A Comparative Survey of Machine Learning and Meta-Heuristic Optimization Algorithms for Sustainable and Smart Healthcare

Hina Firdaus[1], Syed Imtiyaz Hassan[2], Harleen Kaur[3]
School of Engineering Sciences and Technology,
Jamia Hamdard, New Delhi-110062, India
Email: hinafirdaus95@gmail.com[1],
s.imtiyaz@gmail.com[2],
harleen.unu@gmail.com[3]

ABSTRACT

The healthcare industry uses machine learning for diagnosis, prognosis, and surveillance. Health Catalyst believes machine learning (ML) is the life-saving technology that will transform healthcare. This technology challenges the traditional reactive approach to healthcare. In fact, it's the exact opposite; it is the predictive, proactive, and preventive life-saving qualities that make it a critically essential capability in every health system. The third goal of good health and well-being by the United Nations Sustainable Development Goals of 2030 has opened research opportunities across the globe. In this literature review, an intensive study of twenty-one papers published between 2014 and 2017 is carried out. The study has raised various questions regarding the best machine learning technique on different disease dataset and solutions to overcome the problem of optimizing the feature selection for better performance. A comparison of the various algorithms is presented in tabular form based on two categories, namely the use of ML and also the use of ML with meta-heuristic algorithms in disease dataset. A model proposed for future work, which uses meta-heuristic algorithms for feature selections, appears to be better for any dataset. Machine learning in medicine has recently made headlines. Google has developed a machine learning algorithm to help identify cancerous tumors on mammograms. Stanford University is using a deep learning algorithm to identify skin cancer. Deep machine learning algorithm has been used in the literature to diagnose diabetic retinopathy in retinal images. It is clear that machine learning puts another arrow in the quiver of clinical decision making. A dataset is prepared from the surveyed research papers and this is compared for better visualization of algorithms. To make a comparative analysis of the retrieved dataset and choosing a right algorithm for any disease diagnosis, a graph is plotted using ML libraries of python language. The graph shows that the use of Support Vector Machine (SVM) with the optimization algorithms, like EPSO_ABC, and Artificial Neural Networks (ANN), give 100% accuracy. This paper provides an enhanced description of future work in sustainable healthcare. The use of meta-heuristic algorithms with ML techniques will provide better and faster results.

Keywords: *Smart Healthcare, Sustainable healthcare, Machine learning, meta-heuristic optimization algorithm, Nature inspired algorithm*

African Journal of Computing & ICT Reference Format:

Hina Firdaus, Syed Imtiyaz Hassan and Harleen Kaur (2018),
A Comparative Survey of Machine Learning and Meta-Heuristic
Optimization Algorithms for Sustainable and Smart Healthcare,
Afr. J. Comp. & ICT, Vol.11, No.4, pp. 1 - 17.

I. INTRODUCTION

During the second half of the twentieth century, the Artificial Intelligence came into existence to capture human brains in limited domains. This is a result of computer revolution whereby systems developed behave intellectually, reason rationally and have the ability to effectively interpret the environment in real time. Artificial intelligence outperforms every scientist or mathematician in the way of thinking. It has made it possible to simulate complex activities that need professional expertise and make a machine to act like a human. Thus, one of the popular subfields of Artificial Intelligence became Machine Learning (ML). Humans can expand their knowledge to adapt to the changing environment and in doing that, they must "learn". Learning can be simply defined as the acquisition of knowledge or skills through study, experience, or via a process of teaching. Although learning is an easy task for most of the people, to acquire new knowledge or skills from data is too hard and complicated for machines. Moreover, the intelligence level of a machine is directly relevant to its learning capability. The learning of algorithm helped the machine to learn a task from its experience. When we say that the machine learns, we mean that the machine is able to make predictions from examples of desired behavior or past observations and information. A more formal definition of machine learning was given by Tom Mitchell [1]: " A computer program is said to learn from experience E regarding the class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". The ML algorithms run on the different dataset, which is the collection of experiences, by performing tasks in the previous experiences. The automatic ML algorithms learn from the previous incidence to produce a new outcome or make predictions for future benefits.

The applications of machine learning bring a revolution in prediction and empower the algorithms by applying techniques of data mining and big data analytics. Machine learning integrate with various interdisciplinary fields like healthcare, games industry, home assistant, self-driving cars etc. [38] focuses on the necessity of machine learning in healthcare which will help in feature selection and in the processing of large and complex datasets. An analysis of the clinical disease datasets will help the physician plan and provide sustainability in the healthcare industry, leading to better outcomes, accurate examination, lower costs of care, and increased patient satisfaction.

In the year 2015, the United Nations took a step forward towards Sustainable Development Goals (SDG's) in 2030 and good health and well-being was placed as the third goal among the list of 17 goals and 169 targets [28]. India is one of the193 member states. These goals are regulated by NITI Aayog (National Institution for Transforming India) which has a 5 year health plan to understand the focused areas in health sector. A target to end epidemics caused by waterborne diseases, airborne diseases, and communicable diseases (like tuberculosis, malaria, dengue etc) has been set against 2030 under UN SDG's goals [28]. As the Father of Indian nation Mahatma Gandhi once said, [39] "Health is the real wealth and not pieces of gold and silver" [39]. The authors of this present paper believe that this saying is appropriate.

I.II MACHINE LEARNING TECHNIQUES

The pioneer of Artificial Intelligence [33], Arthur Samuel, who coined the term machine learning, remarked thar "machine learning, as a way of programming, gives the computer the ability to learn…."[33]. Machine learning is categorized into three types, namely supervised learning, unsupervised learning, and reinforcement learning.

I.II.I Supervised Learning

The supervised learning [34] assumes the set of the problems using labelled training data. There are two groups in supervised learning, that is Classification and Regression (Figure 1).

Some popular applications of supervised learning are image classification, identity fraud detection, diagnostics (classification algorithm), weather forecasting, population growth prediction, market forecasting (regression algorithm).

I.II.II Unsupervised Learning

Unsupervised learning is not classified or labeled with categorization and classification of data [35]. That is, it uses unlabeled data (Figure 2).

Some real-world applications of unsupervised learning are NASA remote sensing, mini UAV, nano camera fabrication technology, text summarization, ontology extraction, targeted marketing, fracture elicitation, a recommender system, etc [1].

I.II.III Reinforcement Learning

Reinforcement learning is part of human behavioral psychology, which uses an agent to act according to the situation towards maximizing the rewards. [5] The reinforcement learning goals work by setting explicit goals; it works by sensing the environment (figure 3).

Applications of reinforcement learning are vast, and it is used mostly in game development, manufacturing, inventory management, delivery management, power system, finance sector, etc.

I.III NATURE INSPIRED ALOGIRTHM

The meta-heuristic approach for the optimization of a problem can be done using the algorithms inspired from the elements of nature like animals, flowers, plants, micro-organisms, environment, human, and so on. It uses randomization and local search, called nature inspired algorithms. These meta-heuristics algorithms are classified into two broad classes: a) population based genetic algorithms, and b) trajectory based algorithms [40]. There are dozens of nature-inspired algorithms, and selection of the best algorithm for a desired problem is one of the toughest challenges. The meta-heuristic and heuristic algorithms are simulated annealing [41], Genetic algorithm [42], Ant and Bee algorithm [43], Particle swarm optimization [44], firefly algorithm [45], Cuckoo search [46] and Bat algorithm [32] (see figure 4).

The nature inspired algorithm has various applications in automobile industries, healthcare, energy conservation, gaming and many other fields of endeavour.

II. APPLICATION OF MACHINE LEARNING AND META-HEURISTIC ALGORITHM

The literature from 21 sources is reviewed in two sections. The first section is about machine learning algorithms on the different disease dataset. The second section is where the optimization meta-heuristic algorithms are used with the machine learning algorithm.

