Medical Big Data Analytics Using Machine Learning Algorithms

Usha Moorthy, Usha Devi Gandhi

Abstract*: *Artificial intelligence and expert systems plays a key role in modern medicine sciences for disease prediction, surveillance interventions, cost efficiency and better quality of life etc. With the arrival of new web-based data sources and systematic data collection through surveys and medical reporting, there is a need of the hour to develop effective recommendation systems which can support practitioners in better decision-making process. Machine Learning Algorithms (MLA) is a powerful tool which enables computers to learn from data. While many novel developed MLA constantly evolves, there is need to develop more systematic, robust algorithm which can interpret with highest possible accuracy, sensitivity and specificity. The study reviews previously published series on different algorithms their advantages and limitations which shall help make future recommendations for researchers and experts seeking to develop an effective algorithm for predicting the likelihood of various diseases.

Keywords*: *Artificial intelligence, expert systems, Machine Learning Algorithms, disease prediction, future recommendations.

I. INTRODUCTION

The advent of Healthcare Information Systems (HIS) plays an imperative role in the field of medical sciences and technology as it assists medical practitioners in development of accurate methods of disease prediction, high-risk assessment and sustainable health monitoring [1,2]. Healthcare information systems integrate IT with healthcare to meet the growing demands of quality and efficacy in healthcare systems across the globe [3].

Artificial Intelligence and deep learning techniques are currently trending and several studies exclusively focus on analyzing its support towards modern medical decision models [3,4]. With the arrival of new web-based data sources and systematic data collection through surveys and medical reporting, there is a need of the hour to develop effective recommendation systems which can support practitioners in the better decision-making process [5]. Application of these computer-aided systems is advantageous due to its computational speed, accuracy, specificity and sensitivity when compared to traditional statistical modeling.

Revised Manuscript Received on November 08, 2019.

Usha Moorthy, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India, Email: ushmitha@gmail.com
Usha Devi Gandhi *, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India, Email: ushadevi.g@vit.ac.in

Machine Learning Algorithms (MLA) is a powerful tool which enables computers to learn from data [6]. Over past decade, a number of machine learning classifiers have been developed which is broadly classified into the white and black box. While, white box MLA is simple and transparent which includes simple decision tree, black box models which are also known as deep learning models are often difficult to interpret their inner working [7]. Neural Networks are paradigmatic examples of deep learning algorithms [8] which also includes Random Forest model, the Support Vector Machine (SVM) models etc. Along with machine learning algorithms, data mining techniques and statistical analysis provide major support to experts in prediction of disease [9].

Medical data in the form of electronic health records, sensors and monitors analyzed using traditional machine learning algorithms like logistic regression and regression analysis proves to be effective in disease prediction using structured clinical or hospital data. These algorithms were supervised and trained to classify characteristics based on past experiences [10]. However, these models fail to show high levels of accuracy and specificity when applied to big data or data streams especially from wireless sensor networks. With the development of new computational tools such as big data analytics technology, modern machine learning algorithms use unsupervised machine learning approach to select features or attributes automatically from larger datasets to improve the accuracy [11]. Since there is a growing demand for disease prediction in the field of medicine as well as need to develop more powerful analytic tools, efforts are needed to develop a novel algorithm for more evidence-based quality on accurate disease prediction. Therefore, this review present algorithms used for different disease prediction from previous studies and make future recommendations for researchers and experts seeking to develop an effective algorithm for predicting the likelihood of various diseases. In this review article aimed to understand application of different machine learning algorithm in prediction of a particular disease. Also, this article focused on identification of advantages and limitations of the proposed. The methodology used for this study is a review methodology where several journal and peer-reviewed articles were reviewed with regards to machine learning algorithm for medical applications. For this review, peer-reviewed papers from various journals like Science Direct, Elsevier, Taylor & Francis, NCBI, Scopus Journals and works done by other researchers on the similar or corresponding topic were extensively reviewed.

A total of 41 studies were reviewed for this study. In this review, machine learning methods used for the medical application is described. Also, a comparative review of techniques adopted in machine learning technique is adopted and analyzed. The potential usage in the field of medical science, most of the researcher suggested machine learning approaches like ANN, SVM and deep learning methods. The remainder of this paper flows as follows: Review of previous studies conducted in terms of disease management using various machine learning algorithms, future recommendations and conclusion.

II. REVIEW METHOD

This article examined about the role and contribution of machine learning approach in health care application. For systematic review of literature this article evaluated from several research journals and articles. The review has been conducted based on application of several machine learning technique for different diseases. The review is conducted based on the machine learning technique in field of disease related to mental health, heart, liver diseases and other life-threatening diseases such as cancer and diabetes.

A. Information Sources

The review of articles related to machine learning algorithm articles are collected from the databases like Science Direct, IEEE, PLoS One, Springer and Sci Rep. From the reference sources totally 83 articles are collected among those 55 are scrutinized and considered for analysis.

B. Study selection

Among the total selected 83 research articles 55 articles are considered for analysis based on the consideration of following criteria.

Article published between 2014 - 2018 (Among 55 total articles count 6 articles are apart from criteria which are released in the year 2010, 2012, 1998, 2006 and 2008 these articles are used in introduction section due to its strong technical explanation terms.

1. Articles deals with a machine learning approach.
2. Articles are published in the English language.

Based on the above criteria review articles are selected previously selection of review article is based on the key words such as “Machine learning”, “Healthcare”, “Disease Prediction”, “Medical Application”, “Big Data” and “Big Data in Healthcare”.

III. RESULTS

This review article follows the methodology used to the review methodology where several journal and peer-reviewed articles were reviewed with regards to machine learning algorithm for medical applications. For this review, peer-reviewed papers from various journals like Science Direct, Elsevier, Taylor & Francis, NCBI, Scopus Journals and works done by other researchers on the similar or corresponding topic were extensively reviewed. A total of 41 studies were reviewed for this study. In this review, machine learning methods used for the medical application is described. Also, a comparative review of techniques adopted in machine learning technique is adopted and analyzed. The potential usage in the field of medical science, most of the researcher suggested machine learning approaches like ANN, SVM and deep learning methods.

The literature included in review comprises several descriptive articles and studies related to several big data analytics for healthcare applications. Review of articles is conduced based on the consideration of different diseases prediction such as liver, cancer, diabetes, mental health, and heart disease using machine learning algorithm.

