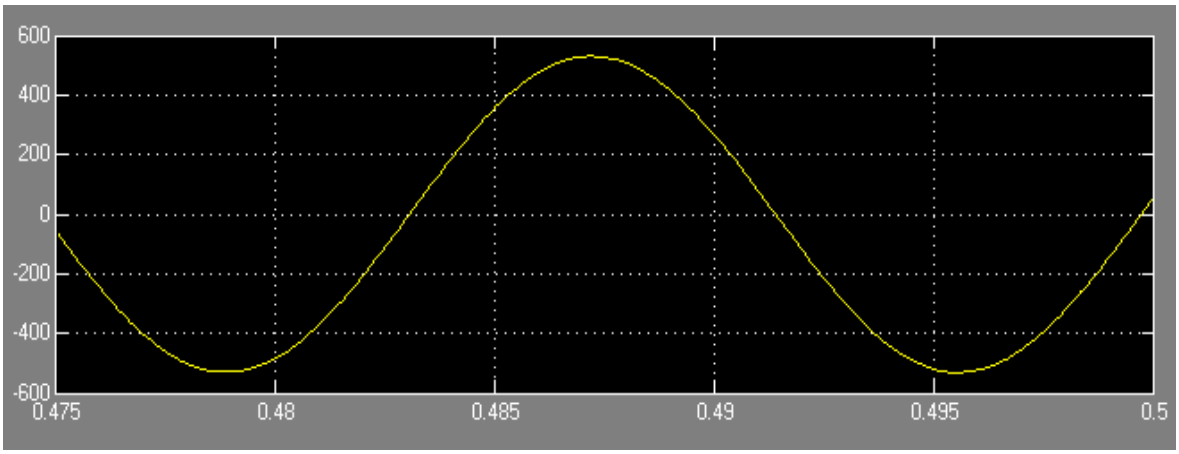


Fig 10 line voltage at load side with D-STATCOM

Fig 5 and Fig 6 shows the results for the voltage and current waveform during the change from inductive to capacitive operation. Fig 7 shows the results for the real and reactive power, which shows that D-STATCOM can be work as a capacitive generating reactive power due to an inductive load. Fig 8 shows the DC link voltage and Fig 9 shows the results for the modulation index. Fig 10 shows the line voltage at load side here by varying an inductive load the voltage will remain approximately constant up to certain limit which shows that a D-STATCOM can make voltage stable and it can compensate the voltage.

7. CONCLUSION

The model of a D-STATCOM was analyzed and developed for use in simulink environment with power system block sets. Here a control system is designed in MATLAB simulink. A D-STATCOM can control reactive power and also regulate bus voltage. It can improve power system performance. Here waveform shows the performance of DSTATCOM in a distribution system. By varying an inductive load at some amount we can observe that D-STATCOM can regulate load side voltage approximately constant which shows the voltage stability of the D-STATCOM.

REFERENCES

[1] Pierre Giroux, Gilbert Sybille, Hoang Le-Huy "*modeling and simulation of distribution STATCOM using Simulink power system blockset"*IECON'01: the 27th annual conference of the IEEE industrial electronics society

[2] Sandeep kaur, G S kochar, D S mahal, Sunita Goyal " *power quality improvement using distributed static synchronous compensator*" international conference on electrical power and energy system

[3] G.sundar ,S.Ramareddy,"*digital simulation of D-STATCOM for voltage fluctuations"* Inter National journal of engineering science and technology Vol.2(5),2010,1131-1135

[4] M. G. Molina, P. E. Mercado "*Control Design and Simulation of DSTATCOM with Energy Storage for Power Quality Improvements"* 2006 IEEE PES Transmission and Distribution Conference and Exposition Latin America, Venezuela

[5] *Power System blockset* for Use with simulink, User's guide , The mathworks Inc,2000

[6] K K Sen, "STATCOM: Theory, Modeling ,Applications," in IEEE PES 1999 wintermeeting proceedings , pp 1177-1183.

[7] Bhim singh, ALka Adya, A P Mittal and J R P Gupta, "*Modeling of DSTATCOM for distribution system*", Int. J. energy technology and policy, Vol. 4, no.1/2,2006,pages :142-160

[8] C Sankaran *" power quality*"

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Design and Implementation of Real-Time Crime Information System for National Drug Law Enforcement Agency (NDLEA) in Nigeria

C.H. Ugwuishiwu & K.C. Ugwu
Department of Computer Science
University of Nigeria
Nsukka, Nigeria
Chilkodili.ugwuishiwu@unn.edu.ng

H.C. Inyiamah
Department of Computer Engineering
Nnamdi Azikiwe University
Awka, Nigeria
drhcinyiama@gmail

ABSTRACT
The problems of narcotic drugs in Nigeria as manifested in the cultivation, trafficking and abuse of drug is of increasing concern to the government, social researchers, and other stakeholders. The impact of illicit drugs continues to threaten the economies and social structures of both producing and consuming countries. This paper presents an implementation framework for a reliable, efficient and mobile real- time crime information system (CIS) that will help to maintain instant flow of information between the general public and NDLEA. This application solved this problem through a just-in-time flow of crime information from an informant who could be an individual, corporate body or any other government agency to the NDLEA office for a necessary action. This kind of information if handled properly will lead to illicit drug supply suppression hence reduce its availability as well as other substances of abuse. It will also help in drug demand reduction which result in a decline in the consumption of illicit drug substances.

Keywords: : NDLEA, Law Enforcement Agency, Crime, Drug Trafficking, illicit drug, information System.

African Journal of Computing & ICT Reference Format
C.H. Ugwuishiwu, H.C. Inyiamah & K.C. Ugwu (2013). Design and Implementation of Real-Time Crime Information System for National Drug Law Enforcement Agency (NDLEA) in Nigeria. Afr J. of Comp & ICTs. Vol 6, No. 1. Pp 129-136.

1. INTRODUCTION

Drugs, alcohol and substance abuse have become an ever-increasing problem in Nigeria with serious health and social implications. The cultivation of hard drugs, its abuse together with trafficking has become a source of worry to the government and the society at large. The proliferation of hard drugs and its subsequent abuses has led to mental illness among young people and the deterioration of the health of the nation's active population. Baniyi(2008) said that drug and alcohol abuse contribute significantly to the incidence of domestic aggression, violent crimes, broken homes and juvenile delinquency. The Federal Republic of Nigeria, in reaction to the adverse effects of drug abuse on the health of users and also in fulfillment of the country's international obligation, as a signatory to the 1988 UN Convention, which recommended separate bodies to lead the onslaught against the ravaging drug menace in many parts of the world, set up the NDLEA through Decree No. 48 of 1989, NDLEA (2011).

This agency as provided in the decree was charged with eliminating the growing, processing, manufacturing, selling, exporting, and trafficking of hard drugs and psychotropic substances. Since the establishment of NDLEA, despite all the adopted strategies geared towards the control of narcotic drug abuse, illicit drug trafficking is still on the increase. Hence, the question on the lips of everyone is why has the NDLEA been visibly inefficient in drug prevention and control? Generally speaking, criminological insights have shown that very little of the crimes committed in the society are known to law enforcement agencies. Consequently, few of the criminals are arrested and prosecuted and the impacts in our society include premature death, crime, mental health disorders, transmission of HIV/AIDS and other blood borne viruses.

According to Gyong and Tanimu (2010), less than half of the crimes committed in America each year are reported to the police, and of those recorded by the police, only about 20% are "cleared by arrest", still an arrest by no means guarantees prosecution and conviction. Similarly, crimes in Nigeria are under-detected, underreported, under-recorded and therefore underestimated. One major cause of these problems is lack of timely flow of crime information and effective collaboration between the NDLEA and the general public.

The illicit drug problem has persisted because we have - not devised a working solution to the problem. Thus, this work is set to design and develop an effective and efficient mobile real-time crime information system for NDLEA that will enable an informant in any part of the country to instantly send any NDLEA's crime related information as soon as it is detected. This will to a large extent bridge the communication gap between the agency and the general public and also reduce illicit drug-related activities to the barest minimum.

This system will enable the agency to generate annual reports that have records of crimes according to regions, states or cities resulting to accurate statistics to draw analysis such as information of a case, the year a crime was committed, a location with highest crime, age range that normally commit maximum number of crime and the type of crime they commit, officer who investigated the crime. These factors will help in fighting crime effectively.
The rest of the work is summarized as follows: Section 2 presents the literature Review; section 3 disuses the system Analysis and Design; section 4 deals with the system Implementation and Discussion of Findings. Section 5 presents the conclusion and recommendation.

2. LITERATURE REVIEW

Montaldo (2013) said that crime occurs when someone breaks the law by an overt act, omission or neglect that can result in punishment. A person who has violated a law, or has breached a rule, is said to have committed a criminal offense. According to National Crime Prevention Strategy (1996), Crime results in the deprivation of the rights and dignity of citizens, and poses a threat to peaceful resolution of differences and rightful participation of all in the democratic process. It inhibits our citizens from communicating with one another freely, from engaging in economic activity and prevents entrepreneurs and investors from taking advantage of the opportunities, which our country offers. For these reasons and many more, Government regards the prevention of crime as a national priority.

Drugs are articles that are intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease in humans or animals, and any articles other than food, water, or oxygen that are intended to affect the mental or body function of humans or animals. Narcotics are any drugs that dull the senses and commonly become addictive after prolonged use. There is no known drug that is not harmful or even poisonous at high doses, and much of the scientific work on drugs has attempted to elucidate the gap between effective and toxic doses. According to United Nations Office on drug and crime (2013), drug trafficking is a global illicit trade involving the cultivation, manufacture, distribution and sale of substances which are subject to drug prohibition laws.

The main examples of illicit drugs are the opiates (mostly heroin), cocaine, cannabis, and ATS (amphetamine-type stimulants) such as amphetamines, methamphetamine and ecstasy. National Drug Law Enforcement Agency (2011) stated that drug problem is as old as man and that no society is insulated from the negative consequences of illicit drugs.

Many analysts are of the opinion that apart from the genocide of Second World War, no other phenomenon has had more debilitating consequences on mankind like the pandemic drug scourge. This view is anchored on the fact that even the much dreaded HIV/AIDS which has yet defied any known cure has narcotic drugs as one of its principal causes. Besides, drugs are known to induce social vices, civil upheavals and other forms of criminalities, therefore the war on drugs should be fought with vigour to reduce crime in the society, *Yahaya (2011).*

UNRISD (1994) observed that drug users not only suffer physical, social and economic problems themselves, but they also impose many direct and indirect costs on society. Of particular concern is the relationship between drug use and crime, especially the violent crime associated with crack cocaine. Women and children who are not themselves drug users may also be affected by problems related to drug abusing men, including HIV infection. This report also noted that despite the expenditure of billions of dollars on controlling supplies, the successful capture of major traffickers and dismantling of their drug empires, and the weaning of peasant growers away from coca and opium via alternative crop incentives, supply and consumption of drugs has continued to grow.

General Assembly of the United Nations (2012) report highlighted that "We will not enjoy development without security, and we will not enjoy security without development" there is a growing recognition that organized crime and illicit drugs are major impediments to development.

Tim and Tivani (2006), describes prevention as 'the cornerstone of drug control', as 'other methods are useless unless the individual comes to the conclusion that drug use is against his/her own interests'. Prevention includes primary approaches (directly stopping people using drugs such as through education or law enforcement) secondary approaches (convincing users to quit), and tertiary approaches (harm reduction approaches, such as preventing needle sharing).

2.1 Some Existing Crime Information System

An information system (IS) is any system that converts data into useful information. Kiirya (2007) described information systems to consist of components that interact to achieve the objective of providing information about day-to-day activities that managers can use to control business operations. An information system contains such elements as hardware, software, personnel, databases, and procedures to accomplish its objectives.

Crime Information System (CIS) is any system that could be used to keep records of crimes or criminal cases which can also help in getting information pertaining to certain crimes. CIS can be used to carry out statistical analysis of crime committed within a region-e.g. South-Eastern region, Nigeria.

Regional Information Sharing Systems (RISS) – The RISS Program is a federally funded program administered by the U.S. Department of Justice (DOJ), Bureau of Justice Assistance (BJA). This crime information system supports law enforcement efforts nationwide to combat illegal drug trafficking, identity theft, human trafficking, violent crime, terrorist activity, and promote officer safety. RISS was established more than 30 years ago, in response to specific regional crime problems and the need for cooperation and secure information sharing among law enforcement agencies. Today, RISS is a national network comprised of six multistate centers designed to operate on a regional basis. RISS operates a secure intranet, known as RISSNET, to facilitate law enforcement communications and information sharing nationwide, Information Sharing Environment *(2012).*

The National Crime Information Center (NCIC) is a USA computerized database of documented criminal justice information available to virtually every law enforcement agency nationwide. It is a nationwide system operated by the Federal Bureau of Investigation (FBI) dedicated to serving and supporting local, state, and federal criminal justice agencies in their mission to uphold the law and protect the public. The NCIC became operational on January 27, 1967, with the goal of assisting law enforcement in apprehending fugitives and locating stolen property. This goal has since expanded to include locating missing persons and further protecting law enforcement personnel and the public.

The NCIC database consists of 18 files. Seven property files contain records for articles, boats, guns, license plates, securities, vehicles, and vehicle and boat parts. The 11 person files are the Convicted Sexual Offender Registry, Foreign Fugitive, Identity Theft, Immigration Violator, Missing Person, Protection Order, Supervised Release, Unidentified Person, U.S. Secret Service Protective, Violent Gang and Terrorist Organization, and Wanted Person Files. In addition, the database contains images that can be associated with NCIC records to assist agencies in identifying people and property items. The Interstate Identification Index, which contains automated criminal history record information, is also accessible through the same network as the NCIC. Commonwealth (2013).

Police Information Management System (PIMS) is an offline project that allows user to store police department's case details, Complaint Details, First Information Report (FIR) Details, etc. This Software Package allows Police Departments to store all the details related to the department and use them whenever necessary. The project reports various cases, FIR report, charge sheet report, Most Wanted Criminals record, payroll, attendance reports and also can upload and view criminal photos and scanned documents. The requirements for this system was collected from Manglaore Police Station, *Free Student Project (2012).*

2.2 Drug Consumption and Crime

There is obviously a relationship between drug consumption and crime, although it is often not clear which is cause and which is effect. In principal consuming areas such as North America and Western Europe, psycho-pharmacological effects, economic-compulsive drives and systemic violence are considered the principal components of the drugs-crime link.

The most harmful psycho-pharmacological effects of drug use, particularly those associated with crack cocaine; involve people becoming irrational, excited, agitated or impulsive. Users may become unable to control their anger and vent it in the form of physical assault, including homicide. In one of the first studies clearly linking violent behaviour and crack cocaine use, it was reported that nearly half the callers to a nationwide cocaine hotline in the United States said they had committed violent crimes or aggressive acts (including child abuse, murder, robbery, rape and physical assault) while using crack, UNRISD (1994).

3. THE CRIME INFORMATION SYSTEM ANALYSIS AND DESIGN

This section of the work includes the architecture of the NDLEA crime information system, the database of the system and the system software requirements.

3.1 Architecture of the crime information system

The NDLEA crime information system architecture defines the key components of this system together with the interactions between these components as shown in figure 1. This system architecture consist of an informant, NDLEA Head Office, Database Management System (DBMS), a web based application and NDLEA's branch Offices at different location in the country.

This system is modeled in a way that when an NDLEA related crime is detected by an informant, he/she sends this information through a web application to NDLEA office with either a mobile phone or computer by opening NDLEA website. This information goes to the NDLEA state office (local level) nearest to the scene of the crime and gives an alert by showing on the screen that there is a new message. This enables NDLEA to instantly initiate the crime investigation process, where the state branch cannot handle the case; they forward it to their head office for a necessary action. The crime information will contain the location of the crime, time it occurred, date of the incident, name of suspect if available, the type of crime committed. This information when sent stores a copy in NDLEA database. This application can equally sends an acknowlegdement message to the crime reporter showing that the information he sent has been received. The NDLEA Database Management System (DBMS) contains all the NDLEA crime information which when properly utilized will help NDLEA to large extent in fighting the war of crime.

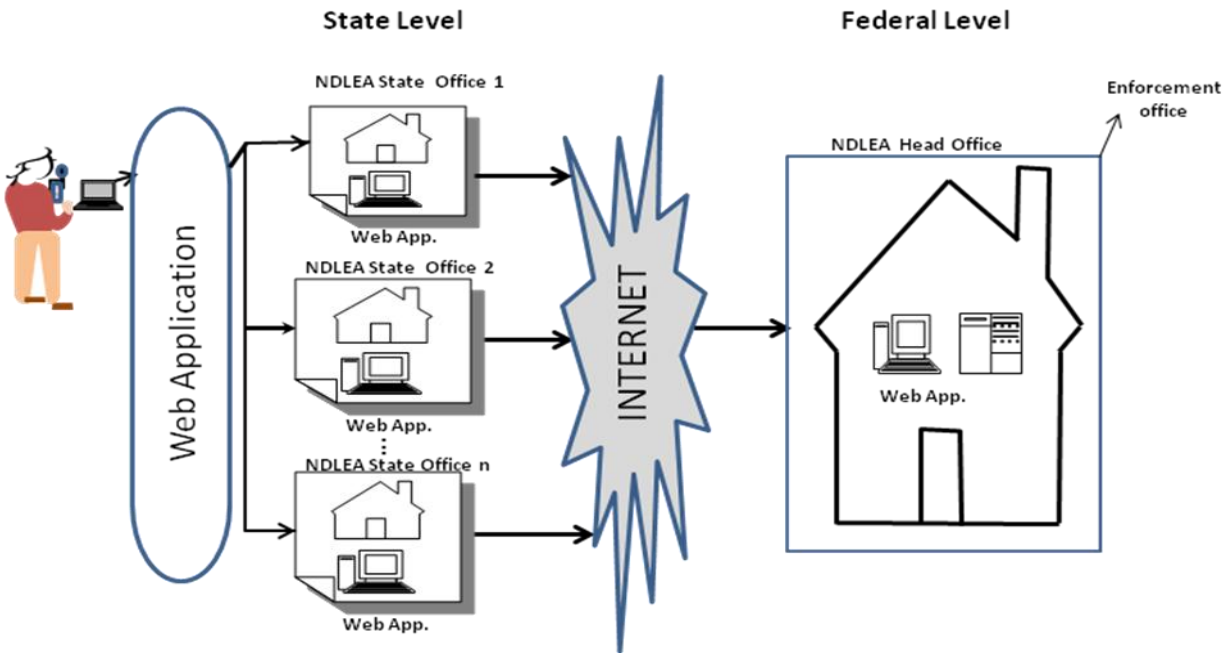


Fig. 1: Architecture of The NDLEA Crime Information System

3.2 Database

Database has the core service for storing, processing and securing data. The database server provides controlled access and rapid transaction processing to meet the requirements of the client. Many tables were used in this work for instance a table that list all the tables in the database of this application, others contain Information on admin page, reported crime details, branch office details, etc.

3.3 Operating System and Programming Platform

Windows Vista, Microsoft SQL Server, Wamp Server and Php, Css, Javascript, Ajax scripting languages were used to implement the application package.

Implementation

This refers to the development, deployment (installation and testing) of all the system components. This involves the transformation of ideas such as algorithms, flowcharts, programming language used, etc. into real process flow of information. The system operates as follows:

A Welcome Screen Form

A welcome screen contains major links to other web pages of the application, some information about NDLEA and also the admin sort code where super administrator (head office admin) can type his password to enter into the admin login page.

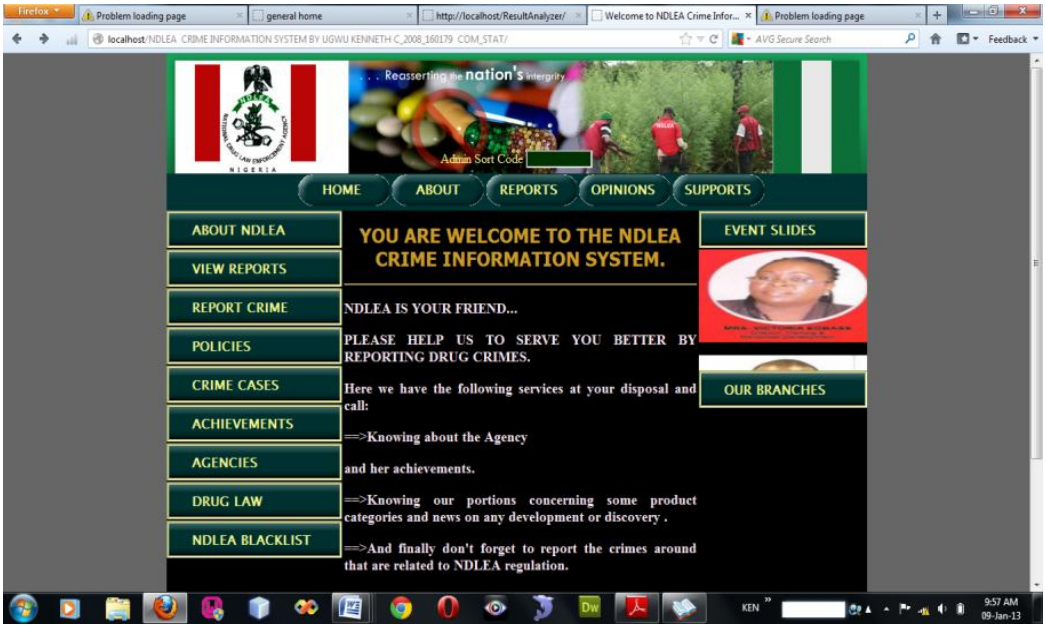
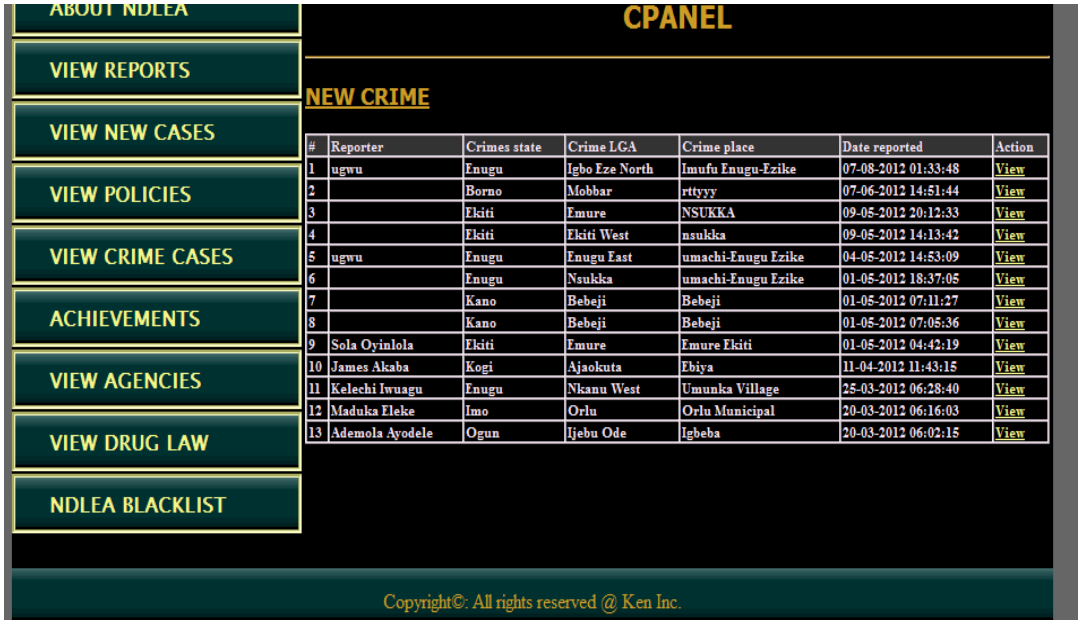


Fig 2.Welcome Screen

Admin Login page/screen- This is where super Admin login to resume his activities using his username and password.

REPORTER' S PROFILE

This page shows the details of the most recently reported crime and it is shown in fig.3. It can be accessed only when the admin logs in. The super Administrator can click on action field (view) this table to confirm, reject or make that information public.



#	Reporter	Crimes state	Crime LGA	Crime place	Date reported	Action
1	ugwu	Enugu	Igbo Eze North	Imufu Enugu-Ezike	07-08-2012 01:33:48	View
2		Borno	Mobbar	rttyyy	07-06-2012 14:51:44	View
3		Ekiti	Emure	NSUKKA	09-05-2012 20:12:33	View
4		Ekiti	Ekiti West	nsukka	09-05-2012 14:13:42	View
5	ugwu	Enugu	Enugu East	umachi-Enugu Ezike	04-05-2012 14:53:09	View
6		Enugu	Nsukka	umachi-Enugu Ezike	01-05-2012 18:37:05	View
7		Kano	Bebeji	Bebeji	01-05-2012 07:11:27	View
8		Kano	Bebeji	Bebeji	01-05-2012 07:05:36	View
9	Sola Oyinlola	Ekiti	Emure	Emure Ekiti	01-05-2012 04:42:19	View
10	James Akaba	Kogi	Ajaokuta	Ebiya	11-04-2012 11:43:15	View
11	Kelechi Iwuagu	Enugu	Nkanu West	Umunka Village	25-03-2012 06:28:40	View
12	Maduka Eleke	Imo	Orlu	Orlu Municipal	20-03-2012 06:16:03	View
13	Ademola Ayodele	Ogun	Ijebu Ode	Igbeba	20-03-2012 06:02:15	View

Fig.3: Recently Reported Crime.

View Report page – This page shows the details of the reported crime. This page shows all the available crime information reported in the database. You can click on view field to view individual crime record. On home page or any other page that contains the application links, you can select any of the options such as About NDLEA, View Reports (all reported crimes),

The crime Report page-
This page is where the reporter fills the details of the crime information to be sent. To send this information, the crime reporter will fill a form called Crime Report Form which has the following fields: reporter's name, reporter's email, reporter's phone, crime state, crime LGA, Crime location and Crime details. Note that fields that are marked asterisk (*) are compulsory to be filled while the rest are optional.

Fig 4. Crime Reports Page

REFERENCES

Other links are Policies (the policies governing the NDLEA can be viewed here). Crime cases (it shows the treated crime cases), achievements (all the achievements made by NDLEA can be viewed), agencies (it shows the working members of NDLEA), Drug Law (the laws regarding to drugs can be viewed), NDLEA Blacklist (it shows the most wanted person by the NDLEA and other similar cases).

4. CONCLUSION

The aim of this research work was achieved by a successful design and implementation of an NDLEA real-time crime information system. This web application enables an informant to instantly communicate NDLEA as soon as crime related to them is detected. From this research, it was observed that there is a strong relationship between drug use and crime, especially the violent crime associated with crack cocaine, it is also noted that these drug users not only suffer physical, social and economic problems themselves, rather they impose many direct and indirect costs on society. It thereby contributes to social disorganization and disintegration. This system when applied properly will to a large extent bridge the communication gap between the agency and the general public which in turn results in reducing the rate and adverse effect of crime in our society.

5. RECOMMENDATION

To ensure the effectiveness of the result of this application, some recommendations
become inevitable:

1. No single body can solve the illicit drug problem. Collaboration of government agencies, corporate body and the general public is needed. We should all sincerely join hands to fight this societal disorganizer. There is strength in unity, and confronting the menace individually is in fact a waste of time and a lost battle.
2. A parallel effort by Governments to show transparency in their actions against drugs trafficking and to gain solid public trust is paramount. For instance, NDLEA officers should not engage in bribery and must be fair to their judgment so that people will appreciate them and fill happy to join and fight crime. So, without public support and trust, all efforts will fail.
3. Government should provide adequate funding to NDLEA just like in any other respectable programme to enable them to carry out their plan of Action effectively and efficiently.
4. The necessary technologies must be made available. For example, intelligence gathering and analysis is at the heart of detection and interdiction. This requires the use of the latest technology in satellite surveillance and secure data transmission. Another example is the latest container scanning equipment which allows detection of drugs and other illegal shipments within unopened containers. This is a powerful tool as a deterrent to would be traffickers
5. Adequate and pertinent training is a prerequisite to efficient intervention in combating the trade in illegal drugs. In law enforcement, forensics, rehabilitation, money laundering, information technology,
6. etc…, personnel who do not have the necessary skills cannot win the fight they are engaged in. The programmes set in place are only as good as the people involved in them. If the persons cannot cope, then all the planning and equipment in the world won't make much difference. Qualified, educated and dedicated people are what make everything work. That is the bottom line.

REFERENCES

[1] Abdullahi S.(2009) Drug Trafficking and its Impact on West Africa retrieved from http://www.giaba.org/media/f/231_09Effects_Of_Drug_Trafficking_For__Parliamentarians,_09-comrs.pdf accessed on 06/12/12.

[2] Baninyi, M. (2008) Pardigms, Pathologies and Practicalities- policing organized crime in England and Wales, Policing, 2 (1) 63-73. Published in National Guidance 2008.

[3] Commonwealth (2013) National Crime Information Center *retrieved* from http://www.mass.gov/eopss/law-enforce-and-cj/cjis/national-crime-information-center-*ncic.html* accessed on 06/12/12.