II.I Literature Review of Machine Learning Techniques on Disease Dataset

The work of Zheng et. al. [6] is on the discovery of vital genotype and phenotype association with Type 2 Diabetes mellitus (T2DM) [6]. The work is on diagnosis, detection and medication cases of T2DM Electronic Health Records (EHR) data using feature engineering and machine learning. Different learning models, such as Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), K- Nearest Neighbors (KNN) [37], Decision Tree- J48 (DT J48) and Random Forest (RF) [3], were used for better comparison. WEKA, an open source software tool, was used for training model. A 4-fold cross-validation was conducted to get the average performance and standard deviation. J48, SVM, RF have the highest performance index, yielding over 95% of accuracy, sensitivity and specificity. In comparison, LR has the highest accuracy of 99% followed with SVM and RF 98%, while the three models KNN, NB, and LR are vulnerable, for instance, to noise in datasets for feature selection.

[8]: This work is on the detection of diabetic nephropathy among the type 2 diabetic patients using SVM. [2] According to the proposed work by Huang, G. M., et al.; early detection of real malfunction is possible using a decision tree based model integrating it with genotype and clinical data of patients implement from a various source of type 2 diabetic patient. 5—fold cross-validation approach used where the data classified according to genders using the algorithms like Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) compared with their attributes for the better result. The training models were compared using WEKA and LibSVM tools. There are four stages in the process: 1. Comparing performance between individuals and a combination of clinical features, 2. Comparing performance between individual and combination of genetic features, 3. Evaluation of performance by integration of clinical and genetic features, 4. Gender-based performance using a decision tree classification for the diabetic nephropathy training datasets. Decision tree gives excellent results where accuracy, specificity, and sensitivity are 85.27%, 83.32% and 85.24%. In the conclusion, the individual use of genetic and clinical features gives a disappointing result while combining genetic and clinical features gives a better result.

[10]: This paper shows the process of embedding the machine learning algorithms with data mining pipelining

in order to extract the knowledge from the vast pool of information. This data mining profiling comprises centre profiling, predictive model targeting, predictive model construction, and model validation. Classification models are Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF). The analysis was done in four stages of data mining pipelining. Method of centre profiling is for selection biasing and is for defined variable performance. An important conclusion derived from the medical point of view is that the period of diabetes disease and the BMI value are the main risk factor for retinopathy and neuropathy while hypertension is the main risk factor for retinopathy and nephropathy. The authors conclude that the data mining technology can integrate with machine learning for better outcome in prognosis and treatment of any disease.

[14]: In the work, an automatic prediction system was developed for cardiovascular and cerebrovascular (strokes etc) events using Heart Rate Variability analysis (HRV). These are high-risk subjects for the patient above 55 years of age in the developed country. A 10-fold cross-validation with several data mining techniques were used for prediction based on various HRV features with Naïve Bayes classifier (NB) and Decision Trees C4.5 (DT C4.5). Others include Random Forest (RF), boosting meta-learning approach (i.e. AdaboostM1 (AB)), SVM and Multi-layer Perceptron Neural Network (MLPNN). In the result, C4.5 and AB achieved the highest performance with chi-square feature selection parameters. RF outperformed by achieving the accuracy of 85.7%, a sensitivity of 71.4%, and a specificity of 87.8%.

[16]: This work is on an anti-diabetic drug failure whereby a prediction system is introduced to keep the point of the exponential increase in type 2 diabetic patient. The method used in the paper is Support Vector Machine (SVM), which is one of the best methods to train large-scale medical dataset. To enhance the effectiveness and ensemble effective of SVM algorithm, E3-SVM method was introduced. The 10-fold cross-validation used is modeled via MATLAB and LibSVM tool. The major shortcomings of the system is with respect to the choice of appropriate k values i.e., data selection parameters. Prediction of k values affects the performance of data at the various levels of data point selection. The accuracy achieved is 80% by SVM for the anti-diabetic drug failure prediction.

[17]: In the work, the authors showed sincere concern on the risk of diabetic neuropathy. The nervous system gets affected when diabetes spread all over the body causing cardiac arrest. Multi-scale Allan vector was applied to determine Heart Rate Variability (HRV), such that the features from ECG recordings were used for machine learning methods and automated detection. The paper introduced a new Graph-Based Machine Learning System (GBMLS) for effective diagnosis of the diabetic neuropathy. The GBLMS method combined with Multi-scale Allen Vector (MAV) gives the best result with high sensitivity of 98% and specificity of 89%, which outperforms other classifiers like Random Forest with 83% and 92% in sensitivity and specificity respectively.

[18]: In the paper, a comparative analysis of healthy patient and Asthma patient is done using an alternative equipment to get the feature vector of Asthma patient under GINA 2010 guidelines. A predictive model was proposed using three machine learning techniques, namely Random forest (RF), AdaBoost Random Forest (AB-RF), and Multi-layer Perceptron Neural Network (MLPNN). The three statistical parameters considered are specificity, sensitivity, and accuracy. Also to avoid over-fitting problem, Leave-One-Out (LOO) validation technique was applied to the training and testing dataset. The best results of the statistical parameter on RF and AB-RF classifier with FEV and FVC1 feature vector are 95.0%, 90.0%, and 92.5% for specificity, sensitivity, and accuracy. The work is a sustainable option of diagnosis in the absence of X-rays or CT- scans for the physicians, as the changed equipment is cheap and available.

[19]: This work is on a surveillance system which is introduced to monitor the effect of dengue hemorrhagic fever (DHF) and Aedes aegypti mosquito infection rate, based on climate and geographical area using the Support Vector Machine (SVM). The nine major areas considered for a dengue epidemic rate were selected within the year 2007—2013. These areas are: temperature, rainfall, humidity, wind speed, Aedes aegypti larvae infection rate, a male mosquito infection rate, a female mosquito infection rate, population density, and morbidity rate. The method takes place in three stages. For the model construction, classification algorithms (like K-Nearest Neighbor (KNN), Decision Tree (DT), Neural Networks (NN), Support Vector Machine (SVM)) were used with different kernels. The 10-fold cross-validation technique was employed to validate the result for SVM effectiveness using the accuracy, sensitivity, and specificity as overall performance metrics. The SVM RBF

kernel defined by the predictive model in two parameters C (parameter of regularization) and σ2 (kernel function parameter) gives the value of 8 and 0.1250. Among all the kernels used in the paper, only SVM-RBF kernel shows better performance with 96.296% accuracy, sensitivity and specificity of 87.47%, which is better among techniques such as SVM-L, SVM-P, KNN, DT, and NN.

[20]: A novel prediction system was developed for early outbreak detection of Influenza-Like Illness (ILI), by detecting Tweets from Twitter micro blogging site. In the study, the authors proved that the use of Naïve Bayes and SVM algorithm will not only give the best results, and that the use of Logistics Regressions and SMO (polynomial kernel and sequential minimal optimization algorithm) also perform well. Assessment was performed with 10-fold cross-validation techniques. The big unstructured dataset is able to analyze with the Map Reduce to generate a meaningful result from the tweets. The Medtex text analysis software was used to extract the features from the text in Twitter messages. Several standard features, like word tokens, stems, and n-grams in presence Twitter username, hashtags, URLs and emoticons, were taken into consideration. However, the classifiers with F-measures of the balanced sheet with cross-validation have higher performances stats with the unbalanced sheet of unseen data validation.

[21]: In the article, an automated diagnosis for Coronary Artery Disease (CAD) which is responsible for cardiac arrests, was discussed. The diagnosis uses novel method like Tunable Q-Wavelet Transform (TQWT) and the heart rate signals from the raw ECG (Electrocardiogram). These features apply to classification algorithms like Least Squared Support Vector Machine (LS-SVM) with different kernels, in which only Morlet Kernel function uses a 3-fold cross validation. It gives accuracy of 99.7%, sensitivity of 99.6%, specificity of 99.8%, and Matthews correlation coefficient of 0.9956, for Q varying between 24 and 30, which gives 100% efficiency. From the work, it is clear that the outcome can also be applied to the diagnosis of heart disease, diabetes, eye disease, and neural diseases.