A. Prediction of Diseases using Classifier

Comparative analysis of different classifier techniques in MRI images Psychosis prediction is presented in several researchers are presented by Salvador et al. [12] and de Wit et al. [13]. Research by Salvador et al. [12] stated that among 383 evaluated patients 128 are subjected to schizophrenia cases and de Wit et al. [13] identified 125 subjects of which 64 were identified with Ultra-High Risk (UHR) of psychosis. de Wit et al. [13] used 99 features. A study by Schnack [14] used CNN to analyze sMRI image in prediction of Schizophrenia, Alzheimer's disease. A study by Chen et al. [15] tested performance of different classifiers in prediction of cerebral infarction in regions of China. Data in the form of EHR, medical image and gene data both structured and unstructured was used. 31,919 hospitalized patient’s records were recovered. For text data CNN-UDRP classifier was used; For structure and text data, Convolutional Neural Network based Multimodal Disease Risk Prediction (CNN-MDRP) was used, and for structured only data NB+ gaussian distribution, KNN and DT (CART) algorithm was used. In research conducted by Chen et al. [16] examined the MRI data for sclerosis disease prediction from multiple clinical data. A total of 1600 subjects’ record was used from CLIMB study. A study by Abos et al. [17] attempted to predict Parkinson’s disease using fMRI images. 133 patients dataset was collected. Supervised machine learning technique was applied, and the study tested performance of SVM.

Table I: Mental Health-related Disease Management

Data input format	Classifiers used	Disease Prediction
sMRI fMRI Clinical/ Hospital data EHR Demographical data	ridge, LASSO, elastic net, L0 norm regularized logistic regressions, a support vector classifier, regularized discriminant analysis, random forests, Gaussian process classifier and Support Vector Regression, CNN, NB+ gaussian distribution, KNN and DT (CART) algorithm	Psychosis, Alzeihmer’s disease, Schizophrenia, cerebral infaraction, multiple sclerosis, Parkinson’s disease

The above-mentioned table1 provides insight into the management of mental health related diseases using different classifiers for different disease prediction. The mental health considered several data input format such as sMRI, fMRI, Clinical/ Hospital data EHR Demographical data. The diseases predicted using classifier are Psychosis, Alzheimer's disease, Schizophrenia, cerebral infarction, multiple sclerosis, Parkinson's disease, Further for mental health disease classifiers such as CART, CNN, KNN, DT and so on.

In recent years, many people are subjected to heart diseases which are considered as a major cause of mortality throughout the world. Hence the article related application of heart disease prediction using big data analytics and machine learning algorithm. To predict cardiovascular heart diseases, Kim et al. [18] conducted a research through the application of multiple supervised machine learning algorithms. The analysis of supervised algorithm with machine learning approach exhibited that ML exhibits significant performance rather than other supervised machine learning algorithm such as scoring or prediction model which uses logistics approaches. Research by Bhatt et al. [19] and Acharya et al. [20] uses ECG signal for CVD diseases prediction and classification of coronary artery diseases using myocardial infarction. For this research UCI machine learning datasets were created based on repository, Hungarian database of heart disease, 7 CAD, 148 MI and 52 Normal subjects echocardiograph (ECG) records with a total of 92273 beats Acharya et al. [20]. Bhatt et al. [19] used a supervised machine learning algorithm, WEKA tool for analysis and used classifiers J48 and Naïve Bayes on each dataset. While a study by Bhatt et al. [19] reported no comparative analysis and prediction outcome for both classifiers. A study by Acharya et al. [20] reported that kNN classifier showed an accuracy of 98.5%, sensitivity of 99.7% and specificity of 98.5% respectively. A study by Masetic and Subasi [21] applied and tested Random Forests algorithm to detect congestive heart failure using ECG signals. Long-term ECG time series collected using Beth Israel Deaconess Medical Center (BIDMC) for prediction of Congestive Heart Failure and PTB Diagnostic through use of ECG databases. The dataset considered involves PhysioNet which contains 13 heartbeats collected from MIT–BIH Arrhythmia database. In collected databases, auto regressive Burg method is applied for feature extraction. Study by Weng et al. [22] predicted cardiovascular risk from routine clinical data. 378,256 patient's data was used. The study tested the following classifiers: Random forest, logistic regression, gradient boosting machines and neural networks. Neural Network showed an accuracy of 67.5%, sensitivity 67.5% and specificity of 70.7%. Study by Lafta et al. [23] tested performance of classifiers ensemble models (FFT-MLE), Neural Networks (NN), Least Square-SVM (LS-SVM) and naive Bayes (NB) in the prediction of heart disease using Tunstall database. Study by Frizzell et al. [24] attempted to predict hospital readmission within 30 days in heart failure patients. Data was obtained from the American Heart Association Get with the Guidelines-Heart Failure (GWTG-HF) registry. 56 477 patients were included. Random forest classifier, logistic regression, Tree augmented Bayesian network (TAN), gradient-boosted classifier and Least Absolute Shrinkage and Selection Operator (LASSO) method was used. Study by Dawes et al. [25] developed and tested a machine-learning survival model using 3D cardiac motion from newly diagnosed pulmonary hypertension patients using Cardiac Magnetic Resonance Images (MRI). The study used supervised machine learning model and obtained dataset from the National Pulmonary Hypertension Service at the Imperial College Healthcare NHS Trust, which included 405 subjects of which 256 confirmed PH patients. Aljaaf et al. [26] presented a multi-level risk prediction model of HF using C4.5 decision tree classifier. It is a supervised automated machine learning technique which can handle both classification and regression. Zheng et al. [27] presented a Computer-Assisted Diagnostic (CAD) system for chronic heart failure. The study extracted features from the cardiac reserve as well as diagnosis heart sound.

Table II: Heart Disease Management

Data input format	Data mining techniques	Prediction
ECG signals Cardiac MRI Demographic	kNN, MLP, Classification and Regression tree, RF, SVM, NN, LR, DT	Coronary heart failure Cardio Vascular disease Quality of pulse Heart disease

From the above, it is evident that Neural Network and SVM classifiers are best to predict heart disease from UCI machine learning repository and similar heart disease datasets. The above table2 presents an overview of heart disease management based on the data input, classifier used and what to expect out of the prediction model.