[4] Diane L. and, James L. (2003) ILLEGAL DRUGS AND DRUG TRAFFICKING retrieved from http://www.parl.gc.ca/Content/LOP/ResearchPublications/bp435- e.htm accessed on 25/12/12

[5] *Free Student Project (2012),* Crime Information Management System *retrieved from* http://www.freestudentprojects.com/studentprojectreport/projectreport/crime-information-management-system/ accessed on 06/12/12

[6] General Assembly of the United Nations (2012): Thematic Debate of the 66th session of the United Nations General Assembly on Drugs and Crime as a Threat to Development On the occasion of the UN International Day against Drug Abuse and Illicit Trafficking retrieved from http://www.un.org/en/ga/president/66/Issues/drugs/drugs-crime.shtml accessed on 06/12/12

[7] George W. B. (2002) The President's National Drug Control Strategy retrieved from http://georgewbush-whitehouse.archives.gov/news/releases/2002/02/20020212-2.html

[8] Gyong J. E. and Tanimu B. (2010) A Sociological Assessment of the National Drug Law Enforcement Agency's (Ndlea) Strategies of Arrest and Detention in Nigeria retrieved from http://maxwellsci.com/print/crjss/v2-127-132.pdf (journal) accessed on 24/12/12

[9] Information Sharing Environment *(2012)* Law Enforcement Information Sharing retrieved from http://www.ise.gov/law-enforcement-information-sharing accessed on 05/12/12

[10] Kiirya G. D. (2007) A Prosecutions Management Information System (PMIS): A Case of the Uganda's Directorate of Public Prosecutions (DPP) retrieved from *dspace.mak.ac.ug/.../3/kiirya-geoffrey-david-cit-masters-report.pdf* accessed on 05/12/12

[11] Montaldo C. (2013) What Is a Crime retrieved from http://crime.about.com/od/Crime_101/a/What-Is-A-Crime.html accessed on 05/12/12

[12] National Crime Prevention Strategy(1996), National Crime Prevention Strategy: Summary retrieved from http://www.info.gov.za/otherdocs/1996/crime1.htm accessed on 0612/12

[13] National Drug Law Enforcement Agency (2011). Brief History of Nigeria's Counter-Narcotic Efforts. Retrieved from http://www.ndlea.gov.ng/v1/?q=content/history accessed on 06/12/12

[14] Public Agenda (2011) Issue Guides: Illegal Drugs-Consider the Choices retrieved from http://www.publicagenda.org/citizen/issueguides/illegal-drugs/considerchoices accessed on 2512/12

[15] Tim W. and Tivani W. (2006), Root Causes of Illicit Drug and Alcohol Problems in Minority Communities in Northern Thailand, and the Shan State of Myanma retrieved from http://www.diakonia.se/documents/public/results_and_highlights/annex_diakonia_thailand_report_jan_2007.pdf accessed on 06/12/12

[16] United Nations Office on drug and crime (2013), Drug trafficking http://www.unodc.org/unodc/en/drug-trafficking/index.html accessed on 06/12/12UNRISD (1994) UNRISD

[17] Briefing Paper No. 2 World Summit For Social Development http://www.unrisd.org/80256B3C005BCCF9/(httpAuxPages)/4C3D0BE90FAD550480256B6400419B57/$file/bp2.pdf accessed on 05/12/12

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Alleviating Classification Problem of Imbalanced Dataset

S.O. Folorunso
Mathematical Sciences Department,
Olabisi Onabanjo University
Ago– Iwoye, Ogun State.
bamidelekeke@gmail.com

A.B. Adeyemo
Department of Computer Science
University of Ibadan
Ibadan, oyo State.
sesanadeyemo@gmail.com.

ABSTRACT

The Class Imbalance problem occurs when there are many more instances of some class than others. i.e. skewed class distribution. In cases like this, standard classifier tends to be overwhelmed by the majority class and ignores the minority class. It is one of the 10 challenging problems of data mining research and pattern recognition. This imbalanced dataset degrades the performance of the classifier as accuracy is tendered towards the majority class. Several techniques have been proposed to solve this problem. This paper aims to improve the true positive rate/ detection of the minority class (GDM) which is the class of interest. This study proposes the use of two under sampling techniques reported in the literature. It involves under sampling the majority class which balances the dataset before classification. These under sampling schemes were evaluated on three learning algorithms (Decision tree both pruned and un- pruned and RIPPER) using Matthew's Correlation Coefficient (MCC) and Kappa Statistics as metrics. The implementation of these under sampling techniques was assessed in the medical domain. The real life dataset collected contained 886 instance of patients with diabetes mellitus disease. The diagnosis was in three classes with the following class distribution TYPE1 containing 62 instances, TYPE2 containing 807 instances and Gestation Diabetes Mellitus (GDM) which is the class of interest containing only 17 instances. This study revealed that, compared with the original dataset and RUS dataset, NCL dataset presents superiority in achieving better true positive rate for the minority class and also high MCC and K-Statistics with the three learning algorithm.

Keywords: **Class Imbalance Problem, Sampling technique, Data reduction, Class Distribution.**

African Journal of Computing & ICT Reference Format
S.O. Folorunso & A.B. Adeyemo (2013). Alleviating Classification Problem of Imbalanced Dataset.
Afr J. of Comp & ICTs. Vol 6, No. 1. Pp 137-144.

1. INTRODUCTION

The class imbalance problem corresponds to the domain for which one class is represented by a large number of examples while the other is represented by few (Japkowicz, 2003). In such cases, standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. This imbalance causes suboptimal classification performance or even worse (Chawla *et al.*, 2004, Fernandez *et al.*, 2011). It is one of the 10 challenging problems of data mining Research (Yang and Wu, 2005) and pattern recognition (Ghanem *et al.*, 2010). When the prediction model is trained on such an imbalance dataset, it tends to show a strong bias towards the majority class, since typical learning algorithms intend to maximize the overall prediction accuracy.

In fact, if 95% of the entire data set belongs to the majority class, the model might ignore the remaining 5% of minority examples and predict that all of the test examples are in the majority class. Even though the accuracy will be 95%, the examples of the minority class will be absolutely misclassified (Hido and Kashima, 2008). The misclassification cost for the minority class, however is usually much higher than that of majority class and should not be ignored (Hido and Kashima, 2008, Thai- Nghe *et al.*, 2009). Domain suffering naturally from class imbalances include detection of oil spill in satellite radar images (Kubat and Matwin, 1997), diagnose diseases in medicine such as rare diseases(cancer) and rare gene mutation(albino), medical diagnosis (Yang and Ma, 2010), network monitoring,

Intrusion detection (Engen, 2010), Earth quakes and nuclear explosion and helicopter (Guo *et al.*, 2008), Risk management (Chawla *et al*., 2004), text classification (Jo and Japkowicz, 2004), Education (the ratio of the number of "pass student" to "fail student") and detection of fraudulent or default banking (Thai- Nghe *et al.*, 2009) Species Distribution Prediction in Ecology and Conservational biology (Johnson *et al.*, 2012), Information Retrieval and Filtering (Lewis and Catlett, 1994), Response Optimization in CRM (Lessmann, 2004), Document Classification (Maneivitz and Yousef, 2002), Image Retrieval (Chen *et al*., 2001), DNA Microarray time series (Pearson *et al*., 2003), Spam-detection and filtering (Kolcz *et al*., 2003) and Sentence boundary detection in speech (Liu *et al*., 2006).

In practical applications, the ratio of the small to large classes can be drastic such as 1:100, 1:1000, or 1 to 10,000 and sometimes even more (Chawla *et al.*, 2004). In a classification problem, algorithm is used to construct a model by learning from training set which contains examples with class labels (Boontarika and Maythapolnum, 2011). Numerous solutions to Class Imbalance Problem had been reported in literature and were developed both at data and algorithmic levels. Almost all the solutions developed were designed for a two- class problem where the imbalance problem observed is that one class is highly under represented but associated with a higher identification importance. At the data level, the objective is to re-balance the class distribution by re-sampling the data space while at the algorithm level, solutions try to adapt existing classifier learning algorithm to strengthen learning with regards to the minor class. The main advantage of the data level techniques is that they are independent of the underlying classifier (Fernandez et al. 2011, Ding, 2011)

At the data level: Re- sampling technique balances the class distribution in the training data by either adding examples to the minority class (Oversampling) or removing examples from the majority class (under sampling) (Yang and Wu, 2006, Ding, 2011) or combination of oversampling and under sampling (Sun *et al.*, 2006, Guo *et al*., 2008, Ding, 2011). The resulting sampled dataset is then made more amenable to traditional algorithm which can then be used to classify the data. This article is limited to only under sampling technique. The problem domain of the disease of diabetes mellitus and is limited to three types which are TYPE1, TYPE2 and Gestation Diabetes Mellitus (GDM) is naturally imbalanced as only a small percentage of patients are likely to have GDM. GDM is the class of interest here as identifying patients with this disease is important. The use of Decision Tree and RIPPER classification algorithm on diabetes Mellitus Dataset has been used by Awokola (2010) to build predictor models and it was found that the classification performance was not optimal.

The study did not also take the imbalanced nature of the dataset into consideration. This paper investigates the impact of class imbalance problem and the use of two of under sampling techniques discussed in the literature (RUS and NCL) to improve the classification performance on the Diabetes Mellitus imbalanced dataset.

2. RELATED WORK

Much research has been carried out with the class imbalance problem and numerous solutions have been reported in the literature. Two of the under sampling techniques reported will be used to alleviate the class imbalance problem of the diabetes mellitus dataset before using decision tree and RIPPER algorithm to classify it. The use of Decision tree and RIPPER algorithm to build a predictor model for the diabetes mellitus disease was first done by Awokola (2010) but the study did not put class imbalance problem into consideration. The dataset is highly skewed as there is disparity in the class distribution. This section discusses the under sampling techniques used and the various types of diabetes mellitus disease.

The under sampling techniques discussed in this section aims to obtain a representative training set with a lower size compared to the original training set and with a similar or even higher classification accuracy for new incoming dataset.

2.1Random under sampling (RUS): This is random elimination of majority class examples. RUS makes no attempt to "intelligently" remove examples from the training data. Instead, RUS simply removes examples from the majority class *at random* until a desired class distribution is achieved. It can discard potentially useful data that could be important for the induction process, and this can make the decision boundary between minority and majority harder to learn (Ding, 2011, Seiffert *et al*., 2010). It creates a subset of the original dataset by the eliminating instances.

2.2 Neighborhood Cleaning Rule (NCL)
With this technique, Wilson's Edited Nearest Neighbor (ENN) (Wilson, 1972) rule is used to identify and remove majority class. The algorithm first finds the three nearest neighbors for each of E_i examples in the training set. If E_i belong to the majority class and it is misclassified by its three Nearest Neighbor (3-NN), then E_i is removed. If E_i belongs to the minority class and it is misclassified by its 3-NN to be the majority class, then removes the three nearest neighbor. To avoid excessive reduction of small classes, only examples from classes misclassified by 2-NN of its 3-NN (Laurikala, 2001).

2.3. DIABETES MELLITUS: or simply diabetes is a group of metabolic diseases in which a person has high blood sugar, either because the pancreas does not produce enough insulin, or because cells do not respond to the insulin that is produced. This high blood sugar produces the classical symptoms of polyuria (frequent urination), polydipsia (increased thirst) and polyphagia (increased hunger). There are three main types of diabetes mellitus (DM).

(a) Type1 DM results from the body's failure to produce insulin, and currently requires the person to inject insulin or wear an insulin pump. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes".

(b) Type2 DM results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. This form was previously referred to as non-insulin-dependent diabetes mellitus (NIDDM) or "adult-onset diabetes".

(c) The third main form, gestational diabetes (GDM) occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. It may precede development of type 2 DM. Other forms of diabetes mellitus include congenital diabetes, which is due to genetic defects of insulin secretion, cystic fibrosis-related diabetes, steroid diabetes induced by high doses of glucocorticoids, and several forms of monogenic diabetes (Sarwar *et al*., 2010).

3. RESEARCH METHODOLOGY

3.1 The Data set

The raw dataset for this study was obtained from the records department of the family medicine clinic of Wesley Guild Unit of Obafemi Awolowo University Teaching Hospital Complex, Ilesha, Osun State, Nigeria. The dataset of outgoing patients suffering from diabetes mellitus was extracted and reviewed. The dataset contained 886 instances of complete record of diabetes mellitus patients from January 2009 to May 2010. This dataset was collected by Awokola (2010). It contained information about patients with three types of diabetes. The dataset contain 886 instances, has 18 attributes and three different classes namely TYPE1, TYPE2 and Gestational Diabetes Mellitus (GDM). The data distribution is 807:62:17 where TYPE2 has 807 instances, TYPE1 has 62 instances and GDM has only 17 instances. This dataset is highly skewed. Table 3.1 below shows information about the variable format that was used.

Table 3.1: Variable Format Used In Diabetes Dataset

S/N	Variable Name	Variable Format	Variable Type
1	Sex	M or F	Categorical
2	Age		Continuous
3	Weight		Continuous
4	Systolic Blood Pressure		Continuous
5	Diastolic Blood Pressure		Continuous
6	All the 13 symptoms	Y or N	Categorical
7	Blood Glucose		Continuous
8	Diagnosis	TYPE1, TYPE2 and GDM	Categorical

3.2 Learning Algorithm

This paper used two learners, all of which were implemented in WEKA (Witten *et al*., 2011), which is an open – source data mining suite. C4.5 (Quinlan, 1993) is a decision tree learner that uses an entropy- based splitting criterion stemming from information theory (Weiss, 2003). Two versions of C4.5 were used in our experiment. C4.5 uses the default WEKA parameters and the only Decision tree used by Awokola (2010), while C4.4 disables pruning and used Laplace smoothing. The WEKA implementation of C4.5 is J48. Repeated Pruning Error Reduction (RIPPER) (Cohen, 1995), is a rule based learner that modifies the incremental reduced error pruning algorithm (Furnkranz and Widmer, 1994), to improve accuracy without sacrificing accuracy. The default WEKA parameters were used for RIPPER.

3.3 Evaluation Metrics: Accuracy is the most common evaluation metrics for most traditional application. But accuracy is not suitable to evaluate imbalance data sets as it places more weight on the majority class than the minority class (Weiss and Provost, 2003, Guo *et al*., 2008). However, it has been observed that for extremely skewed class a distribution, the recall of the minority class is often 0, which means that there are no classification rules generated for the minority class (Guo *et al*., 2008). The following under listed metrics were the most frequently used matrix.

a. Confusion Matrix: In a bi- class problem, the confusion matrix records the result of correctly and incorrectly recognised examples of each class (Galar *et al.*, 2011, Thai-Nghe *et al.*, 2009). Table 3.2 presents the confusion matrix of a bi – class problem. True Positive (TP) shows the number of positive class correctly classified as positive while True Negative (TN) shows the number of negative class correctly classified as negative class. False Positive (FP) shows the number of negative classes that were incorrectly classified as the positive class while false negative (FN) shows the number of positive class that were incorrectly classified as negative class. It could also be extended to multiple class problems.

Table 3.2 Confusion Matrix

	Positive Predict	**Negative Prediction**
Positi	True Positive (TI	False Negative (FN)
Negati	False Positive (F	True Negative (TN)

Accuracy = (TP + TN)/ (TP + FN + FP + TN) eq. (1)
Precision = TP/ (TP + FP) eq. (2)
Recall (True Positive Rate) = TP/ (TP + FN) eq. (3)
FP rate = FP/ (FP + TN) eq. (4)
TP rate = TP/ (TP + FN) eq. (5)
F_ Measure = $\frac{(1+\beta^2)\,Recall * Precision}{\beta^2 * Recall + Precision}$ eq. (6)
G- mean = $\sqrt{Accuracy_+ * Accuracy_-}$ eq. (7)
MS = $Accuracy_+ + Accuracy_-$ eq. (8)

b. Kappa Statistic or Cohen's kappa coefficient
It is used to measure the agreement between predicted and observed categorisation of a dataset, while correcting for an agreement that occurs by chance. Its maximum value is 100% (perfect agreement) and the expected value for random predictor with the column total is 0 (no agreement) (Witten *et al.*, 2011). Cohen's kappa coefficient is a statistical measure of inter- rater agreement or inter-annotator agreement (Carletta, 1996) or qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance.

The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

where Pr(*a*) is the relative observed agreement among raters, and Pr(*e*) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then κ = 1. If there is no agreement among the raters other than what would be expected by chance (as defined by Pr(*e*)), κ = 0.

c. G- Means Criterion
Also known as geometric means and it combines the performance of both positive class and negative class i.e. geometric mean of the accuracies measured separately on each class (Positive and Negative) and depicted by eq. (7). High prediction accuracy on both positive and negative class will give rise to a high G- means value (Ding, 2011). The ordinary mean criterion without the square root is depicted by eq. (8).

b. Matthews Correlation Coefficient (MCC)
This is a strong metric that considers both accuracies and error rates on both classes, since all the four values in the confusion matrix are involved in this formula. A high MCC value means the learner should have high accuracies on positive and negative classes, and also have less misclassification on the two classes. Therefore, MCC can be considered as the best singular assessment metric so far (Ding, 2011).

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{P_c * N_c * P_r * N_r}}$$

Where $P_c = TP + FN$

$N_c = TN + FP$

$P_r = TP + FP$

$N_r = FN + TN$

4. RESULTS

This section reports the different dataset in their original sample, the under sampled datasets from NCL and RUS techniques discussed in section 2.1and 2.2 respectively and the different data distribution. The dataset used is the Diabetes Mellitus Diagnosis dataset collected at Wesley Guild Hospital, Ilesha. It contained 886 instances with 18 different attributes and three different classes namely TYPE1, TYPE2 and GDM. The GDM is the minority class with which this study is trying to improve its detection rate. Table 4.1 depicts the various datasets with their data distribution, and situations when the two sampling techniques were applied to it.

Table 4.1 : Dataset class distribution

Dataset	**Diabetes Mellitus Data Class distribution**
Original dataset	TYPE2 = 807 TYPE1 = 62 GDM = 17
Under Sampling Techniques	**Under Sampled dataset class distribution**
Random Under Sampling (RUS)	TYPE2 = 17 TYPE1 = 17 GDM = 17
Neighborhood Cleaning Rule (NCL)	TYPE2 = 698 TYPE1 = 62 GDM = 17

DECISION TREE LEARNER

This is class for generating a pruned or un-pruned C4.5 decision tree with a Confidence Factor of 0.25. One of the reasons to avoid pruning is that most pruning scheme attempt to minimize the overall error rate. These pruning schemes can be detrimental to the minority class, since reducing the error rate in the majority class, which stands for most of the examples, would result in a greater impact over the overall error rate (Batista *et al.*, 2004, Zadrony and Elkan, 2001). The minimum number of instances per leaf was 2 and the number of folds was 3. The number of seed used for randomizing the data when reduced-error pruning was used is 1.

Table 4.2 and Table 4.3 showed the results of Decision Tree learner (pruned and un- pruned) on the diabetes dataset from both the original dataset and the various sampled datasets. From the table, the result of the original dataset was compared with the various sampled dataset. NCL improves the detection of GDM better than the other methods with a value of 0.638 for the un-pruned and the use of Laplace Transform (C4.4) and 0.689 for the decision tree with WEKA default value (C4.5).

Table 4.2 Decision Tree (C4.4) Learner on Diabetes Mellitus Dataset

Methods	K- Statistic	RMSE	Accuracy	MCC
Original Dataset	0.309	0.2159	0.9018	0.341
Under sampling Technique				
RUS	0.5	0.3994	66.67	0.5
NCL	0.5893	0.1951	0.9305	**0.638**

Table 4.3 Decision Tree (C4.5) WEKA default Learner on Diabetes Mellitus Dataset

Methods	K- Statistic	RMSE	Accuracy	MCC
Original Dataset	0.3186	0.2312	90.513	0.354
Under sampling Technique				
RUS	0.5882	0.3997	72.549	0.59
NCL	0.6358	0.1924	93.9511	**0.689**

RIPPER Learner

Table 4.4 shows the result of RIPPER learner on the diabetes dataset from both the original dataset and the various sampled datasets. From the table, the result of the original dataset was compared with the various sampled dataset. From the single under sampled scheme, NCL improves the detection of GDM better than the other methods with a value of 0.734.

Table 4.4 RIPPER Learner on Diabetes Mellitus Dataset

Methods	K- Statistic	RMSE	Accuracy	MCC
Original Dataset	0.2312	0.2334	89.9549	0.273
Under sampling Technique				
RUS	0.5588	0.3882	70.5882	0.572
NCL	0.6902	0.1805	94.4695	**0.734**

5. DISCUSSION

In this study, two different sampling methods were studied; RUS, NCL to balance the dataset before classification. These sampling methods were to balance the imbalanced class distributions to allow better detection of minority classes that were difficult to identify. The effect of data reduction to the classification ability of the actual data analysis was studied with 3 different classifiers namely C4.4, C4.5 and RIPPER classifier. The classification results obtained from the 2 sampled and the original dataset with the 3 classifiers showed that the NCL method was significantly ($p < 0.05$) better than original dataset and RUS with the three learning algorithm. NCL (a modification of ENN) attempts to avoid the problems caused by noise by applying the ENN algorithm that is designed for noise filtering. NCL also cleans neighbourhood that misclassifies examples belonging to the class of interest (GDM). In addition, NCL removed approximately 25% of the data, while RUS dropped over 50% of the data. The success of NCL is partly due to the ENN algorithm which saves the majority of the data and consequently, manages to keep the accuracies near the original level.

6. ANALYSIS

After performing cross validation over a wide variety of both original and sampled datasets to produce result, we need to determine which classifier is better or which sampling scheme is the best. In this study, pair – t test (two tailed) method with 5% confidence level was used to determine which classifier is better and on which dataset. Landis and Koch (1977) characterized values Kappa Statistic, $K < 0$ as indicating no agreement and, $0 < K > 0.20$ as slight, $0.21 < K < 0.40$ as fair, $0.41 < K < 0.60$ as moderate, $0.61 < K < 0.80$ as substantial, and $0.81 < K < 1$ as almost perfect agreement. The analysis was performed on 2 different sampled dataset with 1 original dataset on 3 different learning algorithms on 10-fold cross validation, giving us 900 results (3 datasets x 3 learning algorithm x 10 fold cross validation x 10 runs = 900 results). Table 4.2, Table 4.3 and Table 4.4 showed the result of the analysis of the sampled dataset with the original dataset against the 3 learning algorithm. From Table 4.2, Table 4.3 and Table 4.4, it could be observed that NCL has a higher kappa statistic value than the original dataset and RUS dataset and it also produced a higher value with the entire learning algorithm (RIPPER (0.6902), Decision Tree (C4.4 = 0.5893 and C4.5 = 0.6358), The result generated in Table 4.2, Table 4.3 and Table 4.4 showed that the NCL under sampling technique is significantly better that the original dataset and RUS dataset.

7. CONCLUSION

Class Imbalance problem is a learning problem. Learning from Imbalance Dataset is an important issue as it degrades classifier performance of classifier. The results showed that the NCL under sampling scheme improves the detection and true positive rate of minority class (GDM). Original dataset was tested on diabetes mellitus dataset collected from Wesley Guilds Hospital, llesha. And various sampling techniques were applied to the dataset to produce different sampled dataset and on different learners, thus providing a diverse test bed. NCL performs better based on domination with Matthew's Correlation Coefficient (MCC) and K- Statistics. NCL forces learning and introduce a bias towards the minority class (GDM). NCL allows for improved identification of difficult minority class (GDM) while keeping the classification ability. There are several areas to be considered further in this line of study. The following suggestions were made based on the findings experience in this course of study namely finding the performance loss/ gain after applying the various sampling technique to the dataset, applying Random Under Sampling (RUS) scheme with varying degree of under sampling and to analyse the ROC curves obtained from the 3 classifiers.

EndNotes:
This paper was initially presented at the iSTEAMS Research nexus Conference, 2013.

REFERENCES

[1] Japkowicz, N. (2003). Class Imbalances: Are we focusing on the right issue? In the Proceeding of ICML workshop: Learning with Imbalanced Data Sets II, pp. 17 – 23.

[2] Chawla, N.V., Japkowicz, N. and Kolcz, A. (2004). Editorial: Special Issue on Learning from Imbalance Data Sets, *SIGKDD Explorations. 6* (1), 1- 6.

[3] Fernandez, A., Garcia, S. and Herrera, F. (2011). Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution. E. Corchado, M. Kurzy´nski, M. Wo´zniak (Eds.): HAIS, Part I, LNAI 6678, (1–10), Springer-Verlag Berlin Heidelberg.

[4] Yang, Q. and Wu, X. (2005). 10 Challenging Problems in Data mining, *International Journal of Information Technology and Decision making.* 5 (4), 597- 604.

[5] Ghanem, A., Venkatesh. S. and West, G. 2010. Multi- Class Pattern Classification in Imbalanced Data, *International Conference on Pattern Recognition*, Pp. 2881- 2884

[6] Hido, S. and Kashima, H. 2008. Roughly Balanced Bagging for Imbalanced Data. "In Proceedings of SIAM Conference on Data Mining (SDM2008), Atlanta, Georgia, USA, April, 2008

[7] Thai-Nghe, N., Andre, B., Lars, S. 2009. Improving Academic Performance Prediction by Dealing with Class Imbalance, *Proc. 9th IEEE International Conference on Intelligent Systems Design and Applications IEEE Computer Society, (ISDA 878-883.*

[8] Kubat, M. and Matwin, S. (1997). Addressing the Curse Imbalanced Training Sets: One-Sided Selection, *In the proceedings of the fourteenth conference on machine learning,*179- 186.

[9] Yang, Y. and Ma, G. 2010. Ensemble- based Active Learning for Classification Problem. J. Biomedical and Engineering, 2010, 3, pp. 1021- 1028. Published online in SciRes. Http:/www. Scrip.org/journal/jbise.

[10] Engen Vegard. 2010. Machine Learning for Network Based Intrusion Detection.
a. Ph. D thesis, Bounemouth University, 2010.

[11] Guo, X., Yin, Y., Dong, C., Yang, G. and Zhou, G. (2008). On the Class Imbalance Problem, *Fourth International Conference on Natural Computation*, IEEE Computer Society.pp.192 – 201

[12] Jo, T. and Japkowicz, N. 2004. A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence, Volume 20, Number 1, pp. 18- 36, 2004.

[13] Johnson, R. A., Chawla, N. V., Hellman, J. J. (2012). Species Distribution Modelling and Prediction: A Class Imbalance Problem. NASA Conference on Intelligent Data Understanding (CIDU), Boulder, CO.

[14] Lewis, D. and Catlett, J. (1994). Training Text Classifiers by Uncertainty Sampling. In the Proceedings of the 14th ACM SIGIR Conference on Research and Development in Information Retrieval.

[15] Lessmann, S. 2004. Solving imbalanced classification problems with support vector machines. In International Conference on Artificial Intelligence. pp. 214–220.

[16] Maneivitz, L. M. and Yousef, M. (2002). One – Class SVMs for Document Classification. Journal of Machine Learning Research, 2, 139 – 154

[17] Chen, Y., Zhou, X. and Huang, T. (2001). One class SVM for learning in Image Retrieval. In the Proceedings of IEEE International Conference on Image Processing (ICIP' 01 oral).

[18] Pearson, R. K., Gonye, G. E. and Schwaber, J. S. (2003). Imbalance Clustering of Microarray Time-Series. *Workshop on Learning from Imbalanced Datasets II, ICML*, Washington DC.

[19] Kolcz, A., Chowdhury, A. and Alspector, J. (2003) Data duplication: an imbalance problem? Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC.

[20] Liu, X. and Zhou, Z. (2006). The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study. The 6th International conference on Data Mining ICDM'06. Pp. 970 – 974.

[21] Boontarika, L. and Maythapolnun, A. (2011). An Empirical Study of Multiclass with Class Imbalance Problems, *Proc. of Business and Information (BAI), 8*,(1).

[22] Ding, Z. (2011). Diversified Ensemble Classifier for Highly imbalanced Data Learning and their application in Bioinformatics, Ph. D thesis, *College of Arts and science, Department of Computer Science, Georgia State University*, 2011. Http://digitalarchive.gsu.edu/cs_diss/60

[23] Sun, Y. Kamel, M. S. and Wang, Y. (2006). Boosting for Learning Multiple Classes with Imbalanced Class Distribution, IEEE *Proc. of the Sixth International Conference on Data Mining (ICDM'), 9*, 7695-2701.

[24] Awokola, E. O. (2010). Rapid Diagnosis and Treatment of Diabetes Mellitus Using Artificial Neural Networks. Dissertation submitted to the Department of Computer Science, University of Ibadan, Ibadan Oyo State, Nigeria.

[25] Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V. and Napolitano, A. (2010). RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans. 40 (1). 185 – 197.

[26] Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man, and Communications 2(3), 408–421.

[27] Laurikkala, J. (2001). Improving Identification of Difficult Small Classes by Balancing Class Distribution, *University of Tampere, Tech. Rep. A* (2).

[28] Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio E, Ingelsson E, Lawlor DA, Selvin E, Stampfer M, Stehouwer CD, Lewington S, Pennells L, Thompson A, Sattar N, White IR, Ray KK, Danesh J. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet. 2010; 375:2215-2222

[29] Witten, I. H., Frank E. and Hall M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Third Edition. Morgan Kaufmann Publishers. www.mkp.com. ISBN: 978-0-12-374856-0

[30] Quinlan, J. R. (1993) C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993

[31] Weiss, G.M., Provost, F.J. (2003). Learning when training data are costly: The effect of class distribution on tree induction. Journal of Artificial Intelligence Research Vol 19, pp 315 – 354.

[32] Cohen, W. C. (1995). Fast Effective Rule Induction. In the proceedings of the 12th International Conference on Machine Learning, 115-123.

[33] J. Furnkranz and G. Widmer. (1994). Incremental reduced error pruning," in Proceedings of International Conference of Machine Learning. Pp. 70-77.

[34] Batista, G.E.A.P.A., Prati, R. C. & Ronard, M. C. (2004). A study of the Behaviour of Several Methods for Balancing Machine Learning Training Data, *SIGKDD Explorations*, *6*(1), 20-29.

[35] Zadrozny, B., and Elkan, C. (2001). Learning and Making Decisions When Costs and Probabilities are Both Unknown. In KDD, pp. 204–213.

[36] Landis, J.R.; and Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics* **33** (1): 159–174. doi:10.2307/2529310. JSTOR 2529310. PMID 843571.

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Towards the Design of an Oil and Gas Production Sub-System Using 3-Tier Architecture

T. Aransiola & E.O. Nwachukwu
Department of Computer Science
University of Port Harcourt, Nigeria.
tunde_aransiola@yahoo.com, enoch.nwachukwu@uniport.edu.ng

ABSTRACT

In the last few decades, the internet and portable media devices have evolved considerably that has necessitated application developers to adopt design methodology that will enable their applications to be deployed not only on LAN servers but also on web servers. This design methodology requires existing applications to be easily extended to varieties of clients including mobile devices with minimal redesign of existing applications. The integration of application logic with the user interface (UI) makes it very difficult to extend an existing 2-Tier client/server application. The 3-Tier architecture separates the UI from application logic and data. This separation gives flexibility and independence in application design. It also makes it simpler to manage the application. The focus of this study is on Oil and Gas Production Sub-system using 3-Tier design Architecture. I have implemented this system using .NET framework 3.5 to design the middleware and the server pages. The performance of this architecture is tested in the 3-Tier web-based environment using the Internet server running on a windows server to host the pages. The 3-Tier architecture is scalable, maintainable and appropriate for enterprise development. Using the windows mobile device emulator, the functionalities of the system were extended to the mobile environment using the same middleware.

Keywords: Oil, Gas, LAN, Architecture, Production, .NET Framework and Environment.

African Journal of Computing & ICT Reference Format
T. Aransiola & E.O. Nwachukwu (2013). Oil and Gas production Sub-System Using 3-Tier Architecture.
Afr J. of Comp & ICTs. Vol 6, No. 1. Pp 145- 148..

1. INTRODUCTION

Designing distributed applications involves making decisions about its logical and physical architectures, the technology and infrastructure used to implement its functionality. In order to make these decisions, you must have sound understanding of the business processes that the application will perform its functional requirements and the levels of scalability, availability, security and maintainability required. Also, with the increase in mobile devices, clients and rich interactive web and mobile applications, the 3-tier Internet client-server architecture comes to the fore and it has become a widely accepted tenet of distributed application design that components that perform similar functions be grouped into tiers and developer should divide their application into components providing presentation, business, and data services [1]. This partitioned view of an application development helps in building online systems with high reliability, high performance, and high interaction [2].