[22]: In the work, prediction of heart disease is found from the BagMOOV novel ensemble method. This framework is based on enhancing bagging approach for the multi-purpose weighted voting scheme for the prediction and analysis. The assessment was done using 10-fold cross-validation and ANOVA (Analysis of Variance) methods. Along with the BagMOOV approach, a decision support system (DSS) was introduced for heart disease diagnosis using novel ensemble method. In the work, five different datasets used are SPECT, SPECTF, Heart disease and Statlog, and Eric datasets with different attributes in the class labels denoted as class 0 and 1 for distinguishing features. The BagMOOV ensemble algorithm gives accurate and efficient results in all the datasets compared to the other state-of-art techniques. Supporting the work, the decision support system uses 138 patients samples with the accuracy of 84.78%, 73.47% sensitivity, 91.01% specificity, and 81.30% F-measure achieved.

[23]: This work dwells on the study of a psychiatry solution of mood disorder, a psychological behavioral trait of the human being, using machine learning algorithms. To perform training, the three algorithms of machine learning, namely Least Average Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM), and Relevance Vector Machine (RVM), were designed using MATLAB for prediction of possibility in suicide attempter. The Leave-One-Out (LOO) cross-validation technique was used for assessment of training and testing datasets. Among the three algorithms, RVM had the best performance rate because 103 out of 144 patients guessed correct that they are likely to resort to suicide or not, with the accuracy of 72%, the sensitivity of 72.1%, specificity of 71.3%, and chi-squared is $p<0.0001$. Using the confusion matrix validation, the accuracy is 71.4% for RVM.

[24]: In the paper, a survivability kit was developed for prediction of some common epidemic diseases like Colds-Flu Gripe, Dengue, Malaria, Cholera, Leptospirosis, Chikungunya, Chickenpox, and Diarrhea. To perform the study, data were collected from the hospital of Nasik, Maharashtra (India) from 316 patients. Algorithms like Decision Tree J48(DT J48), Multi-layer Perceptron Neural Network (MLPNN), Support Vector Machine (SMO), K-Nearest Neighbor (LWL), and Naïve Bayes (NB) were assessed by 10-fold cross-validation and were implemented in WEKA software.

[15]: The paper studied breast cancer risk, its prediction and diagnosis, with the use of four machine learning algorithms viz. Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), and K-Nearest Neighbor (KNN). The original Wisconsin breast cancer datasets from UCI repository were trained using WEKA tool. For assessment, a 10-fold cross-validation technique was used. The effectiveness of the classifier was determined

using error finding methods such as Kappa statistics, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE). SVM outperforms the other algorithms used with 97.13% accuracy, and also with lowest error rate of 0.02.

[25]: The work predicts chronic illness from the hospital database in China between the year 2013 and 2015 using Machine learning algorithms. The datasets used are vividly classified as Structured data (S-data), Text data (T-data), as well as Structured and Text data (S&T-data). Prediction of cerebral infraction disease in S-data used Naïve Bayes (NB), Decision Tree (DT). There is a 50% accuracy gained from K-Nearest Neighbor (K-NN). The CNN-based unimodal disease risk prediction (CNN-UDRP) used for T-data shows accuracy of 94.2% and recall of 98.08%. The CNN based multimodal disease risk prediction (CNN-MDRP) for S&T-data shows accuracy and recall of 94.8% and 99.923% respectively. C++ language was used for processing machine learning and deep learning algorithm. The 10-fold cross-validation was applied on the training and testing datasets. It was proved from the experiment that the novel approach of CNN-MDRP outperforms CNN-UDRP. The structure of disease dataset have a major impact on deciding the accuracy of the model.

II.II Literature review of Machine learning techniques with meta-heuristic algorithms on disease dataset

[7]: This work studied the Support Vector Machine (SVM) algorithm with the Fruit-fly Optimization Algorithm (FOA) in various medical datasets such as Wisconsin breast cancer dataset, Pima Indians diabetes dataset, Parkinson's dataset, and thyroid disease diagnosis, got from UCI repository. The ML SVM technique is hybridized with Particle Swarm Optimization Algorithm-based SVM (PSO-SVM), Genetic Algorithm-based SVM (GA-SVM), Bacterial Forging Optimization-based SVM (BFO-SVM), and Grid Search Technique-based SVM (Grid-SVM), and implemented with tools like MATLAB and LibSVM. 10-fold cross-validation technique was used. The SVM-FOA gives the highest accuracy as 96.9%, 77.46%, 77.46%, and 96.38% in Wisconsin dataset, Pima dataset, Parkinson dataset, and thyroid dataset respectively.

[9]: In the work, diagnosis of diabetes was conducted using K-Means clustering algorithm based on the outlier detection, followed with Genetic Algorithm (GA) for feature selection with Support Vector Machine (SVM) as a classifier to classify the dataset of Pima Indians Diabetes from UCI repository. The 10-fold cross-validation technique was used for assessment. SVM with K-Means and Genetic Algorithm model gives an accuracy of 98.82%. The work arrived at the following conclusion: (a) the minimum and maximum classification accuracy are 98.43% and 99.21% respectively in SVM and average accuracy is 98.79%. (b) The K-means outlier detection percentage is 33.46%. Out of 768 instances, 511 samples were selected and 257 samples included as outlier, (c) The minimum number of the attribute selected is 3 and maximum is 6, (b) attributes such as Pregnancies, PG Concentration, and Age are difficult attributes in the dataset. Overall, findings in the paper indicate a 2.08% increment in accuracy of SVM classification model over the changed K-means algorithm.

[26]: This work is on predicting malaria transmission using Support Vector Machine and Firefly Algorithm (SVM-FFA) to show which of the two has a better performance in prediction. The work relates to malaria epidemy which is widespread in the state of Rajasthan leading to death and illness; lack of primary healthcare makes the situation worse. The four model systems designed were SVM-FFA, Auto-Regressive Moving Average (ARMA), Artificial Neural Networks (ANN) and SVM. The models were developed using LibSVM library in MATLAB, and the ARMA model was defined using the IBM-SPSS software for better insight in-depth prediction. The R2 statistic and NMSE parameter were used for training and testing such that use of R2 gave accurate result for SVM-FFA in predicting malaria incidences. In conclusion, it was established that the novel approach of SVM-FFA is best among all the models.

[13]: The paper dwells on classification of UCI’s different disease dataset, where SVM is hybridized with Endocrine-Based Particle Swarm Optimization (EPSO) and Artificial Bee Colony (ABC) algorithm of the Evolutionary Algorithm. The different dataset have diseases like Ecoli, Breast, heart, Parkinsons, CTGs, SPECT from UCI repository. Dev C++ language along with LibSVM are the tools used for development of the model. The 5-fold cross validation used 80-20 rule for training and testing sets. EPSO_ABC-SVM technique in Parkinson's disease, compared with other techniques and dataset, gives 100% accuracy with 4.0/22 features. This paper has a promise with respect to application in diagnosis of clinical datasets. A more intensive research

with other unsupervised algorithms, combined with the EPSO_ABC technique, is expected for better performance measures and accuracy in different disease dataset.

[47]: In the paper, a heart disease diagnosis system based on a novel approach of Interval Type-2 Fuzzy Logic System (IT2FLS) was presented. To enhance the novel approach, the rough datasets use chaos firefly algorithm to optimize the attributes and to reduce the computational burden on IT2FLS. This approach compared with other ML techniques like SVM, NB, ANN. The model is systematic and executes in three stages: Normalization, Chaos Firefly Algorithm and Rough Sets based Attribute Reduction (CFARS-AR). A building of IT2FLS with two popular datasets of UCI repository (like Heart disease and SPECTF datasets) was used. Two different tools were used for processing, in the stage of CFARS-AR MATLAB was used, and for the rest ML algorithm, WEKA was used. Performance of the techniques was tested with metrics like accuracy, sensitivity, and specificity. The algorithm in the paper gives good result in heart disease dataset which executes 88.3% accuracy with CFARS-AR and 87.2% for SPECTF datasets. In conclusion, this paper shows that the attribute reduction concept is interesting, although it has to prove whether this method will work on unstructured data and the huge dataset with more attributes or not. This literature has a defect in the performance rate because training time taken by chaos firefly algorithm and IT2FLS is very slow.