B. Data Mining techniques for Lung and Liver Disease Prediction

Study by Le-Dong et al. [28] predicted interstitial lung disease based on combined data obtained from Pulmonary Function Test (PFT) and Hi-Res CT. The dataset consisted of 323 subjects of which 244 patients were identified with systemic sclerosis. Data exploration and cluster analysis were performed, and the following classifiers were evaluated for performance; Decision tree, logistic regression, SVM, NN, RF, kNN and SVM with z score.

Table III: Lung Disease

Data Type Format	Classifiers used	Prediction
HER Hospital data Tele-monitoring	DT, LR, SVM, NN, RF, kNN SVM with z score, NBC, ABN, LASSO, RPT, CIT, ID3 C4.5, NB and BN (K2).	Asthma COPD

The above table3 summarizes the machine learning classifiers used for prediction of various lung diseases. The data format used for prediction of lung diseases is EHR, Hospital data and Tele-Monitoring with several classification techniques such as DT, LR and so on.

The prediction of lung diseases such as COPD and Asthma through classifier techniques. Study by Spathis and Vlamos [29] predicted Asthma and Chronic Obstructive Pulmonary Disease (COPD). Dataset consist of 132 subjects and the study reported that Random Forest outperforms other algorithms with precision rate of 97.7% for COPD and 80.3% for asthma. Study by Yip et al. [30] interpreted Non-Alcoholic Fatty Liver Disease (NAFLD) from clinical and laboratory data. Dataset consisted of 922 subjects and 23 parameters.

Table IV: Liver disease management

Data type format	Classifiers used	Prediction
Clinical and laboratory data Liver function test Real-time tissue elastography (RTE)	LR, RR, AdaBoost, DT, (ADT), GA, PSO, MLR, SVM, NB, RF and kNN	non-alcoholic fatty liver disease Liver fibrosis Liver function

The above table 4, concise the different classifiers used for liver disease prediction and function. From the analysis of articles, it is observed that data format of three types is used for prediction of liver function. Study by Hashem et al. [31] predicted advanced liver fibrosis in chronic hepatitis patients. The dataset consisted of 39,567 patient's records and the different classifiers developed and tested were Alternating Decision Tree (ADT), genetic algorithm, particle swarm optimization, and multilinear regression models. Study by Saritha et al. [32] classified liver function data using medical data of Liver Function Test (LFT). The dataset consisted of 3750 records of patients with 41250 values. The algorithm developed was a separation of points by planes. The algorithm showed an accuracy of 85.1% in diagnosing hepatitis or liver disorder. Study by Chen et al. [33] attempted to measure hepatic fibrosis using Real-Time Tissue Elastography (RTE). 513 subjects who underwent liver biopsies were included in the study.

C. Prediction of Mortality Diseases

In worldwide, cancer is considered as leading mortality rate since several types of research are considered Study by Guadagni et al. [34] predicted breast cancer from web based source. Routinely collected clinical data were used for analysis. The study evaluated the performance of multiple kernel learning approach combining SVM and Random optimization but did not externally validate the result due to the limitation of data. A study by Murty and Babu [35] predicted lung cancer using NB. The dataset from UCI Machine Learning Repository of Lung Cancer Patients and Michigan Lung Cancer included 32 subjects of which 16 were identified as lung cancer patients. The initial 57 attributes were later developed to 7130 attributes like the number of subjects, and cancer patients were increased to 96 and 86 respectively.

Table V: Cancer disease management

Data type format	Classifiers used	Prediction
Clinical, demographical data Ultrasound Web based source Open source data repository	multiple kernel learning approach combining SVM and Random optimization, NB, RBF Neural Network, MLP, C4.5 (J48) algorithm, trained NN, SVM, RF	Lung, breast, prostate, colorectal, reoccurrence, survivability.

Similar to other disease prediction models prediction of cancer diseases is performed through other classifier techniques. In a review of existing research process cancer region are predicted in lung, breast and other regions of the human body.

Study by López et al. [36] predicted type 2 diabetes using Random Forest algorithm and kNN. The dataset consisted of 677 subjects. The performance of the algorithms was evaluated against Support Vector Machines and Logistic Regression. The study found that Random Forest outperforms other algorithms in terms of accuracy and stability.

Study by Alghamdi et al. [37] predicted diabetes mellitus using ensemble Machine learning approach. The dataset included medical records of cardiorespiratory fitness of 32,555 patients and 62 attributes from Henry Ford FIT dataset

Table VI: Diabetes Mellitus Management

Data type format	Classifiers used	Prediction
Clinical and hospital records, cardiorespiratory fitness dataset, open source data repository	RF, kNN, SVM, NB tree, LM tree, NB, J47	Diabetes Mellitus

In an analysis of diabetes mellitus prediction classifiers such as RF, kNN, SVM, NB tree, J47 and LM tree are used for diseases prediction.

IV. DISCUSSION

The systematic review of research articles related to disease prediction through big data analytics leads to a certain conclusion for processing. From the review of article [12 13 14 15] Structured machine learning technique was followed and the images from MRI was smoothened using Gaussian kernel with Full Width at Half Maximum (FWHM); The studies tested the following classifiers: ridge, LASSO, elastic net, L0 norm regularized logistic regressions, a support vector classifier, regularized discriminant analysis, random forests, Gaussian process classifier and Support Vector Regression. Salvador et al. [12] reported a similar accuracy rate among different classifiers when adequate feature type was selected. However, de Wit et al. [13] identified Support Vector Regression SVR classifications with accuracy ranging from 67 to 73% for complex models. In [16] The study tested the following classifiers, multiple (k) SVM classifiers; Decision tree such as Random Forest; fuzzy c means clustering; Gaussian distribution; multi-kernel learning. The study used Clustering technique based on data partitioning and concluded that fuzzy c means clustering helps to increase subgroup hence achieve higher accuracy in combination with SVM when compared to other models. In disease prediction CNN-based Unimodal Disease Risk Prediction (CNN-UDRP) algorithm showed an accuracy rate of 94.8%. CNN-MDRP algorithm showed an accuracy rate of 94.80%. For Structured data, although DT showed the highest accuracy rate of 63%, overall NB classifier showed better performance in disease prediction 17.

