International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 4, August 2020, pp. 4217~4225
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i4.pp4217-4225

A comprehensive study on disease risk predictions in machine learning

G. Saranya, A. Pravin
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, India

Article Info

Article history:

Received Oct 4, 2019
Revised Feb 22, 2020
Accepted Feb 29, 2020

Keywords:

Breast cancer
Disease prediction
Heart disease
Machine learning
Prediction models

ABSTRACT

Over recent years, multiple disease risk prediction models have been developed. These models use various patient characteristics to estimate the probability of outcomes over a certain period of time and hold the potential to improve decision making and individualize care. Discovering hidden patterns and interactions from medical databases with growing evaluation of the disease prediction model has become crucial. It needs many trials in traditional clinical findings that could complicate disease prediction. A Comprehensive study on different strategies used to predict disease is conferred in this paper. Applying these techniques to healthcare data, has improvement of risk prediction models to find out the patients who would get benefit from disease management programs to reduce hospital readmission and healthcare cost, but the results of these endeavors have been shifted.

Corresponding Author:

G. Saranya,
Research Scholar, Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology,
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai-600 119, Tamilnadu, India.
Email: saranyag3@srmist.edu.in

1. INTRODUCTION

Machine learning is one of the most prevalent methods used in multiple computer engineering areas and has been commonly used in the processing of natural language, picture processing, pattern recognition, cyber security, and multiple areas. One of the most dynamic researches of machine learning is healthcare industry [1]. As healthcare firms are trying to collect records of patients, estimates show that there are about 1 trillion bytes of information, which is increasing every day. These data need to be properly extracted to obtain precious information [2].

Sometimes patients fail to define their medical problems correctly and the results of laboratory research can lead to some degree of mistake. Specialists find it difficult to make decisions about the illnesses because they may not have skills in all areas. To address this issue, it is necessary to develop a disease prediction system that combines medical knowledge with an integrated system to produce the biggest results and can help society [3].

Several earlier investigations tried to use patient laboratory tests [4-6] and drugs [7] to predict the occurrence of disease. Such prototypes were also used to define unknown risk factors, often while enhancing sensitivity and specificity of detection simultaneously. Recent studies have been effective in predicting disease through multiple methods, including supporting vector machines [8-10], logistical regression [11], random forests [3, 12], neural networks [4, 13], and time series modeling techniques [14].

In summary, the paper focuses on various disease prediction models of machine learning. The paper is structured in the following way. Section 2 explains the basic concept of machine learning, prediction models and types of prediction models. Section 3 explains the survey of disease prediction models. Section 4 explains the comparative study of heart diseases. Section 5 summarizes a comparative study of breast cancer. Section 6 gives the conclusion and future enhancement.

2. INTRODUCTION TO MACHINE LEARNING

Machine learning (ML) is an artificial intelligence (AI) branch that helps analyze the data structure and fit the information into models correctly. It is one of the computer science fields, and differs from other computing technology by the way of training the computers based on data inputs and it uses statistical analysis to get the proper output. For this reason, ML is used in automated decision-making models like Facial recognition, Recommendation engines, OCR and Self driving car applications.

Machine learning methods are categorized into three classifications according to the training processes used. The categories are supervised machine learning, unsupervised machine learning and reinforcement learning. In supervised learning, the data samples with category labels are used in training. Classification and regression models are examples of supervised learning. The algorithms used in this approach are decision tree, naïve bayes etc. In unsupervised learning, the data samples are directly used in training without category label. Clustering techniques and encoders are the basic examples of unsupervised approach. Reinforcement learning is a mixture of prior two approaches. It uses agent that finds the correct action to achieve the overall goal of the application [15].

3. PREDICTION MODELS

A prediction model is characterized as a model that provides a way to evaluate the individual danger of a patient for the outcome of a disease. The question of when, what and how to use these models arises with the growth of such prediction models. These models can be taught over time, providing the demands of the company, to react to new information or views.

Two types of prediction models exist. They are models of classification predicting class outcomes, and models of regression predicting the relationship between a response variable Y and a predictor variable X. Various basic and advanced algorithm, listed in Figure 1, Figure 2 conducts data analysis and statistical analysis and determine information trends and patterns. While machine learning and prediction analysis can provide an opportunity for any implementation, the haphazard implementation of these options will drastically impede their ability to provide insight into the demands of the organization without considering how they fit into everyday operations. Organizations need to guarantee that they have the architecture in place to support these alternatives, as well as high-quality information to feed them and assist them learn, to make the most of prediction analytics and machine learning [16].