III. RESULT

In the previous section, 21 research papers that were published between the year of 2014 and 2017 were reviewed. The review is arranged in two categories. The first category contains 16 literature while the second category contains 5. In the present section, 1) machine learning techniques were compared with various dataset based on accuracy and best performance (Table 1); 2) results on the use of machine learning techniques with meta-heuristic optimization algorithms [41] on various disease dataset are presented (Table 2).

A graph (figure 5) was plotted using the Spyder software in Python language. The features were selected on the basis of best performing algorithm and accuracy percentage. It was shown in the reviewed literature that the out-performance is achieved by the optimization algorithms EPSO_ABC-SVM [13] whose score is full 100% accuracy. The neural networks [36] exponentially increase the process time on complex dataset but gives a good result. Similarly in [24], Artificial Neural Networks (ANN) show 100%. accuracy on the various clinical dataset obtained from a hospital. Apart from all the top performing algorithms for labelled dataset, having the accuracy above 90% is really a good accomplishment. Logistic regression records 99% [6], LS-SVM shows 99.7% [21]. SVM, when optimized using Genetic algorithms, give 99.21% [9]. SVM gives 97.3% on breast cancer dataset [15], SVM-FOA also shows 96.9% [7], ensemble of SVM-RBF shows 96.296% [19], CNN-MDRP shows 94.8% [25], and AdaBoost ensemble with Random forest give 92.5% [18].

IV. DISCUSSION

From results extracted from the first category of ML techniques in the disease dataset as shown in Table 1, an intensive study of 16 research papers depicts the good performance of classification algorithms. Surprisingly, logistic regression [6] obtain 99% accuracy on an Electronic Health Record. Rest of the algorithms like SVM on breast cancer dataset [15] gives 97.13%, ensemble methods SVM-RBF gives 96.296% on dengue hemorrhagic fever (DHF) dataset [19]. In [24] various classification algorithms were used on the acute illness dataset and the Artificial Neural Network (ANN) gave an outstanding accuracy of 100%. The processing speed of ANN is however slower compared with SMO, DT, MLPN, etc.

The level of accuracy obtained when ML techniques are used with meta-heuristic optimization algorithms (such as Genetic algorithm, the evolutionary algorithm, nature-inspired algorithm, etc) is remarkable (see Table 2). Use of Endocrine-Based Particle Swarm Optimization (EPSO) and Artificial Bee Colony (ABC) algorithm with Support Vector Machine [13] gives 100% accuracy in a different clinical disease dataset. The evolutionary algorithm, Genetic Algorithm (GA) with Support Vector Machine (SVM) [9], gives 99.21% on a diabetes dataset. However, in [12] and [7], classification algorithms like SVM perform error-free and better with the Fire-fly Optimization algorithm. Another different disease dataset compared in [7] shows that the Wisconsin dataset gives 96.9% accuracy using SVM-FOA techniques.

V. FUTURE SCOPE AND CONCLUSION

In this section, a case is made for a novel research area in machine learning, namely the heterogeneous technique of meta-heuristic optimization algorithm and ML. The framework (Figure 6) is designed in a manner where the appropriate use of archival data can be made by choosing a proper healthcare dataset. The established selection of datasets will undergo cleaning stage of pre-processing and visualization. In most of the literature studied, the supervised learning algorithms are more often used than the other learning strategies such as unsupervised and reinforcement learning. In supervised learning, SVM algorithm has more accuracy in the result than the other classification and regression algorithms.

From the above Figure 6, we can also remark that the use of real-time data is not encouraging with unsupervised learning, since it will be time-consuming to train them. The work on supervised learning can be done on an unlabeled dataset but it won't give the accurate result. The accuracy which is achieved by the algorithm will not be the endmost evaluation criterion. The real challenge is to use a technique which can work in any situation on any dataset with a proper performance evaluation.

It is unlikely that in this proposed framework, the performance of the other techniques will be measured by using the statistical performance measures like accuracy, recall, sensitivity, and specificity. A meta-heuristic algorithm (Figure 7) will be combined with the algorithm for parameter optimization, and results will be compared on the basis of performance stats. This set of framework solution is designed for proposed problem statement so that the real groundwork will bring a novel approach towards the study and prediction of diseases and improvement of the sustainability in healthcare.

In the generation of automation, humans are becoming oblivious about the most precious gift we possess i.e. our health. The biggest challenge is sustainable ICT [4]. Since the collection of health records is a tedious and time-consuming task, most of the data available in public domain have missing values, noises, huge differences in data, dissimilarity etc, which are not appropriate options for conducting a full-fledged research work. The challenge of getting right datasets for the disease prediction is very much important, as the data available in open source websites (like Kaggle, electronic health records, CDC, hospitals data, government data etc), have so many abnormalities. Such abnormalities include empty records, gibberish values, noises, although these data have to be integrated and used for better outcomes. Machine learning techniques can apply nature-inspired algorithms for optimization [29] such as ant colony optimization [30], Firefly optimization [31] and bat algorithms [32]. On the other hand the achieved result can have various applications such as in Mobile Clinical Decision Support System (MCDSS) or a web-based application. MDS using fuzzy logic have even been shown to have better result in diagnosis [48, 49] which can help both the doctor and the patient to keep track of disease and to stay safe from the symptoms to avoid unnecessary dangerous conditions. Furthermore, communicable diseases are epidemic in some regions. Good procedure will help the government authorities to get the medicine and equipment ready before the outbreak so as to reduce the number of death caused yearly due to deadly epidemiological communicable disease.

The present paper has provided an enlightened overview of machine learning and its various optimization techniques in disease datasets to avoid any kind of epidemic. The study will help in directing future work on sustainability. Ultimately, sustainability to healthcare will be achieved exclusively when our datasets can efficiently visualize and analyze with no flaws.

ACKNOWLDGEMENT

We would like to thank the anonymous reviewer(s) who gave valuable suggestions and helped in improving the technical content of the paper. The Editorial Office of this journal (African Journal of Computing & ICT) is also acknowledged for fundamental improvement in the grammatical construction of the paper. We would like to extend thanks to Dr. Shandar Ahmed from JNU for giving insight about ML. Also we would like to extend warm gratitude to Ms. Taqdees Fatimah for her help in structuring the paper.

REFERENCES

[1] Carbonell, J; Michalski, R; and Mitchell, T. "An Overview of Machine Learning", Machine Learning, pp. 3-23, 1983.

[2] "An Overview of Diabetic Neuropathy", WebMD, 2018. [Online]. Available: https://www.webmd.com/diabetes/diabetes-neuropathy#1. [Accessed: 11- Sep- 2017].

[3] De'Ath, G. Boosted trees for ecological modeling and prediction. Ecology. 2007; 88(1):243–51. PMID: 17489472

[4] "Information and communications technology", En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Communications_technology. [Accessed: 17- Nov- 2017].

[5] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. Vol. 1. No. 1. Cambridge: MIT Press, 1998.

[6] Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., Yang, G. and Chen, Y., 2017. A machine learning-based framework to identify type 2 diabetes through electronic health records. International Journal of Medical Informatics, 97, pp.120-127.

[7] Shen, L., Chen, H., Yu, Z., Kang, W., Zhang, B., Li, H., Yang, B. and Liu, D., 2016. Evolving support vector machines using fruit fly optimization for medical data classification. Knowledge-Based Systems, 96, pp.61-75.

[8] Huang, G.M., Huang, K.Y., Lee, T.Y. and Weng, J.T.Y., 2015, December. An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. In BMC bioinformatics (Vol. 16, No. 1, p. S5). BioMed Central.

[9] Santhanam, T. and Padmavathi, M.S., 2015. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science, 47, pp.76-83.