Most of the functionality performed by client programs in the 2-tier client-server architecture is moved to the middle-tier to make the client thin, and the back-end database is shielded so that each end user can access the database via the middle-tier by using a web browser from any client.

1.1 Research Objectives

The aim of this work is to design Oil and Gas Production Subsystem using 3-Tier Architecture.
The objectives of this thesis are:

1. To design the conceptual data model of an oil and gas production system.
2. To design and implement a web-based Oil and Gas production subsystem using 3-Tier architecture.
3. Extend the functionalities of the system to mobile device environment.

With the advent and spread of internet technologies, more and more applications and going web and more developers are leveraging on the 3-Tier architecture to developing tools required to implement IT solutions for the enterprise. This work is of immense importance because it provides a model by which developers can create flexible and reusable applications by partitioning application into tiers. We limited the scope of the work is to the design of a data model and implementation of an oil and gas production subsystem using the 3-Tier architecture. The proposed system is meant to be implemented in a web environment with functions extended to mobile environment.

2. ANALYSIS, DESIGN & IMPLEMENTATION

3-Tier, the most common occurrence of a multi tier architecture, logically separates the function of an application into a User Interface (UI), business logic and the data source component. This architecture emerged to overcome the limitations of the 2-tier architecture.

The basic web model is a 2-tier model whereby the browser makes a request from a web server which then processes the request and returns the desired response. This approach improves scalability and divides the UI from the data layer, however it does not partition the application layer so that it can be reused [3]. To update a portion of the application, the entire application must be updated because the tiers are not separated. In order to understand the limitations of the 2-tier design model in a basic web application, one must look inside the code implementation and recognize that there are three inherent different components for a database application.

In a large-scale application development that requires high activity, transaction volumes and frequent maintenance, 2-tier design architecture does not suffice. The 3-tier architecture is suitable to build such high-ended application like Oil and Gas Production system. With this architecture, the web browser acts like a client, the middleware contains the business logic and the database server handles the data access functions. The client only accepts data input and displays data but has no part in processing that produce the result. In a situation where the database server is overwhelmed with requests, the server can be clustered or upgraded.

Partitioning application into distinct layers can be seen to allow for parallel development of the different layers of the application, ease of updates and maintenance since it is easier to upgrade a specific component without updating the whole application [4]. Ease of application development and deployment and also cost effectiveness are the main strengths and benefits of the 3-tier architecture [5]. Main advantages of 3-tier architecture when compared to 2-tier are increased performance, flexibility, maintainability, reusability and scalability.

The goal this research is not to implement a full-fledged Oil and Gas Production system but to implement the production sub-system in a web and mobile environment using the 3-tier architecture. The Oil and Gas production business is largely data-driven; to develop an efficient solution, accurate and up-to-date information about the enterprise must be maintained. In order to achieve this, key-based conceptual data model was adopted. The key-based data model gives the guidelines for constructing the fully-attributed model.

The key-based conceptual data model is business and user driven. The content and structure of the model are controlled by the business requirements rather than technology. The emphasis is on requirements rather than solution. The terms used in the model is stated in the language of the business and not that of system development. This provides a context for discussions on what is important to the business. The key-based conceptual data model, usually in the form of an entity relationship diagram (ERD), is developed in order to understand and captures the key business entities and the relationships between these entities. Attributes are not added to key-based conceptual data models as these are higher levels model [6].

When developing a conceptual data model the system architect must ask probing questions and think of possible exceptions that could arise. Significant problems can arise when conceptual data models are not developed. Developing this model at the requirements gathering phase of a project is invaluable from both a project management and development perspective. It will help to identify the breadth and depth of the subject area and establish the scope of the project.

The conceptual data model serves as a basis for developing the fully-attributed model. The fully-attributed model is a third normal form data model that includes all the entities, attributes and relationships needed by the project. In the context of the relational database, the data model is used to create the relational tables. It is worth mentioning that, the conceptual data model is independent of implementation, and so does not require that the implementation be done in a particular database or programming language. The data model is a clear unambiguous specification of what is wanted in the enterprise.

2.1 The Production Allocation

This is the core of the oil and gas production system. Back allocation using data generated by periodic well testing and daily station production measurements provide the basic requirements for data reporting and volume proration and serves as the basis for assessing the performance of producing hydrocarbon reservoirs. However, poor well testing and production measurements can result in improper allocation, which negatively affect hydrocarbon reserves accuracy and reservoir performance. For each day the station is on production, the following data are captured:

(a) Active completion strings whose status is "*OPEN*" are listed from which the ones on productions for the day is/are selected.
(b) The daily station gas utilized (Mscf).
(c) For each selected Completion String in (a), the following data are captured:
- Bean Size (nth/64)
- BSW (%)
- Hours of production for each segment of 6-Hours production period. Hours and minutes or minutes of production are entered in decimal.
- Sand Cut (pptb).
- Temperature (^{0}F).
- Closed Tubing Head Pressure (psi).
- Flowing Tubing Head Pressure (psi).
- Flow Line Pressure(psi).
- Salinity (kppm)
- Lift Gas, Vented gas from Associated & Non-Associated Gas (Mscf)

(d) The Daily Station Oil, Gas, Water and Condensate produced are captured. Oil produced, Water and Condensate produced in barrels (bbl), Gas produced in (Mscf).

2.2 The Allocation Equation

LiftGas = $\sum_{str=1}^{n}(LiftGasAG + LiftGasNAG)$ ….i

Actual Gas Produced = Gas produced – LiftGas ……..ii

StringRate = $(\frac{\sum HoursOnProduction}{24})$ x$StringTestRate$ ……..iii

Allocation = DailyProduction x$(\frac{StringRate}{\sum StringsTestRate})$… …..iv

The allocation equation applies to daily production of Oil, Gas, Condensate, Sand and Water produced. The procedure allocates station production to, Daily String production and Reservoir production.

Table 1: Units of Production Data Measurements

Units	Description
bbl	Barrel
BSW	Basic Sediment & Water
^{0}F	Degree Farenheight
kppm	Thousand parts per million
Mscf	Thousand Standard Cubic Feet
pptb	Pound per thousand barrel
psi	Unit of Pressure

3. CONCLUDING REMARKS

We have presented the design model for an Oil and Gas Production Sub-System Using 3-Tier Architecture. It is expected that the implementation of this architecture will improve production.

REFERENCES

[1] Fong, Joseph and Hui, Rieman (1999): "Application of middleware in 3-Tier Client/Server database design methodology. J. Braz. Comp. Soc. [online]. www.scielo.br/scielo.php?pid=S0104-65001999000200005&script...

[2] Orfali, R. , Harkey D and Edwards J. (1999): "Client Server Survival Guide, Wiley, 3rd Ed.

[3] Gallaugher, J. & Ramanathan, S. "Choosing a Client/Server Architecture. A Comparison of Two-Tier and Three-Tier Systems." Information Systems Management Magazine 13, 2 (Spring 1996): 7-13

[4] Myerson J., (2002); The Complete book of Middleware, Auerbach Philadephia.

[5] Yung Wang and Mario Apicella (1998); "A New three-tier web-based computing model Internet" Infoworld December 7, 1998.

[6] All Fusion Data Modeler, Methods Guide r7.2; https://supportcontent.ca.com

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Design of Clinical Decision Support System for Chronic Heart Disease Diagnosis Using Case Base Reasoning

Prem Pal Singh Tomar & Ranjit Singh
Faculty of Engineering
Dayalbagh Educational Institute, Agra, India
singhppst@rediffmail.com, rsingh_dei@yahoo.com

ABSTRACT

The present research work is to design and develop "Multimedia Based Clinical Decision Support System for Diagnosis of Chronic Heart Diseases (CHLD-MMCDSS)" to improve the Quality of Life (QOL) of industrial/operational workers suffering from semistructured/ unstructured Chronic Heart Diseases and to facilitate Medical Diagnosis. A CHLD-MMCDSS is designed and developed to improve the Quality of Life (QOL) of industrial / operational workers suffering from semistructured / unstructured Chronic Heart Diseases and its main components: Model Base Management (MBM) System, Medical Data Base Access and Management (MDBAM) System, Central Medical Vision Navigator (CMVN) Board, Clinical Vision Technology (CVT) Base comprising of Case Base Reasoning Algorithm based Case Base Reasoning Desk (CBR-Desk) Multi Media Medical Communication (MMMC) Desk, Chronic Disease Queries Support (CDQS) Server and Medical Decision Exchange (MDE) Server based Dialog Management (DiM) System and Clinical Decision Making and User (CDMU) Desk are analyzed and discussed. The Cardio Informatics Portal of CHLD-MMCDSS is designed and developed for Medical diagnosis of industrial/operational workers suffering from Chronic Heart disease. Medical Diagnosis Capsule (MDiC) of Model Base Management (MBM) System provides the diagnosis using four main Chronic Heart Disease Application Modules i.e. Coronary Artery (CAD) Module, Rheumatic Valvular (RHE) Module, Chronic Cor Pulmonale (CCP) Module, and Congenital (CON) Module for four Chronic Heart diseases. The designed user friendly CHLD-MMCDSS is developed using Microsoft Visual Studio and SQL Server.

Keywords: Clinical, Decision Support System (DSS), Case-based Reasoning, Heart Disease and Quality of Life. Chronic Heart Disease Diagnosis Using Case Base Reasoning.

African Journal of Computing & ICT Reference Format
Prem Pal Singh Tomar & Ranjit Singh (2013). Design of Clinical Decision Support System for Chronic Heart Disease Diagnosis Using Case Base Reasoning. Afr J. of Comp & ICTs. Vol 6, No. 1. Pp 149-160.

1. INTRODUCTION

Samore, Kim and Stephen [1] measured the added value of Clinical Decision Support Systems (CDSS) when coupled with a community intervention to reduce inappropriate prescribing of antimicrobial drugs for acute respiratory tract infections. Sim et al. [2] described the research and policy challenges for capturing research and practice-based evidence in machine-interpretable repositories and made recommendations for accelerating the development and adoption of CDSSs for evidence-based medicine. Jeste et al. [3] reviewed studies that compared the effects of multimedia (video- or computer-based) educational aids with those of routine procedures to inform healthcare consumers about medical evaluations or management and concluded that multimedia educational aids hold promise for improving the provision of complex medical information to patients and caregivers.

A French computer-assisted hypertension and cardiovascular risk education program ISIS was developed by Consoli et al [4] to provide patients at cardiovascular risk with a modern interactive educational tool combining rigorously scientific information with the aesthetic attractiveness of multimedia communication. Degoulet, Jean and Safranh [5] provided the end-user with a virtual and unified device that conceals the complexity of the underlying information system and proposed a reference set of evaluation criteria that includes functional, technical organizational, medical, cultural and ethical, economic and industrial components. According to authors workstations for health care professionals will provide access to distributed healthcare information systems. Kindler et al. [6] illustrated the current state of the International Computer Database for Radiation Accident Case (ICDRAC) multimedia component. According to authors, Images plays an increasing role as complementary information beside text and numerical data in medicine.

The systematic evolution of a hypermedia-based group decision support system architecture to support collaborative Medical Decision-Making (MDM) was presented by Rao and Turoff [7]. This GDSS was also used by designers and researchers in the GDSS medical informatics arena. Krause and Glowinski [8] presented a requirement analysis of the aspects that should be covered by the formal specification. The discussion was focused on the specification of a Medical Decision Support System. Mangiameli, West and Rampal [9] examined the model section decision for a Medical Diagnostic Decision Support System. Schnipper et al [10] developed a "Smart Forms" to facilitate documentation-based Clinical Decision Support Systems (CDSS) integrated within Electronic Medical Records (EMR) for coronary artery disease and diabetes mellitus. Goggin et al [11] reviewed the progress of the implementation of Clinical Decision Support Systems in the field of otology.

The system provided greater accuracy to physicians across a wide range of medical specialties. Lomotan et al [12] investigated the use of a new guideline-based, Computerized Clinical Decision Support system for asthma in a pediatric pulmonology clinic of a large academic medical center that included patterns of computer use in relation to patient care, and themes surrounded the relationship between asthma care and computer use. Kumar, Sathyadevi and Sivanesh [13] focused on medical diagnosis by learning pattern through the collected data of diabetes, hepatitis and heart diseases and developed intelligent medical DSS to help the physicians. Ruiz et al [14] presented a CDSS for melanoma diagnosis using an image set of the skin lesion to be diagnosed. Hoeksema et al [15] evaluated the accuracy of a computerized CDSS designed to support assessment and management of pediatric asthma in a subspecialty clinic. Trivedi et al [16] designed and developed a computerized DSS for the treatment of major depressive disorder by using evidence-based guidelines, transferring the knowledge gained from the Texas Medication Algorithm Project (TMAP) that provided support in diagnosis, treatment, follow-up, and preventive care with ensured accuracy and reliability.

2. DESIGN OF CHLD-MMCDSS

The CHLD-MMCDSS has been designed to have quick and interactive communication capability, instant content support, clinical decision making ability, anonymity of medical decision member's identity, exhaustive resource content and systems support. The CHLD-MMCDSS helps to retrieve, analyze and display structured medical data from large multi-dimensional or relational databases, provides access to a model base, provides access to medical documents and unstructured data and facilitates multimedia based communication and clinical decision making. CHLD-MMCDSS has been designed with symbiotic approach to achieve the aim of providing the necessary global medical information, resources visualization, processing and analysis, for real life interaction among System Users and Healthcare Workers to enable speedy Medical Diagnosis. The salient features of the CHLD-MMCDSS are: quick, interactive, work as training tool for new teachers in the field of medicine, and reduces the keeping of unnecessary documents, instant knowledge support, multimedia based communication, vast resources of query and decision exchange support.

CHLD-MMCDSS is comprised of eight main components– Model Base Management (MBM) System, Medical Data Base Access and Management (MDBAM) System, Central Medical Vision Navigator (CMVN) Board, Clinical Vision Technology (CVT) Base, Multi Media Medical Communication (MMMC) Desk, Dialog Management (DiM) System and Clinical Decision Making and User (CDMU) Desk. The proposed CHLD-MMCDSS is based on a central black board problem solving architecture, where Clinical Vision Technology Base (CVT-Base) comprised of Case Base Reasoning Algorithm based Case Base Reasoning Desk (CBR-Desk) associated to different Disease Application Modules of two Clinical Informatics Portals: Cardio Informatics (CI) Portal, interact through a Central Medical Vision Navigator (CMVN) Board using Multi Media Medical Communication (MMMC) Desk to make speedy Medical Diagnosis. The basic components of the CHLD-MMCDSS architecture are presented in Figure 1. It underlines the four main Chronic Heart Disease Application Modules i.e. Coronary Artery (CAD) Module, Rheumatic Valvular (RHE) Module, Chronic Cor Pulmonale (CCP) Module, and Congenital (CON) Module related to the four crucial types of heart problems, of Cardio Informatics (CI) Portal
and each module is customized with Medical Data Base Access and Management (MDBAM) System, and Clinical Vision Technology Base (CVT-Base).

The interaction between the System Users Healthcare Workers in the world over and the global database takes place at Central Medical Vision Navigator (CMVN) Board through Clinical Vision Technology Base (CVT-Base) for Medical Diagnosis with the tasks of defining, scanning, managing and prioritizing the Multi Futuristic Medical Decision Parameters (MFMDP). The System Users and Medical Decision Makers coordinate the use of Clinical Vision Technology (CVT) Base, and Dialog Management (DiM) System.

3. MODEL BASE MANAGEMENT SYSTEM

The Model Base Management (MBM) System accepts the data from Medical Data Base Access and Management (MDBAM) System; and interacts with the four main Chronic Heart Disease Application Modules of Cardio Informatics (CI) Portal related to the four crucial types of Heart problems of industrial/operational workers suffering from Chronic Heart Diseases.

MBM System has one Capsule: Medical Diagnosis Capsule (MDiC). Medical Diagnosis Capsule (MDiC) provides the diagnosis of a Chronic Heart disease on the basis of symptoms and diagnostic test values using Clinical Vision Technology (CVT) Base comprising of Case Base Reasoning Algorithm based Case Base Reasoning Desk (CBR-Desk). This diagnosed Chronic Heart disease is displayed on Central Medical Vision Navigator (CMVN) Board through Dialog Management (DiM) System for improving the Quality of Life (QOL) of industrial/operational workers.

4. MEDICAL DATA BASE ACCESS AND MANAGEMENT SYSTEM

CHLD-MMCDSS utilizes a Medical Data Base Access and Management (MDBAM) System for reducing computation time and effort for supportive repetitive interaction processes. The main task of the MDBAM System is the simplification, preparation, pre–processing, control and verification of the bulk of input medical data required by the four main Chronic Heart Disease Application Modules of Cardio Informatics (CI) Portal for taking medical decisions and answering the queries. The queries are arranged in Medical Data Base (MDB) as relational database. For Cardio Informatics (CI) Portal, the fields of the various medical decisions and queries entered in the Medical Data Base (MDB) include: Questions, Answers, Textual Descriptions, Website links, Video, Audio, Pictures, Animations and Visual Impressions, and Graphics.

The relational database, which is comprised of heart diseases related problem analysis report as reported by various medical experts and other medical agencies after pre-processing provides reliable medical data for necessary Medical Diagnosis. The MDBAM System is comprised of the seven main entities: Permission Database, Symptom and Question Database, Diseases and Queries Database, Clinical Queries and Answers Database, Patient Symptoms and Diagnosis Detail Database, Multimedia Medical Database, and Disease Management Database

5. CENTRAL MEDICAL VISION NAVIGATOR BOARD

The designed CHLD-MMCDSS is based on a central black board problem solving architecture where Case Base Reasoning Desk (CBR-Desk) based Clinical Vision Technology Base (CVT-Base) connected to different Disease Application Modules of Clinical Informatics Portal: CI Portal, interact through a Central Medical Vision Navigator (CMVN) Board using Multi Media Medical Communication (MMMC) Desk to make speedy Medical Diagnosis and Disease. In this CMVN-Board common globally shared database, and specialized Clinical Vision Technology Base (CVT-Base) sources act upon a central black board problem solving architecture, according to a strategy aiming at building a decision solution, both cooperatively and opportunistically.

The CMVN-Board for improving the Quality of Life (QOL) of industrial/operational workers suffering from semistructured/unstructured Chronic Heart Diseases is based on independent and interactive medical knowledge based agents as Medical Data Completion, Data Analysis, Medical Information Visualization, and Prediction and Control functions. It stores a representation of the four main Chronic Heart Disease Application Modules related to the four crucial types of Heart problems of Cardio Informatics (CI) Portal structures and services; and provides facility for viewing generation of prioritized course of action chart in real time as a result of simulation. It attracts the attention of the set of System Users and its visual presentation assist cognition and early medical decision making and speedy Medical Diagnosis. The CMVN-Board is well suited for real-time systems and is basically communication medium that enables information to be shared among the System Users. The modular and distributed computing environment allows a kind of parallel processing and reasoning with the integration of Clinical Vision Technology (CVT) Base comprising of Case Base Reasoning Algorithm based Case Base Reasoning Desk (CBR-Desk) System Users and Medical Decision Makers.

CMVN-Board supports Medical Decision Making using Diagnosis Protocols (DP). When a new problem case of industrial/operational workers suffering from semi structured/unstructured Chronic Heart Diseases comes to Clinical Decision Making and User (CDMU) Desk for Medical Diagnosis, it is displayed on CMVN Board as Diagnosis Protocol: DP# 0. The symptoms of a new case are entered on the Case Base Reasoning (CBR) Desk of Clinical Vision Technology Base (CVT-Base). This Case-Based Reasoning algorithm based Case Base Reasoning (CBR) Desk carries out the work of matching. Upon getting exact match same diagnostic result is displayed, while in case no exact match is found the nearest neighbour is looked for. The diagnostic result of Case-Based Reasoning algorithm based Case Base Reasoning

(CBR) Desk is displayed on CMVN Board as Diagnosis Protocol: DP# 1 communicated through Multi Media Medical Communication (MMMC) Desk on CMVN Board, the System Users Healthcare Workers can take a decision about the semi structured/unstructured Chronic Heart Diseases of industrial/operational workers and hence the Quality of Life (QOL) of industrial/operational workers will be improved.

6. CLINICAL VISION TECHNOLOGY BASE

The Clinical Vision Technology (CVT) Base comprising of Case Base Reasoning (CBR) Algorithm based Case Base Reasoning Desk (CBR-Desk) provides an effective dimension to deal with complex semi structured/unstructured Chronic Heart Diseases of industrial/operational workers. The Clinical Vision Technology (CVT) Base is equipped with complex knowledge discovery and artificial intelligence technologies to achieve high levels of mutual affinity and early medical decision making and speedy Medical Diagnosis in a virtual meeting environment.

Case Base Reasoning (CBR) Algorithm based Case Base Reasoning Desk (CBR-Desk) of Clinical Vision Technology (CVT) Base provides an effective Medical Diagnosis dimension to deal with complex medical information from the four main Chronic Heart Disease Application Modules of CI Portal related to the four crucial types of Heart problems of industrial/operational workers suffering from Chronic Heart Diseases.

6.1 Case Base Reasoning Desk

Case Based Reasoning (CBR) Desk processes Medical Decision problems currently encountered by referring to previous experiences stored in the Case Base. Depending on experiences accumulated by individual, he/she will take different measures in their reasoning principles and thinking procedures against Medical Decision problems. CBR Desk is inspired by the way human's reasoning e.g. solve a new Medical Decision problem by applying previous experiences adapted to the current situation. A case (an experience) normally contains a Medical Decision problem to be solved, a Medical Diagnosis or its classification, an appropriate solution and its results. For a new Medical Diagnosis problem case, a CBR Desk matches the problem part of the case against cases in the Case Base of CBR Desk and retrieves the solutions of the most similar cases that are suggested as solution after adapting it to the current situation. The great advantage of working by using previous cases is that any new cases which come up, and the solutions of which has to be "REvised", will be saved in the Case Base of CBR Desk for future use, hence automatically updating the Case Base according to the time and changing situations and improving the ability of the CBR Desk to solve Medical Diagnostic problems, as more and more new Medical Diagnostic problems it solves.

Flow chart of the CBR Desk operating procedures incorporating four major procedures i.e. REtrieve, REuse, REvise and REtain [18][19], is illustrated in Figure 2. Case Base Reasoning Desk (CBR-Desk) of Clinical Vision Technology (CVT) Base provides the formulation of CBR Model. The formulation of CBR Model is based on following steps:

Step 1: Assignment of Weights to the Diagnostic Symptoms:
The Diagnostic Symptoms of Chronic Heart Diseases are grouped together and assigned different weight values (Table 1).

Those Diagnostic symptoms which are common among all Disease Application Modules of Cardio Informatics (CI) Portal should be given the least weight value. Those Diagnostic Symptoms which are common among a few Disease Application Modules of CI Portal are given a higher weight value and the unique ones are given the maximum value.

Step 2: Acquisition of Diagnostic Symptoms from the Subjects:
The details about the Diagnostic Symptoms are collected from the System Users or Subjects using a specially graphically designed interactive interface in Microsoft Visual Basic .Net 2005. The collected Diagnostic Symptoms by the system are represented as a New Case object (as shown in Figure 3).

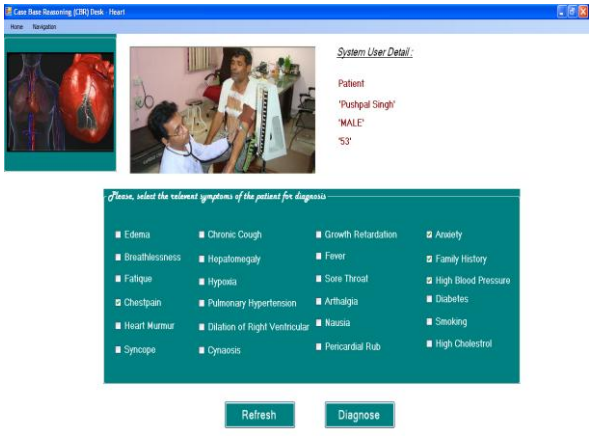


Figure 3: Interactive Interface for acquisition of High Blood Presure Diagnostic Symptom in CBR Model

Step 3: Case REtrieval and Similarity Measures:
Every Diagnostic Symptom in the input new case is matched to its corresponding Diagnostic Symptom in the stored or old cases available in the Case Base of the Case Base Reasoning (CBR) Desk. The Nearest-Neighbor Approach or kNN [20], computes the similarity value with all its available cases in the Case Base. This is based on the matching of a weighted sum of the Diagnostic Symptoms. Compare each attribute of the new case against that of the old ones in Case Base. The case with the highest similarity takes first priority. The CBR Desk will use the result of the nearest match found and use adaptation process to "REvise" this result according to the demands of the novel situation. The system uses the Nearest-Neighbor Approach that finds the closest matches of the cases already stored in the Case Base to the new case using a distance calculation, which determines how similar two cases are by comparing their Diagnostic Symptoms (Table 2).

Let T is new case and S1 and S2 are old cases then using the Nearest-Neighbor Approach, we get:

(T, S1)=63/74=0.851
(T, S2)=46/74=0.621

So, S1 is the Nearest Neighbor.

Then the presence of the symptoms in the new and the old case is listed in the next two columns. Local similarity is given in medical decision support system. The total of all the weights is calculated by adding them which is 91. Then the sum of weight*similarity is calculated by adding all the products of weight*similarity. In the first comparison the sum is 63 while in the second comparison it is 46.

The sum of weights in the first comparison is: 74
The nearest neighbor value is: 63/74= 0.851
In the second comparison:
The sum of weights is: 46
The sum of weights in the first comparison is: 74
The nearest neighbor value is: 46/74 = 0.621
Therefore, the first comparison, which is case S1, is the Nearest-Neighbor for the new case T.

The above represents a typical Nearest-Neighbor Approach that describes a situation for which T (Target case) and S (Source case) are two cases compared for similarity. CBR Desk normalizes the computed similarity value, after using Nearest-Neighbor Approach and checks for the Normalize similarity value (N_{sim}) that lies within the range 0 to 1, where 1 represents to a perfect match and 0 indicates a total mismatch between new and old case or Similarity Value is less then threshold value.

Step 4: REuse of the old case:
If N_{sim} =1 (i.e. the exact match has been found in the Case Base) then the old case available in the Case Base is Reused as diagnostic result to the input new case. The procedure of case Reuse is to locate cases that match the new case from the selected cases of the past. CBR Desk finds a case that exactly matches the input new case and goes directly to the solution. The probability of finding exact match increases as CBR Desk adds more new cases.

Step 5: Graphical Solution Display: The CBR Desk informs the System User and Medical Decision Makers by graphically displaying the proposed diagnostic result as a resultant value of the chronic heart disease. The diagnostic result of Case-Based Reasoning algorithm based Case Base Reasoning (CBR) Desk is displayed on Central Medical Vision Navigator (CMVN) Board as Diagnosis Protocol: DP # 1 (Figure 4).

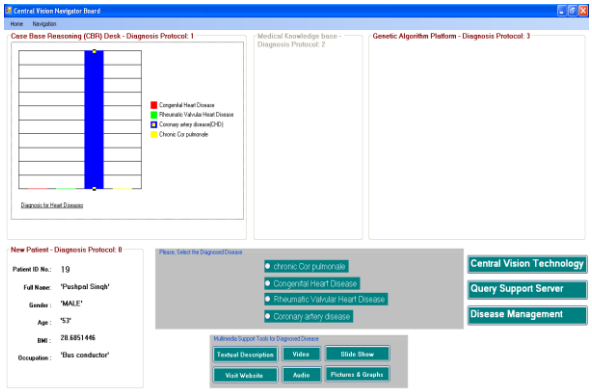


Figure 4: Diagnostic Result for Perfect Match Case Using CBR Model on CMVN Board

Step 6: Multimedia Support Tools: Multimedia support tools like textual description video, audio, slide show, pictures and graphs, and website link for diagnosed disease are used to encourage a Medical Decision Process and Medical Decision Making.

Step 7: Partial Similarity Computation:
If in Step 4, N_{sim}<1 (i.e. no exact match has been found in the available Case Base), then Similarity algorithm computes the partial similarity of the New case with each disease cases in the Case Base. The higher similarity between contexts represents the smaller knowledge gap and instantiation uncertainty and the higher the chances for a successful solution to a new case. Extract most similarity case object value for each chronic disease of Disease Application Modules of CI Portal.

www.ajocict.net

The cases, for which the similarity values are below threshold, are discarded. A relevant case object i is obtained for each chronic disease by the CBR Desk. Each case object i is decomposed as that of new case object into Diagnostic Symptoms, Diagnosis, Case History and Environment; and is compared. Then the selection of the best cases for the chronic diseases of Disease Application Modules of CI Portal takes place. These diagnostic results of Case-Based Reasoning algorithm based Case Base Reasoning (CBR) Desk are Displayed on Central Medical Vision Navigator (CMVN) Board as Diagnosis Protocol: DP # 1 (Figure 5). The next Step 8 explains the case REvision of the best case as diagnostic result.

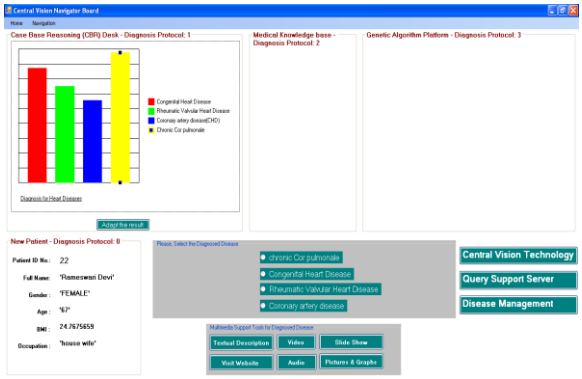


Figure 5: Diagnostic Results for Partial Match Cases Using CBR Model on CMVN Board

Step 8: Case Adaptation:

To REvise the obtained best selected case as diagnostic result, the System User and Medical Decision Makers uses the "Case Adaptation". Introspective Learning Approach i.e. Case Based Adaptation in which the case based knowledge itself provides a source from which training data for the adaptation task can be assembled. The adaptation processes recognize differences between the new and REtrieved cases from Case Base of the CBR (Adaptation) as shown in Figure 6, and appropriately refine the REtrieved Diagnostic result to reflect these differences (Figure 7). The first and foremost step is to compare the Diagnostic Symptoms and segregate the distinguishing ones for the purpose of Medical Diagnosis. The old Diagnostic result of the similar case is transformed into a new Diagnostic result using Transformational Approach [17].