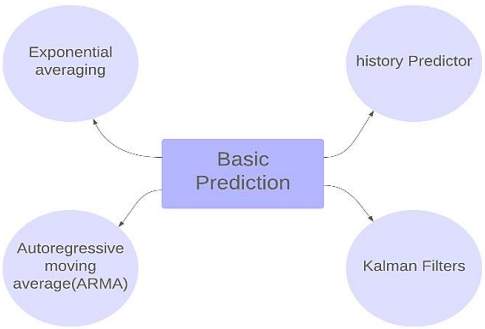

Figure 1. Basic prediction techniques

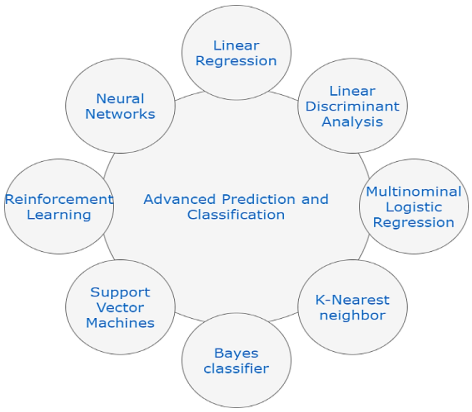

Figure 2. Advanced prediction techniques

4. EXISTING TECHNIQUES FOR DISEASE PREDICTION

Mingyu Pak, Miyoung Shin [17], explored few environmental variables linked to type-2 diabetes disease and selected some variables to develop an analytical prototype of prediction of disease. For the choice of important variables, they first pre-processed all external environmental factors into numerical data and then estimated the maximum/minimum probability ratios of all the sorted exogenous variables. The top-n positioned variables were then chosen as input variables for the forecast model depending on ansan/ansung cohort 2 data collected from the Korean National Institute of Health (KNIH), the disease risk factor prediction model was created using SVM. Their prediction model showed the output of 65.97 percent precision and showed very identical performance only with genetic factors with particular environmental variables to the model [17].

The research shown in [18] focuses on disease forecasting from medical information supplied by New York's Presbyterian Hospital. Since these are medical information, predicting computerized predictions is generally unique and simpler than forecasting user text inputs. Nicolae Dragu et al discussed-on forecast of serious contagious diseases from web material sources, which is also a specific origin where open clinical terms are used [19]. Many attempts have been made to forecast selective diseases [20, 21]. For instance, the authors in [20] deal with the prediction by mining content of coronary heart disease. There are also an important amount of study works on medicinal services debates that have been carried out. Lav petrov used NLP to evaluate and break down user remarks to predict disease and concentrate uncommon responses to drugs [22].

Ryan McDonald et al, use a terminology-laden interface (i.e. clients need to explore a long list of hints). It's an awkward job from the user's view, and the operation is also tedious. Also, if the customers do not find a certain indication, they are forced to prevent that symptom that is not in any manner desired [23]. Client's directed text input [24, 25]. They rely on negligible symptom-disease relationship system [26, 27] in any situation and use complete content database. These frameworks begin to search for accurate term match at the input of the client from each input line in the database. For example, if a client with symptoms equivalent is not specified in the database, then we cannot match the information correctly. If the user input should contain a greater number of semi-technical terms than anticipated, it will degrade its performance. The system used is especially strong and restricted to particular data types.

Xiaoyan Wang et al, proposed an automated system for disease prediction that based on the client's driven feedback. Their structure requires input from the client such as the names of symptoms and certain significant parameters and provides a list of likely illnesses (maximum infections are more likely to occur). The accuracy of the automated disease prediction scheme (ADPS) evaluated the solution of the current scheme with a standard 14.35 percent greater accuracy in examination [28]. S Manimekala et al, suggested that the Automatic Disease Prediction technique determine the most probable disease based on the feedback of the patient that facilitates early detection. The model uses data mining algorithm apriori-frequent pattern (A-FP) to identify illnesses through health data mining focused on the signs of input. This method is used for finding medical datasets from which to create association rules. The goal is to identify from the health dataset relevant and frequent illnesses [29].