[10] Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L. and Bellazzi, R., 2017. Machine Learning Methods to Predict Diabetes Complications. Journal of diabetes science and technology, p.1932296817706375.

[11] Santillana, M., Nguyen, A.T., Dredze, M., Paul, M.J., Nsoesie, E.O. and Brownstein, J.S., 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS computational biology, 11(10), p. e1004513.

[12] Ch, S., Sohani, S.K., Kumar, D., Malik, A., Chahar, B.R., Nema, A.K., Panigrahi, B.K. and Dhiman, R.C., 2014. A support vector machine-firefly algorithm based forecasting model to determine malaria transmission. Neurocomputing, 129, pp.279-288.

[13] Lin, K.C. and Hsieh, Y.H., 2015. Classification of medical datasets using SVMs with hybrid evolutionary algorithms based on endocrine-based particle swarm optimization and artificial bee colony algorithms. Journal of medical systems, 39(10), p.119.

[14] Melillo, P., Izzo, R., Orrico, A., Scala, P., Attanasio, M., Mirra, M., De Luca, N. and Pecchia, L., 2015. Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. PloS one, 10(3), p. e0118504.

[15] Asri, H., Mousannif, H., Al Moatassime, H. and Noel, T., 2016. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, pp.1064-1069.

[16] Kang, S., Kang, P., Ko, T., Cho, S., Rhee, S.J. and Yu, K.S., 2015. An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. Expert Systems with Applications, 42(9), pp.4265-4273.

[17] Jelinek, H.F., Cornforth, D.J. and Kelarev, A.V., 2016. Machine learning methods for automated detection of severe diabetic neuropathy. J Diabetic Complications Med, 1(108), p.2.

[18] Emanet, N., Öz, H.R., Bayram, N. and Delen, D., 2014. A comparative analysis of machine learning methods for classification type decision problems in healthcare. Decision Analytics, 1(1), p.6.

[19] Kesorn, K., Ongruk, P., Chompoosri, J., Phumee, A., Thavara, U., Tawatsin, A. and Siriyasatien, P., 2015. Morbidity rate prediction of dengue hemorrhagic fever (DHF) using the support vector machine and the Aedes aegypti infection rate in similar climates and geographical areas. PloS one, 10(5), p. e0125049.

[20] Zuccon, G., Khanna, S., Nguyen, A., Boyle, J., Hamlet, M. and Cameron, M., 2015. Automatic detection of tweets reporting cases of influenza like illnesses in Australia. Health information science and systems, 3(S1), p. S4.

[21] Patidar, S., Pachori, R.B. and Acharya, U.R., 2015. Automated diagnosis of coronary artery disease using tunable-Q wavelet transform applied on heart rate signals. Knowledge-Based Systems, 82, pp.1-10.

[22] Bashir, S., Qamar, U. and Khan, F.H., 2015. BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. Australasian physical & engineering sciences in medicine, 38(2), pp.305-323.

[23] Passos, I.C., Mwangi, B., Cao, B., Hamilton, J.E., Wu, M.J., Zhang, X.Y., Zunta-Soares, G.B., Quevedo, J., Kauer-Sant'Anna, M., Kapczinski, F. and Soares, J.C., 2016. Identifying a clinical signature of suicidality among patients with mood disorders: a pilot study using a machine

learning approach. Journal of affective disorders, 193, pp.109-116.

[24] Rane, A.L., 2015, January. Clinical decision support model for prevailing diseases to improve human life survivability. In Pervasive Computing (ICPC), 2015 International Conference on (pp. 1-5). IEEE.

[25] Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5, pp.8869-8879.

[26] Yang, X.S., 2009, October. Firefly algorithms for multimodal optimization. In International symposium on stochastic algorithms (pp. 169-178). Springer, Berlin, Heidelberg.

[27] Jain, R. and Rao, B., 2016, April. India's Focus On Medical Diagnostic Laboratories: Indian Planning and Programmes. In Proceedings of International Academic Conferences (No. 3505995). International Institute of Social and Economic Sciences.

[28] "Goal 3: Good health and well-being". UNDP. Retrieved 13 April 2017. http://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-3-good-health-and-well-being.html

[29] Fister Jr, I., Yang, X.S., Fister, I., Brest, J. and Fister, D., 2013. A brief review of nature-inspired algorithms for optimization. arXiv preprint arXiv:1307.4186.

[30] Dorigo, Marco, 1992 Optimization, learning and natural algorithms. Ph.D. Thesis, Politecnico di Milano, Italy.

[31] Yang, X.S., 2005, Engineering optimizations via nature-inspired virtual bee algorithms. volume 3562, pages 317–323.

[32] Yang, X.S., 2010, A new meta-heuristic bat-inspired algorithm. Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), pages 65–74.

[33] Samuel, A.L., 1959. Some studies in machine learning to use the game of checkers. IBM Journal of research and development, 3(3), pp.210-229.

[34] Mohri, M; Talwalkar, A; and Rostamizadeh, A., Foundations of machine learning. Cambridge (Massachusetts): MIT Press, 2012.

[35] "Unsupervised learning", En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Unsupervised_learning. [Accessed: 10- Nov- 2017].

[36] Jordan, Michael I.; Bishop, Christopher M. (2004). "Neural Networks". In Allen B. Tucker. Computer Science Handbook, Second Edition (Section VII: Intelligent Systems). Boca Raton, FL: Chapman & Hall/CRC Press LLC. ISBN 1-58488-360-X.

[37] Singha, S.K. and Hassan, S.I., 2017. Enhancing the Classification Accuracy of Noisy Dataset By Fusing Correlation Based Feature Selection with K-Nearest Neighbour.

[38] Hassan, S.I., 2017. Designing a flexible system for automatic detection of categorical student sentiment polarity using machine learning. International Journal of u-and e-Service, Science and Technology, 10(3), pp.25-32.

[39] Somoskovi, A., Ahmedov, S. and Salfinger, M., 2013. It is health that is real wealth & not pieces of gold & silver. The Indian journal of medical research, 137(3), p.437.

[40] Yang, Xin-She, Nature-Inspired Optimization Algorithms. Elsevier, 2014.

[41] Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P., 1983. Optimization by simulated annealing. science, 220(4598), pp.671-680.

[42] Holland, J; "Genetic Algorithms", Scientific American, vol. 267, no. 1, pp. 66-72, 1992.

[43] Dorigo, M; Caro, G; and Gambardella, L; "Ant Algorithms for Discrete Optimization", Artificial Life, vol. 5, no. 2, pp. 137-172, 1999.

[44] Eberhart, R; and Kennedy, J; "A new optimizer using particle swarm theory", MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science.

[45] Yang, X; "Firefly Algorithms for Multimodal Optimization", Stochastic Algorithms: Foundations and Applications, pp. 169-178, 2009.

[46] Yang, X; and Deb, Suash; "Cuckoo Search via L&#x00E9;vy flights", 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), 2009.

[47] Long, N; Meesad, P; and Unger, H. "A highly accurate firefly based algorithm for heart disease prediction", Expert Systems with Applications, vol. 42, no. 21, pp. 8221-8231, 2015.

[48] Awotunde, J.B., Matiluko, O.E. and Fatai, O.W., "Medical diagnosis system using fuzzy logic". African Journal of Computing & ICT, 7(2), pp.99-106, 2014.

[49] Alam, S; Hassan, S; and Uddin, M; "Fuzzy Models and Business Intelligence in Web-Based Applications", Applications of Soft Computing for the Web, pp. 193-221, 2017.

AUTHORS BIODATA

Ms. Hina Firdaus has completed her post graduation M.Tech in Computer Science and Engineering, Jamia Hamdard, New Delhi (India) in July 2018. Her primary area of research is Machine learning, Computer Graphics, data mining, and ICT for sustainable development.