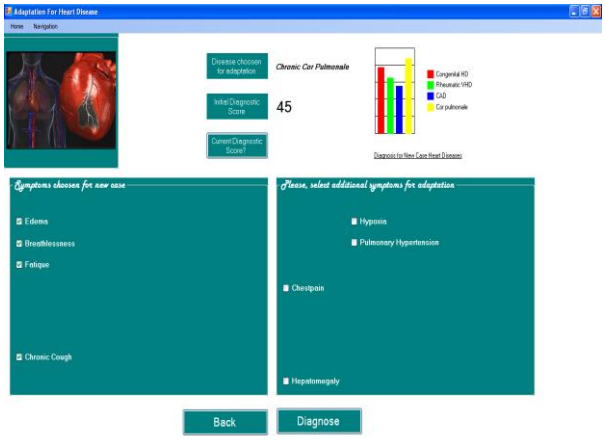


Figure 6: Adaptation Process- Differences between the New Case and Retrieved Case in CBR Model

Figure 7: Adaptation Process - Refining the Diagnostic Result in CBR Model

Figure 8: Case Verification of the Diagnostic Result in CBR Model

Step 9: REvised Diagnostic Result Verification:
After adaptation of the best REtrieved case, it is used for conducting Case Verification. Case Verification, REvises and verifies the accuracy of the answer using Diagnostic Score (Figure 8) and through the professional knowledge of Super Specialists, Medical Professionals and Experts.

Step 10: Case Indexing and REtaining of Diagnostic Result:
After REvision and verification of the new adapted case, it is now represented as New Case Diagnostic result and the New Case Diagnostic result and its components are used for indexing and finally REtained in the Case Base of CBR Desk for the future use. The new object case REtaining process comes into action for Diagnostic result storage so that it can be successfully used for future REtrieval and REuse. This Case Retaining enhances its completeness and consolidates the self-learning mechanism of the CBR Desk. Next time when the similar case is encountered, this case will become one of the important cases for reference.

8. MULTI MEDIA MEDICAL COMMUNICATION DESK

The Multi Media Medical Communication (MMMC) Desk adds another important dimension i.e. use of Multimedia technologies in Clinical Decision Support System to generate an excellent Multimedia environment by supporting alternate decisions through text, graphics, audio, images, animation, visual impressions and website links. Multi Media Medical Communication (MMMC) Desk allows visual simulations, visual output of simulations and mitigates uncertainty by providing rich information and generates a totally new level of cognitive-style thinking in Clinical Decision Making for more speedy, effective and efficient Medical Diagnosis.

The Multi Media Medical Communication (MMMC) Desk enhances the ability of System Users and Healthcare Workers with different Clinical Decision Making tasks, to use medical multimedia information intelligently and efficiently that make it robust, interactive, and permits System User's and Medical Decision Maker's to update, and relationally search medical multimedia information.

9. DIALOG MANAGEMENT SYSTEM

The purpose of Dialog Management (DiM) System is to facilitate the user-friendly interface for Human-Computer Interaction (HCI) and is designed for System Users and Healthcare Workers with a variety of Medical Decision Making needs of industrial/operational workers suffering from Chronic Heart Diseases. The DiM System captures the System Users and Healthcare Workers' preference, degree of expertise, skills and then receives and interprets their input, which is communicated to Central Medical Vision Navigator (CMVN) Board after computing the results from Model Base Management (MBM) System, Clinical Vision Technology (CVT) Base, and finally presents the output in the form of charts, text, graphs along with suitable form and visual displays in Central Medical Vision Navigator (CMVN) Board for effective and efficient Medical Diagnosis.

The DiM System capabilities are broadly classified into two Servers: Chronic Disease Queries Support (CDQS) Server and Medical Decision Exchange (MDE) Server due to variety of System Users and Healthcare Workers with different Medical Decision Making tasks. While Chronic Disease Queries Support (CDQS) Server allows ad-hoc retrieval of the information regarding Chronic Heart Diseases from Medical Data Base of Medical Data Base Access and Management (MDBAM) System, the second Medical Decision Exchange (MDE) Server supports the four main Chronic Heart Disease Application Modules of CI Portal to receive the processed data from Model Base Management (MBM) System, Clinical Vision Technology (CVT) Base. The output results are transmitted in real-time to Central Medical Vision Navigator (CMVN) Board having central black board problem solving architecture, to the Clinical Decision Making and User (CDMU) Desk, for analysis and speedy Medical Diagnosis.

1) 1.10 CLINICAL DECISION MAKING AND USER DESK

Clinical Decision Making and User (CDMU) Desk is responsible for detecting and elaborating on the nature of the Medical Problem; generating possible Medical Decision Solutions, evaluating potential and formulating strategies for implementing Clinical Decision Solutions. CDMU-Desk is primarily intended to be used by System Users and Healthcare Workers and plays an important role in the Medical Diagnosis of Chronic Heart Diseases through Chronic Disease Queries Support (CDQS) Server and Medical Decision Exchange (MDE) Server of the DiM System to improve the Quality of Life (QOL) of industrial/operational workers suffering from semistructured / unstructured Chronic Heart Diseases.

10. CONCLUDING REMARKS

"Multimedia Based Clinical Decision Support System for the Management of Chronic Heart Diseases (CHLD-MMCDSS)" is designed and developed with symbiotic approach to achieve the aim of furnishing the query support and decision exchange support functions, necessary for retrieving extensive global medical information resources, for real life interaction among System Users and Healthcare Workers, to enable them to make speedy Medical Diagnosis to improve the Quality of Life (QOL) of industrial / operational workers suffering from semistructured/unstructured Chronic Heart Diseases and its main components are analyzed and discussed.

In particular, emphasis is focused on the Medical Diagnosis Capsule (MDiC) and Disease Management Capsule (DmaC) based Model Base Management (MBM) System associated with four main Chronic Heart Disease Application Modules Module of CI Portal, Medical Data Base Access and Management (MDBAM) System, Central Medical Vision Navigator (CMVN) Board, Clinical Vision Technology (CVT) Base comprising of Case Base Reasoning Algorithm based Case Base Reasoning Desk (CBR-Desk), Multi Media Medical Communication (MMMC) Desk, Chronic Disease Queries Support (CDQS) Server and Medical Decision Exchange (MDE) Server based Dialog Management (DiM) System and Clinical Decision Making and User (CDMU) Desk.

The Clinical Vision Technology (CVT) Base, and Multi Media Medical Communication (MMMC) Desk provide the necessary global medical information, resources visualization, processing and analysis, for real life interaction among System Users and Healthcare Workers to enable speedy Medical Diagnosis.

The system incorporates the best of formal, incremental and systematic paradigms and accounts for dynamic interaction among System Users, Medical System Decision Makers, Subject, and Medical Experts to provide the alternative quality medical decisions or mixes of scenarios and decision plan. CHLD-MMCDSS helps in managerial Medical Decision Making by comparing the effectiveness of various alternate Multi Futuristic Medical Decision Parameters (MFMDP) and policies with the objective of identifying the semistructured/unstructured Chronic Heart Diseases of industrial / operational workers. The CHLD-MMCDSS can test existing as well as new medical cases and can answer several queries related with semistructured/unstructured Chronic Heart Diseases. The central black board problem solving architecture support at Central Medical Vision Navigator (CMVN) Board with the support from DiM System is very user friendly.

This is the part of my Ph.D. research work titled “Multimedia based Decision Support System for Diagnosis and Management of Chronic Heart and Lung Diseases” CHLD-MMCDSS at DEI, Agra.

REFERENCES

[1] Samore, M. H., B. Kim and C.A. Stephen, (2005), “Clinical Decision Support and Appropriateness of Antimicrobial Prescribing, A Randomized Trial”, *JAMA*, **294(18)**, 57-64.

[2] Sim, I., P. Gorman, R. A. Greenes, R. B. Haynes, B. Kaplan, H. Lehmann and P. C. Tang, (2001), “Clinical Decision Support Systems for the practice of evidence-based medicine”, *J Am Med Inform Assoc.*, **8(6)**, 527-534.

[3] Jeste, D.V., L.B. Dunn, D. P. Folsom and D. Zisook, (2008), “Multimedia educational aids for improving consumer knowledge about illness management and treatment decisions”, *Journal of Psychiatric Research*, **42**, 1–21.

[4] Consoli, S.M., M.B. Said, J. Jean, J. Menard, P.F. Plouin and G. Chatellier, (1995), “Benefits of a computer-assisted education program for hypertensive patients compared with standard education tools”, *Patient Education and Counseling*, **26**, 343-347.

[5] Degoulet, P., F.C. Jean and C. Safranh, (1995), “The health care professional multimedia workstation: development and integration issues”, *International Journal of Bio-Medical Computing*, **39**, I19-125.

[6] Kindler, H., R.U Peter., A. F. Baranov, T. M. Fliedner and D.Densow, (1998), “Providing dermatological photographs using the multimedia extension of the international computer Database for radiation accident case histories”, *International Journal of Medical Informatics*, **51**, 39–50.

[7] Rao, G.R. and M. Turoff, (2000), “A hypermedia-based Group Decision Support System to support collaborative medical decision-making”, *Decision Support Systems*, **30**, 187–216.

[8] Krause, P. and A. Glowinski, (1993), “Formal Specifications and Medical Decision Support Systems”, *Applied Artificial Intelligence*, **7**, 237-256.

[9] Mangiameli, P., D. West and R. Rampal, (2004), “Model selection for Medical Diagnosis Decision Support Systems”, *Decision Support Systems*, **36**, 247– 259.

[10] Schnipper, J.L., J. A. Linder, M.B. Palchuk , J. S. Einbinder , Q. Li , A. Postilnik and B. Middleton, (2008) , “ "Smart Forms" in an Electronic Medical Record: documentation-based Clinical Decision Support to improve disease management”, *J Am Med Inform Assoc.*, **15(4)**, 513-523.

[11] Goggin, L. S., H. E. Robert and D. A. Marcus, (2007), “Clinical Decision Support Systems and computer-aided diagnosis in otology”, *Otolaryngol Head Neck Surg.*, **136 (4)**, 21-26.

[12] Lomotan, E.A., L.J. Hoeksema, D.E. Edmonds, G. Ramírez-Garnica, R.N. Shiffman and L.I. Horwitz, (2012), “Evaluating the use of a computerized clinical decision support system for asthma by pediatric pulmonologists”, *Int J Med Inform.*, **81(3)**,157-65.

[13] Kumar, D.S., G. Sathyadevi and S. Sivanesh, (2011), “Decision Support System for Medical Diagnosis Using Data Mining”, *International Journal of Computer Science Issues*, **8(3)**, 147-153.

[14] Ruiz, D., V. Berenguer, A. Soriano and B. Sanchez, (2011), "A decision support system for the diagnosis of melanoma: A comparative approach", *Expert Systems with Applications*, **38(12)**, 15217–15223.

[15] Hoeksema, L.J., A. B. Asaad, E.A. Lomotan, D.E. Edmonds, G. Ramírez-Garnica, R.N. Shiffman and L.I. Horwitz,(2011), "Accuracy of a computerized clinical decision-support system for asthma assessment and management", *J Am Med Inform Assoc.*, **18(3)**, 243-250.

[16] Trivedi, M.H., J.K. Kern, B.D. Grannemann, K.Z. Altshuler and P. Sunderajan, (2004), "A computerized clinical decision support system as a means of implementing depression guidelines", *Psychiatr Serv.*, **55(8)**, 879-885.

[17] Susan, C., N. Wiratunga and R. C. Rowe,(2006), "Learning Adaptation Knowledge to Improve Case-Based Reasoning", *Artificial Intelligence*, **170(16-17),** 1175-1192.

[18] Slade, S., (1991), "Case-based reasoning: a research paradigm" *AI Mag.*,**12(1)**,42-55.

[19] Aamodt, A., and E. Plaza, (1994), "Case-based reasoning: foundational issues, methodological variations, and system approaches", *AI Comm.*, **7(1)**, 39-59.

[20] Mitchell, T., (1997), "Decision Tree Learning", in T. Mitchell, Machine Learning, the McGraw-Hill Companies, Inc., 52-78.

APPENDIX

Table 1: Diagnostic Symptoms Weight Values in CBR Model

S.No.	Symptoms	Weight	S.No.	Symptoms	Weight
1.	Edema	1	13.	Growth retardation	3
2.	breathlessness	1	14.	Fever	3
3.	Fatigue	2	15.	Sore throat	4
4.	Angina	2	16.	Arthalgia	4
5.	Heart murmur	3	17.	Nausea	4
6.	Syncope	3	18.	Pericardial rub	4
7.	Chronic cough(COPD)	2	19.	Anxiety	4
8.	Hepatomegaly	2	20.	Family history	4
9.	Hypoxia	3	21.	High BP	4
10	Pulmonary hypertension	3	22.	Diabetes (blood sugar)	4
11.	Hypertrophy & dilation of RV	3	23.	Smoking	4
12.	Cyanosis	3	24.	High cholesterol	4

Table 2: Nearest-Neighbor Calculation in CBR Model

S.No.	Symptoms	Weight	New Case T	Old Case S1	Loc. Simil.	Wei*Sim		Old Case S2	Loc. Simil.	Wei*Sim
1.	Edema	1	yes	yes	1	1		yes	1	1
2.	Breathlessness	1	yes	yes	1	1		no	0	0
3.	Fatigue	2	yes	yes	1	2		yes	1	2
4.	Angina	2	no	yes	0	0		yes	0	0
5.	Heart murmur	3	yes	yes	1	3		yes	1	3
6.	Syncope	3	no	yes	0	0		yes	0	0
7.	Chronic cough	2	yes	yes	1	2		yes	1	2
24.	High cholesterol	4	yes	no	0	0		no	0	0
	Sum	74				63				46

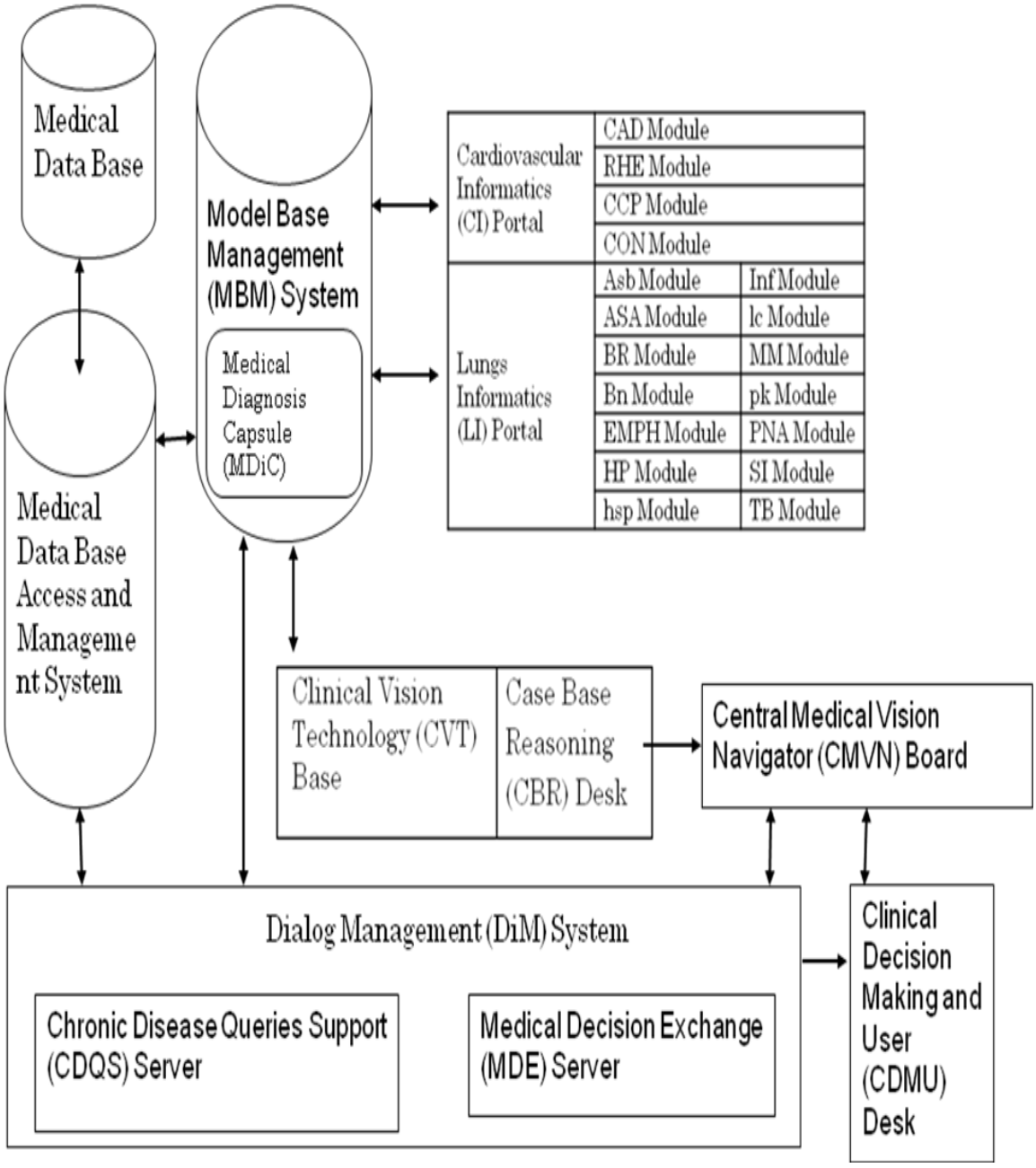


Figure 1: The Basic Components of the CHLD-MMCDSS Architecture

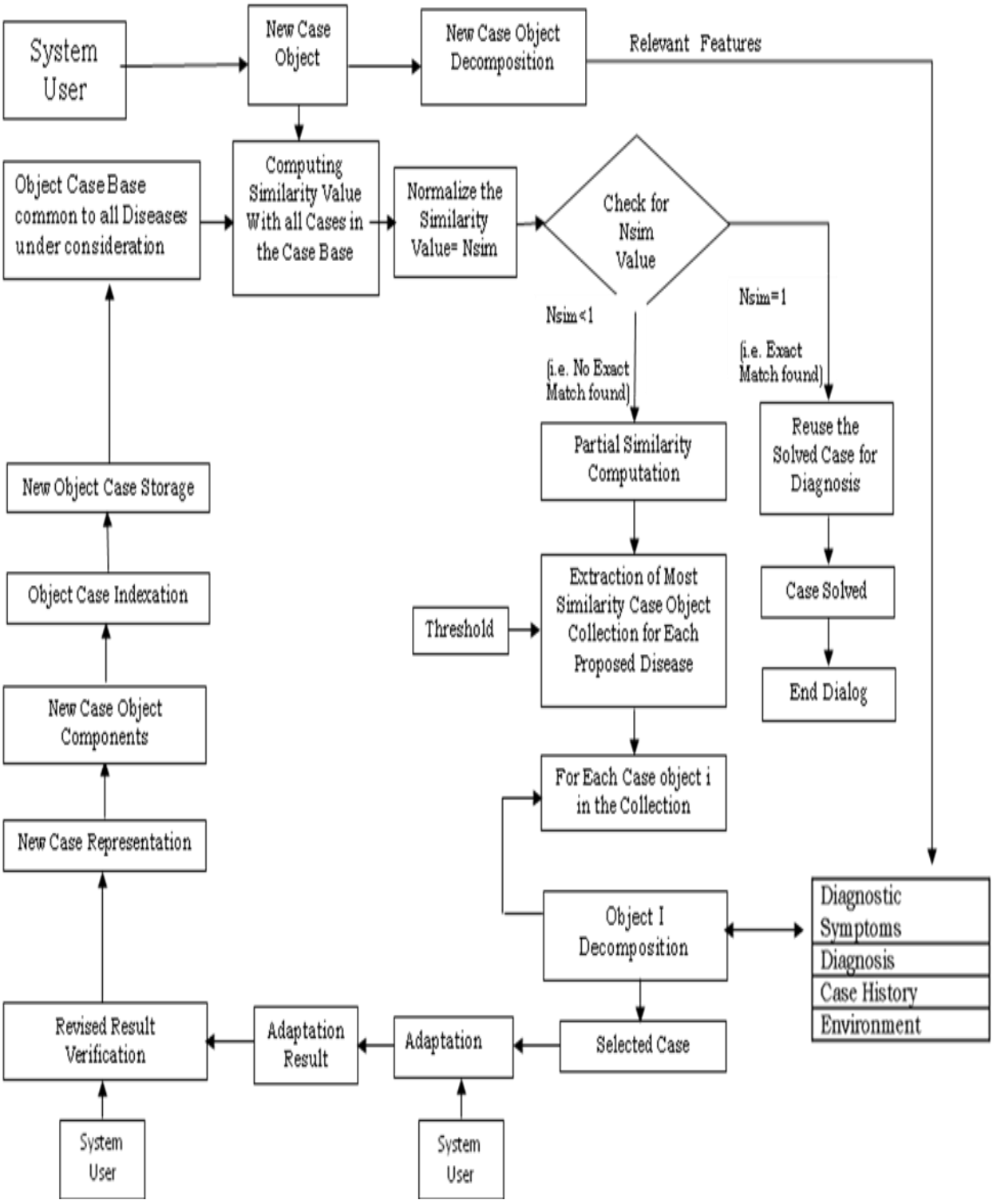


Figure 2: Flow Chart of the CBR Desk Operating Procedures

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Recursive Algorithm for Statistical Analysis of Bivariate Data Using Abstract Data Type Binary Tree

O.E. Oguike
Department of Computer Science
University of Nigeria
Nsukka, Nigeria
osondu.oguike@unn.edu.ng

ABSTRACT

Bivariate data can be defined as a collection of n pairs of data. Bivariate numeric data can be very useful in statistical analysis because it can be used to compute some important statistics, like correlation coefficient, regression analysis parameters, covariace etc. Though the recursive algorithms that will use binary tree to store and manipulate bivariate data will have double recursion, it has become a matter of academic concern to consider how the binary tree can be used to store and manipulate bivariate data. This paper develops novel algorithms for the operations of a binary tree that has two data items at the node, and it uses these novel algorithms to develop novel recursive algorithms that will compute some important bivariate statistics from the bivariate data stored in a binary tree.

Keywords- Bivariate data, correlation, regression analysis, covariance, statistical analysis, binary tree, recursive algorithm, abstract data type.

African Journal of Computing & ICT Reference Format
O.E. Oguike (2013). Recursive Algorithm for Statistical Analysis of Bivariate Data Using Abstract Data Type Binary Tree. Afr J. of Comp & ICTs. Vol 6, No. 1. Pp 161-168.

1. INTRODUCTION

Binary tree can be used to store, access and manipulate a collection of bivariate data. Though the traditional binary tree [3] requires that the nodes store a single data item, the binary tree that will be used to store bivariate data will have two data items at the node of each subtree. This implies that the operations of this binary tree will be extended from the traditional binary tree that each node stores single data item, to a binary tree that each node stores a pair of data item. Furthermore, the extension can continue in order to store multivariate collection of data. Data stored in a binary tree are stored in such a way that when they are accessed, the data will be in ascending or descending order. Figure 1 shows a collection of bivariate data in a tree.

Though tree is a traditional abstract data type with its access operations, it can be regarded as a nature-inspired model of computation. It depicts a natural tree, with its branches forming the left and right subtrees, and its fruits are the data that are stored at the nodes of the tree. The famous cocoa tree found in the western part of Nigeria is an example of such natural tree.

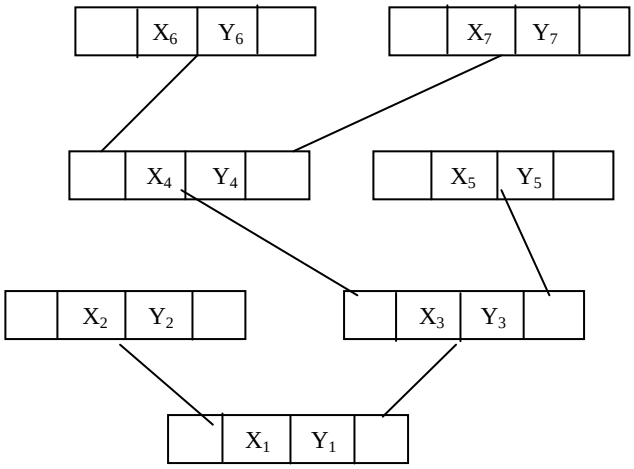


Figure 1: Bivariate Data in a Binary Tree Structure

The importance of bivariate statistical analysis cannot be overemphasized; this is because the various statistics that we can obtain from bivariate data are very useful in statistical analysis. Example, the bivariate regression analysis statistics help us to develop prediction model that can be used to predict a dependent variable, given an independent variable, the bivariate correlation coefficient can be used to establish the type of relationship that exist between two variables, while the statistic, covariance can be used to establish the extent of variation of bivariate data.

2. SURVEY OF RELATED LITERATURE

On the definition of abstract data type, the authors in [5] defined it as an abstract model of data objects with a formal encapsulation of logical architecture and valid operations of the data object. According to them, ADT encapsulates a data object and presents the user with an interface through which data can be accessed. In [6], the authors opined that a type might be viewed as a set of clothes (or a suit of armor) that protects an underlying untyped representation from arbitrary or unintended use. It provides a protective covering that hides the underlying representation and constrains the way objects may interact with other objects.

According to them, in an untyped system, untyped objects are *naked* in that the underlying representation is exposed for all to see. Even on parallel machines, writing on the importance of abstraction, the author in [7] said that abstraction is essential to enable clean separation of application code from the *assembly-level mechanisms* coordinating parallel execution, thus encapsulating the issues of communication, synchronization, and distribution. Still on the importance of data abstractions, [8] maintained that it improves the modularity of programs by encapsulating implementation details, and by providing a clear dilineation between design and implementation. It has been noted in [10] that the abstract data type facility (ADT) in a programming language allows the user to create a collection of representing types, and functions defined on these types, hiding the representation of the types, and the implementation of the functions, and allowing only the use of the functions defined.

Though it has been argued that the recursive approach to transversing a binary tree is not very efficient [9], but one should also consider other factors for the choice of recursive approach, which include, easy to implement, understand, debug and very structured when implemented. The statistical formulae that compute the various bivariate statistics have been expressed in different, but equivalent forms in [1],[2],[4].

3. DESIGN OF ALGORITHM FOR THE OPERATIONS OF ABSTRACT DATA TYPE TREE

The algorithms that will be used to store and access bivariate data in the abstract data type tree need to be defined. The structure of this tree will be defined in a java class called stattree as follows:

Stattree leftree
Double datax
Double datay
Stattree rightree

The algorithms for the operations of the abstract data type, tree can be defined as follows:

stattree emptytree()
1. Determine emptytree
1.2 emptytree = null;
2. ***Display emptytree***

The algorithm empytree takes nothing as parameter, but it returns a tree that does not have bivariate data, and does not have the left and right subtrees.

boolean istreeempty(stattree t)
1. Request t
2. Determine istreeempty
2.1 IF t = null THEN
2.1.1 istreeempty = true;
ELSE
2.1.2 istreeempty = false
3. Display istreeempty

The algorithm, istreeempty takes a tree as parameter, and it returns true if the tree is empty, otherwise it returns false.

stattree makenode(double x, double y, stattree l, stattree r)
1. Request data
1.1 Request x
1.2 Request y
1.3 Request l
1.4 Request r
2. Determine makenode
2.1 Declare and allocate, temp of the type stattree
2.2 temp.leftree = l;
2.3 temp.datax = x;
2.4 temp.datay = y;
2.5 temp.rightree = r;
2.6 makenode = temp
3. Display makenode;

The algorithm, makenode makes a new node of the defined tree structure. It takes four parameters, two real numbers that make up the bivariate data and two trees that make up the left and the right subtrees. The algorithm puts the parameters in their positions in the new node and returns the new node.

```
double treedatax(stattree t)
1. Request t
2. Determine treedatax
   2.1   IF istreeempty(t) THEN
         2.1.1   treedatax = "Empty tree does not have any data";
         ELSE
         2.1.2   treedatax = t.datax
3. Display treedatax
```

The algorithm, treedatax takes a tree as parameter and it returns the first part of the bivariate data at the node of the tree.

```
double treedatay(stattree t)
1  Request t
2  Determine treedatay
   2.1   IF istreeempty(t) THEN
         2.1.1   treedatay = "Empty tree does not have any data";
         ELSE
         2.1.2   treedatay = t.datay
3.  Display treedatay
```

The algorithm, treedatay takes a tree as parameter and it returns the second part of the bivariate data at the node of the tree.

```
stattree leftsubtree(stattree t)
1  Request t
2  Determine leftsubtree
   2.1   IF istreeempty(t) THEN
         2.1.1   leftsubtree = t
         ELSE
         2.1.2   leftsubtree = t.leftree
3.  Display leftsubtree
```

The algorithm leftsubtree takes a tree as parameter and it returns the left subtree of that tree if the tree is not empty, otherwise it returns the tree.

```
stattree rightsubtree(stattree t)
1  Request t
2.  Determine rightsubtree
   2.1  IF istreeempty(t) THEN
         2.1.1   rightsubtree = t
         ELSE
         2.1.2   rightsubtree = t.rightree
3.  Display rightsubtree
```

The algorithm rightsubtree takes a tree as parameter and it returns the right subtree of that tree if the tree is not empty, otherwise it returns the tree.

```
stattree insertree(double x, double y,  stattree t)
1.  Request data
    1.1   Request x
    1.2   Request y
    1.3   Request t
2.  Determine insertree
    2.1   IF istreeempty(t) THEN
          2.1.1   .insertree = makenode(x, y, null, null);
          ELSE
          2.1.2   IF (x < treedatax(t)) THEN
                  2.1.2.1   insertree = makenode(treedatax(t), treedatay(t), insertree(x,y,leftsubtree(t)),rightsubtree(t));
                  ELSE
                  2.1.2.2   insertree = makenode(treedatax(t),treedatay(t), leftsubtree(t),insertree(x,y,rightsubtree(t)));
    3   Display insertree
```

The algorithm insertree takes three parameters, which are the two real numbers that form the pair of data and a tree, the algorithm insertree makes new node and put the node at the correct position in the binary tree.