The SVM classifier was tested for both clinical and MRI data which showed the following performance: Accuracy rate of 70%, sensitivity of 62% and 71%, specificity of 65% and 68% respectively in predicting multiple sclerosis. The study found that SVM had an accuracy rate of 80% 18 19. The study Acharya et al. [20] Masetic and Subasi [21] Weng et al [22] Lafta et al [23] Frizzell et al [24] compared the performance of Random Forests with other classifiers like Artificial Neural Networks (ANN), C4.5, SVM, Artificial Neural Networks (ANN) and k-Nearest Neighbors (k-NN) and reported that Random Forests shows good accuracy in classifying a subject into normal or CHF when compared to other models. Data consisted of structured records of six patients with a total of 7,147 different time series records. The study used MATLAB tool and found that Ensemble (FFT-MLE) model showed better prediction time and accuracy rate (88.75-95.30%) when compared to other models. Study by Pouriyeh et al. [38] predicted heart disease using ensemble model. Cleveland data set for heart diseases, containing 303 instances, 76 attributes were used for the purpose of study. The traditional classifiers like Naive Bayes (NB), K-Nearest Neighbor (K-NN), Radial Basis Function (RBF), Decision Tree (DT), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Single Conjunctive Rule Learner (SCRL) and ensemble approach (i.e hybrid approach) such as boosting, bagging and stacking was tested and compared. The study found that boosting and SVM classifier offered a better result than the above-mentioned techniques. When compared with traditional prediction models using validated EHR, machine learning algorithms did not improve prediction rate of hospital readmissions cases. The study Le-Dong et al [28] tested 3 nested prediction models which included the clinical, hemodynamic, functional predictors, MR volumetry markers and 3D three-dimensional motion. The study reported that their model 3 which included all above mentioned multivariable prediction outperformed other models. Dataset was obtained from the Cleveland Clinic Foundation heart disease which includes 297 instances and 13 attributes. The proposed model was tested using a 10-fold cross-validation procedure. The proposed model when compared with other decision tree models such as Decision Tree with Reduced Error Pruning Method and A combination of Naïve Bayes, Decision Tree and SVM showed accuracy of 86.53%, specificity 95.5% and sensitivity of 86.5% [29]. The CAD model used Least Squares SVM (LS-SVM) classifier which was when compared to other classifiers such as ANN, and Hidden Markov Models (HMM) showed accuracy, sensitivity and specificity of 95.39%, 96.59% and 93.75% respectively [30].A comparative study by Acharya [39] on heart disease prediction, used Heart Disease Data Set from UCI machine learning repository. The study tested performance of the following classifiers: NB, KNN, ANN, Adaboost, SVM and Random Forest method whereas the experimental results show that the SVM and ANN provided better results for heart disease prediction. Study by Papini et al. [40] estimated quality of pulses using PPG. The dataset consists of 8 min recordings of 42 patients (29 pediatric, 13 adults); two recordings of 1 h each from two different pediatric patients in an intensive care unit. The study tested the performance of Multi-Layer Perceptron (MLP) neural network classifier and reported a sensitivity of above 96%. Study by Le-Dong et al. [41] attempted to predict asthma based on patient's telemonitoring data. The dataset consisted of 7001 daily telemonitoring records of adult asthma patients for 7 days. Classifiers such a naive Bayesian classifier, adaptive Bayesian network, and SVM was tested for performance and found that adaptive Bayesian network showed accuracy, sensitivity and specificity of 100% each respectively.The study Chen et al. [33] reported that SVM with z score showed accuracy 84%, sensitivity 60% and specificity of 96% respectively. Study by Wiemken et al. [42] analysed statistical and machine learning algorithm in prediction re-hospitalisation within 30 days among pneumonia patients. Datasets were obtained from hospital which included 3249 patients suffering from pneumonia. The study tested the following classifiers: LR, LASSO regression, RF, RPT, CIT and NB. The study reported that it is a challenge to predict re-hospitalization among pneumonia patients using statistical and MLA. Study by Saleh et al. [43] predicted COPD using machine learning algorithm and compared its performance against classifiers decision tree ID3 and C4.5, naive Bayes and Bayesian network (K2). WEKA tool was used for validation. The Pouriyeh et al [44] study concluded that neural network or deep learning methods have increased accuracy rate when compared to other classifiers. The different classifiers tested were Logistic regression, ridge regression, AdaBoost and decision tree models. Out of the above-mentioned classifiers Ridge regression achieved 92% sensitivity, 90% specificity and 87% accuracy. In a study Finkelstein and Jeong [41] four classifiers, SVM, NB, Random Forest and KNN were tested against traditional Liver Fibrosis Index (LFI) method and linear regression models. The study concluded that Random Forest showed maximum accuracy rate of 82.87%, SVM with a sensitivity of 92.97% and NB showed specificity of 82.50%. Overall the study proved that machine learning algorithm outperformed traditional prediction models in liver disease prediction.

Table VII: Comparison of the existing literature

Author and Year	Source of data	Feature Selection/ Data Mining application
[66]	FHS (Framingham Heart Study)	Decision tree, Naïve Bayes, Supp vector machine (SVM) and Artificial ne network (ANN)
[44]	Cleveland heart disease dataset 2016 with records of 297 patients, 13 features used	Feature selection using Relief, Minimal-Redundancy-Maximal-Relevance (mRMR), And Least Absolute Shrinkage and Selection Operator (LASSO)
[45]	UCI Machine Learning Repository Dataset (303 records) and Statlog Database for Heart disease prediction (270 records) and 13 similar features. After removal of missing values, finally, 566 instances was included	Waikato Environment for Knowledge Learning (WEKA)- Naïve Bayes, Support Vector Machine, Simple Logistic Regression, Random Forest & Artificial Neural Network (ANN)
[46, 47]	Cleveland Heart Disease database: 303 records and 19 attributes used	Naïve Bayesian, decision tree with Information Gain