Cristinel Ababei et al, discussed the overall context of computer frameworks on multicore processor systems, they provided a discussion on the most common prediction and classification methods. They introduced some prediction systems from simple to more complicated, while highlighting one frequent basic theme: each of these systems misuses the prior history of the variable of interest [30]. Ankita Dewan et al, in conjunction with the back-propagation method, they suggested a successful genetic algorithm for predicting heart disease and created a model that could identify and extract unknown data (patterns and relationships) linked to coronary heart disease from archived database records of heart disease [31]. Table 1 shows the different methods and their advantages and disadvantages. The merits and demerits are specified in terms of efficiency, prediction accuracy and how easily a model could be implemented.

Table 1. Merits and demerits of prediction techniques

Methods	Pros	Cons
KNN	Simple, high accuracy	Local information structure sensitivity. Calculated slowly.
Linear Regression	Accuracy of good prediction	Training needs. When fresh information arrive, coefficients may need to be updated constantly.
SVM	Practically efficient, very popular	The size of the training information can improve the number of bias functions. Selection of parameters depends on information.
Bayes classifier	high accuracy, efficient	Training needs
LDA	Fast, Implementation is comparatively simple	Training needs

4.1. Description of existing techniques for heart disease prediction

P. K. Anooj [32], suggested a decision-making scheme based on fuzzy rules to assess the level of risk of heart disease. He first pre-processed the information for missing values in his method. He then performed fuzzy sets generation and created a fuzzy decision-making system. The technique selects the required attribute based on the number of events in the database to create weighted fuzzy laws. These weighted fuzzy rules are then used to create a decision-based scheme of assistance. He tested his suggested scheme on three distinct dataset kinds that are collected from V.A. Cleveland dataset. Medical centre with 303 cases of training data for 202 documents and test information for 101 documents. P. K. Anooj contrasted his model with models based on neural networks and obtained the greatest precision [32].

Tsipouras M.G., et al. [33] suggested an automated system based on fuzzy modeling and information mining to predict heart disease. His model includes measures such as inferring information decision tree, extracting guidelines, formulating crisp model, transforming crisp model into fuzzy model, and optimizing it. The information gathered from Invasive Cardiology, University Department, Ioannina Hospital. The technique provides 80% precision in awareness and 65% precision in specificity [33].

Chaitrali S. Dangare et al. [34], used heart disease prediction data mining method. He first pre-processed the missing information using the mean mode technique and used the perceptron multi-layer model to map the information. To analyse the database of heart disease, Naive bays, neural networks and decision tree were used. He gathered 303 documents from the repository of Cleveland heart disease and used it as a training set and gathered 270 documents as test information from the repository of Stat log Heart Disease. The information set consists of attribute, input and key predictable. His model provides 100% accuracy for neural networks, Decision tree with 99.62 and Naïve bayes with 90.74% precision [34].

Mai Shouman et al. [35], by incorporating decision tree and k-means clustering, suggested a technique for diagnosing patients with heart disease. For k-means, the method utilizes centroid selection technique and decision tree is used to determine the clusters. Thirteen distinct characteristics are gathered from the Cleveland Clinic Foundation Heart Disease. Compared to the current decision tree, the integrated model of k-means and decision tree obtained greater outcomes of 83.9% [35].

S. Pal et al [36], suggested using information mining to predict heart disease. He surveyed 3 distinct classifiers such as CART, ID3 and tree of Decision. The information set from the Cleveland Clinic Foundation shows that 83.49 percent of the classification and regression tree (CART) precision was much better than the decision tree and ID3 (Iterative Dichotomized 3) [36]. H. A. Huijer et al. [37], designed a decision support service to find unrest transition through decision trust measure, trust-based SVM and trust-based multilevel SVM to discover agitation transition. 240 Samples are gathered via human body sensors. The patient experiences distinctive tension inventory of state-quality scale (T-STAI), used to calculate uncomfortable adults. The technique provides 91.4 percent precision when compared to traditional vector supporting machine with 90.9 percent precision [37].