4. DESIGN OF RECURSIVE ALGORITHM FOR BIVARIATE STATISTICAL ANALYSIS

The algorithms that will calculate the following bivariate statistics, covariance, correlation and bivariate regression analysis parameters will be designed. The following algorithms will use the algorithms for the operations of binary tree to design the algorithms, which will be used to compute the three bivariate statistics mentioned above.

```
Double sumpro(stattree t)
1.  Request t
2.  Determine sumpro
    2.1   IF istreeempty(t) THEN
          2.1.1   sumpro = 0
          ELSE
          2.1.2   sumpro          =
                  treedatax(t)*treedatayt)  +
                  sumpro(leftsubtree(t))    +
                  sumpro(rightsubtree(t))
3.  Display sumpro
```

The algorithm sumpro takes a tree as parameter and it returns the sum of the product of the bivariate data in the tree.

```
Double sumx(stattree t)
1.  Request t
2   Determine sumx
    2.1   IF istreeempty(t) THEN
          2.1.1   sumx = 0
          ELSE
          2.1.2   sumx = treedatax(t) +
          sumx(leftsubtree(t))        +
          sumx(rightsubtree(t))
3   Display sumx
```

The algorithm sumx takes a tree as parameter and it returns the sum of datax attribute of the bivariate data in the tree.

```
Double sumy(stattree t)
1.  Request t
    2.  Determine sumy
        2.1 IF istreeempty(t) THEN
            2.1.1   sumy = 0
            ELSE
            2.1.2   sumy = treedatay(t) +
                    sumy(leftsubtree(t))  +
                    sumy(rightsubtree(t))
3   Display sumy
```

The algorithm sumy takes a tree as parameter and it returns the sum of datay attribute of the bivariate data in the tree.

```
Double sumsqx(stattree t)
1.  Request t
2.  Determine sumsqx
    2.1 IF istreeempty(t) THEN
        2.1.1 sumsqx = 0
        ELSE
        2.1.2 sumsqx = treedatax(t)*treedatax(t) +
              sumsqx(leftsubtree(t))             +
              sumsqx(rightsubtree(t))
3.  Display sumsqx
```

The algorithm sumsqx takes a tree as parameter and it returns the sum of square of the datax attribute of the bivariate data in the tree.

```
Double sumsqy(stattree t)
1.  Request t
2.  Determine sumsqy
    2.1 IF istreeempty(t) THEN
    2.1.1   sumsqy = 0
    ELSE
    2.1.2   sumsqy = treedatay(t)*treedatay(t) +
            sumsqy(leftsubtree(t))             +
            sumsqy(rightsubtree(t))
3.  Display sumsqy
```

The algorithm sumsqy takes a tree as parameter and it returns the sum of square of the datay attribute of the bivariate data in the tree.

```
int count(stattree t)
1.  Request t
2   Determine count
    2.1 IF istreeempty(t) THEN
        2.1.1 count = 0
        ELSE
        2.1.2 count = 1 + count(leftsubtree(t)) +
                          count(rightsubtree(t))
3.  Display count
```

The algorithm count takes a tree as parameter and it returns the number of pairs of data in the tree.

5. DESIGN OF ALGORITHM FOR THE BIVARIATE REGRESSION ANALYSIS PARAMETERS

One of the parameters of bivariate regression analysis is the beta parameter. For any given n collection of bivariate data, Xi,Yi, the beta parameter is defined, using this statistical formula:

$$\left(\frac{\left(n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i \right)\left(\sum_{i=1}^{n} Y_i \right) \right)}{\left(n\sum_{i=1}^{n} X_i{}^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right)} \right)$$

Using the statistical algorithms defined in the previous section, the algorithm that will compute the bivate regression parameter, beta can be defined as follows:

Double beta(stattree t)
1 Request t
2 Determine beta
2.1 top = count(t)*sumpro(t) – sumx(t)*sumy(t)
2.2 bottom = count(t)*sumsqx(t) – sumx(t)*sumx(t)
2.3 beta = top/bottom
3. Display beta

The second bivariate regression parameter is called alpha, and it is defined statistically as:

$$\frac{\sum_{i=1}^{n} Y_i}{n} - \beta \frac{\sum_{i=1}^{n} X_i}{n}$$

Using the recursive statistical algorithms defined in the previous section, the algorithm that will compute the bivariate regression parameter, alpha can be defined as follows:

Double alpha(stattree t)
1. Request t
2. Determine alpha
2.1 Alpha = (sumy(t)/count(t)) – beta(t)*sumx(t)/count(t)
3 Display alpha

6. DESIGN OF ALGORITHM FOR THE BIVARIATE CORRELATION COEFFICIENT

The correlation coefficient is used to establish the type of relationship, if any, between the two variables. The statistical formulae that can be used to compute the correlation coefficient is given below as:

$$\frac{\left(\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i \right)\left(\sum_{i=1}^{n} Y_i \right)}{n} \right)}{\sqrt{\left(\sum_{i=1}^{n} X_i{}^2 - \frac{\left(\sum_{i=1}^{n} X_i \right)^2}{n} \right)\left(\sum_{i=1}^{n} Y_i{}^2 - \frac{\left(\sum_{i=1}^{n} Y_i \right)^2}{n} \right)}}$$

Using the recursive statistical algorithms defined in the previous section, the algorithm that will compute the bivariate correlation coefficient can be defined as follows:

Double correlate(stattree t)
1 Request t
2 Determine correlate
2.1 top = sumpro(t)-sumx(t)*sumy(t)/count(t)
2.2 first = sumsqx(t)-sumx(t)*sumx(t)/count(t)
2.3 second = sumsqy(t)-sumy(t)*sumy(t)/count(t)
2.4 down = $\sqrt{first * second}$
2.5 correlate = top/down
3. Display correlate

7. DESIGN OF ALGORITHM FOR THE BIVARIATE STATISTICAL ANALYSIS, COVARIANCE

This is another important bivariate statistic, it is defined as :

$$\frac{\left(\sum_{i=1}^{n} XiYi - \frac{\left(\sum_{i=1}^{n} Xi\right)\left(\sum_{i=1}^{n} Yi\right)}{n}\right)}{(n-1)}$$

Using the recursive statistical algorithms defined in the previous section, the algorithm that will compute the bivariate statistic covariance can be defined as follows:

Double covariance(stattree t)
1 Request t
2 Determine covariance
2.1 IF count(t) <= 1 THEN
2.1.1 covariance = "Covariance is not defined.
ELSE
2.1.2 covariance = (sumpro(t) – sumx(t)*sumy(t)/count(t)/(count(t) – 1)
3. Display covariance

8. IMPLEMENTATION OF THE ALGORITHMS

A structured approach to programming was used to implement the recursive algorithms, using Java programming language. All the algorithms for the operations of the abstract data type, tree were implemented as java methods in a java class called statree, thereby encapsulating the operations of the abstract data type, tree. Furthermore, all the recursive statistical algorithms were implemented as java methods in another java class called statistical. Finally, all the algorithms for the various bivariate statistics were implemented as java methods in another class called bivariate A test program was written in another java class called test, which provided the options that the user would use to read in bivariate data into the abstract data type tree and compute and display the relevant bivariate statistics. The four java classes interacted with each other using the concept of composition and inheritance. Figure 2 below shows the various classes and the way they interacted with each other.

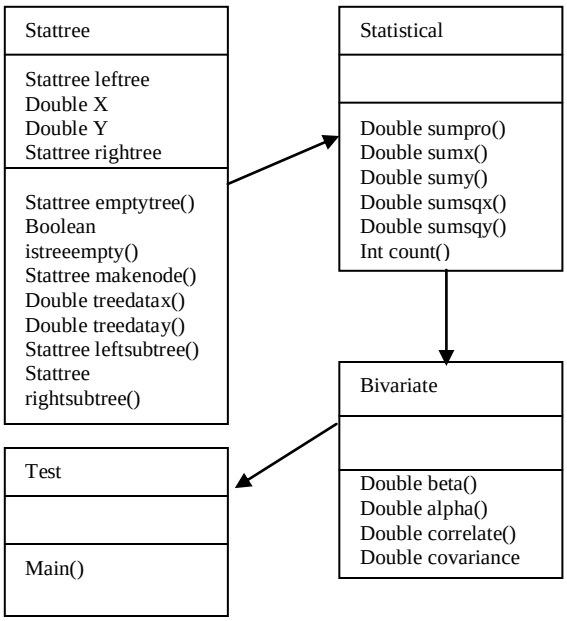


Figure 2 Class Diagram Showing Interaction Between the Classes..

9. RESULTS OF THE IMPLEMENTED ALGORITHMS

Sample test data was used to test the implemented algorithms. Table1 below shows a sample test data that was used to test the implemented algorithm.

Table1. Sample Test data

X	Y
5	20
10	30
12	34
15	38
23	45
25	46
30	50
32	51
35	55
40	59
42	61
45	65
47	67
50	70
51	71
55	75
56	77
58	79
60	81
61	82
65	87
67	88
70	91
75	96
77	98
80	100

The test program was used to read the collection of bivariate data and store them in a tree, using the implemented algorithms for the operations of a tree. The test program also tested all the implemented algorithms and the following results were obtained: bivariate regression parameter, beta is 1.0168, bivariate regression parameter, alpha is 19.617.

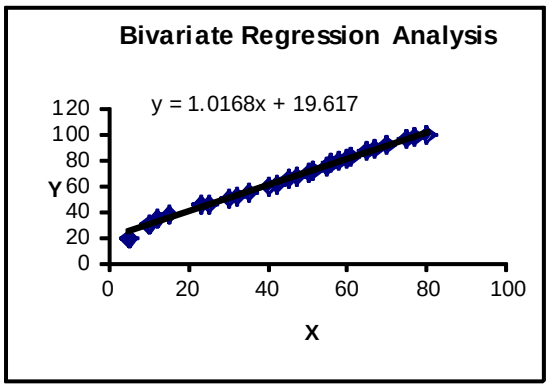


Figure 3: Confirmation of Result

When the two parameters were combined, the line of best fit was obtained as, Y = 1.0168X + 19.617. Other results obtained from the implemented algorithms were bivariate correlation coefficient, which was obtained as 0.997926, and covariance, which was obtained as 480.36. The results were validated using an alternative approach (Excel spreedsheet), and figure 3 above confirms the result for the regression analysis.

9. CONCLUSION

This paper has been able to use a binary tree to store and access a collection of bivariate data. Novel recursive algorithms have been developed using the operations of the binary tree. The paper has been able to use the operations of the abstract data type, tree to develop recursive algorithms that can be used for statistical analysis. It has also been able to use the recursive algorithms to compute some bivariate statistics that are of interest. All the algorithms were implemented and tested and the resulted were validated using an alternative approach (Excel spreedsheet).

REFERENCES

[1] Robert V. Hogg and Allen T. Craig, "Introduction to Mathematical Statistics", Macmilllan Publishing Co, 1978.

[2] Oguike Osondu Everestus and Monica N. Agu, "A First Course in Computer Programming: Statistical, Mathematical & String Processing Examples, using Pascal and Basic, Africom Limited.

[3] R. Harrison, "Abstract Data Types in MODULA-2", John Wiley & Sons, 1989.

[4] Kisbor Shridharbhai Trivedi, "Probability and Statistics with Reliability, Queuing and Computer Science Applications.", John Wiley & Sons, Inc., 2002.

[5] Yingxu Wang, Xinming Tan, Wuhan, Cyprian F. Ngolah, Philip C.-Y. Sheu, "The Formal Design Models of a Set of Abstract Data Types (ADTs)", International Journal of Software Science and Computational Intelligence, 2(4), 72-100, October-December 2010

[6] Luca Cardelli and Peter Wegner, "On Understanding Types, Data Abstraction, and Polymorphism", Computing Surveys, Vol 17 n. 4, pp 471-522, December 1985

[7] D. M. Goodeve S. A. Dobson, J. M. Nash, J. R. Davy, P. M. Dew, M. Kara and C. P. Wadsworth, "Toward a Model for Shared Data Abstraction with Performance1.", Journal of Parallel and Distributed Computing **49,** 156–167 (1998).

[8] John W. Baugh, Jr., **"**Data Abstraction in Engineering Software Development", *Journal of Computing in Civil Engineering*, Vol. 6, No. 3, July 1992, pp. 282-301

[9] Vinu V Das, "A new Non-Recursive Algorithm for Reconstructing a Binary Tree from its Traversals," International Journal of Algorithms, Computing and Mathematics Volume 2, Number 1, February 2009 ©Eashwar Publications

[10] Ramesh Subrahmanyam, Data "Abstraction and General Recursion.", Technical Reports , Department of Computer & Information Science, University of Pennsylvania, 1991, http://repository.upenn.edu/cis reports/322

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Analysis and Evaluation of Network-Based Intrusion Detection and Prevention System in an Enterprise Network Using Snort Freeware

[1]O.B. Lawal
Computer Science Department,
Olabisi Onabanjo University Consult, Ibadan, Nigeria
lawal5@yahoo.com

A. Ibitola
Department of Computer and Information Science
Lead City University
Ibadan, Nigeria

O.B. Longe
Department of Computer Science
University of Ibadan
Ibadan, Nigeria
longeolumide@fulbrightmail.org

[1]Corresponding Author: lawal5@yahoo.com

ABSTRACT

Information Systems and Networks are subject to electronic attacks. Attempts to breach information security are rising every day, along with the availability of the Vulnerability Assessment tools that are widely available on the internet, for free, as well as for a commercial use. Tools such as SubSeven, BackOrifce, Nmaps, L0ftCrack, can all be used to scan, identify, probe, and penetrate systems on the network. Firewalls are put in place to prevent unauthorized access to the Enterprise Networks. Unfortunately, firewalls alone are not enough to protect our systems. This paper describes the characteristics of Network-Base IDPS technologies, outlines the necessity of the implementation of Intrusion Detection Systems in the enterprise environment and a brief evaluation of Snort® Freeware technology.

Keywords: Intrusion detection, Intrusion Prevention, IDPS, Network, Firewalls, Snort Freeware

African Journal of Computing & ICT Reference Format
O.B. Lawal, A. Ibitola & O.B. Longe (2013). Analysis and Evaluation of Network-Based Intrusion Detection and Prevention System in an Enterprise Network Using Snort Freeware. Afr J. of Comp & ICTs. Vol 6, No. 1. Pp 169-184

1. INTRODUCTION

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices [1]. Intrusion prevention is the process of performing intrusion detection and attempting to stop detected possible incidents. Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, attempting to stop them, and reporting them to security administrators. In addition, organizations use IDPSs for other purposes, such as identifying problems with security policies, documenting existing threats, and deterring individuals from violating security policies. IDPSs have become a necessary addition to the security infrastructure of nearly every organization [1].

IDPSs typically record information related to observed events, notify security administrators of important observed events, and produce reports. Many IDPSs can also respond to a detected threat by attempting to prevent it from succeeding[1]. They use several response techniques, which involve the IDPS stopping the attack itself, changing the security environment (e.g., reconfiguring a firewall), or changing the attack's content [2]. Snort® is an open source network intrusion prevention and detection system (IDS/IPS) developed by Sourcefire [27]. Combining the benefits of signature, protocol, and anomaly-based inspection, Snort is the most widely deployed IDS/IPS technology worldwide [27]. With millions of downloads and nearly 400,000 registered users, Snort has become the de facto standard for IPS [27].

2. INTRUSION DETECTION AND PREVENTION PRINCIPLES

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible *incidents*, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices. Incidents have many causes, such as malware (e.g., worms, spyware), attackers gaining unauthorized access to systems from the Internet, and authorized users of systems who misuse their privileges or attempt to gain additional privileges for which they are not authorized. Although many incidents are malicious in nature, many others are not; for example, a person might mistype the address of a computer and accidentally attempt to connect to a different system without authorization [5].

Endorf [8] defines an *intrusion detection system* (IDS) has software that automates the intrusion detection process. An *intrusion prevention system* (IPS) is software that has all the capabilities of an intrusion detection system and can also attempt to stop possible incidents.

IDS and IPS technologies offer many of the same capabilities, and administrators can usually disable prevention features in IPS products, causing them to function as IDSs [1]. Accordingly, for brevity the term *intrusion detection and prevention systems* (*IDPS)* is used throughout the rest of this paper to refer to both IDS and IPS technologies.

IDS and IPS technologies offer many of the same capabilities, and administrators can usually disable prevention features in IPS products, causing them to function as IDSs [1]. Accordingly, for brevity the term *intrusion detection and prevention systems* (*IDPS)* is used throughout the rest of this guide to refer to both IDS and IPS technologies. Any exceptions are specifically noted.

2.1 Uses of IDPS Technologies

IDPSs are primarily focused on identifying possible incidents [5]. For example, an IDPS could detect when an attacker has successfully compromised a system by exploiting a vulnerability in the system. The IDPS could then report the incident to security administrators, who could quickly initiate incident response actions to minimize the damage caused by the incident. The IDPS could also log information that could be used by the incident handlers. Many IDPSs can also be configured to recognize violations of security policies. For example, some IDPSs can be configured with firewall ruleset-like settings, allowing them to identify network traffic that violates the organization's security or acceptable use policies. Also, some IDPSs can monitor file transfers and identify ones that might be suspicious, such as copying a large database onto a user's laptop [4].

Many IDPSs can also identify reconnaissance activity, which may indicate that an attack is imminent [4]. For example, some attack tools and forms of malware, particularly worms, perform reconnaissance activities such as host and port scans to identify targets for subsequent attacks. An IDPS might be able to block reconnaissance and notify security administrators, who can take actions if needed to alter other security controls to prevent related incidents. Because reconnaissance activity is so frequent on the Internet, reconnaissance detection is often performed primarily on protected internal networks [2].

In addition to identifying incidents and supporting incident response efforts, organizations have found other uses for IDPSs, including the following:

- **Identifying security policy problems.** An IDPS can provide some degree of quality control for security policy implementation, such as duplicating firewall rulesets and alerting when it sees network traffic that should have been blocked by the firewall but was not because of a firewall configuration error.
- **Documenting the existing threat to an organization.** IDPSs log information about the threats that they detect. Understanding the frequency and characteristics of attacks against an organization's computing resources is helpful in identifying the appropriate security measures for protecting the resources. The information can also be used to educate management about the threats that the organization faces.
- **Deterring individuals from violating security policies.** If individuals are aware that their actions are being monitored by IDPS technologies for security policy violations, they may be less likely to commit such violations because of the risk of detection.

Because of the increasing dependence on information systems and the prevalence and potential impact of intrusions against those systems, IDPSs have become a necessary addition to the security infrastructure of nearly every organization [9].

2.3 Key Functions of IDPS Technologies

There are many types of IDPS technologies, which are differentiated primarily by the types of events that they can recognize and the methodologies that they use to identify incidents [7]. In addition to monitoring and analyzing events to identify undesirable activity, all types of IDPS technologies typically perform the following functions:

- **Recording information related to observed events.** Information is usually recorded locally, and might also be sent to separate systems such as centralized logging servers, security information and event management (SIEM) solutions, and enterprise management systems.
- **Notifying security administrators of important observed events.** This notification, known as an *alert*, occurs through any of several methods, including the following: e-mails, pages, messages on the IDPS user interface, Simple Network Management Protocol (SNMP) traps, syslog messages, and user-defined programs and scripts. A notification message typically includes only basic information regarding an event; administrators need to access the IDPS for additional information [1].
- **Producing reports.** Reports summarize the monitored events or provide details on particular events of interest.

Some IDPSs are also able to change their security profile when a new threat is detected [7]. For example, an IDPS might be able to collect more detailed information for a particular session after malicious activity is detected within that session. An IDPS might also alter the settings for when certain alerts are triggered or what priority should be assigned to subsequent alerts after a particular threat is detected [7].

IPS technologies are differentiated from IDS technologies by one characteristic: IPS technologies can respond to a detected threat by attempting to prevent it from succeeding. They use several response techniques, which can be divided into the following groups:

- **The IPS stops the attack itself.** Examples of how this could be done are as follows:
  - Terminate the network connection or user session that is being used for the attack
  - Block access to the target (or possibly other likely targets) from the offending user account, IP address, or other attacker attribute
  - Block all access to the targeted host, service, application, or other resource.
- **The IPS changes the security environment.** The IPS could change the configuration of other security controls to disrupt an attack. Common examples are reconfiguring a network device (e.g., firewall, router, switch) to block access from the attacker or to the target, and altering a host-based firewall on a target to block incoming attacks. Some IPSs can even cause patches to be applied to a host if the IPS detects that the host has vulnerabilities.
- **The IPS changes the attack's content.** Some IPS technologies can remove or replace malicious portions of an attack to make it benign. A simple example is an IPS removing an infected file attachment from an e-mail and then permitting the cleaned email to reach its recipient. A more complex example is an IPS that acts as a proxy and *normalizes* incoming requests, which means that the proxy repackages the payloads of the requests, discarding header information. This might cause certain attacks to be discarded as part of the normalization process.

Another common attribute of IDPS technologies is that they cannot provide completely accurate detection [1] When an IDPS incorrectly identifies benign activity as being malicious, a *false positive* has occurred. When an IDPS fails to identify malicious activity, a *false negative* has occurred. It is not possible to eliminate all false positives and negatives; in most cases, reducing the occurrences of one increases the occurrences of the other. Many organizations choose to decrease false negatives at the cost of increasing false positives, which means that more malicious events are detected but more analysis resources are needed to differentiate false positives from true malicious events. Altering the configuration of an IDPS to improve its detection accuracy is known as *tuning* [1].

Most IDPS technologies also offer features that compensate for the use of common evasion techniques [1]. *Evasion* is modifying the format or timing of malicious activity so that its appearance changes but its effect is the same. Attackers use evasion techniques to try to prevent IDPS technologies from detecting their attacks. For example, an attacker could encode text characters in a particular way, knowing that the target understands the encoding and hoping that any monitoring IDPSs do not. Most IDPS technologies can overcome common evasion techniques by duplicating special processing performed by the targets. If the IDPS can "see" the activity in the same way that the target would, then evasion techniques will generally be unsuccessful at hiding attacks [10].

2.3 Common Detection Methodologies

IDPS technologies use many methodologies to detect incidents. Sections 2.3.1 through 2.3.3 discuss the primary classes of detection methodologies: signature-based, anomaly-based, and stateful protocol analysis, respectively. Most IDPS technologies use multiple detection methodologies, either separately or integrated, to provide more broad and accurate detection [1].

2.3.2 Signature-Based Detection

According to [10], a *signature* is a pattern that corresponds to a known threat. *Signature-based detection* is the process of comparing signatures against observed events to identify possible incidents.5 Examples of signatures are as follows:

- A telnet attempt with a username of "root", which is a violation of an organization's security policy
- An e-mail with a subject of "Free pictures!" and an attachment filename of "freepics.exe", which are characteristics of a known form of malware
- An operating system log entry with a status code value of 645, which indicates that the host's auditing has been disabled.

Signature-based detection is very effective at detecting known threats but largely ineffective at detecting previously unknown threats, threats disguised by the use of evasion techniques, and many variants of known threats. For example, if an attacker modified the malware in the previous example to use a filename of "freepics2.exe", a signature looking for "freepics.exe" would not match it [10].

Signature-based detection is the simplest detection method because it just compares the current unit of activity, such as a packet or a log entry, to a list of signatures using string comparison operations. Signature-based detection technologies have little understanding of many network or application protocols and cannot track and understand the state of complex communications [10]. For example, they cannot pair a request with the corresponding response, such as knowing that a request to a Web server for a particular page generated a response status code of 403, meaning that the server refused to fill the request. They also lack the ability to remember previous requests when processing the current request. This limitation prevents signature-based detection methods from detecting attacks that comprise multiple events if none of the events contains a clear indication of an attack [10].

2.3.2 Anomaly-Based Detection

Anomaly-based detection is the process of comparing definitions of what activity is considered normal against observed events to identify significant deviations [1]. An IDPS using anomaly-based detection has *profiles* that represent the normal behavior of such things as users, hosts, network connections, or applications. The profiles are developed by monitoring the characteristics of typical activity over a period of time. For example, a profile for a network might show that Web activity comprises an average of 13% of network bandwidth at the Internet border during typical workday hours [1].

The IDPS then uses statistical methods to compare the characteristics of current activity to thresholds related to the profile, such as detecting when Web activity comprises significantly more bandwidth than expected and alerting an administrator of the anomaly. Profiles can be developed for many behavioral attributes, such as the number of e-mails sent by a user, the number of failed login attempts for a host, and the level of processor usage for a host in a given period of time [1]. The major benefit of anomaly-based detection methods is that they can be very effective at detecting previously unknown threats [1]. For example, suppose that a computer becomes infected with a new type of malware. The malware could consume the computer's processing resources, send large numbers of e-mails, initiate large numbers of network connections, and perform other behavior that would be significantly different from the established profiles for the computer [10].

An initial profile is generated over a period of time (typically days, sometimes weeks) sometimes called a *training period*. Profiles for anomaly-based detection can either be static or dynamic [1]. Once generated, a static profile is unchanged unless the IDPS is specifically directed to generate a new profile. A dynamic profile is adjusted constantly as additional events are observed. Because systems and networks change over time, the corresponding measures of normal behavior also change; a static profile will eventually become inaccurate, so it needs to be regenerated periodically. Dynamic profiles do not have this problem, but they are susceptible to evasion attempts from attackers [8]. For example, an attacker can perform small amounts of malicious activity occasionally, then slowly increase the frequency and quantity of activity. If the rate of change is sufficiently slow, the IDPS might think the malicious activity is normal behavior and include it in its profile. Malicious activity might also be observed by an IDPS while it builds its initial profiles.

Inadvertently including malicious activity as part of a profile is a common problem with anomaly-based IDPS products. (In some cases, administrators can modify the profile to exclude activity in the profile that is known to be malicious.) [8]. Another problem with building profiles is that it can be very challenging in some cases to make them accurate, because computing activity can be so complex. For example, if a particular maintenance activity that performs large file transfers occurs only once a month, it might not be observed during the training period; when the maintenance occurs, it is likely to be considered a significant deviation from the profile and trigger an alert. Anomaly-based IDPS products often produce many false positives because of benign activity that deviates significantly from profiles, especially in more diverse or dynamic environments [8].

Another noteworthy problem with the use of anomaly-based detection techniques is that it is often difficult for analysts to determine why a particular alert was generated and to validate that an alert is accurate and not a false positive, because of the complexity of events and number of events that may have caused the alert to be generated [1].

2.3.3 Stateful Protocol Analysis

Stateful protocol analysis is the process of comparing predetermined profiles of generally accepted definitions of benign protocol activity for each protocol state against observed events to identify deviations [1]. Unlike anomaly-based detection, which uses host or network-specific profiles, stateful protocol analysis relies on vendor-developed universal profiles that specify how particular protocols should and should not be used. The "stateful" in stateful protocol analysis means that the IDPS is capable of understanding and tracking the state of network, transport, and application protocols that have a notion of state. For example, when a user starts a File Transfer Protocol (FTP) session, the session is initially in the unauthenticated state. Unauthenticated users should only perform a few commands in this state, such as viewing help information or providing usernames and passwords [8]. An important part of understanding state is pairing requests with responses, so when an FTP authentication attempt occurs, the IDPS can determine if it was successful by finding the status code in the corresponding response. Once the user has authenticated successfully, the session is in the authenticated state, and users are expected to perform any of several dozen commands. Performing most of these commands while in the unauthenticated state would be considered suspicious, but in the authenticated state performing most of them is considered benign [8].

2.4 Types of IDPS Technologies

There are many types of IDPS technologies [1]. For the purposes of this paper, they are divided into the following four groups based on the type of events that they monitor and the ways in which they are deployed:

i. Network-Based, which monitors network traffic for particular network segments or devices and analyzes the network and application protocol activity to identify suspicious activity [1]. It can identify many different types of events of interest. It is most commonly deployed at a boundary between networks, such as in proximity to border firewalls or routers, virtual private network (VPN) servers, remote access servers, and wireless networks.

ii. Wireless, which monitors wireless network traffic and analyzes its wireless networking protocols to identify suspicious activity involving the protocols themselves. It cannot identify suspicious activity in the application or higher-layer network protocols (e.g., TCP, UDP) that the wireless network traffic is transferring. It is most commonly deployed within range of an organization's wireless network to monitor it, but can also be deployed to locations where unauthorized wireless networking could be occurring [10].

iii. Network Behavior Analysis (NBA), which examines network traffic to identify threats that generate unusual traffic flows, such as distributed denial of service (DDoS) attacks, certain forms of malware (e.g., worms, backdoors), and policy violations (e.g., a client system providing network services to other systems) [1]. NBA systems are most often deployed to monitor flows on an organization's internal networks, and are also sometimes deployed where they can monitor flows between an organization's networks and external networks (e.g., the Internet, business partners' networks).

iv. Host-Based, which monitors the characteristics of a single host and the events occurring within that host for suspicious activity. Examples of the types of characteristics a host-based IDPS might monitor are network traffic (only for that host), system logs, running processes, application activity, file access and modification, and system and application configuration changes. Host-based IDPSs are most commonly deployed on critical hosts such as publicly accessible servers and servers containing sensitive information [8].

Some forms of IDPS are more mature than others because they have been in use much longer [10]. Network-based IDPS and some forms of host-based IDPS have been commercially available for over ten years (which form the base of this paper) [10]. Network behavior analysis software is a somewhat newer form of IDPS that evolved in part from products created primarily to detect DDoS attacks, and in part from products developed to monitor traffic flows on internal networks. Wireless technologies are a relatively new type of IDPS, developed in response to the popularity of wireless local area networks (WLAN) and the growing threats against WLANs and WLAN clients [6].

3. NETWORK-BASED IDPS

A network-based IDPS monitors network traffic for particular network segments or devices and analyzes network, transport, and application protocols to identify suspicious activity [9].

3.1 Networking Overview

TCP/IP is widely used throughout the world to provide network communications. TCP/IP communications are composed of four layers that work together [1] [3] [6]. When a user wants to transfer data across networks, the data is passed from the highest layer through intermediate layers to the lowest layer, with each layer adding more information. The lowest layer sends the accumulated data through the physical network; the data is then passed up through the layers to its destination. Essentially, the data produced by a layer is encapsulated in a larger container by the layer below it. The four TCP/IP layers, from highest to lowest, are shown in Figure 3-1.

Application Layer. This layer sends and receives data for particular applications, such as Domain Name System (DNS), Hypertext Transfer Protocol (HTTP), and Simple Mail Transfer Protocol (SMTP).
Transport Layer. This layer provides connection-oriented or connectionless services for transporting application layer services between networks. The transport layer can optionally ensure the reliability of communications. Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) are commonly used transport layer protocols.
Internet Protocol (IP) Layer (also known as Network Layer). This layer routes packets across networks. IPv4 is the fundamental network layer protocol for TCP/IP. Other commonly used protocols at the network layer are IPv6, Internet Control Message Protocol (ICMP), and Internet Group Management Protocol (IGMP).
Hardware Layer (also known as Data Link Layer). This layer handles communications on the physical network components. The best known data link layer protocol is Ethernet.

Figure 3.1 - TCP/IP Layers [1]

The four TCP/IP layers work together to transfer data between hosts. Network-based IDPSs typically perform most of their analysis at the application layer. They also analyze activity at the transport and network layers both to identify attacks at those layers and to facilitate the analysis of the application layer activity (e.g., a TCP port number may indicate which application is being used). Some network-based IDPSs also perform limited analysis at the hardware layer [3].

3.1.1 Application Layer

The application layer enables applications to transfer data between an application server and client. An example of an application layer protocol is Hypertext Transfer Protocol (HTTP), which transfers data between a Web server and a Web browser. Other common application layer protocols include Domain Name System (DNS), File Transfer Protocol (FTP), Simple Mail Transfer Protocol (SMTP), and Simple Network Management Protocol (SNMP). There are hundreds of unique application layer protocols in common use, and many more that are not so common. Regardless of the protocol in use, application data is generated and then passed to the transport layer for further processing [3].