[48]	University of California Irvine (UCI) machine learning repository: 313 instances with 14 attributes used	Classification Association Rules (CARs) using associative algorithm Like Apriro and FP-Growth
[49]	m Cleveland database of repository and 13 attribu used	WEKA- decision tree (J48); Logistic Modeling Tree (LMT); RF
[50]	Structured data – Strategic selection of 500 data from UCI Machine Learning Repository based on Pima Indian population : 768 instances, 8 attributes	Associative classification
[51]	Structured data - Pima Indian diabetes dataset: 768 instances, 8 attributes	Feature selection was done using Boruta wrapper algorithm
[52]	Structured data - Tokyo Women"s Medical University Hospital free datasets 779 patients and 164 variables	Supervised Learning Approaches
[54]	Structured data – Pima Indians Diabetes from UCI repository with 768 instances, 8 attributes	Logistic Regression, SVM, Naïve Bayes, Decision Tree and Random Forest
[55]	Structured data – strategic selection of 10,000 patients record from HCUP Natic Inpatient Sample database	Logistic Regression And Svm
[56]	Structured data - Pima Indian Diabetes Dataset attributes and 768 Instances	C4.5, SVM, nearest neighbor (k-NN), Prototype NN (PNN), and Binary Logistic Regression (BLR).
[57]	Structured data – Strategic selection of 10814 type 2 diabetes patient data collected based on self-monitored glucose (SMBG) from cl trial	Support vector Machine
[58]	Structured data - UCI ma learning repository with 9 Attributes	Naive Bayes, J48(C4.5) JRip ,Neural networks, Decision trees, KNN, Fuzzy logic and Genetic Algorithms
[59]	Structured data - UCI based machine learning Dataset.	Machine Learning Models

In an analysis of cancer prediction, the different classifiers tested were NB Bayesian, RBF Neural Network, MLP, DT and C4.5 (J48) algorithm using WEKA tool. The study reported that NB outperforms other prediction models.Study by Bryanton et al. 62 predicted lung cancer recurrence within two years posts radiotherapy using trained MLA neural network. Dataset was obtained from Ottawa hospital after inclusion, exclusion criteria CT and PET records of 161 patients were used for testing. Trained neural network classifier showed accuracy, sensitivity and specificity of 79%, 60% and 88% respectively. Another research by Bychkov et al. [63] predicted colorectal cancer using digitalized tumour samples and machine learning. The dataset consists of Tumor Tissue Microassay (TMA) of 420 colorectal cancer patients. The study tested performance of convolutional neural networks and Long Short-Term Memory networks. The study concluded that novel deep learning models or neural networks could act as biomarkers in colorectal cancer. A survey study by Sharma et al. [64] analyzed different MLA techniques in predicting the survivability rate of cancer patients. The different classifiers reviewed were a neural network, SVM, naïve bayes and decision tree and concluded that naïve bayes is a suitable to model for prediction of various cancer when compared to other classifier models. Research by Xiao et al. [65] predicted prostate cancer using random forest algorithm. Dataset consisted of clinical, demographical data, transrectal ultrasound findings of 941 patients with prostate cancer. The tested classifier Random Forest showed accuracy, sensitivity and specificity of 83.10%, 65.64% and 93.83% respectively. The researcher tested performance of ensemble-based prediction model RF, NB tree and Logistic Model (LM) Tree with that of RF and NB tree. The study found that accuracy of ensemble model increased by 92% in prediction of diabetes mellitus. Research by Alehegn and Joshi [66] used ensemble model in prediction of diabetes mellitus. A total of 768 instances from Pima Indian Diabetes Data Set were used for analysis. The study used WEKA and JAVA tool to test the performance of ensemble hybrid model combining KNN, NB, Random forest, and J48 with individual classifiers. The study reported that ensemble models show high accuracy when compared to individual machine learning algorithms.

Table VIII: Summary of applications and advantages of the techniques.

Author and Year	Feature Selection/ Data Mining application	Advantages
[67]	Decision tree, Naïve Bayes, Support vector machine (SVM) and Artificial neural network (ANN)	Helps identify features causing heart disease and features for predicting deaths due to heart disease
[46]	Feature selection using Relief, Minimal-Redundancy-Maximal-Relevance (mRMR), and Least Absolute Shrinkage and Selection Operator (LASSO)	Assist the doctors to diagnosis heart patients efficiently. Feature selection should be used before classification to improve the classification accuracy of classifiers and reduce the computation time.
[47]	Waikato Environment for Knowledge Learning (WEKA)- Naïve Bayes, SVM, Simple Logistic Regression, Random Forest & ANN	Bigger dataset and higher accuracy; Real-time application using Android-based apps; Useful for both patients and doctors in any part of globe Application to other disease prediction as well

[49]	Naïve Bayesian, decision tree with Information Gain	The study recommends use of Selective Naïve Bayes classifier, i.e., removal of unnecessary and irrelevant attributes and including only relevant information using C4.5
[50]	Classification Association Rules (CARs) using an associative algorithm like Apriori and FP-Growth And performance evaluation using Naive bayes, ZeroR, OneR, J48, IBk and k-nearest neighbor	hybrid technique for CARs help to achieve high accuracy (99.19%) in predicting heart disease
[51]	WEKA- decision tree (J48); Logistic Modeling Tree (LMT); RF	The LMT algorithm appears to perform better on data sets with numerous numerical attributes
[52]	Supervised Learning Approaches	Better prediction than traditional single classifiers
[54]	Logistic Regression, SVM, Naïve Bayes, Decision Tree and Random Forest	help diabetic patients lead a better life. Early prediction shall help in a significant reduction of effects
[55]	Logistic Regression And Svm	help Clinicians make a better decision about their patient's disease status.
[57]	Support vector Machine	The advantage for resource planning and admission schedules. Planning bed usage, the requirement for specialists, health insurance schemes, reimbursement from private sector, planning discharge dates
[58]	C4.5, SVM, nearest neighbour (k-NN), Prototype NN (PNN), and Binary Logistic Regression (BLR).	Prediction of diabetes
[59]	Machine Learning Models	hypoglycemia prediction models provide a forecast for time left for treatment initiation
[60]	Naive Bayes, J48(C4.5) JRip ,Neural networks, Decision trees, KNN, Fuzzy logic and Genetic Algorithms	Low sample size (n=15);

A. Main Findings

From the above studies we can conclude that for prediction of mental health-related diseases using MRI image, SVM classifier has outperformed other types of machine learning classifiers in terms of accuracy, sensitivity and specificity. In case of lung disease prediction, it is understood that neural network and SVM outperformed other prediction models. Further, we infer that ensemble models combining various classifiers are used in prediction of diabetes mellitus. For the prediction of cancer in human anatomy it is identified that Naïve Bayes and neural network are better prediction models for cancer disease. Neural Network and SVM classifiers are best to predict heart disease from UCI machine learning repository and similar heart disease datasets.