Latha Parthiban et al. [38], outlined a prediction technique for smart heart disease prediction. The method is executed using coactive neuro fuzzy inference system, neural network, and genetic algorithm. The dataset is gathered from UCI and the prototype is simulated using Neuro Solution Software. The mean square error of CANFIS was 0.000842 [38]. N. Deepika et al. [39], suggested a heart attack patient classification model. He pre-processed his information sets for missing values and then implemented the same width binning interval approach. Then numerical parameters are transformed into categorical parameters and frequent patterns are mined based on pruning-classification rule algorithm linked to heart disease. His model used an effective forecast of particular class label [39].

T. Turner [40] proposed the idea of diagnosing heart disease by combining naïve bayes and k-means clustering with distinct choice of centroids. Cleveland clinic foundation collects the data set. The precision of the embedded k-k-is 84.5 percent compared to the traditional algorithm [40]. Uzma Ansari et al. [41] used weighted associative classifier to develop a model for predicting heart attacks. The data set is gathered from the ML database of Irvine University of California (UCI). He used 2 class labels 1 in his model for "No heart disease" and other one for "Heart disease" rather than getting 5 class labels with 1 for no heart disease and 4 for four heart disease kinds. He used 80% of confidence value and 25% of support value. The prototype proposed achieves precision of 81.51 percent. He concludes that the measured associative classifier is the easiest way to acquire efficient important patterns from information setting for cardiac disease [41]. The below Table 2 shows the comparative surveyof heart disease techniques which we have discussed above.

Table 2. The comparative survey of heart disease

Authors	Year	Data set	Methodology	Accuracy
Latha Parthiban	2007	University of California Irvine (UCI)	Neural network and Genetic algorithm	MSE -.000842
Tsipouras M. G	2008	Invasive Cardiology, Hospital of Loannina	Fuzzy Modelling and Data Mining	80%
H. A. Huijer	2010	Sensor Data	SVM	91.4%
N. Deepika	2011	University of California Irvine (UCI)	Pruning-Classification Association rule	Predicts effectively
Uzma Ansari	2011	University of California Irvine (UCI)	Weighted Associative Classifier	81.51%
P. K. Anooj	2012	Cleveland, Hungarian Institute of Cardiology, University Hospital, Zurich, Switzerland	Weighted Fuzzy rules	57.85%
Chaitrali S.	2012	Cleveland heart disease database	Naïve bayes, Neural network and Decision tree	100%
Mai Shouman	2012	Cleveland heart disease database	Centroid Selection Technique	83.9%
T. Turner	2012	Cleveland heart disease database	Integrated k- means and Naïve bayes	84.5%
S. Pal	2013	Cleveland heart disease database	CART, ID3, Decision Tree	83.49%

4.2. Description of existing techniques for breast cancer prediction

A fuzzy model was created by Yassi et al. [42] to differentiate between normal and malicious breast cancer. The technique brought disorder into the hierarchical cluster of partial swarm enhancement of multispecies, prompting the improvement of chaotic hierarchical cluster-based multispecies swarm enhancement of particles (CHCMSPSE). CHCMSPSE helps to distinguish the form of cancer of the breast and to enhance the fuzzy rules. The model also discovers fuzzy rules very correctly. The dataset is gathered from the machine learning database of Irvine University of California (UCI). Thus, for worldwide search capability, the technique utilizes 11 chaotic maps. Sinusoidal chaotic map acquired 99 percent precision from those maps because it matched with the position of the issue. The model achieves more than 90% precision [42].

In defining 5, 10 and 15 years of specific breast cancer sustainability, Lundin M et al. [43] provided a prototype for accessing the precision of ANN. The data source is collected from City Hospital of Turku and Turku University Central Hospital with 951 instances. In that training set of 651 instances and a validation set of 300 instances. This prototype compares the outcomes of artificial neural network and logistical regression. The precision of breast cancer specific survival for 5 years reported as 0.909,10 years reported as 0.086 and 15 years reported as 0.883 [43].

Delen D et al. [44] implemented a data mining technique comparison approach that involves logistical regression, decision tree, and artificial neural network to predict breast cancer development. The model utilizes over 200,000 cases of an enormous information repository. Thus, logistical regression precision is 89.2%, decision tree precision (C5) 93.6% and artificial neural network 91.2%. 10-fold cross-validation for information testing, unbiased estimation measurement and 3 techniques prediction. Research indicates that the selection trees is the best determinant method for defining breast cancer growth relative to the artificial neural network and logistic regression [44].