3.1.2 Transport Layer

The transport layer is responsible for packaging data so that it can be transmitted between hosts. Most applications that communicate over networks rely on the transport layer to ensure reliable delivery of data [1][6]. Generally, this is accomplished by using TCP. When loss of some application data is not a concern (e.g., streaming audio, video), or the application itself ensures reliable delivery of data, UDP is typically used. UDP is connectionless; one host simply sends data to another host without any preliminary negotiations. Each TCP or UDP packet has a source port number and a destination port number. One of the ports is associated with a server application on one system; the other port is associated with a corresponding client application on the other system. Client systems typically select any available port number for application use, whereas server systems usually have a static port number dedicated to each application. Although UDP and TCP ports are very similar, they are distinct from each other and are not interchangeable [1].

3.1.3 Network Layer

The network layer, also known as the IP layer, is responsible for handling the addressing and routing of data that it receives from the transport layer. After the network layer has encapsulated the transport layer data, the resulting logical units are referred to as *packets*. Each packet contains a *header*, which is composed of various *fields* that specify characteristics of the transport protocol in use; optionally, packets may also contain a *payload*, which holds the application data. The IP header contains a field called IP Version, which indicates which version of IP is in use. Typically this is set to 4 for IPv4; but the use of IPv6 is increasing, so this field may be set to 6 instead. Other significant IP header fields are as follows:

- **Source and Destination IP Addresses.** These are the "from" and "to" addresses that are intended tindicate the endpoints of the communication. Example of IP addresses are 10.3.1.70 (IPv4) and 1000::2F:8A:400:427:9BD1 (IPv6).
- **IP Protocol Number.** This indicates which network or transport layer protocol the IP payload contains. Commonly used IP numbers includes 1 (ICMP), 6(TCP), 17 (UDP) and 50 (Encapsulating Security Payload [ESP]).

The network layer is also responsible for providing error and status information involving the addressing and routing of data [7]; it does this with ICMP. ICMP is a connectionless protocol that makes no attempt to guarantee that its error and status messages are delivered. Because it is designed to transfer limited information, not application data, ICMP does not have ports; instead, it has message types, which indicate the purpose of each ICMP message. Some message types also have message codes, which can be thought of as subtypes [3].

3.1.4 Hardware Layer

As the name implies, the hardware layer, also called the data link layer, involves the physical components of the network, including cables, routers, switches, and network interface cards (NIC) [1][7]. The hardware layer also includes various hardware layer protocols, with Ethernet being the most widely used. Ethernet relies on the concept of a media access control (MAC) address, which is a unique six-byte value (such as 00-02-B4-DA-92-2C) that is permanently assigned to a particular NIC. Each *frame*, the logical unit at the hardware layer, contains two MAC addresses, which indicate the MAC address of the NIC that just routed the frame and the MAC address of the next NIC to which the frame is being sent. As a frame passes through networking equipment (such as routers and firewalls) on its way between the original source host and the final destination host, the MAC addresses are updated to refer to the local source and destination. Several separate hardware layer transmissions may be linked together within a single network layer transmission [3].

In addition to the MAC addresses, each frame also contains an EtherType value, which indicates the protocol that the frame's payload contains (typically IP or Address Resolution Protocol [ARP]). When IP is used, each IP address maps to a particular MAC address. (Because multiple IP addresses can map to a single MAC address, a MAC address does not necessarily uniquely identify an IP address) [1].

3.2.2 Typical Components

A typical network-based IDPS is composed of sensors, one or more management servers, multiple consoles, and optionally one or more database servers (if the network-based IDPS supports their use). All of these components are similar to other types of IDPS technologies, except for the sensors [1]. A network-based IDPS sensor monitors and analyzes network activity on one or more network segments. The network interface cards that will be performing monitoring are placed into *promiscuous mode*, which means that they will accept all incoming packets that they see, regardless of their intended destinations. Most IDPS deployments use multiple sensors, with large deployments having hundreds of sensors. Sensors are available in two formats:

Appliance. An appliance-based sensor is comprised of specialized hardware and sensor software. The hardware is typically optimized for sensor use, including specialized NICs and NIC drivers for efficient capture of packets, and specialized processors or other hardware components that assist in analysis. Parts or all of the IDPS software might reside in firmware for increased efficiency. Appliances often use a customized, hardened operating system (OS) that administrators are not intended to acess directly [7].

Software Only. Some vendors sell sensor software without an appliance. Administrators can install the software onto hosts that meet certain specifications. The sensor software might include a customized OS, or it might be installed onto a standard OS just as any other application would [1] [3].

3.2.2 Network Architectures and Sensor Locations

Organizations should consider using management networks for their network-based IDPS deployments whenever feasible. If an IDPS is deployed without a separate management network, organizations should consider whether or not a VLAN is needed to protect the IDPS communications [11]. In addition to choosing the appropriate network for the components, administrators also need to decide where the IDPS sensors should be located. Sensors can be deployed in one of two modes [11]:

Inline. An *inline sensor* is deployed so that the network traffic it is monitoring must pass through it, much like the traffic flow associated with a firewall. In fact, some inline sensors are hybrid firewall/IDPS devices, while others are simply IDPSs [1]. The primary motivation for deploying IDPS sensors inline is to enable them to stop attacks by blocking network traffic. Inline sensors are typically placed where network firewalls and other network security devices would be placed—at the divisions between networks, such as connections with external networks and borders between different internal networks that should be segregated. Inline sensors that are not hybrid firewall/IDPS devices are often deployed on the more secure side of a network division so that they have less traffic to process. Figure 3-2 shows such a deployment. Sensors can also be placed on the less secure side of a network division to provide protection for and reduce the load on the dividing device, such as a firewall [1][7].

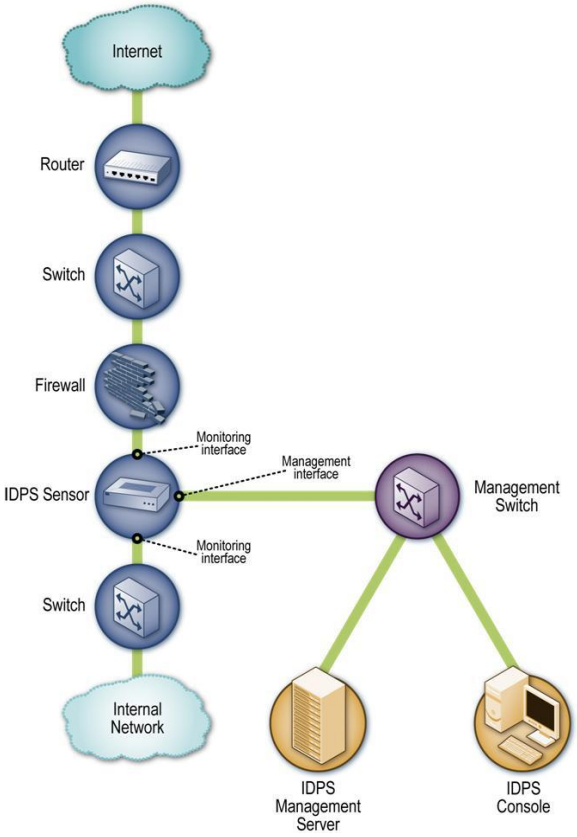


Figure 3.2 - Inline Network-Based IDPS Sensor Architecture Example [1]

Passive. A *passive sensor* is deployed so that it monitors a copy of the actual network traffic; no traffic actually passes through the sensor [1]. Passive sensors are typically deployed so that they can monitor key network locations, such as the divisions between networks, and key network segments, such as activity on a demilitarized zone (DMZ) subnet. Passive sensors can monitor traffic through various methods, including the following:

- **Spanning Port.** Many switches have a *spanning port*, which is a port that can see all network traffic going through the switch. Connecting a sensor to a spanning port can allow it to monitor traffic going to and from many hosts. Although this monitoring method is relatively easy and inexpensive, it can also be problematic. If a switch is configured or reconfigured incorrectly, the spanning port might not be able to see all the traffic [12].

– **Network Tap.** A *network tap* is a direct connection between a sensor and the physical network media itself, such as a fiber optic cable. The tap provides the sensor with a copy of all network traffic being carried by the media. Installing a tap generally involves some network downtime, and problems with a tap could cause additional downtime. Also, unlike spanning ports, which are usually already present throughout an organization, network taps need to be purchased as add-ons to the network.

– **IDS Load Balancer.** An *IDS load balancer* is a device that aggregates and directs network traffic to monitoring systems, including IDPS sensors. A load balancer can receive copies of network traffic from one or more spanning ports or network taps and aggregate traffic from different networks (e.g., reassemble a session that was split between two networks). The load balancer then distributes copies of the traffic to one or more listening devices, including IDPS sensors, based on a set of rules configured by an administrator. The rules tell the load balancer which types of traffic to provide to each listening device. Common configurations include the following:

Send all traffic to multiple IDPS sensors. This could be done for high availability or to have multiple types of IDPS sensors perform concurrent analysis of the same activity.

Dynamically split the traffic among multiple IDPS sensors based on volume. This is typically done to perform load balancing so that no sensor is overwhelmed with the amount of traffic and corresponding analysis [1][3].

Split the traffic among multiple IDPS sensors based on IP addresses, protocols, or other characteristics. This could be done for load balancing purposes, such as having one IDPS sensor dedicated to Web activity and another IDPS sensor monitoring all other activity. Splitting traffic could also be done to perform more detailed analysis of certain types of traffic (e.g., activity involving the most important hosts).

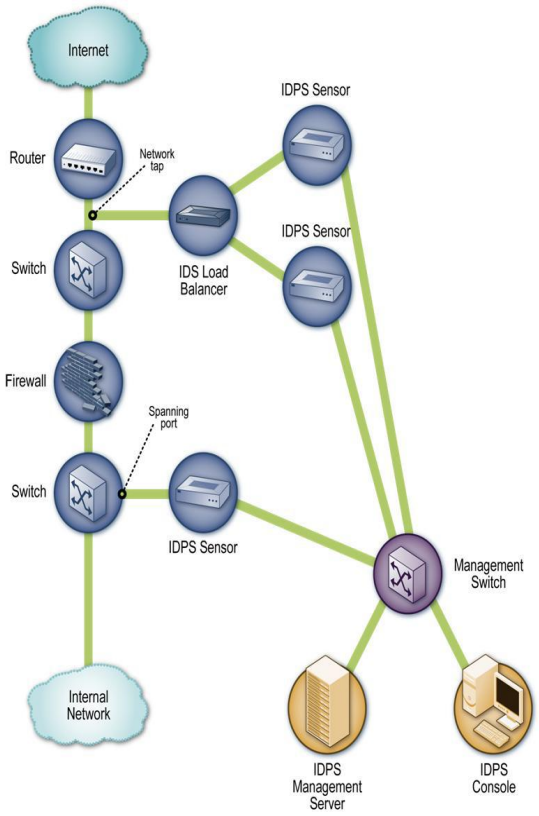


Figure 3.3 - Passive Network-Based IDPS Sensor Architecture Example [1]

3.3 Security Capabilities

Network-based IDPS products provide a wide variety of security capabilities. These are described below [14].

3.3.1 Information Gathering Capabilities

Some network-based IDPSs offer limited information gathering capabilities, which means that they can collect information on hosts and the network activity involving those hosts. Examples of information gathering capabilities are as follows:

- **Identifying Hosts.** An IDPS sensor might be able to create a list of hosts on the organization's network arranged by IP address or MAC address. The list can be used as a profile to identify new hosts on the network [14].
- **Identifying Operating Systems.** An IDPS sensor might be able to identify the OSs and OS versions used by the organization's hosts through various techniques. For example, the sensor could track which ports are used on each host, which could indicate a particular OS or OS family (e.g., Windows, Unix). Another technique is to analyze packet headers for certain unusual characteristics or combinations of characteristics that are exhibited by particular OSs; this is known as *passive fingerprinting*. Some sensors can also identify application versions (as described below).
- **Identifying Applications.** For some applications, an IDPS sensor can identify the application versions in use by keeping track of which ports are used and monitoring certain characteristics of application communications. For example, when a client establishes a connection with a server, the server might tell the client what application server software version it is running, and vice versa. Information on application versions can be used to identify potentially vulnerable applications, as well as unauthorized use of some applications [12].
- **Identifying Network Characteristics.** Some IDPS sensors collect general information about network traffic related to the configuration of network devices and hosts, such as the number of hops between two devices. This information can be used to detect changes to the network configuration.

3.3.2 Logging Capabilities

Network-based IDPSs typically perform extensive logging of data related to detected events [1]. This data can be used to confirm the validity of alerts, to investigate incidents, and to correlate events between the IDPS and other logging sources. Data fields commonly logged by network-based IDPSs include the following:

- Timestamp (usually date and time)
- Connection or session ID
- Event or alert type
- Rating (e.g., priority, severity, impact, confidence)
- Network, transport, and application layer protocols
- Source and destination IP addresses
- Source and destination TCP or UDP ports, or ICMP types and codes
- Number of bytes transmitted over the connection
- Decoded payload data, such as application requests and responses
- State-related information (e.g., authenticated username)
- Prevention action performed (if any) .

Most network-based IDPSs can also perform packet captures. Typically this is done once an alert has occurred, either to record subsequent activity in the connection or to record the entire connection if the IDPS has been temporarily storing the previous packets [14].

3.3.3 Detection Capabilities

Network-based IDPSs typically offer extensive and broad detection capabilities [15]. Most products use a combination of signature-based detection, anomaly-based detection, and stateful protocol analysis techniques to perform in-depth analysis of common protocols; organizations should use network-based IDPS products that use such a combination of techniques. The detection methods are usually tightly interwoven; for example, a stateful protocol analysis engine might parse activity into requests and responses, each of which is examined for anomalies and compared to signatures of known bad activity. Some products also use the same techniques and provide the same functionality as network behavior analysis (NBA) software [14]. This section discusses the following aspects of detection capabilities: (1) Types of events detected, (2) Detection accuracy, (3) Tuning and customization and (4) Technology limitations.

3.3.3.1 Types of Events Detected

The types of events most commonly detected by network-based IDPS sensors includes:

Application layer reconnaissance and attacks (e.g., banner grabbing, buffer overflows, format string attacks, password guessing, malware transmission). Most network-based IDPSs analyze several dozen application protocols. Commonly analyzed ones include Dynamic Host Configuration Protocol (DHCP), DNS, Finger, FTP, HTTP, Internet Message Access Protocol (IMAP), Internet Relay Chat (IRC), Network File System (NFS), Post Office Protocol (POP), rlogin/rsh, Remote Procedure Call (RPC), Session Initiation Protocol (SIP), Server Message Block (SMB), SMTP, SNMP, Telnet, and Trivial File Transfer Protocol (TFTP), as well as database protocols, instant messaging applications, and peer-to-peer file sharing software [1].

Transport layer reconnaissance and attacks (e.g., port scanning, unusual packet fragmentation, SYN floods). The most frequently analyzed transport layer protocols are TCP and UDP [16].

Network layer reconnaissance and attacks (e.g., spoofed IP addresses, illegal IP header values). The most frequently analyzed network layer protocols are IPv4, ICMP, and IGMP. Many products are also adding support for IPv6 analysis. The level of IPv6 analysis that network-based IDPSs can perform varies considerably among products.

Unexpected application services (e.g., tunneled protocols, backdoors, hosts running unauthorized application services). These are usually detected through stateful protocol analysis methods, which can determine if the activity in a connection is consistent with the expected application protocol, or through anomaly detection methods, which can identify changes in network flows and open ports on hosts.

Policy violations (e.g., use of inappropriate Web sites, use of forbidden application protocols). Some types of security policy violations can be detected by IDPSs that allow administrators to specify the characteristics of activity that should not be permitted, such as TCP or UDP port numbers, IP addresses, Web site names, and other pieces of data that can be identified by examining network traffic [1].

Some IDPSs can also monitor the initial negotiation conducted when establishing encrypted communications to identify client or server software that has known vulnerabilities or is misconfigured. This can include application layer protocols such as secure shell (SSH) and Secure Sockets Layer (SSL), and network layer virtual private networking protocols such as IP Security (IPsec) [3].

3.3.3.2 Detection Accuracy

Historically, network-based IDPSs have been associated with high rates of false positives and false negatives [20]. Most of the early technologies relied primarily on signature-based detection, which by itself is accurate only for detecting relatively simple well-known threats. Newer technologies use a combination of detection methods to increase accuracy and the breadth of detection, and generally the rates of false positives and false negatives have declined. Another common problem with network-based IDPSs' accuracy is that they typically require considerable tuning and customization to take into account the characteristics of the monitored environment [17].

Ideally, network-based IDPSs would be able to interpret all network activity just as the endpoints do [21]. For example, different types of Web servers can interpret the same Web request in different ways. Stateful protocol analysis techniques often attempt to do this by replicating the processing performed by common types of clients and servers. This allows the sensors to improve their detection accuracy slightly. Many attackers employ client and server-specific processing characteristics, such as handling character encodings, in their attacks as evasion techniques. Organizations should use network-based IDPSs that can compensate for the use of common evasion techniques [14].

3.3.3.3 Tuning and Customization

As mentioned in Section above, network-based IDPSs usually require extensive tuning and customization to improve their detection accuracy. Examples of tuning and customization capabilities are thresholds for port scans and application authentication attempts, blacklists and whitelists for host IP addresses and usernames, and alert settings. Some products also provide code editing features, which is usually limited to signatures but in some cases may allow access to additional code, such as programs used to perform stateful protocol analysis [1].

Some network-based IDPSs can use information regarding the organization's hosts to improve detection accuracy [1][3]. For example, an IDPS might allow administrators to specify the IP addresses used by the organization's Web servers, mail servers, and other common types of hosts, and also specify the types of services provided by each. This allows the IDPS to better prioritize alerts; for example, an alert for an Apache attack directed at an Apache Web server would have a higher priority than the same attack directed at a different type of Web server. Some network-based IDPSs can also import the results of vulnerability scans and use them to determine which attacks would likely be successful if not blocked. This allows the IDPS to make better decisions on prevention actions and prioritize alerts more accurately [1].

3.3.3.4 Technology Limitations

Although network-based IDPSs offer extensive detection capabilities, they do have some significant limitations. Three of the most important are analyzing encrypted network traffic, handling high traffic loads, and withstanding attacks against the IDPSs themselves. These limitations are discussed below [28].

Network-based IDPSs cannot detect attacks within encrypted network traffic, including virtual private network (VPN) connections, HTTP over SSL (HTTPS), and SSH sessions. As previously mentioned, some network-based IDPSs can do some analysis of the setup of encrypted connections, which can identify that the client or server software has known vulnerabilities or is misconfigured [28]. Network-based IDPSs may be unable to perform full analysis under high loads. Passive IDPS sensors might drop some packets, which could cause some incidents to go undetected, especially if stateful protocol analysis methods are in use [1]. For inline IDPS sensors, dropping packets under high loads causes disruptions in network availability; also, delays in processing packets could cause unacceptably high latency. To avoid this, organizations using inline IDPS sensors should select ones that can recognize high load conditions and either pass certain types of network traffic through the sensor without performing full analysis (i.e., partial or no analysis) or drop low-priority traffic to reduce load [1].

IDPS sensors are susceptible to various types of attacks [2]. Attackers can generate unusually large volumes of traffic, such as distributed denial of service (DDoS) attacks, and anomalous activity (e.g., unusually fragmented packets) to attempt to exhaust a sensor's resources or cause it to crash. Another attack technique, known as *blinding*, generates network traffic that is likely to trigger many IDPS alerts in a short period of time; typically, the network traffic is specially crafted to take advantage of typical configurations of IDPS sensors. In many cases, the blinding traffic is not intended to actually attack any targets. An attacker runs the "real" attack separately at the same time as the blinding traffic [2].

3.3.4 Prevention Capabilities

Network-based IDPS sensors offer various prevention capabilities [1], including the following:

Ending the Current TCP Session. A passive sensor can attempt to end an existing TCP session by sending TCP reset packets to both endpoints; this is sometimes called *session sniping*.

Performing Inline Firewalling. Most inline IDPS sensors offer firewall capabilities that can be used to drop or reject suspicious network activity.

Throttling Bandwidth Usage. If a particular protocol is being used inappropriately, such as for a DoS attack, malware distribution, or peer-to-peer file sharing, some inline IDPS sensors can limit the percentage of network bandwidth that the protocol can use. This prevents the activity from negatively impacting bandwidth usage for other resources.

Altering Malicious Content. As described earlier some inline IDPS sensors can sanitize part of a packet, which means that malicious content is replaced with benign content and the sanitized packet sent to its destination. A sensor that acts as a proxy might perform automatic normalization of all traffic, such as repackaging application payloads in new packets [1].

Reconfiguring Other Network Security Devices. Many IDPS sensors can instruct network security devices such as firewalls, routers, and switches to reconfigure themselves to block certain types of activity or route it elsewhere. This can be helpful in several situations, such as keeping an external attacker out of a network and quarantining an internal host that has been compromised

Running a Third-Party Program or Script. Some IDPS sensors can run an administrator-specified script or program when certain malicious activity is detected. This could trigger any prevention action desired by the administrator, such as reconfiguring other security devices to block the malicious activity. Third-party programs or scripts are most commonly used when the IDPS does not support the prevention actions that administrators want to have performed.

4. ADVANTAGES OF NETWORK BASED INTRUSION DETECTION SYSTEMS:

The followings are merits of NIDS [19].

1. Lower Cost of Ownership: Network based IDS can be deployed for each network segment. An IDS monitors network traffic destined for all the systems in a network segment. This nullifies the requirement of loading software at different hosts in the network segment. This reduces management overhead, as there is no need to maintain sensor software at the host level.

2. Easier to deploy: Network based IDS are easier to deploy as it does not affect existing systems or infrastructure. The network-based IDS systems are Operating system independent. A network based IDS sensor will listen for all the attacks on a network segment regardless of the type of the operating system the target host is running.

3. Detect network based attacks: Network based IDS sensors can detect attacks, which host-based sensors fail to detect. A network based IDS checks for all the packet headers for any malicious attack. Many IP-based denial of service attacks like TCP SYN attack, fragmented packet attack etc. can be identified only by looking at the packet headers as they travel across a network. A network based IDS sensor can quickly detect this type of attack by looking at the contents of the packets at the real time.

4. Retaining evidence: Network based IDS use live network traffic and does real time intrusion detection. Therefore, the attacker cannot remove evidence of attack. This data can be used for forensic analysis. On the other hand, a host-based sensor detects attacks by looking at the system log files. Lot of hackers are capable of making changes in the log files so as to remove any evidence of an attack.

5. Real Time detection and quick response: Network based IDS monitors traffic on a real time. So, network based IDS can detect malicious activity as they occur. Based on how the sensor is configured, such attack can be stopped even before they can get to a host and compromise the system. On the other hand, host based systems detect attacks by looking at changes made to system files. By this time critical systems may have already been compromised.

6. Detection of failed attacks: A network based IDS sensor deployed outside the firewall (as shown in picture1 above) can detect malicious attacks on resources behind the firewall, even though the firewall may be rejecting these attempts. This information can be very useful for forensic analysis. Host based sensors do not see rejected attacks that could never hit a host inside the firewall.

5. TYPES OF INTRUSION DETECTION MECHANISMS

Any enterprise or organisation transacting over the network should require intrusion detection and preventing mechanism. Below I provide a list of vendors that offer Intrusion Detection products and services. Products vary from freeware to commercially available [24].

Freeware:

- Snort - http://www.snort.org/
- Shadow

Commercially Available:

- RealSecure from ISS - http://www.iss.net/customer_care/resource_center/product_lit/
- NetProwler from Symantec- http://enterprisesecurity.symantec.com/products/products.cfm?ProductID=50&PID=5863267
- NFR - http://www.nfr.com/

6. SNORT OVERVIEW

Snort® is an open source network intrusion prevention and detection system (IDS/IPS) developed by Sourcefire [26]. Combining the benefits of signature, protocol, and anomaly-based inspection, Snort is the most widely deployed IDS/IPS technology worldwide. With millions of downloads and nearly 400,000 registered users, Snort has become the de facto standard for IPS [24].

The Snort® open source intrusion detection and prevention technology was created in 1998 by Martin Roesch [24], the founder of Sourcefire®. Snort uses a rule-driven language which combines the benefits of signature, protocol and anomaly-based inspection methods. With its dramatic speed, power and performance, Snort quickly gained momentum. With nearly 4 million downloads to date, Snort has become the single most widely deployed intrusion detection and prevention technology in the world [24].

The wide availability of open source Snort brings many advantages.
- Because the source code is open, development can occur at a markedly accelerated pace compared to proprietary models

- A vast community of security experts that continually reviews, tests, and proposes improvements to the code

- Security engineers and specialists the world over write Snort rules for new and evolving threats every hour of the day, often in record time.

As a result, the Snort open source community has a well-earned reputation for extraordinary organization and dedication.

6.1 Uses
Snort's open source network-based intrusion detection system (NIDS) has the ability to perform real-time traffic analysis and packet logging on Internet Protocol (IP) networks. Snort performs protocol analysis, content searching, and content matching. The program can also be used to detect probes or attacks, including, but not limited to, operating system fingerprinting attempts, common gateway interface, buffer overflows, server message block probes, and stealth port scans [25].

Snort can be configured in three main modes: sniffer, packet logger, and network intrusion detection. In sniffer mode, the program will read network packets and display them on the console. In packet logger mode, the program will log packets to the disk. In intrusion detection mode, the program will monitor network traffic and analyze it against a rule set defined by the user. The program will then perform a specific action based on what has been identified [25].

6.2 Required Software
Before installing Snort you need to verify that you have a number of software packages installed [27]. These are: Libpcap, PCRE, Libnet and Barnyard.

If you're downloading Snort binaries the only requirements are WinPcap and Barnyard for Windows users [27].

- **Libpcap** In the field of computer network administration, pcap (packet capture) consists of an application programming interface (API) for capturing network traffic. Unix-like systems implement pcap in the libpcap library; Windows uses a port of libpcap known as WinPcap [25].

Monitoring software may use libpcap and/or WinPcap to capture packets traveling over a network. libpcap and WinPcap also support saving captured packets to a file and reading files containing saved packets. Snort uses these files to read network traffic and analyze it [25].

- **Perl Compatible Regular Expressions** (PCRE) is a regular expression C library inspired by Perl's external interface, written by Philip Hazel. The PCRE library is incorporated into a number of prominent open-source programs such as the Apache HTTP Server, the PHP and R scripting languages, and Snort [25].

- **Libdnet** is a generic networking API that provides access to several protocols.

- **Barnyard** is an output system for Snort. Snort creates a special binary output format called unified. Barnyard2 reads this file, and then resends the data to a database back-end. Unlike the database output plugin, Barnyard2 manages the sending of events to the database and stores them when the database temporarily cannot accept connections [26].

- **DAQ** is the Data-Acquisition API that is necessary to use Snort version 2.9.0 and above [24].

6.3 Snort Rules
Once you've downloaded and installed Snort, you must download and maintain a ruleset in order for Snort to have the latest detection capabilities [27].

Sourcefire VRT Certified Rules
Sourcefire Vulnerability Research Team (VRT) Rules are the official rules of snort.org. Each rule is developed and tested using the same rigorous standards the VRT uses for Sourcefire customers. These rules are distributed under the VRT Certified Rules License Agreement. This license agreement allows you to study and modify VRT rules but restricts commercial redistribution [26].

There are two ways Snort users can obtain these rules:

- Subscribers: Real-time access to VRT Certified Rules Updates requires a paid subscription.
- Registered Users: Registered users of Snort.org are able to download and use VRT rules free of charge 30 days after their initial release date [24].

6.4 Keeping your Snort Rules Updated
Users may opt to manually download and updates rules files, however most Snort users automate the process using PulledPork, an open source perl script. If you plan on using PulledPork to manage VRT Rules updates you'll need to login to snort.org and generate an Oinkcode to properly configure PulledPork [24].

6.5 Installing Snort
32/64bit Windows Intrusion Detection System (WinIDS) Guided Install

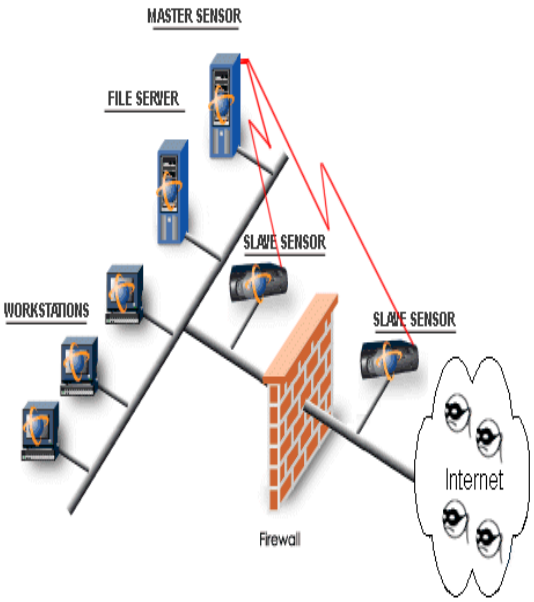


Fig 3: Snort Installation Scenario (Michael E. Steele @ winsnort.com) [27]

6.5.1 Pre-installation Tasks

- Downloading and extracting the 'WinIDS - Core Software Pack', and the 'WinIDS - (32/64bit) Software Support Pack [27].
- Download The 'WinIDS - Core Software Pack' to a temporary location.

Depending on the processors architecture being used (32bit or 64bit)

- Open an Explorer window and navigate to the location of the 'winids-csdp-xx.xx.xx.zip' file, right-click the 'winids-csdp-xx.xx.xx.zip' file, highlight and left-click 'Extract all...', in the 'Files will be extracted to this folder:' dialog box type 'd:\temp' (less the outside quotes), left-click and uncheck the 'Show extracted files when complete' radio box, left-click extract, in 'Password:' dialog box type 'w1nsn03t.c0m' (less the outside quotes), and left-click 'OK'.

Right-click the 'winids-sdp-xnn-xx.xx.xx.zip' file, highlight and left-click 'Extract all...', in the 'Files will be extracted to this folder:' dialog box type 'd:\temp' (less the outside quotes), left-click and uncheck the 'Show extracted files when complete' radio box, left-click extract, in 'Password:' dialog box type 'w1nsn03t.c0m' (less the outside quotes), left-click 'OK', and eXit the explorer window [27].

6.5.2 Installing WinPcap

The full installation guide can be found at http://winsnort.com/index.php?module=Pages&func=display&pageid=49 [27].

7. RECOMMENDATIONS

Implementing the following recommendations should facilitate more efficient and effective intrusion detection and prevention system use for any enterprise or organisation [1],[3].