Dr. Syed Imtiyaz Hassan works as an Assistant Professor at the Department of Computer Science and Engineering, Jamia Hamdard, New Delhi (India). His professional experience spans over more than 17 years of teaching, research, and project supervision. He has supervised more than 80 students for interdisciplinary research and industrial projects. Over the years, he has published many research papers with national and international journals of repute. In addition to these, he is also in the Editorial Boards and Reviewers' Panels of various journals. His primary area of research is Computational Sustainability. To meet the objectives of Computational Sustainability, he explores the role of Nature Inspired Computing, Machine Learning, Data Science, Mobile Crowd Sensing, and IoT for developing Smart & Sustainable Software Systems.

Dr. Harleen Kaur is a faculty member at the School of Engineering Sciences and Technology at Jamia Hamdard, New Delhi, India. She recently worked as Research Fellow at United Nations University (UNU) in IIGH-International Centre for Excellence, Malaysia to conduct research on funded projects from South-East Asian Nations (SEAN). She is currently working on an Indo-Poland bilateral international project funded by the Ministry of Science and Technology, India, and the Ministry of Polish, Poland. In addition, she is working on a national project, catalyzed and supported by the National Council for Science and Technology Communication (NCSTC), the Ministry of Science and Technology, India. Her key research areas include data analytics, big data, applied machine learning and predictive modeling. She is the author of various publications and has authored/ edited several reputed books. She is a member of various international bodies and is a member of the editorial board of international journals on data analytics and machine learning. She is the recipient of Ambassador for Peace Award (UN Agency) and honors and is funded researcher by external groups.

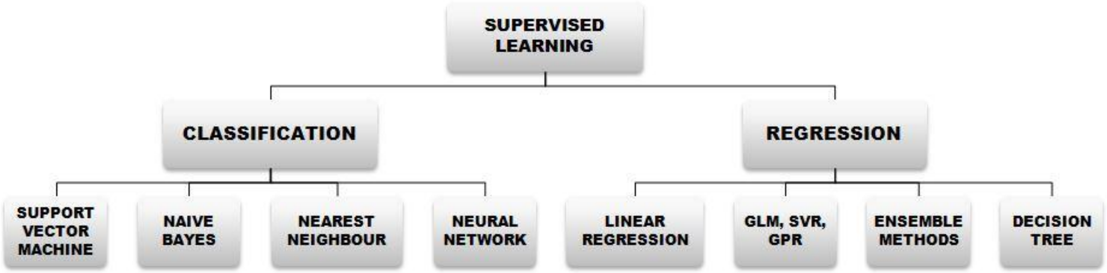


Figure 1. Supervised Learning algorithm classification

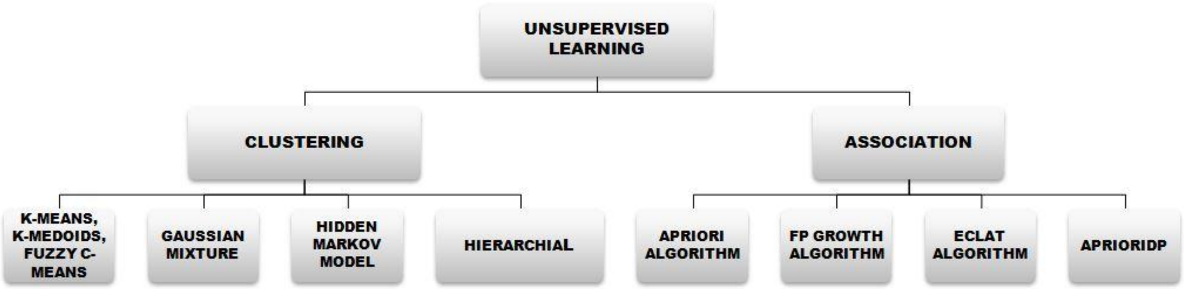


Figure 2. Unsupervised Learning algorithm classification

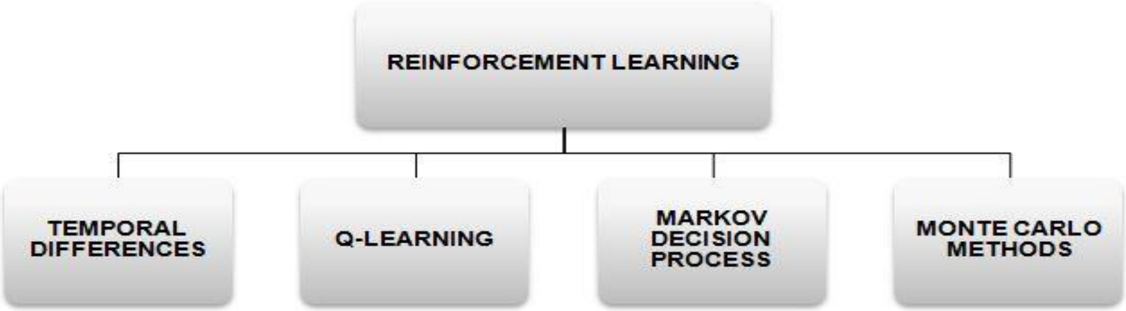


Figure 3. Reinforcement Learning Algorithm classification

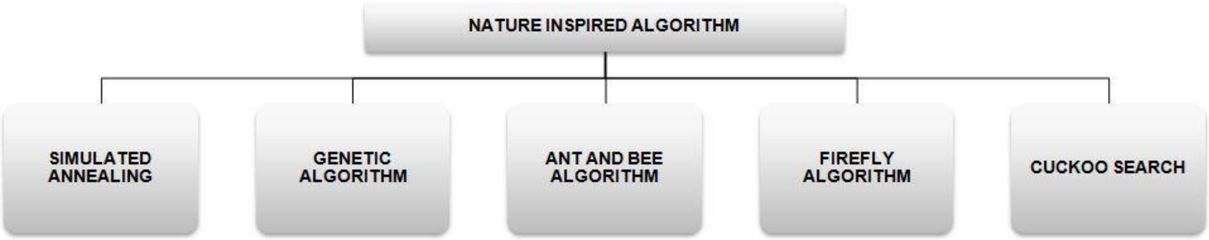
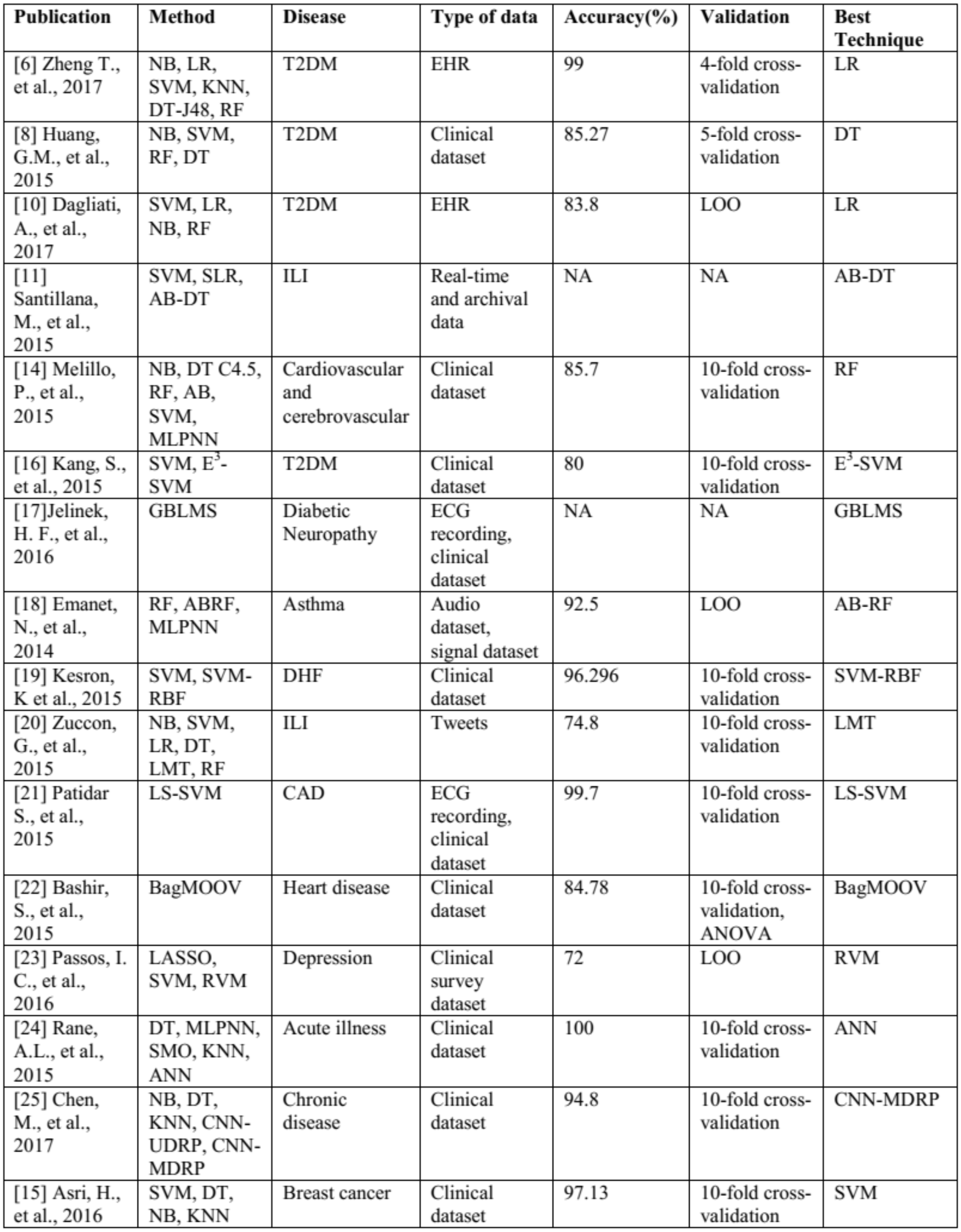