Bellaachia Abdelghani et al. [45] used information mining techniques to present a model for predicting breast cancer development. The pre-classification process is carried out in three areas: recovery of essential status, recovery of survival time and cause of death. Three techniques of machine learning, namely: neural network propagated back, naïve bayes and C4.5 for classification performance. The data set is gathered from the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results. There are 151.886 records in the dataset, with 16 characteristics. The model provides roughly 87% precision [45].

H. Koyuncu et al. [46] implemented a biomedical pattern based on artificial neural network rotational forest algorithm (RF-ANN). Multilayer perceptron was used as the classifier of base and the model used RF algorithm as the classifier of ensemble. Using the main component assessment, different function sets are gathered from the information set. The precision of RF-ANN is therefore 98.05 percent [46].

S. M. Jamarani H et al. [47] proposed a method for recognizing the disease and helping radiologists predict breast cancer. The model combines decomposition of artificial neural network (ANN) and sub-band picture based on multiwavelet. The technique is studied using mammographic database by mammographic image analysis society (MIAS). The highest output of biorthogonal Geronimo, Hardin and massopust multiwavelet 2 (BiGHM2) was among the various kinds of multiwavelet. Thus, BiGHM2 accomplished precision in the operating characteristic curve of the receiver around 0.96 [47].

T. Nguyen et al. [48] suggested an automatic wavelet-based technique for classifying medical information and a type-2 fuzzy logic. They carried out execution from the UCI database for machine learning on 2 medical datasets: Cleveland heart disease and Wisconsin breast cancer. The outcome demonstrates that, compared to other machine learning methods, the advantage of interval type-2 fuzzy logic scheme is better [48]. Z. Mahmud et al. [49] suggested a method to use age, marital status and therapy among Malaysian women to find out about cervical cancer. The records of patients with cervical cancer are gathered from the medical center of Kebangsaan Malaysia University (UKM). The model has four phases, with 444 records of patients impacted by cervical cancer, and finding out the age and marital status of women's medical therapy. They discovered that the 46-year-old females are more likely to develop cervical cancer. So, it is suggested that Malaysian women undergo testing prior to the age of 45 and they also found that Chinese women under the age of 57 have more likelihood of being diagnosed with radiotherapy in the original phase of cervical cancer [49].

M. Seera et al. [50] suggested using hybrid smart classification to classify medical information to predict cancer. The model has a random forest, classified trees, and a regression tree and a min-max neural network. The technique of random forest is used to create a classification and regression Tree model ensemble. Fuzzy min-max is used for teaching purposes. The tree of classification and regression is used to extract the rule. The precision of this model for cancer forecast was 98.84% [50].

W. Kuo et al. [51] suggested a novel technique for breast tumour prediction in clinical ultrasonic images using the decision tree. The model concentrated on pictures from the United States. The decision tree

uses the C5.0 algorithm. Machine learning with decision tree algorithm help predict breast tumour illness with 93.33% responsiveness precision and 96.67% specificity precision [51].

Seon-Hak [52] constructed a prototype using rough set structures for hierarchical classification. The model is based on the framework of hierarchical granulation to find out the laws of classification and therefore suggested a discovery of laws. The technique is validated against the Wisconsin breast cancer (WBC) data gathered dataset. The model still generates excellent efficiency when loaded with simple rules and brief conditionals. Thus, by creating minimal classification rules, the model was effective in decreasing the number of dimensions. His model makes it simpler for us to analyze the information system [52]. The below Table 3 shows the comparative study of breast cancer techniques which we have discussed above.

Table 3. The comparative survey of breast cancer

Authors	Year	Data set	Methodology	Accuracy
Lundin M	1999	Turku City Hospital and Central Hospital of Turku University	ANN	5yrs-0.909, 10yrs-0.086 and 15yrs0.883 G
W. Kuo	2001	Machine learning repository	C5.0 algorithm	96.67%
Seon-Hak Seo	2001	Wisconsin breast cancer (WBC), UCI diabetes for Pima Indians	Hierarchical Classification	83.49%
Delen D	2005	Machine learning repository	ANN and logistic regression	93.6%
Jamarani S. M. h	2005	Machine learning repository	(ANN), decomposition of multiwavelet-based picture sub-band	ROC-0.96
Bellaachia Abdelghani	2006	National Cancer Institute (NCI) surveillance, epidemiology and end results	Data Mining Methods	87%
Uzma Ansari	2013	University of California Irvine (UCI)	Rotation forest algorithm (RF-ANN)	98.05%
M. Seera	2014	Wisconsin breast cancer (WBC), UCI diabetes for Pima Indians	Random forest, classified trees and regression tree (CART) and fuzzy neural network min-max	98.84%
A. Yassi	2014	University of California Irvine (UCI)	hierarchical-clustering	90%
T. Nguyen	2015	University of California Irvine (UCI)	Wavelets and interval type-2 fuzzy logic system (IT2FLS)	97.40%