B. Future Recommendation

From the review, it is understood that many studies have used long follow up duration of more than five years 13,25. Study by Schnack 14 reported that white box algorithms such as simple decision tree is easy to interpret but shows large variance; however, black box algorithms such as ANN and RF results are difficult to interpret. The same study reported that Gaussian distribution helps to identify subjects lying outside of normal distribution hence easy to identify these outlying subjects as 'patients' however it does not help to determine the type of disease nor recommends for personalized treatment. When considering a large unbalance data amongst various features of numerous modalities and different input data types, the multiple kernels are utilized [14]. Several studies have reported their limitations such as performing an evaluation using smaller datasets [12,23,38], not externally validated [34,37,41] interference with results 17. Studies have also reported that complex models are more prone to overfitting and result in poor interpretation [13,25]. Complexity in prediction based on structured data obtained from clinical records [68–70]. As the size of dataset increases, the accuracy rate decreases [22]. Absence of comparative analysis and prediction on clinical outcomes using real patient data retrieved from the hospital and medical research institutions 19. Comparison with a limited number of classifiers41, imbalanced dependent variable 41. Inappropriate choice of subject [30]. The developed model may not be applicable in other countries [30]. Insufficient dataset to compare and test [30,34] and not performing an evaluation on continuous datasets [35].

V. CONCLUSION

In this review, the overview of research in big data analytics as specific towards machine learning approach is provided. Moreover, the experimental and theoretical features in large-scale data-intensive fields specifically; medical application for disease prediction is discussed. Big data analytics leads to several challenges compared with conventional machine learning approaches. The challenges emerged through big data were adaptability, usability and scalability. Due to drawbacks related to big data provides significant transformation towards machine learning approaches in order to address several technical challenges which rely on real-world impacts. The challenges and opportunities lead to promotion towards future direction of research. The machine learning performs effectively for a huge volume of medical data w2ith consequent data representation through unsupervised data collection technique. This technique offers an effective tool to deal with big data analytics which involves examination of a large number of data which can be unsorted and unsupervised learning approach. Further, a comparison between SVM, ANN and Deep Learning reveals that SVM and ANN have a higher rate of accuracy as compared to deep learning which exhibits a higher rate of accuracy only in linear datasets. This review article presented an in-depth explanation about machine learning algorithm along with an examination of issues related to machine learning algorithm.

Further, this article presented certain remedies to overcome certain challenges related to machine learning algorithm. In future, we planned to adopt certain solution to resolve certain issues for an uncertain and incomplete dataset for prediction of Parkinson telecommunication dataset. Further focusing on recent trend will be interesting and big data leads to several issues were machine learning, and big data are considered as a focused area for recent trends.

REFERENCES

1. Jamison D, Breman J, Measham A. Disease Control Priorities in Developing Countries. In: The International Bank for Reconstruction and Development. New York: Oxford University Press; 2006.
2. Winters-Miner LA. Seven ways predictive analytics can improve healthcare: Medical predictive analytics have the potential to revolutionize healthcare around the world. Elsevier. 2014.
3. Omachonu VK, Einspruch NG. Innovation in Healthcare Delivery Systems: A Conceptual Framework. Innov J Public Sect Innov J. 2010;15:1–20.
4. Fatima M, Pasha M. Survey of Machine Learning Algorithms for Disease Diagnostic. J Intell Learn Syst Appl. 2017;09:1–16.
5. Assistant Secretary for Planning and Evaluation. Improving Data for Decision Making: HHS Data Collection Strategies for a Transformed Health System. 2011.
6. Bhatt C, Dey N, Ashour AS. Internet of Things and Big Data Technologies for Next Generation Healthcare. London: Springer; 2017.
7. IBM. Decision Tree Models. 2012.
8. Srinivas S, Sarvadevabhatla RK, Mopuri KR, Prabhu N, Kruthiventi SSS, Babu RV. A Taxonomy of Deep Convolutional Neural Nets for Computer Vision. 2016;
9. Danjuma K, Osofisan A O. Evaluation of Predictive Data Mining Algorithms in Erythemato-Squamous Disease Diagnosis. 2015;10.
10. Dodek PM, Wiggs BR. Logistic regression model to predict outcome after in-hospital cardiac arrest: validation, accuracy, sensitivity and specificity. Resuscitation. 1998;36:201–8.
11. L'Heureux A, Grolinger K, Elyamany HF, Capretz MAM. Machine Learning With Big Data: Challenges and Approaches. IEEE Access. 2017;5:7776–97.