Figure 4. Various types of Nature Inspired Algorithm

Table 1. Publications which use ML method on various disease dataset



Publication	Method	Disease	Type of data	Accuracy(%)	Validation	Best Technique
[6] Zheng T., et al., 2017	NB, LR, SVM, KNN, DT-J48, RF	T2DM	EHR	99	4-fold cross-validation	LR
[8] Huang, G.M., et al., 2015	NB, SVM, RF, DT	T2DM	Clinical dataset	85.27	5-fold cross-validation	DT
[10] Dagliati, A., et al., 2017	SVM, LR, NB, RF	T2DM	EHR	83.8	LOO	LR
[11] Santillana, M., et al., 2015	SVM, SLR, AB-DT	ILI	Real-time and archival data	NA	NA	AB-DT
[14] Melillo, P., et al., 2015	NB, DT C4.5, RF, AB, SVM, MLPNN	Cardiovascular and cerebrovascular	Clinical dataset	85.7	10-fold cross-validation	RF
[16] Kang, S., et al., 2015	SVM, E^3-SVM	T2DM	Clinical dataset	80	10-fold cross-validation	E^3-SVM
[17]Jelinek, H. F., et al., 2016	GBLMS	Diabetic Neuropathy	ECG recording, clinical dataset	NA	NA	GBLMS
[18] Emanet, N., et al., 2014	RF, ABRF, MLPNN	Asthma	Audio dataset, signal dataset	92.5	LOO	AB-RF
[19] Kesron, K et al., 2015	SVM, SVM-RBF	DHF	Clinical dataset	96.296	10-fold cross-validation	SVM-RBF
[20] Zuccon, G., et al., 2015	NB, SVM, LR, DT, LMT, RF	ILI	Tweets	74.8	10-fold cross-validation	LMT
[21] Patidar S., et al., 2015	LS-SVM	CAD	ECG recording, clinical dataset	99.7	10-fold cross-validation	LS-SVM
[22] Bashir, S., et al., 2015	BagMOOV	Heart disease	Clinical dataset	84.78	10-fold cross-validation, ANOVA	BagMOOV
[23] Passos, I. C., et al., 2016	LASSO, SVM, RVM	Depression	Clinical survey dataset	72	LOO	RVM
[24] Rane, A.L., et al., 2015	DT, MLPNN, SMO, KNN, ANN	Acute illness	Clinical dataset	100	10-fold cross-validation	ANN
[25] Chen, M., et al., 2017	NB, DT, KNN, CNN-UDRP, CNN-MDRP	Chronic disease	Clinical dataset	94.8	10-fold cross-validation	CNN-MDRP
[15] Asri, H., et al., 2016	SVM, DT, NB, KNN	Breast cancer	Clinical dataset	97.13	10-fold cross-validation	SVM

Table 2. Publications which use ML method with Meta-Heuristic optimization technique on various disease dataset

Publication	**Method**	**Disease**	**Type of data**	**Accuracy(%)**	**Validation**	**Best Technique**
[12]Ch, Sudheer et al., 2017	SVM-FFA, ARMA, ANN, SVM	Malaria	Clinical dataset	NA	NA	SVM-FFA
[13] Lin, K.C., et al. 2015	EPSO_ABC-SVM	Various diseases	Clinical dataset	100	5-fold cross-validation	EPSO_ABC-SVM
[47] Long, N.C., et al. 2017	CFARS-AR, IT2FLS, SVM, NB, ANN	Heart disease	Clinical dataset	88.3, 87.2	NA	IT2FLS
[9] Santhanam, T. et al., 2015	SVM, GA	Diabetes	Clinical dataset	99.21	10-fold cross-validation	SVM
[7] Shen, L., et al., 2016	SVM-FOA, PSO-SVM, BFO-SVM, GA-SVM, Grid-SVM	Disease	Clinical dataset	96.9, 77.46, 77.46, 96.38	10-fold cross-validation	SVM-FOA

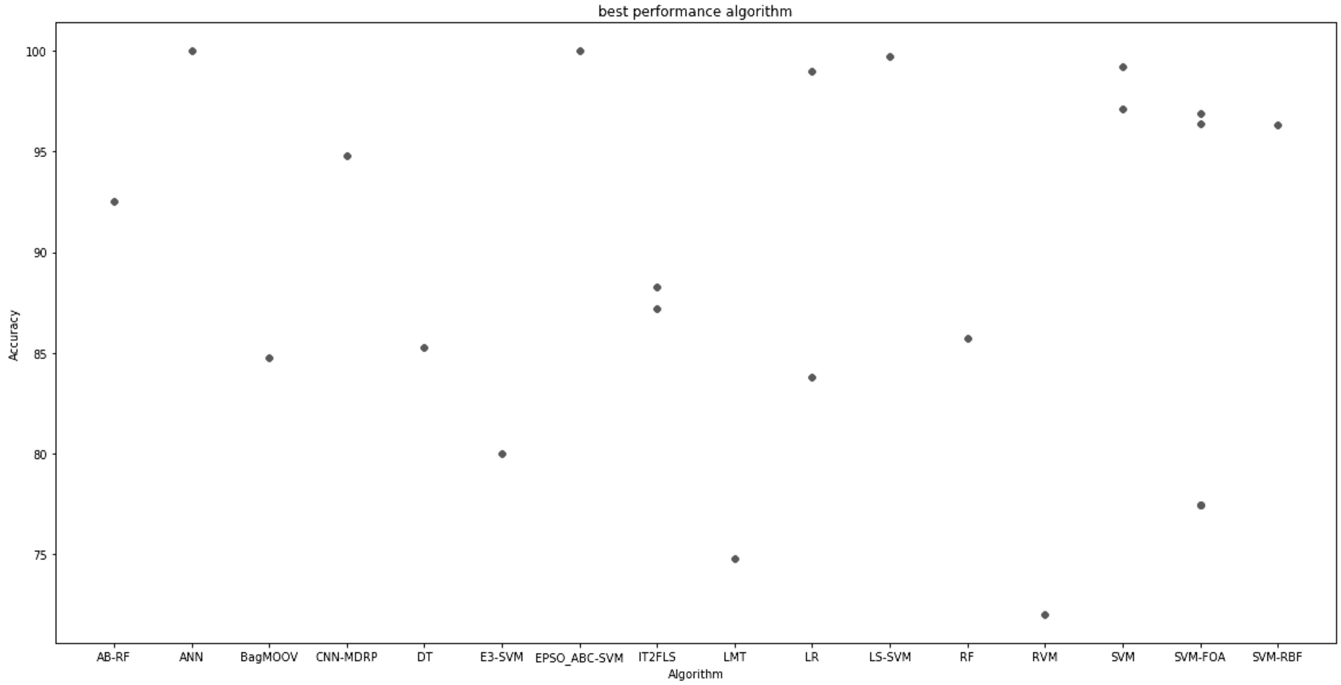


Figure 5. A scatter graph comparing the best performing algorithms using metric evaluation of accuracy in the proposed survey. Accuracy of 100% is achieved by EPSO_SVM-ABC and ANN algorithms.