5. CHALLENGES AND RESEARCH OPPORTUNITIES

Reviews of data mining methods, classification methods, smart methods and choice of features for disease prediction were discussed here. As the selection of characteristics enables us to eradicate unnecessary data, large-dimensional data must be compressed without the loss of data, which improves the effectiveness of classification. But the difficulty of subset attribute selection is high, which is complicated as it requires complex interdependence on a wide range of factors. We could incorporate guidelines and feature selection for better results in the classifiers in the future. Additionally, new feature selection method such as ant colony optimization, etc. is possible to test to improve quality, and you can attempt experimenting with algorithm potential for most medical datasets that include distinct features such as noisy information, sparsity, missing value, etc. to enhance model accuracy.

6. CONCLUSION

This paper's primary focus is to discuss various prediction models and techniques used for predicting heart disease and breast cancer. The technique also sheds light on the significance in medical dataset of various classification techniques for disease prediction. The dataset that we have discussed in so many current methods is linked to heart and breast cancer. As a classifier, the different machine learning methods are used to construct a value-effective model for predicting disease. It is therefore well recognized by the exhaustive study that the extraction of the necessary data from the clinical repository helps us to promote excellently-informed testing and choices.

REFERENCES

[1] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *J. Am Med Inform Assoc*, vol. 18, no. 5, pp. 601–606, 2011.

[2] R. Snyderman, "Personalized health care: from theory to practice," *Biotechnology Journal*, vol. 7, no. 8, pp. 973–979, Aug. 2012.