12. Salvador R, Radua J, Canales-Rodríguez EJ, Solanes A, Sarró S, Goikolea JM, et al. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. Hu D, editor. PLoS One. 2017;12:e0175683.
13. de Wit S, Ziermans TB, Nieuwenhuis M, Schothorst PF, van Engeland H, Kahn RS, et al. Individual prediction of long-term outcome in adolescents at ultra-high risk for psychosis: Applying machine learning techniques to brain imaging data. Hum Brain Mapp. 2017;38:704–14.
14. Schnack HG. Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). Schizophr Res. 2017;
15. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease Prediction by Machine Learning over Big Data from Healthcare Communities. IEEE Access [Internet]. 2017;5:8869–79. Available from: http://mmlab.snu.ac.kr/~mchen/min_paper/2017/2017-IEEE-Access-1-DiseasePrediction.pdf
16. Zhao Y, Healy BC, Rotstein D, Guttmann CRG, Bakshi R, Weiner HL, et al. Exploration of machine learning techniques in predicting multiple sclerosis disease course. Ramagopalan S V., editor. PLoS One. 2017;12:e0174866.
17. Abos A, Baggio HC, Segura B, García-Díaz AI, Compta Y, Martí MJ, et al. Discriminating cognitive status in Parkinson's disease through functional connectomics and machine learning. Sci Rep. 2017;7:45347.
18. Kim HC, Jo I-J, Sung JM, Chang H-J. Abstract P202: Cardiovascular Risk Prediction Using Machine-learning Methods in the Middle-aged Korean Population. Circulation. 2017;135:AP202.
19. Bhatt A, Dubey SK, Bhatt AK, Joshi M. Data Mining Approach to Predict and Analyze the Cardiovascular Disease. In 2017. p. 117–26.
20. Acharya UR, Fujita H, Adam M, Lih OS, Sudarshan VK, Hong TJ, et al. Automated characterization and classification of coronary artery disease and myocardial infarction by decomposition of ECG signals: A comparative study. Inf Sci (Ny). 2017;377:17–29.
21. Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. Comput Methods Programs Biomed. 2016;130:54–64.
22. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Liu B, editor. PLoS One. 2017;12:e0174944.
23. Lafta R, Zhang J, Tao X, Li Y, Abbas W, Luo Y, et al. A Fast Fourier Transform-Coupled Machine Learning-Based Ensemble Model for Disease Risk Prediction Using a Real-Life Dataset. In: Kim J, Shim K, Cao L, Lee J, Lin X, Moon Y, editors. Advances in Knowledge Discovery and Data Mining PAKDD 2017 Lecture Notes in Computer Science. Springer, Cham; 2017. p. 654–70.
24. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure. JAMA Cardiol. 2017;2:204.
25. Dawes TJW, de Marvao A, Shi W, Fletcher T, Watson GMJ, Wharton J, et al. Machine Learning of Three-dimensional Right Ventricular Motion Enables Outcome Prediction in Pulmonary Hypertension: A Cardiac MR Imaging Study. Radiology. 2017;283:381–90.
26. Aljaaf AJ, Al-Jumeily D, Hussain AJ, Dawson T, Fergus P, Al-Jumaily M. Predicting the likelihood of heart failure with a multi level risk assessment using decision tree. In: 2015 Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE). IEEE; 2015. p. 101–6.
27. Zheng Y, Guo X, Qin J, Xiao S. Computer-assisted diagnosis for chronic heart failure by the analysis of their cardiac reserve and heart sound characteristics. Comput Methods Programs Biomed. 2015;122:372–83.
28. Le-Dong N-N, Hua-Huy T, Ngoc HMN, Martinot J-B, Xuan A-TD. Detection of Interstitial Lung Disease in Systemic Sclerosis Using a Machine Learning Approach Based on Pulmonary Function Tests. Am J Respir Crit Care Med. 2017;195:A2531.
29. Spathis D, Vlamos P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. Health Informatics J. 2017;
30. Yip TC-F, Ma AJ, Wong VW-S, Tse Y-K, Chan HL-Y, Yuen P-C, et al. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. Aliment Pharmacol Ther. 2017;46:447–56.
31. Hashem S, Esmat G, Elakel W, Habashy S, Abdel Raouf S, Elhefnawi M, et al. Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients. IEEE/ACM Trans Comput Biol Bioinforma. 2017;1–1.
32. Saritha B, Manaswini N, Hiranmayi D, Ramana SV, Priyanka R, Eswaran K. Classification of Liver Data using a New Algorithm. Int J Eng Technol Sci Res. 2017;4:330–4.
33. Chen Y, Luo Y, Huang W, Hu D, Zheng R, Cong S, et al. Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B. Comput Biol Med. 2017;89:18–23.
34. Guadagni F, Zanzotto FM, Scarpato N, Rullo A, Riondino S, Ferroni P, et al. RISK: A Random Optimization Interactive System Based on Kernel Learning for Predicting Breast Cancer Disease Progression. In: Rojas I, Ortuño F, editors. Bioinformatics and Biomedical Engineering IWBBIO 2017 Lecture Notes in Computer Science. Springer, Cham; 2017. p. 189–96.
35. Murty NVR, Babu PMSP. A Critical Study of Classification Algorithms for LungCancer Disease Detection and Diagnosis. nternational J Comput Intell Res. 2017;13:1041–8.
36. López B, Torrent-Fontbona F, Viñas R, Fernández-Real JM. Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction. Artif Intell Med. 2017;
37. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. Liu B, editor. PLoS One. 2017;12:e0179805.
38. Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. In: 2017 IEEE Symposium on Computers and Communications (ISCC). Greece: IEEE; 2017. p. 204–7.
39. Acharya A. Comparative Study of Machine Learning Algorithms for Heart Disease Prediction. Helsinki Metropolia University of Applied Sciences; 2017.