Organizations should ensure that all IDPS components are secured appropriately [1]**.**
Securing IDPS components is very important because IDPSs are often targeted by attackers who want to prevent the IDPSs from detecting attacks or want to gain access to sensitive information in the IDPSs, such as host configurations and known vulnerabilities. All components' operating systems and applications should be kept fully up-to-date, and all software-based IDPS components should be hardened against threats. Specific protective actions of particular importance include creating separate accounts for each IDPS user and administrator, restricting network access to IDPS components, and ensuring that IDPS management communications are protected appropriately, such as encrypting them or transmitting them over a physically or logically separate network.

Administrators should maintain the security of the IDPS components on an ongoing basis, including verifying that the components are functioning as desired, monitoring the components for security issues, performing regular vulnerability assessments, responding appropriately to vulnerabilities in the IDPS components, and testing and deploying IDPS updates. Administrators should also back up configuration settings periodically and before applying updates to ensure that existing settings are not inadvertently lost.

Organizations should consider using multiple types of IDPS technologies to achieve more comprehensive and accurate detection and prevention of malicious activity.
The four primary types of IDPS technologies— network-based, wireless, NBA, and host-based— each offer fundamentally different information gathering, logging, detection, and prevention capabilities. Each technology type offers benefits over the others, such as detecting some events that the others cannot and detecting some events with significantly greater accuracy than the other technologies. In many environments, a robust IDPS solution cannot be achieved without using multiple types of IDPS technologies. For most environments, a combination of network-based and host-based IDPS technologies is needed for an effective IDPS solution. Wireless IDPS technologies may also be needed if the organization determines that its wireless networks need additional monitoring or if the organization wants to ensure that rogue wireless networks are not in use in the organization's facilities. NBA technologies can also be deployed if organizations desire additional detection capabilities for denial of service attacks, worms, and

other threats that NBAs are particularly well-suited to detecting. Organizations should consider the different capabilities of each technology type along with other cost-benefit information when selecting IDPS technologies [3].

Organizations planning to use multiple types of IDPS technologies or multiple products of the same IDPS technology type should consider whether or not the IDPSs should be integrated.
Direct IDPS integration most often occurs when an organization uses multiple IDPS products from a single vendor, by having a single console that can be used to manage and monitor the multiple products. Some products can also mutually share data, which can speed the analysis process and help users to better prioritize threats [1]. A more limited form of direct IDPS integration is having one IDPS product provide data for another IDPS product (but no data sharing in the opposite direction). Indirect IDPS integration is usually performed with security information and event management (SIEM) software, which is designed to import information from various security-related logs and correlate events among them. SIEM software complements IDPS technologies in several ways, including correlating events logged by different technologies, displaying data from many event sources, and providing supporting information from other sources to help users verify the accuracy of IDPS alerts [1].

Before evaluating IDPS products, organizations should define the requirements that the products should meet.
Evaluators need to understand the characteristics of the organization's system and network environments, so that a compatible IDPS can be selected that can monitor the events of interest on the systems and/or networks. Evaluators should articulate the goals and objectives they wish to attain by using an IDPS, such as stopping common attacks, identifying misconfigured wireless network devices, and detecting misuse of the organization's system and network resources. Evaluators should also review their existing security policies, which serve as a specification for many of the features that the IDPS products need to provide. In addition, evaluators should understand whether or not the organization is subject to oversight or review by another organization. If so, they should determine if that oversight authority requires IDPSs or other specific system security resources. Resource constraints should also be taken into consideration by evaluators. Evaluators also need to define specialized sets of requirements for the following:

- Security capabilities, including information gathering, logging, detection, and prevention
- Performance, including maximum capacity and performance features
- Management, including design and implementation (e.g., reliability, interoperability, scalability, product security), operation and maintenance (including software updates), and training, documentation, and technical support
- Life cycle costs, both initial and maintenance costs.

When evaluating IDPS products, organizations should consider using a combination of several sources of data on the products' characteristics and capabilities.
Common product data sources include test lab or real-world product testing, vendor-provided information, third-party product reviews, and previous IDPS experience from individuals within the organization and trusted individuals at other organizations. When using data from other parties, organizations should consider the fidelity of the data because it is often presented without an explanation of how it was generated. There are several major challenges in performing in-depth hands-on IDPS testing, such as the considerable resources needed and the lack of a standard test methodology and test suites, which often make it infeasible. However, limited IDPS testing is helpful for evaluating security requirements, performance, and operation and maintenance capabilities [1], [3].

8. CONCLUSION

IDS are becoming the logical next step for many organizations after deploying firewall technology at the network perimeter. IDS can offer protection from external users and internal attackers, where traffic doesn't go past the firewall at all.

However, the following points are very important to always keep in mind. If all of these points are not adhered to, an IDS implementation along with a firewall alone can not make a highly secured infrastructure.

1. Strong identification and authentication:
An IDS uses very good signature analysis mechanisms to detect intrusions or potential misuse; however, organizations must still ensure that they have strong user identification and authentication mechanism in place.

2. Intrusion Detection Systems are not a solution to all security concerns:
IDS perform an excellent job of ensuring that intruder attempts are monitored and reported. In addition, companies must employ a process of employee education, system testing, and development of and adherence to a good security policy in order to minimize the risk of intrusions.

3. An IDS is not a substitute for a good security policy:
As with other security and monitoring products, an IDS functions as one element of a corporate security policy. Successful intrusion detection requires that a well-defined policy must be followed to ensure that intrusions and vulnerabilities, virus outbreaks, etc. are handled according to corporate security policy guidelines.

4. Human intervention is required:
The security administrator or network manager must investigate the attack once it is detected and reported, determine how it occurred, correct the problem and take necessary action to prevent the occurrence of the same attack in future.

Lastly, Tight integration between host and network based IDS is very much necessary. As shown in Fig. 3.2 and 3.3, it is advised to use network based IDS inside and outside the firewall or between each firewall in a multi-layered environment and host based IDS on all critical or key hosts. Also, as shown in Picture1, it is important although not always necessary to have an integrated deployment of host based and network based Intrusion Detection Systems.

As security continues to move to the center stage, managers and network administrators alike are beginning to focus their attention on intrusion-detection technology. While modern-day IDSes are far from bulletproof, they can add significant value to established information-security programs. With vendors working on eliminating the shortcomings of Intrusion Detection Systems, the future looks brighter for this technology.

9. REFERENCES

[1] Bace, Rebecca, *Intrusion Detection*, Macmillan Technical Publishing, 2000.
[2] Bejtlich, Richard, *The Tao of Network Security Monitoring: Beyond Intrusion Detection*, Addison-Wesley, 2004.
[3] Crothers, Tim, *Implementing Intrusion Detection Systems: A Hands-On Guide for Securing the Network*, 2002.
[4] Endorf, Carl et al, *Intrusion Detection and Prevention*, McGraw-Hill Osborne Media, 2003.
[5] Kruegel, Chris et al, *Intrusion Detection and Correlation: Challenges and Solutions*, Springer, 2004.
[6] Nazario, Jose, *Defense and Detection Strategies Against Internet Worms*, Artech House Publishers, 2003.
[7] Northcutt, Stephen and Novak, Judy, *Network Intrusion Detection: An Analyst's Handbook, Third Edition*, New
[8] Riders, 2003.
[9] Rash, Michael et al, *Intrusion Prevention and Active Response: Deployment Network and Host IPS*, Syngress, 2005.
[10] NIST SP 800-86, *Guide to Integrating Forensic Techniques into Incident Response*, http://csrc.nist.gov/publications/nistpubs/.
[11] NIST SP 800-94, *Guide to Intrusion Detection and Prevention Systems (IDPS)*: Recommendations of the National Institute of Standards and Technology by Karen Scarfone and Peter Mell.
[12] NIST SP 800-95, *Guide to Web Services Security (DRAFT)*, available at http://csrc.nist.gov/publications/drafts.html.
[13] NIST Special Publication (SP) 800-61, *Computer Security Incident Handling Guide*, available at http://csrc.nist.gov/publications/nistpubs/.
[14] Distributed Intrusion Detection System (DShield), http://dshield.org/indexd.html
[15] IETF Intrusion Detection Exchange Format (idwg) Working Group, http://www.ietf.org/html.charters/OLD/idwg-charter.html
[16] An Introduction to Intrusion Detection Systems, http://www.securityfocus.com/infocus/1520
[17] Evaluating Intrusion Prevention Systems, http://www.cioupdate.com/article.php/3563306
[18] Intrusion Detection System Overview, http://www.webopedia.com/TERM/I/intrusion_detection_system.html
[19] Recommendations for Deploying an Intrusion-Detection System, http://searchsecurity.techtarget.com/tip/1,289483,sid14_gci781471,00.html
[20] http://www.sans.org/reading_room/whitepapers/detection/understanding-intrusion-detection-systems_337
[21] http://www.iana.org/assignments/version-numbers
[22] http://www.iana.org/assignments/protocol-numbers.
[23] http://www.iana.org/assignments/icmp-parameters.
[24] http://www.iana.org/assignments/ethernet-numbers
[25] http://www.snort.org/start/rules
[26] http://en.wikipedia.org/wiki/Snort_(software)
[27] http://www.sourcefire.com/security-technologies/open-source/snort
[28] 27.http://winsnort.com/index.php?module=Pages&func=display&pageid=49. Written by Michael E. Steele.
[29] 28.http://www.juniper.net/solutions/literature/white_papers/200063.pdf

Vol 6. No. 2, June 2013

African Journal of Computing & ICT

www.ajocict.net

Soft Computing Techniques for Weather and Climate Change Studies

A.B. Adeyemo
Computer Science Department,
University of Ibadan
Ibadan, Nigeria
sesanadeyemo@gmail.com

ABSTRACT

The Weather is a continuous, data-intensive, multidimensional, dynamic process that makes weather forecasting a formidable challenge. Weather forecasting involves predicting how the present state of the atmosphere will change. Climate is the long-term effect of the sun's radiation on the rotating earth's varied surface and atmosphere. The Day-by-day variations in a given area constitute the weather, whereas climate is the long-term synthesis of such variations. A simple, long-term summary of weather changes, however, is still not a true picture of climate. To obtain this requires the analysis of daily, monthly, and yearly patterns. This paper presents the use of Self Organizing Maps (SOM) and Co-Active Neuro Fuzzy Inference System (CANFIS) soft computing techniques for weather and climate change studies using historical data collected from the city of Ibadan, Nigeria between 1951 and 2009. The results show that the soft computing techniques can be used for knowledge discovery in weather prediction and climate change studies.

Keywords: Weather Forecasting, Climate Change, Self Organizing Maps (SOM), Co-Active Neuro-Fuzzy Inference System (CANFIS)

African Journal of Computing & ICT Reference Format
A.B. Adeyemo (2013). Soft Computing Techniques for Weather and Climate Change Studies.
Afr J. of Comp & ICTs. Vol 6, No. 1. Pp 185-199

1. INTRODUTION

Weather can be described as the state of the atmosphere at a given time and place. Most weather activities takes place in the troposphere which is the lowest layer of the atmosphere. Weather is measured and described in a variety of ways by meteorologists who are scientists that study and predict weather. Weather forecasting entails predicting how the present state of the atmosphere will change. It has been one of the most scientifically and technologically challenging problems around the world in the last century. This is due mainly to two factors: first, it's used for many human activities and secondly, due to the opportunism created by the various technological advances that are directly related to this concrete research field, like the evolution of computation and the improvement in measurement systems [3]. To make an accurate prediction is one of the major challenges facing meteorologist all over the world. Since ancient times, weather prediction has been one of the most interesting and fascinating domain. Scientists have tried to forecast meteorological characteristics using a number of methods, some of these methods being more accurate than others [6]. Some methods of forecasting are:

- Persistence Method: This method assumes that the conditions at the time of the forecast will not change. For example, if it is sunny and 67degrees today, the persistence method predicts that it will be sunny and 67 degrees the next day. The method works well when weather patterns change very little and features on the weather maps move very slowly. However, if weather conditions change significantly from day to day, the persistence method usually breaks down and is not the best forecasting method to use although it's the simplest method. While it is assumed that it only works for shorter-term forecasts (e.g. a forecast for a day or two), actually one of the most useful roles of the persistence forecast is predicting long range weather conditions or making climate forecasts.
- Climatology Method: This method involves averaging weather statistics accumulated over many years to make the forecast. For example to predict the weather for a city on a particular day involves assembling all the weather data that has been recorded for that day over the years and taking its average. This method works well only when the weather pattern is similar to that expected for the chosen time of year. If the pattern is unusual for the given time of year, the climatology method will often fail.

- Analog Method: This method is a slightly more complicated method of producing a forecast. It involves examining the days forecast scenario and remembering a day in the past when the weather scenario looked similar (an analog). The forecaster would predict that the weather in this forecast will behave the same as it did in the past. The method is difficult to use because it is virtually impossible to find a perfect analog.
- Numerical Weather Prediction: This method uses the power of computers to make a forecast. Complex computer programs, also known as forecast models that run on supercomputers provide predictions on many atmospheric variables such as temperature, pressure, wind, and rainfall. A forecaster examines how the features predicted by the computer will interact to produce the day's weather. Numerical Weather Prediction models are complex mathematical models and require a lot of computational power to solve these equations. Also the time taken to produce results limits the methods ability to provide very short-term forecasts. Even though the method may be costly and time consuming, it provides more accurate results for forecasts of both short and longer time steps ranging from one hour and beyond.

Modern weather forecasts are made by obtaining present weather conditions by ground observations, observations from ships and aircraft, radiosondes, doppler radar, and satellites. This information is sent to meteorological centers where the data is collected, analyzed, and a variety of charts, maps, and graphs are plotted using high-speed computers which are then used to develop surface and upper-air maps with the help of meteorologists who correct the maps for errors. The computer is also used to predict the future state of the maps.

Climate is the long-term effect of the sun's radiation on the rotating earth's varied surface and atmosphere. The Day-by-day variations in a given area constitute the weather, whereas climate is the long-term synthesis of such variations. Weather is measured by thermometers, rain gauges, barometers, and other instruments, but the study of climate relies on statistics which can be handled efficiently by computers. A simple, long-term summary of weather changes, however, is still not a true picture of climate. To obtain this requires the analysis of daily, monthly, and yearly patterns [7].

Climate change is a significant and lasting change in the statistical distribution of weather patterns over periods ranging from decades to millions of years. It may be a change in average weather conditions or the distribution of events around that average (e.g., more or fewer extreme weather events). The term is sometimes used to refer specifically to climate change caused by human activity, as opposed to changes in climate that may have resulted as part of Earth's natural processes. Climate change today is synonymous with anthropogenic global warming. While the term global warming has been used to refer to surface temperature increases, climate change includes global warming and everything else that increasing greenhouse gas amounts will affect [16].

Weather prediction and Climate change/prediction problems differ with respect to the time scale being considered. While weather predictions are on a much shorter time scale such as next few hours, days or weeks, Climatic forecasts are on a longer time scale such as hundreds or thousands of years. Therefore a variation of 5^{o}C change in temperature from one day to the next day is not as significant as a 5^{o}C in global climatic temperature. The mathematical models for explaining and predicting future weather and climatic conditions are complex non-linear dynamical systems which are currently being processed with the aid of powerful super computer systems running massively parallel algorithms. However, a lot of historical data whose origins coincides with the advent of modern weather forecasting, which started with the invention of the barometer in 1644, has been collected over the years. With this database of weather/climatic data available, data mining techniques which have proven to be efficient at solving complex non-linear problems can be applied to both weather prediction and climate change/forecasting problems.

In this work rainfall and weather data for the city of Ibadan were analyzed using Self Organizing Maps (SOM) and Co-Active Neuro Fuzzy Inference System (CANFIS) soft computing techniques for knowledge discovery and prediction of weather parameters such as wind-speed, sunshine irradiation, minimum and maximum temperature, and rainfall intensity using historical data.

2. SELF-ORGANIZING MAPS

Clustering groups data into sets in such a way that the intra-cluster similarity is maximized while the inter-cluster similarity is minimized [18]. Self-Organizing Maps (SOM) are competitive networks that provide a "topological" mapping from the input space to the clusters [11]. The SOM working algorithm is a variant of multidimensional vectors clustering of which the Kmeans clustering algorithm is an example of this type of algorithm [14]. ***Competitive learning*** is an adaptive process in which the neurons in a neural network gradually become sensitive to different input categories, sets of samples in a specific domain of the input space. A kind of a division of labor emerges in the network when different neurons specialize to represent different types of inputs. The specialization is enforced by competition among the neurons: when an input $\mathbf{x}$ arrives, the neuron that is best able to represent it wins the competition and is allowed to learn it even better, as described in Kohonen [14]: If there exists an ordering between the neurons, i.e., the neurons are located on a discrete lattice, the self-organizing map, the competitive learning algorithm can be generalized: if not only the winning neuron but also its **neighbors** on the lattice are allowed to learn, neighboring neurons will gradually specialize to represent similar inputs, and the representations will become **ordered** on the map lattice. This is the essence of the SOM algorithm [10]). The neurons represent the inputs with reference vectors $\mathbf{m}_i$, the components of which correspond to synaptic weights. One reference vector is associated with each neuron called **unit** in a more abstract setting. The unit, indexed with *c*, whose reference vector is nearest to the input $\mathbf{x}$ is the winner of the competition:

$$c = c(\mathbf{x}) = \arg\min_i \{\|\mathbf{x} - \mathbf{m}_i\|^2\} \quad \text{.....Equation 1}$$

Usually Euclidean metric is used, although other choices are possible as well. The winning unit and its neighbors adapt to represent the input even better by modifying their reference vectors towards the current input. The amount the units learn will be governed by a neighborhood kernel *h*, which is a decreasing function of the distance of the units from the winning unit on the map lattice. If the locations of units *i* and *j* on the map grid are denoted by the two-dimensional vectors $\mathbf{r}_i$ and $\mathbf{r}_j$, respectively,

then

$$h_{ij}(t) = h(\|\mathbf{r}_i - \mathbf{r}_j\|; t)$$

where *t* denotes time.

During the learning process at time *t* the reference vectors are changed iteratively according to the following adaptation rule, Where

$\mathbf{x}(t)$ is the input at time *t* and

$$c = c(\mathbf{x}(t))$$

is the index of the winning unit:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)] \quad \text{.....Equation 2}$$

In practice the neighborhood kernel is chosen to be wide in the beginning of the learning process to guarantee global ordering of the map, and both its width and height decrease slowly during learning. The learning process consisting of winner selection by Equation 1 and adaptation of the synaptic weights by Equation 2, can be modeled with a neural network structure, in which the neurons are coupled by inhibitory connections[10].

A problem with the clustering methods is that the interpretation of the clusters may be difficult. Most clustering algorithms prefer certain cluster shapes, and the algorithms will always assign the data to clusters of such shapes even if there were no clusters in the data. Therefore, if the goal is not just to compress the data set but also to make inferences about its cluster structure, it is essential to analyze whether the data set exhibits a clustering tendency.

The results of the cluster analysis need to be validated, as well [10]. Another potential problem is that the choice of the number of clusters may be critical: for example in Kmeans clustering different kinds of clusters may emerge when *K* is changed. Good initialization of the cluster centroids may also be crucial; some clusters may even be left empty if their centroids lie initially far from the distribution of data. Clustering can be used to reduce the amount of data and to induce a categorization. In exploratory data analysis, however, the categories have only limited value as such. The clusters should be illustrated somehow to aid in understanding of what they are like [10]. The SOM is a special case in that it can be used at the same time both to reduce the amount of data by clustering, and for projecting the data nonlinearly onto a lower-dimensional display.

Generally, standard SOMs are applied to feature values of numeric type which usually uses an Euclidean function to calculate the distances between input vectors and reference vectors. During the learning, the update of reference vectors is performed by incremental or arithmetic operations. Unfortunately, these calculations are not practical on categorical values. Although categorical data has been discussed in some clustering algorithms, it is not directly addressed in SOMs due to the limitation of learning laws. A traditional approach is to translate categories to numeric numbers in data preprocess and then perform standard SOMs on the transformed data [5].

3. CO-ACTIVE NEURO FUZZY INFERENCE SYSTEM (CANFIS)

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a huge number of highly interconnected artificial neurons working in unison to solve specific problems. The artificial neuron is an information processing unit that is fundamental to the operation of a neural network. ANNs, like people, learn by example. An ANN is configured for a particular application, such as pattern recognition or data classification, through a learning process. Fuzzy logic is a form of logic used in systems where variables can have degrees of truthfulness or false hood represented by a range of values between 1 (true) and 0 (false). With fuzzy logic, the outcome of an operation can be expressed imprecisely rather than as a certainty. For example, instead of being true or false, an outcome might have such meaning as probably true, possibly true, and probably false.

A hybrid neuro-fuzzy system is a fuzzy system that uses a learning algorithm based on gradients or inspired by the neural networks theory (heuristical learning strategies) to determine its parameters (fuzzy sets and fuzzy rules) through the patterns processing (input and output). A neuro-fuzzy system can be interpreted as a set of fuzzy rules.

This system can be totally created from input output data or initialized with the *à priori* knowledge (fuzzy rules). The resultant system by fusing fuzzy systems and neural networks has as advantages of learning through patterns and the easy interpretation of its functionality. There are several different ways to develop hybrid neuro-fuzzy systems; therefore, being a recent research subject, each researcher has defined its own particular models. These models are similar in its essence, but they present basic differences. Examples of these systems include: Fuzzy Adaptive learning Control Network (FALCON), Adaptive Neuro Fuzzy Inference System (ANFIS), Generalized Approximate Reasoning based Intelligent Control (GARIC), Fuzzy Inference and Neural Network in Fuzzy Inference Software (FINEST), FUzzy Net (FUN), Evolving Fuzzy Neural Network (EFuNN), Self Constructing Neural Fuzzy Inference Network (SONFIN) and Neuro-Fuzzy Control (NEFCON) [1].

ANFIS is perhaps the first integrated hybrid neuro-fuzzy model and belongs to the class of rule-extracting systems using a decompositional strategy, where rules are extracted at the level of individual nodes within the neural network. After extraction, rules are aggregated to form global behavior descriptions [17]. The ANFIS architecture (figure 1) consists of a five-layer structure where in the first layer, the node output is the degree to which the given input satisfies the linguistic label associated to the membership functions. The parameters in the first layer are referred to as premise parameters. In the second layer, each node function computes the firing strength of the associated rule. In general, any T-norm operators that perform fuzzy AND can be used as the node function in this layer. Each node i in third layer calculates the ratio of the ith rule firing strength for the sum of firing strength of all rules. The fourth layer is the product of the normalized firing level and the individual rule output of the corresponding rule. Parameters in this layer are referred to as consequent parameters. The single node function of the fifth layer computes the overall system output as the sum of all incoming signals. Only Layer 1 and Layer 4 contain modifiable parameters. [17]

www.ajocict.net

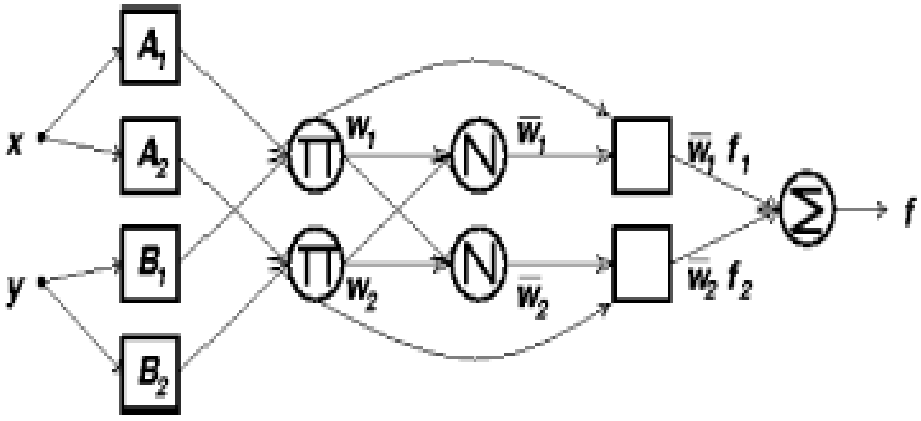


Figure 1: ANFIS system (Implementing Tsukamoto fuzzy inference system

The architecture of ANFIS is a one-output fuzzy inference system based on an adaptive network. CANFIS is a generalized form of ANFIS. CANFIS allows more than one output with the advantage of non-linear rule formations. The CANFIS model (figure 2) integrates fuzzy inputs with a modular neural network to quickly solve poorly defined problems [9].

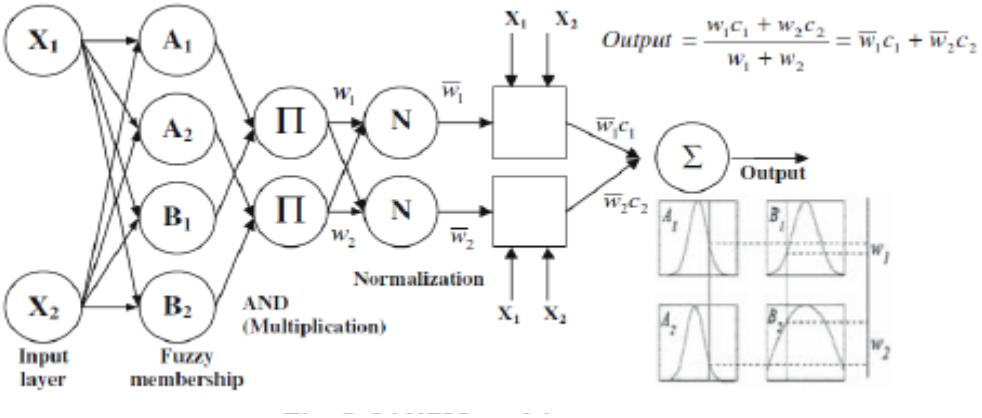


Figure 2: The CANFIS system

The fundamental concepts of CANFIS can be found in Heydari and Talaee [9] and in J.S.R. Jang, C.T. Sun, and E. Mizutani in Neuro-Fuzzy and Soft Computing, published by Prentice Hall in 1997 where it was originally proposed. CANFIS also uses a five layer modular network structure that uses two fuzzy structures: the Tsukamoto model and the Takagi–Sugeno–Kang (TSK) model and uses the Bell and Gaussian Membership Functions. The CANFIS model is more computationally intensive than most other models.

MATERIALS AND METHODS

For knowledge discovery from the rainfall data using SOM clustering, the rainfall data from Ibadan, Oyo State, Nigeria, covering the period between 1951 and 2008 was used. The rainfall data was clustered using NeuroXL Clusterize software. The attributes of the dataset are: year, month and volume of rainfall in millimeters. This is presented in table 1.

Table 1: Attributes of Ibadan Rainfall Dataset

Attribute	Type	Data Coding	Description
Year	Numeric	Numeric (Ordinal)	Year considered e.g. 1951, 1952, 1953 etc
Month	Categorical	Numeric (Ordinal)	Month considered e.g. Jan, Feb, Mar etc
Rainfall	Numeric	Numeric (Continous)	Monthly rainfall (mm)

Clusters were generated and the arithmetic mean and standard deviation of each cluster group was computed. As a measure of central tendency the mean describes the central location of data (that is the center of gravity of the data), it is usually used with other statistical measures such as the standard deviation because it can be affected by extreme values in the data set and therefore be biased. The standard deviation describes the spread of the data and is a popular measure of dispersion. It measures the average distance between a single observation and its mean. For weather prediction, the weather data was collected from Ibadan, Oyo State, Nigeria from Ibadan Synoptic Airport through the Nigerian Meteorological Agency, Oyo State office which covered a period of 120 months from January 2000 to December 2009 [8] was used. The data attributes of the weather dataset are presented in table 2.

Table 2: Attributes of Meteorological Dataset

Attribute	Type	Description
Year	Numeric	Year considered
Month	Numeric	Month considered
Wind speed	Numeric	Wind run in km
Radiation	Numeric	The amount of solar irradiation
MinTemp	Numeric	The monthly Minimum Temperature
MaxTemp	Numeric	Maximum Temperature
Rainfall	Numeric	Total monthly rainfall

The CANFIS program was used for the weather parameters prediction program because CANFIS can support multiple output variables, while ANFIS supports one output variable. The NeuroSolutions implementation of the CANFIS program was used. NeuroSolutions incorporates a number of standard parameters which can be used to evaluate the performance of a neural network model. Those relevant to the study as presented in NeuroSolutions are:

a. Mean Squared Error: The mean squared error (MSE) and the normalized mean squared error (NMSE) can be used to determine how well the network output fits the desired output. It however does not necessarily reflect whether the two sets of data move in the same direction. The mean squared error is simply two times the average cost, For instance, by simply scaling the network output, the MSE can be changed without changing the directionality of the data.

b. The correlation coefficient (r): The value of the correlation coefficient ranges between [-1, 1]. When r =1 there is a perfect positive correlation. When r = -1, there is perfectly linear negative correlation. When r =0 there is no correlation between the two quantities investigated. Intermediate values describe partial correlations.

c. Learning curve: The learning curve that shows how the mean square error evolves with the training iteration is a quantity that can be used to check the progress of learning. The difficulty of the task and how to control the learning parameters can be determined form the learning curve. When the learning curve is flat, the step size is increased to speed up leaning, when the learning curve oscillates up and down the step size is deceased. In the extreme, the error curve increases steadily upwards, showing that learning is unstable, at this point the network is reset.

4. RESULTS AND DISCUSSION

Rainfall data clustering using SOM

Three SOM software used were used and the one the gave the best result was selected. These are: NeuroXL Clusterizer, a proprietary Artificial Neural Network clustering software implemented in Microsoft Excel; NNClust, a SOM software implemented in Microsoft Excel and Pittnet Neural Network Educational Software, an open source SOM software coded using C++. The NNClust SOM clustering software was trained using a starting learning rate of 0.9 and was trained over 100 epochs. The software was programmed to normalize (scale) the data automatically between 1 and -1, and only works with numeric values. Non numeric values are treated as missing values which are replaced by the column mean. The software was set to use a square SOM grid of 10x10, which also determines the maximum number of clusters to be generated. The software generated only nine clusters from the dataset and increasing the training cycle did not improve the results. The Pitnett Neural Network software was trained using a starting learning rate of 0.9 and was set to train over 100 epochs, although the software stops training as soon as the maximum number of clusters have been generated. The software was programed to normalize (scale) the data automatically between 0 and 1. The network requires the number of clusters expected to be specified apriori. This number is used in conjunction with the number of input signals to determine the SOM grid size. This number was set to twenty. The software generated 18 clusters, and the mean and standard deviation of these clusters were computed. Table 4 presents the analyses of the 18 clusters, while figure 5 presents the chart of the clusters and rainfall average.