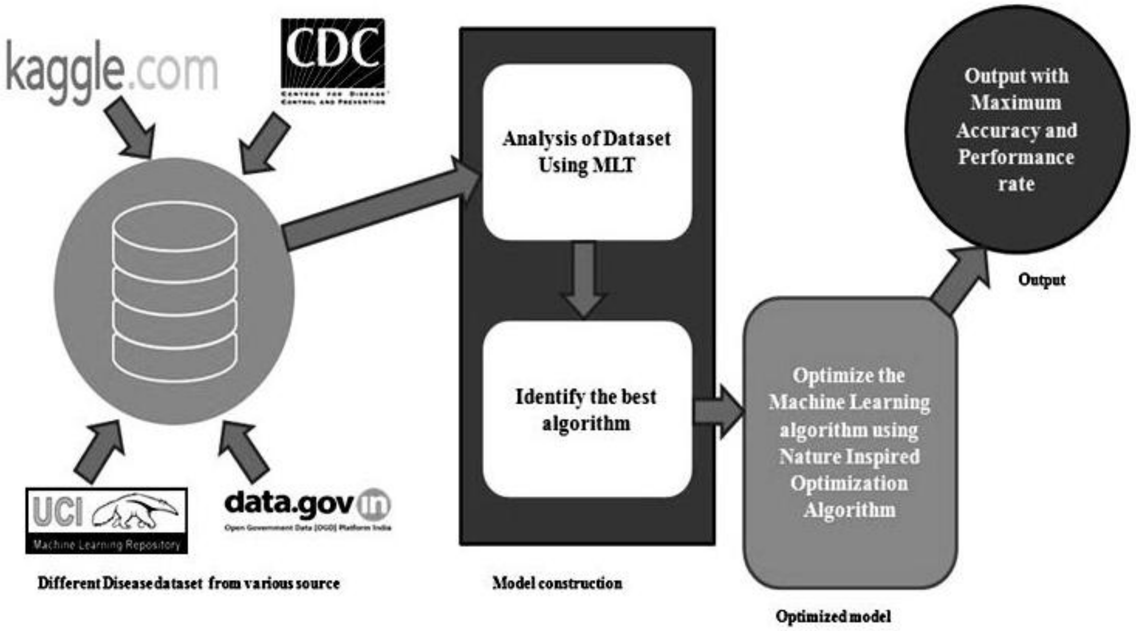

kaggle.com
CDC
Centers for Disease Control and Prevention
UCI
Machine Learning Repository
data.gov.in
Open Government Data (OGD) Platform India
Different Disease dataset from various source
Analysis of Dataset Using MLT
Identify the best algorithm
Model construction
Optimize the Machine Learning algorithm using Nature Inspired Optimization Algorithm
Optimized model
Output with Maximum Accuracy and Performance rate
Output

Figure 6. Optimized sustainable healthcare framework using machine learning techniques and nature-inspired optimization algorithm

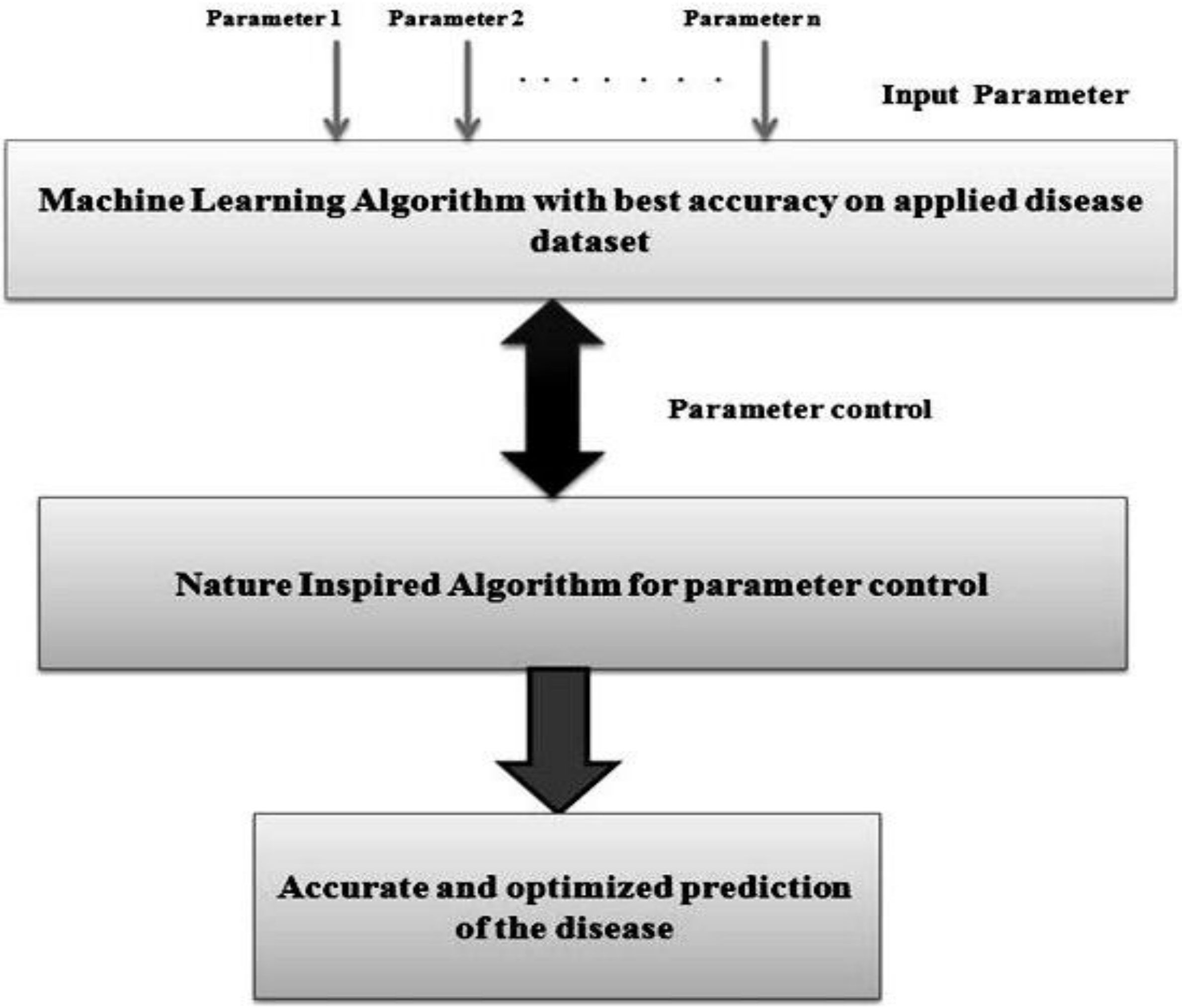

Parameter 1
Parameter 2
. . . .
Parameter n
Input Parameter
Machine Learning Algorithm with best accuracy on applied disease dataset
Parameter control
Nature Inspired Algorithm for parameter control
Accurate and optimized prediction of the disease

Figure 7. Framework of Machine learning and nature inspired algorithm working by doing parameter control