40. Papini GB, Fonseca P, Aubert XL, Overeem S, Bergmans JWM, Vullings R. Photoplethysmography beat detection and pulse morphology quality assessment for signal reliability estimation. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2017. p. 117–20.
41. Finkelstein J, Jeong I cheol. Machine learning approaches to personalize early prediction of asthma exacerbations. Ann N Y Acad Sci. 2017;1387:153–65.
42. Wiemken TL, Furmanek SP, Mattingly WA, Guinn BE, Cavallazzi R, Fernandez-Botran R, et al. Predicting 30-day mortality in hospitalized patients with community-acquired pneumonia using statistical and machine learning approaches. J Respir Infect. 2017;1:50–56.
43. Saleh L, Mcheick H, Ajami H, Mili H, Dargham J. Comparison of Machine Learning Algorithms to Increase Prediction Accuracy of COPD Domain. In: Mokhtari M, Abdulrazak B, Aloulou H, editors. Enhanced Quality of Life and Smart Living ICOST 2017 Lecture Notes in Computer Science. Switzerland: Springer, Cham; 2017. p. 247–54.
44. Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. In: 2017 IEEE Symposium on Computers and Communications (ISCC) [Internet]. IEEE; 2017. p. 204–7. Available from: http://ieeexplore.ieee.org/document/8024530/
45. Masih N, Ahuja S. Prediction of Heart Diseases Using Data Mining Techniques. Int J Big Data Anal Healthc [Internet]. 2018;3:1–9. Available from: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJBDAH.2018070101
46. Haq AU, Li JP, Memon MH, Nazir S, Sun R. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Mob Inf Syst [Internet]. 2018;2018:1–21. Available from: https://www.hindawi.com/journals/misy/2018/3860146/
47. Nashif S, Raihan MR, Islam MR, Imam MH. Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. World J Eng Technol [Internet]. 2018;06:854–73. Available from: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/wjet.2018.64057
48. [48] Kumar Sen S. Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms [Internet]. International Journal Of Engineering And Computer Science. 2017. Available from: https://www.researchgate.net/publication/317486673_Predicting_and_Diagnosing_of_Heart_Disease_Using_Machine_Learning_Algorithms
49. [49] Nikhar S, Karandikar AM. Prediction of Heart Disease Using Machine Learning Algorithms. Int J Adv Eng Manag Sci [Internet]. 2016;2:617–21. Available from: https://media.neliti.com/media/publications/239484-prediction-of-heart-disease-using-machin-4b2e96d4.pdf
50. Singh J, Kamra A, Singh H. Prediction of heart diseases using associative classification. In: 2016 5th International Conference on Wireless Networks and Embedded Systems (WECON) [Internet]. IEEE; 2016. p. 1–7. Available from: http://ieeexplore.ieee.org/document/7993480/
51. Patel J, Tejalupadhyay S, Patel S. Heart Disease prediction using Machine learning and Data Mining Technique [Internet]. 2016. 129–137 p. Available from: https://www.researchgate.net/publication/309210947_Heart_Disease_prediction_using_Machine_learning_and_Data_Mining_Technique
52. Nnamoko N, Hussain A, England D. Predicting Diabetes Onset: An Ensemble Supervised Learning Approach. In: 2018 IEEE Congress on Evolutionary Computation (CEC) [Internet]. IEEE; 2018. p. 1–7. Available from: https://ieeexplore.ieee.org/document/8477663/
53. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach [Internet]. Applied Computing and Informatics. 2018 [cited 2019 Mar 6]. Available from: https://linkinghub.elsevier.com/retrieve/pii/S221083271830365X
54. Rahman T, Farzana SM, Khanom AZ. Prediction of Diabetes Induced Complications Using Different Machine Learning Algorithms [Internet]. BRAC University; 2018. Available from: http://dspace.bracu.ac.bd/xmlui/bitstream/handle/10361/10945/15101128_CSE.pdf?sequence=1&isAllowed=y
55. Joshi TN, Chawan PM. Logistic Regression And Svm Based Diabetes Prediction System. Int J Technol Res Eng [Internet]. 2018;5:4347–50. Available from: https://www.researchgate.net/publication/326416823_LOGISTIC_REGRESSION_AND_SVM_BASED_DIABETES_PREDICTION_SYSTEM
56. Santhanam T, Padmavathi MS. Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. Procedia Comput Sci [Internet]. 2015;47:76–83. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1877050915004536
57. Morton A, Marzban E, Giannoulis G, Patel A, Aparasu R, Kakadiaris IA. A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay among Diabetic Patients. In: 2014 13th International Conference on Machine Learning and Applications [Internet]. IEEE; 2014. p. 428–31. Available from: http://ieeexplore.ieee.org/document/7033154/
58. Radha P, Srinivasan B. Predicting Diabetes by cosequencing the various Data Mining Classification Techniques. Int J Innov Sci Eng Technol [Internet]. 2014;1:334–9. Available from: http://ijiset.com/v1s6/IJISET_V1_I6_55.pdf
59. Sudharsan B, Peeples M, Shomali M. Hypoglycemia Prediction Using Machine Learning Models for Patients With Type 2 Diabetes. J Diabetes Sci Technol [Internet]. 2015;9:86–90. Available from: http://journals.sagepub.com/doi/10.1177/1932296814554260
60. Kumar V, Velide L. A Data Mining Approach For Prediction And Treatment Ofdiabetes Disease. Int J Sci Invent Today. 2014;3:73–9.
61. Singh A. Diabetes Disease Prediction System Using C4.5 and FCM algorithm. Int J Contemp Technol Manag [Internet]. 2018;7:1–7. Available from: http://www.ijctm.in/Admin/upload/ijctm-07-09-1201865.pdf
62. Bryanton M, Pathak R, Russa D La, Holmes O, Gotfrit R, Cook G, et al. Predicting lung cancer recurrence in patients within two years of curative radiotherapy via a trained machine learning algorithm. J Nucl Med. 2017;58:113.
63. Bychkov D, Turkki R, Haglund C, Linder N, Lundin J. Abstract 5718: Outcome prediction in colorectal cancer using digitized tumor samples and machine learning. Cancer Res. 2017;77:5718–5718.
64. Sharma A, Karthik GS, Mittal N, Sindhu VL, Pradeep KR. A Survey on Predictive Analysis of Cancer Survivability Rate Using Machine Learning Algorithm. 2017;271–8.
65. Xiao L-H, Chen P-R, Gou Z-P, Li Y-Z, Li M, Xiang L-C, et al. Prostate cancer prediction using the random forest algorithm that takes into account transrectal ultrasound findings, age, and serum levels of prostate-specific antigen. Asian J Androl. 2017;19:586.
66. Alehegn M, Joshi R. Analysis and prediction of diabetes diseases using machine learning algorithm : Ensemble approach. Int Res J Eng Technol. 2017;4:426–36.
67. Masih N, Ahuja S. Prediction of Heart Diseases Using Data Mining Techniques: Application on Framingham Heart Study. Int J Big Data Anal Healthc. 2018;3:1–9.
68. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. IEEE Access [Internet]. 2017;5:8869–79. Available from: http://ieeexplore.ieee.org/document/7912315/
69. Shi F, Zhang Q, Chen J, Karimi HR. Macroscopic Expressions of Molecular Adiabatic Compressibility of Methyl and Ethyl Caprate under High Pressure and High Temperature. Abstr Appl Anal. 2014;2014:1–10.
70. Raja M, Ali MS. An analysis of consumer perception towards retail brands in Big Bazaar Chennai. Indian J Appl Res. 2014;4:1–3.
71. PM Kumar PM, Gandhi UD A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. Computers & Electrical Engineering. 2018; 65: 222-235
72. Gokulnath V, Usha Devi G. A Review on Classification Algorithms of Medical Diagnostics Research Journal of Pharmaceutical, Biological and Chemical Sciences. 2017; 2656-2660
73. Usha Moorthy, Usha Devi Gandhi. A Survey of Big Data Analytics Using Machine Learning Algorithms, IGI Global. 2018; 95-123
74. V Vijayakumar, MK Priyan, G Ushadevi, R Varatharajan, Gunasekaran Manogaran, Prathamesh Vijay Tarare. E-health cloud security using timing enabled proxy re-encryption, Mobile Networks and Applications, 2019; 3: 1034-1045.

AUTHORS PROFILE

Usha Moorthy is pursuing her PhD in the School of Information Technology and Engineering, Vellore Institute of Technology University. She received her Master of Science in Software Engineering from VIT University her current research interests include big data analytics, Data Mining and Cryptography. She has published a number of international journals and conferences.

Usha Devi Gandhi is working as an Associate Professor in the School of Information Technology and Engineering, Vellore Institute of Technology University. She received her Bachelor of Engineering and Master of Engineering degree from the Anna University. Her current research interest includes big data analytics and wireless networks. She has published a number of international journals and conferences. She is a member of CSI and IEEE.