Table 4: Analysis of the 18 clusters from Pitnett

SNo	Cluster	Year	Month	Min Rainfall	Max Rainfall	Rainfall Average	Rainfall SD
1	1	1971 - 1978	3,4	199.8	350.1	270.6333313	75.47989326
2	2	1975 - 1994	8,9,10,11,12	58.5	235.8	172.5448282	53.81027918
3	4	1974 - 1996	4,5,6,7,8	156.6	312.4	240.9055549	44.76238685
4	5	1989 - 2008	6,7,8,9,10	146.1	377	232.3911102	54.96919143
5	6	1951 - 1972	4,5,6,7	237	399.8	310.8666662	60.64680031
6	7	1977 - 2008	8,10,11,12	0	71.5	18.40169495	23.75895108
7	8	1951 - 1976	8,9,10,11,12	0	152.4	42.46027352	44.08310212
8	9	1965 - 1998	8,9,10,11,12	0	195.8	91.58064508	58.98524609
9	10	1963 - 1999	8,9,10	255.5	415	314.3277808	50.33098414
10	11	1964 - 1960	9,10	266.4	324.8	298.0666707	29.36363096
11	12	1982 - 2008	1,2,3,4,5	7	194	46.51124946	54.05845832
12	14	1997, 2003	4,6	242.6	343.5	293.0999985	71.27636566
13	15	1951 - 1975	1,2,3,4,5,6,7	0	129.8	30.90303033	34.458354
14	16	2000 - 2008	4,5	129.9	206.3	173.5333337	24.30319025
15	17	1951 – 1972	7,8,9,10	51.1	256.7	167.2795464	65.28443524
16	18	1992 – 2008	5,6,7,8,9,10	9	215.6	111.080001	50.7121805
17	19	1952 – 1972	5,6,7,8	118.1	289.1	209.7083346	52.38667157
18	20	1951 – 1974	3,4,5,6,7	28.4	209.8	151.3382366	41.06354325

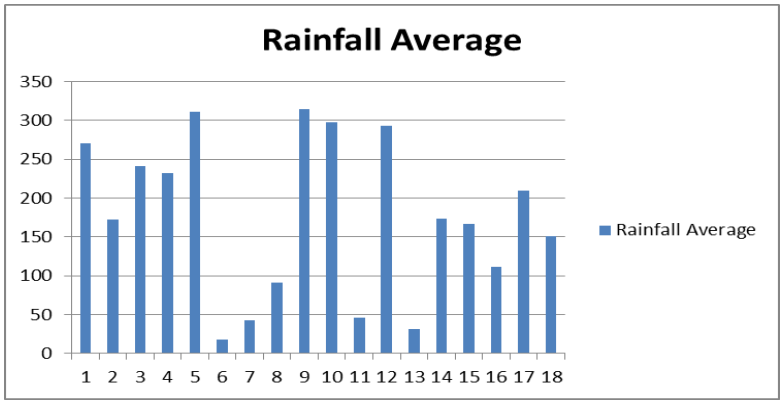


Figure 5: Chart of clusters and rainfall average from Pitnett

The NeuroXL Clusterizer network was trained using a starting learning rate of 0.9 and was trained over 100 epochs. The software can be set to either scale (normalize) the data or not. The number of clusters to be retrieved from the data can also be set and different values were used. The mean and standard deviation of these clusters were computed. From the results the clustering run which gave 20 clusters was selected, since it gave clusters which had the minimum standard deviation values. The analyses of the twenty clusters are presented in table 5 while figure 6 shows a chart of the clusters and rainfall average.

Table 5: Analysis of the twenty clusters from NeuroXL Clusterizer

Clusters	Years	Months	Min Rainfall	Max Rainfall	Rainfall Av	Rainfall SD
1	1970 -1994	1,2,3,8,9.11.12	13.6	34.8	21.91111111	5.623599541
2	1994 – 2008	1,2,3	0	15.1	3.02962963	4.993288943
3	1994 – 2008	11,12	0	9	1.077777778	2.664042827
4	1976 – 1993	1,2,4,11,12	0	10.2	1.523809524	2.809288559
5	1951 1974	1,2,3,6,7,8,9,11,12	19.3	55.6	34.57692308	10.59335595
6	1951 1979	2,3,4,5,7,8,11	56.4	102.5	76.54565217	14.24529094
7	1951 -1976	1,2,3,11,12	0	18.3	3.980645161	5.451119263
8	2003 -2008	1,3,4,11,12	18.7	36.6	28.47777778	5.92574421
9	1975 – 2006	1,2,3,4,7,8,911,12	37.5	61.1	49.15483871	6.728686782
10	1981 – 2008	2,3,4,5,7,8,11	60.2	84.1	69.84615385	6.213870341
11	1951 – 1982	3,4,5,7,8,9,10	103.1	150.6	127.79375	13.05282893
12	1967 – 2004	2,3,4,5,7,11	83.6	105.9	95.35769231	6.886053911
13	1977 – 2008	3,4,6,7,8,9,10,12	101.6	127.6	112.1684211	7.512881725
14	1982 -2008	4,5,6,8,9,10	124.7	159.5	142.35	10.57428408
15	1951 – 2008	2,3,4,5,6,7,8,9,10	147.8	186.8	168.3053333	10.61280325
16	1952 – 2006	2,4,5,6,7,8,9,10,11	183.3	223.8	203.1868852	10.76423825
17	1951 – 2008	4,5,6,7,8,9,10	224.8	266.8	246.3188679	11.926555
18	1957 – 2008	4,5,6,7,8,9,10	269.3	312.4	288.1685714	12.17958203
19	1960 – 2003	4,5,6,8,9,10	320.2	364.3	340.9583333	14.0279888
20	1951 – 1999	6,7,8,10	363.5	415	385.7625	19.3775966

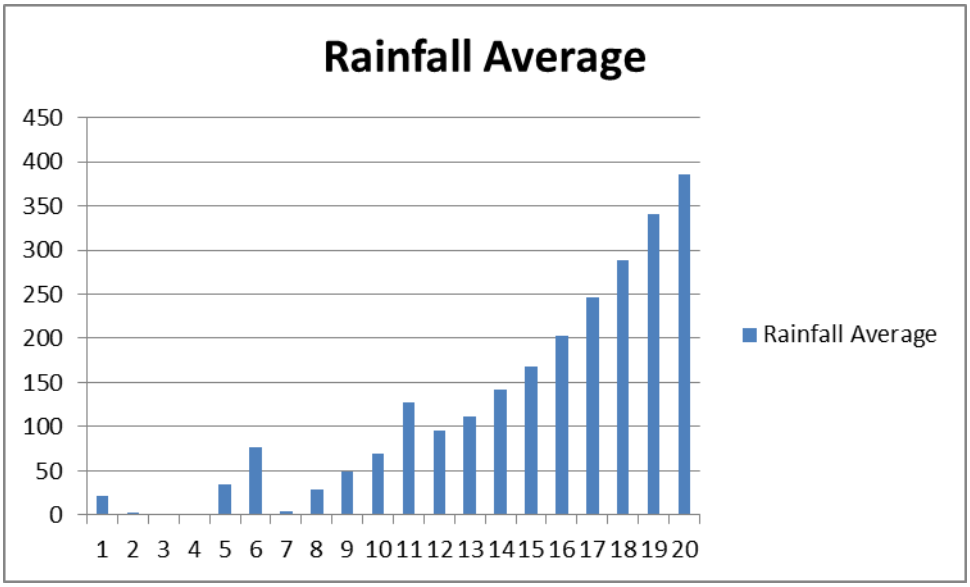


Figure 6: Chart of clusters and rainfall average from NeuroXL Clusterizer

Some known facts about the climatic condition of Ibadan [15] are: Ibadan has tropical wet and dry climate (by the Koppen Climate Classification), it has a lengthy wet season (between March to October) with a break in August, which divides the wet season into two. There are two peaks for rainfall in the wet season which is in June and September. The dry season is from November to February, while the driest month is December [16]. The weather condition over the country had been stable relatively between 1971 and 2000, when departures from the normal conditions could be noticed (Nigeria Climate Review, 2010). Some significant events that took place due to excessive rainfall levels were the floods disasters recorded in Ibadan in 1951, 1955, 1960, 1963, 1969, 1978, 1980, 2011 and 2012. Table 3 presents some of the known average weather data (temperature, rainfall, humidity and sunshine) for Ibadan, while figures 3 present a chart of the known average rainfall and precipitation for Ibadan.

Table 3: Climate data for Ibadan (Source: Wikipedia, 2013)

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Record high °C (°F)	37 (99)	39 (102)	38 (100)	38 (100)	35 (95)	33 (91)	31 (88)	31 (88)	36 (97)	33 (91)	34 (93)	35 (95)	39 (102)
Average high °C (°F)	33 (91)	34 (93)	34 (93)	33 (91)	32 (90)	29 (84)	28 (82)	27 (81)	29 (84)	30 (86)	32 (90)	33 (91)	31 (88)
Average low °C (°F)	21 (70)	22 (72)	23 (73)	23 (73)	22 (72)	22 (72)	21 (70)	21 (70)	22 (72)	22 (72)	22 (72)	21 (70)	22 (72)
Record low °C (°F)	10 (50)	12 (54)	18 (64)	18 (64)	18 (64)	18 (64)	16 (61)	16 (61)	17 (63)	18 (64)	14 (57)	14 (57)	10 (50)
Rainfall mm (inches)	8 (0.31)	23 (0.91)	76 (2.99)	125 (4.92)	145 (5.71)	163 (6.42)	132 (5.2)	74 (2.91)	170 (6.69)	152 (5.98)	43 (1.69)	10 (0.39)	1,121 (44.13)
Avg. precipitation days	1	2	5	9	11	12	12	10	15	12	4	1	94
% humidity	76	71	75	78	82	86	88	88	86	84	80	76	81
Mean monthly sunshine hours	170	198	170	170	170	141	85	57	85	141	198	198	1,783

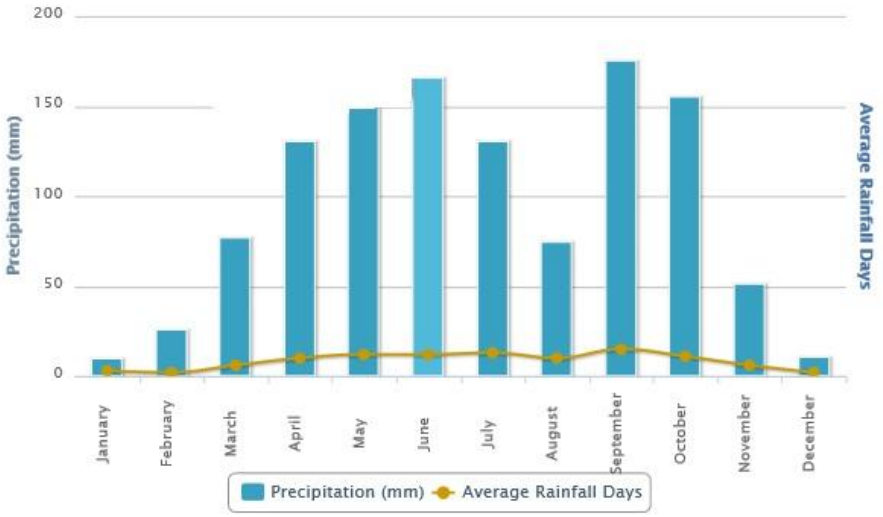


Figure 3: Average Rainfall and Precipitation for Ibadan (Source: http://www.worldweatheronline.com/Ibadan-weather-averages/Oyo/NG.aspx)

The clusters generated by the NeuroXL Clusterizer software gave the best performance because they had the lowest standard deviation values. It was also observed that the clusters represented rainfall intensity bands (groups) which corresponded to some known climatic and weather events in Ibadan between 1951 and 2008. For example the data records in cluster #20 (table 4) contains periods of peak rainfalls recorded in Ibadan, which also resulted in flooding disasters in the town. None of the clusters generated by the other software were able to drill down to this level.

Table 4: Cluster #20 of NeuroXL Clusterizer

Year	Month	Rainfall_mm
1951	7	369.1
1960	6	385.3
1963	8	363.5
1965	6	399.8
1965	7	369.3
1980	8	415
1993	7	377
1999	10	407.1

Also the data in cluster #2, #3, and #7 corresponds to the dry season period. Table 5 shows the data records in cluster #3. Cluster #2 represents dry season months which occurred in January, February and March between 1994 and 2008, while cluster #3 represents dry season months which occurred in November and December between 1994 and 2008. Cluster #7 contains data which represents the dry season months which occurred between November and March, 1951 to 1976.

Table 5: Cluster #3 of NeuroXL Clusterizer

Year	Month	Rainfall_mm
1994	12	0
1995	12	3.4
1996	11	0
1996	12	0
1999	12	0
2000	11	0
2000	12	0
2001	12	0
2002	12	0
2003	12	0
2004	11	0
2004	12	0
2005	11	0
2005	12	0
2006	12	0
2007	12	7
2008	11	0
2005	8	9

Weather prediction using CANFIS

The Neurosolutions CANFIS network used for the weather parameters prediction program had seven input neurons in its input layer and five output neurons in its output layer. The inputs were the year, month, wind-speed, radiation, minimum temperature, maximum temperature and rainfall while the output was the predicted wind-speed, minimum temperature, maximum temperature and rainfall intensity. The network was trained for 1000 epochs (cycles) using Levenberg Marquardt learning. Batch training was used and the training stopping criteria was on increasing MSE on the cross validation samples. The best network result selected used the Takagi–Sugeno–Kang (TSK) model with five (5) Bell shaped membership functions. Table 6 presents the network training report, Figure 4 presents the learning curve, table 7 presents the network testing report, figure 5 presents the sensitivity analysis result which shows the input variables contribution to the output result, and table 8 presents the predicted network output for rainfall in year 2010, while figure 6 presents this information in graphical form. Figure 7 shows the column chart (extracted from figure 6) for the predicted rainfall for 2010. It is noted that the rainfall pattern follows the monthly profile shown in figure 3. Figure 8 presents a chart of the historical weather parameter (windspeed, minimum and maximum temperature, and rainfall averages) for Ibadan for 2010 downloaded from www.Tutiempo.net/en (Weather station: 652080; Latitude: 7.43; Longitude: 3.9; Altitude: 228).

Table 6: Network training report

Best Networks	Training	Cross Validation
Epoch #	1000	46
Minimum MSE	0.014810273	0.046169199
Final MSE	0.014810273	0.078883972

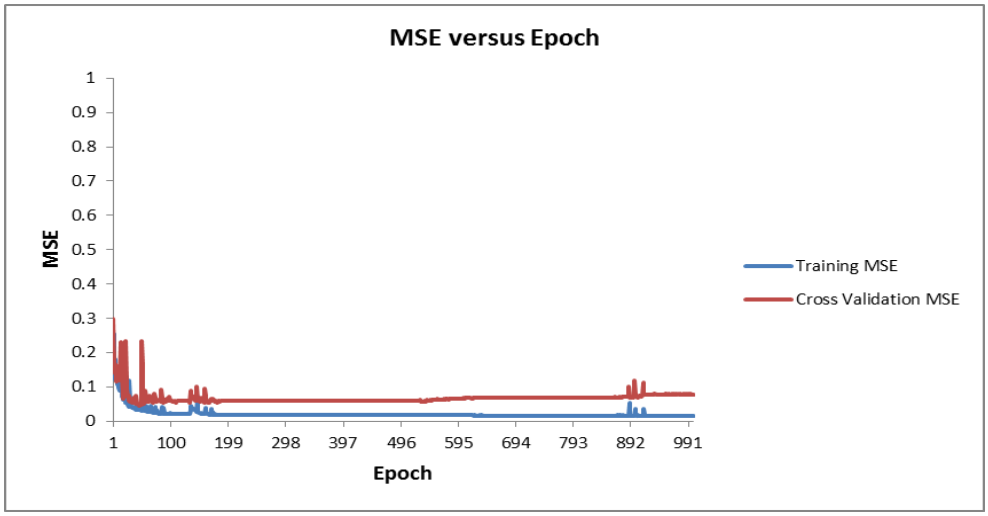


Figure 4: Network learning curve

Table 7: Network testing report

Performance	Wind	Radiation	MinTemp	MaxTemp	Rainfall
MSE	619.5845617	5.120205621	12.22381097	3.838489961	11595.28926
NMSE	1.972268028	1.37819717	20.5391322	1.108895982	1.07732946
R	0.750344634	0.519604583	-0.299922647	0.543774968	0.712360315

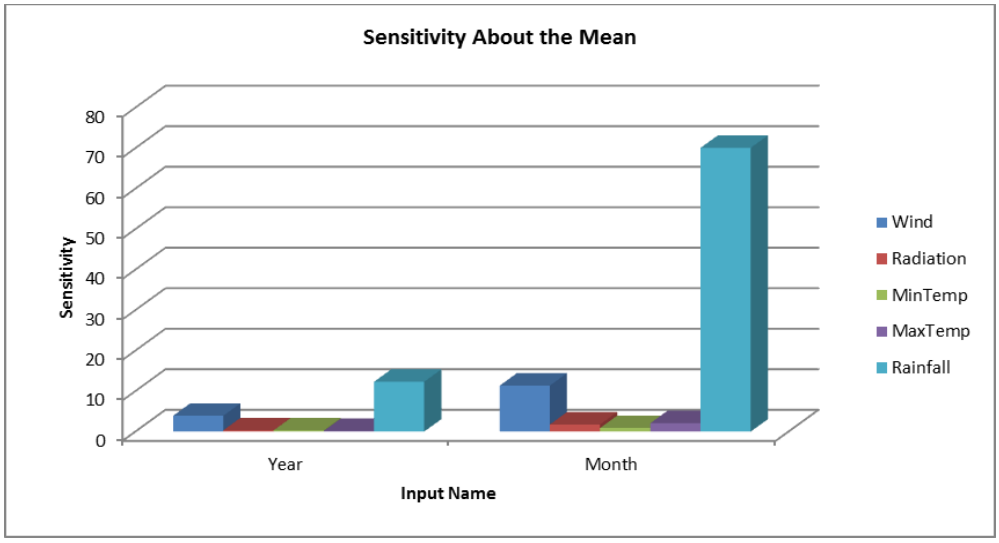


Figure 5: Sensitivity analysis result

Table 8: Predicted network output

Year	Month	Wind	Radiation	MinTemp	MaxTemp	Rainfall
2010	1	149.4943	9.768377	24.98293	33.76771	18.19179
2010	2	161.5787	10.2203	25.55385	33.27231	38.88475
2010	3	206.1897	10.55984	26.82792	32.44958	67.3745
2010	4	173.7567	10.53395	26.93608	31.41716	82.03233
2010	5	156.6988	10.58878	28.79529	29.56042	110.219
2010	6	154.0258	10.36948	30.9407	27.44971	134.1238
2010	7	165.9713	4.785612	30.88352	25.95928	15.42722
2010	8	148.2829	10.60675	25.2678	29.87004	57.98322
2010	9	145.3576	10.31745	21.5661	30.43777	210.909
2010	10	138.3491	8.763394	21.98761	30.14349	174.9529
2010	11	128.6779	13.77576	24.64287	29.83381	-10.7456
2010	12	122.3363	15.84211	25.05434	30.11157	-40.6949

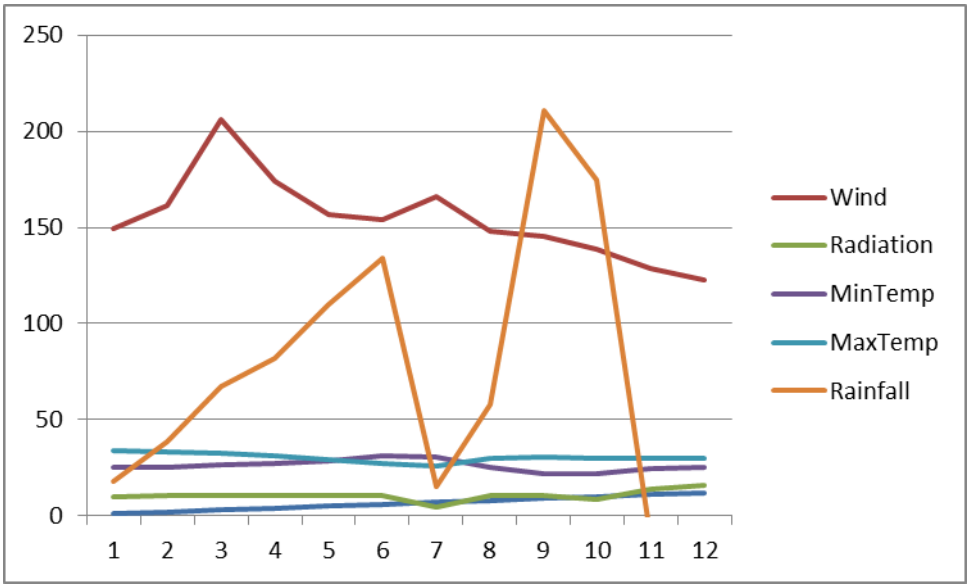


Figure 6: Predicted network output

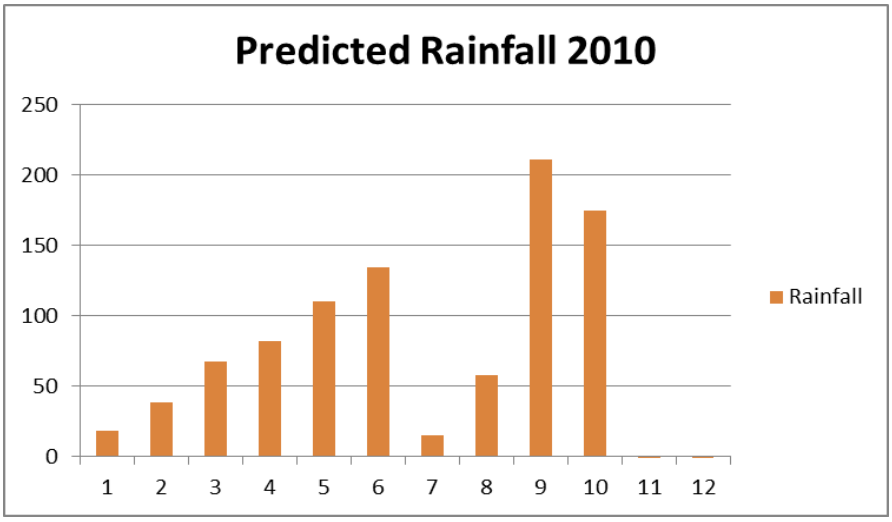


Figure 7: Predicted Rainfall for 2010

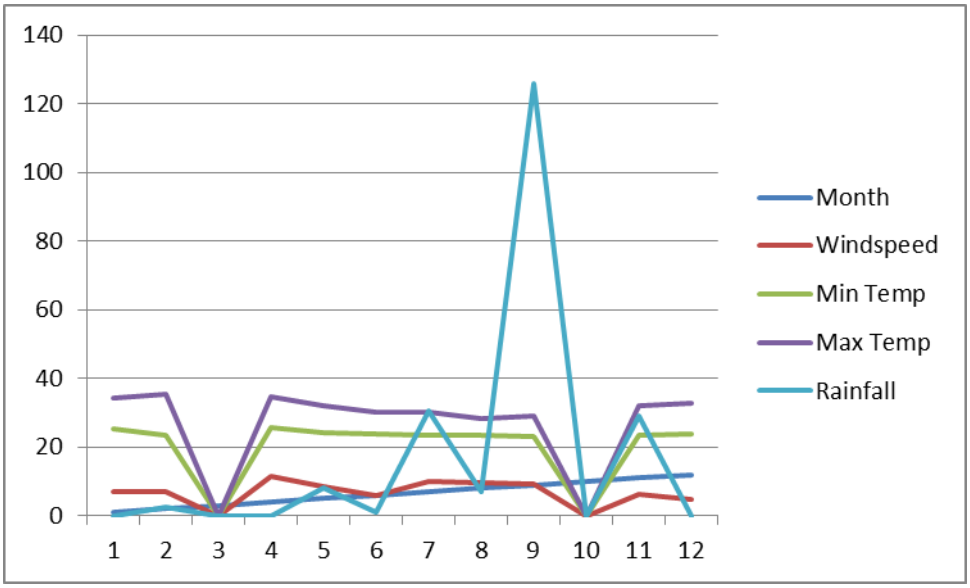


Figure 8: Ibadan average rainfall 2010 (www.Tutiempo.net/en)

5. CONCLUSION

This work presents the use of soft computing techniques (SOM and CANFIS) for knowledge discovery and prediction of rainfall and weather parameters. Clusters were generated from the rainfall data which are representative of some of the known climatic events in the town. The prediction of weather parameters carried out with the CANFIS technique shows that it can be used for long range weather forecasts (in this study of 12 months ahead), on a time scale within the current weather pattern (of within 20 to 30 years interval) of the data set used. A shift in the weather pattern forecasted using historical data may be indicative of climate change activities.

ACKNOWLEDGEMENT
The author acknowledges the contributions of Mrs A. K. Ojo and Mr Olaiya who collected the data used for this work.

REFERENCES

[1] Abraham A. and Baikunth N., (2000), "Designing Optimal Neuro-Fuzzy Systems for Intelligent Contro"l, In Proceedings of The Sixth International Conference on Control, Automation, Robotics and Vision, (ICARCV 2000 - Singapore),(CD ROM Proceeding), Wang J.L. (Editor)

[2] Abraham A., (2004), Adaptation of Fuzzy Inference System Using Neural Learning, Computer Science Department, Oklahoma State University, USA, ajith.abraham@ieee.org, http://ajith.softcomputing.net

[3] Casas D. M, Gonzalez A.T, Rodrígue J. E. A., Pet J. V., 2009, "Using Data-Mining for Short-Term Rainfall Forecasting", *Notes in Computer Science*, Volume 5518, 487-490

[4] Chang C., Ding Z., (2004), "Categorical data visualization and clustering using subjective factors", Data & Knowledge Engineering, Published by Elsevier B.V.

[5] Chen N. and Marques N. C., (2005), "An Extension of Self-Organizing Maps to Categorical Data", Proceedings of the 12th Portuguese conference on progress in Artificial Intelligence , pp 304 - 313, (Sringer-Verlag Berlin, Heidelberg ©2005)

[6] Elia G. P., 2009, "A Decision Tree for Weather Prediction", Universitatea Petrol-Gaze din Ploiesti, Bd. Bucuresti 39, Ploiesti, Catedra de Informatică, Vol. LXI, No. 1

[7] Fairbridge R. W., 2007, "Climate" *Microsoft® Student 2008 [DVD]*, Redmond, WA: Microsoft Corporation, 2007.

[8] Folorunsho O., Adeyemo A. B., (2012), "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", I.J. Information Engineering and Electronic Business, 2012, 1, 51-59, Published Online February 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijieeb.2012.01.07, Copyright © 2012 MECS

[9] Heydari M and Talaee P. H., (2011), "Prediction of flow through rockfill dams using a neuro-fuzzy computing technique", The Journal of Mathematics and Computer Science (TJMCS), Vol.2, No.3 (2011), pp 515-528, Available online at http://www.TJMCS.com

[10] Kaski S., (1997), "Data exploration using self-organizing maps", Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82, Espoo 1997, 57 pp. Published by the Finnish Academy of Technology. ISBN 952-5148-13-0. ISSN 1238-9803. UDC 681.327.12:159.95:519.2

[11] Kohonen T, (1999), "The Self-Organizing Map (SOM)", Helsinki University of Technology, Laboratory of Computer and Information Science, Neural Networks Research Centre, Quinquennial Report (1994-1998), (Downloaded from http://www.cis.hut.fi/research/reports/quinquennial/ January 2006).

[12] Nigeria Climate Review, 2010, Nigerian Meteorological Agency, www.nimetng.org

[13] Pelczer I. J. and Cisneros H. L., (2008), "Identification of rainfall patterns over the Valley of Mexico", 11th International Conference on Urban Drainage, Edinburgh, Scotland, UK, 2008

[14] Statsoft Electronic Statistics Textbook, (2002), Copyright, 1984-2003, (http://www.statsoftinc.com/txtbook/glosd.html# Data Mining), Downloaded June 2002.

[15] Wikipedia, 2011, "Climate change" *From Wikipedia - the free encyclopedia*, retrieved from http://en.wikipedia.org/wiki/Climate_change in August 2011

[16] Wikipedia, (2013), "Rain" *From Wikipedia - the free encyclopedia*, retrieved from http://en.wikipedia.org/wiki/Rain in February 2013

[17] Zanchettin C., Minku L. L., Ludermir T. B., (2010), "Design of Experiments in Neuro-Fuzzy Systems", International Journal of Computational Intelligence and Applications, Vol. 9, No. 2 (2010) 137–152

[18] Zengyou He, Xiaofe I Fe, Shengchun Deng, (2003), "Clustering Mixed Categorical and Numeric Data", Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, P. R. China

The African Journal of Computing & ICTs.

Call for Papers

INTRODUCTION

The African Journal of Computing and ICT was established by the Nigeria Computer section of the Institute of Electrical and Electronics Engineers (IEEE) Inc, New York, USA. Afr. J of Comp and ICT publishes both online and print editions in April, August, December. **The Journal solicits original research articles from interested authors for her upcoming editions.** The goal of the journal is to contribute to rapid IT developments in Nigeria, Africa and the world through the publication of topical and relevant papers from theoretic/academic and practical / industrial points of view. Submitted papers should contain substantial ideas which have the potential of aiding an invention or a technological breakthrough for general and specialized use .

SCOPE OF ARTICLES

Papers are welcome from pure/applied computing and ICT, and allied areas including but not limited to software engineering, computer communication/networks, human computer interaction, data mining, discrete/data structures, design & analysis of algorithms, web security, systems architecture, ubiquitous computing, systems design, information systems, bioinformatics, information management, information science and e-systems (e-health, e-learning, e-government).

The journal considers original research papers, case reports and review papers. Authors must clearly justify the importance of their submission to the readership of the journal. Prospective authors need not be a member of the (Computer Chapter) IEEE Nigeria Section before his/ her paper is published. **All papers are peer refereed.** The AJCICT Publication guideline can be found at www.ajocict.net.

PAGINATION FEES AND INFORMATION FOR SUBSCRIBERS AND ADVERTISERS

Authors of accepted papers will be charged a pagination fee of $150 (N20000). The subscription fee for yearly volumes of the journal to institutions and organizations is US 200 (N30000 Naira) excluding postage. Copies of individual volumes can be obtained at a cost of $50 (N8000) Correspondence on paper publications, book review, subscription and advertisement should be directed to the Managing Editor, African Journal of ICT. Request for information pertaining to the Computer Chapter of the IEEE Nigeria Section should be sent to the Chair of the Chapter, Prof. 'Dele Oluwade bamideleoluwade@computer.org.

General information on the aim and objective as well as membership of IEEE, IEEE Computer Society and the IEEE Nigeria Section can be found on the following websites respectively: http://www.ieee.org ; http:// computer.org ; http://www.ieee.org/nigeria.