[3] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *J. Am Med Inform Assoc*, vol. 18, no. 5, pp. 601–606, 2011.
[4] N. Razavian and D. Sontag, "Temporal convolutional neural networks for diagnosis from lab tests," *arXiv:1511.07938v4*, 2015.
[5] R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad, "Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis," *Journal of the American Medical Informatics Association: JAMIA*, vol. 22, no. 4, pp. 872–880, Jul. 2015.
[6] N. Tangri, L. A. Stevens, J. Griffith, H. Tighiouart, O. Djurdjev, D. Naimark, A. Levin, and A. S. Levey, "A predictive model for progression of chronic kidney disease to kidney failure," *JAMA*, vol. 305, no. 15, pp. 1553–1559, 2011.
[7] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," *Proceedingsofthe1st MachineLearningforHealthcareConference, ser. Proceedings of Machine Learning Research*, pp. 301–318, Aug 2016.
[8] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 1114– 1120, Jul. 2010.
[9] Wu jionglin M. S., *et al*., "Prediction Modeling Using EHR Data: Challenges, and a Comparison of Machine Learning Approaches," *Journal of the medical care section, American public health association*, vol. 48, no. 6, pp. S106-S113, 2010.
[10] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making*, vol. 10, no. 16, pp. 1-7, 2010.
[11] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors," *Big Data*, vol. 3, no. 4, pp. 277–287, Dec. 2015.
[12] A. V. Lebedev, E. Westman, G. J. P. Van Westen, M. G. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, B. Vellas, S. Lovestone, and A. Simmons, "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness," *NeuroImage: Clinical*, vol. 6, pp. 115–125, 2014.
[13] N. Tangri, L. A. Stevens, J. Griffith, H. Tighiouart, O. Djurdjev, D. Naimark, A. Levin, and A. S. Levey, "A predictive model for progression of chronic kidney disease to kidney failure," *JAMA*, vol. 305, no. 15, pp. 1553–1559, 2011.
[14] A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad, "Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis," *Journal of the American Medical Informatics Association: JAMIA*, vol. 22, no. 4, pp. 872–880, Jul. 2015.
[15] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J. Adv. Signal Process*., vol. 2016, Dec. 2016.
[16] B. Shickel, P. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Informatics*, vol. 22, no. 5, pp. 1589-1604, 2018.
[17] Mingyu Pak, Miyoung Shin, "Developing Disease Risk Prediction Model Based on Environmental Factors," *The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014)*, 2014.
[18] Xiaoyan Wang, Amy Chused, Nomie Elhadad, Carol Friedman, and Marianthi Markatou: "Automated Knowledge Acquisition from Clinical Narrative Reports," *AMIA 2008 Symposium Proceedings*, pp. 783-787, 2008.
[19] Nicolae Dragu, Fouad Elkhoury, Takunari Ralph and A. Morelli Nicolas di Tada, "Ontology-Based Text Mining for Predicting Disease Outbreaks," *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*, 2010.
[20] Kumar Sen, Shamsher Bahadur Patel and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy," *International Journal of Engineering and Computer Science*, vol. 2, no. 9, pp. 2663-2671, 2013.
[21] Saba Bashir, Usman Qamar, Farhan Hassan Khan, "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting Received," *Australasian Physical & Engineering Sciences in Medicine*, vol. 38, pp. 305-323, 2015.
[22] lav Petrov, Dipanjan Das and Ryan McDonald, "A Universal Part of-Speech Tagset," *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012*), 2012.
[23] Mayo clinic, [Online]. Available: www.mayoclinic.org [Accessed 17/10/2015]
[24] Isabel, [Online]. Available: www.isabelhealthcare.com [Accessed 12/10/2015]
[25] Patient, [Online]. Available: www.patient.co.uk [Accessed 11/10/2015]
[26] [Online]. Available: www.symptomchecker.isabelhealthcare.com [Accessed 30/10/2015]
[27] Md. Tahmid Rahman Laskar, Md. Tahmid Hossain, Abu Raihan Mostofa Kamal, Nafiul Rashid, "Automated Disease Prediction System (ADPS): A User Input-based Reliable Architecture for Disease Prediction," *International Journal of Computer Applications*, vol. 133, no. 15, pp. 24-29, Jan. 2016.
[28] Xiaoyan Wang, Amy Chused, Nomie Elhadad, Carol Friedman, and Marianthi Markatou, "Automated Knowledge Acquisition from Clinical Narrative Reports," *AMIA 2008 Symposium Proceedings*, pp. 783-787, 2008.

[29] S Manimekalai, R Suguna, S Arulselvarani, "An Intelligent Automatic Multi-Disease Prediction Technique using Data Mining Algorithms and Big Data," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 10, Oct. 2018
[30] Cristinel Ababei, Milad Ghorbani Moghaddam, "A Survey of Prediction and Classification Techniques in Multicore Processor Systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 1, no. 1, Sep. 2017.
[31] Ankita Dewan, and Meghna Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2015.
[32] P. K. Anooj, "Clinical Decision Support system: Risk level prediction of heart disease using weighted fuzzy rules," *Journal of King Saud University- Computer and Information Sciences*, vol. 24, pp. 27-40, 2012.
[33] Tsipouras M.G., Exarchos T.P., Fotiadis D.I., Kotsia A.P., Vakalis K.V., Naka K.K., Michalis L.K., "Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling," *IEEE T. Inf. Techno. B.*, vol. 12, no. 4, pp. 447–458, 2008.
[34] Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44-48, Jun. 2012.
[35] Mai Shouman, Tim Turner and Rob Stocker, "Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients," *Proceedings of the International Conference on Data Mining*, 2012.
[36] V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques," *Caribbean Journal of Science and Technology*, vol. 1, pp. 208-217, 2013.
[37] G. E. Sakr, I. H. Elhajj and H. A. Huijer, "Support vector machines to define and detect agitation transition," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 98-108, Dec. 2010.
[38] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm," *International Journal of Biological and Life Science*, vol. 15, pp. 157–160, 2007.
[39] K. Chandra shekar and N. Deepika, "Association rule for classification of Heart Attack Patients," *International Journal of Advanced Engineering Science and Technologies*, vol. 11, no. 2, pp. 253-257, 2011.
[40] M. Shouman, T. Turner and R. Stocker, "Integrating naïve bayes and K- Means clustering with different initial centroid selection methods in the diagnosis of heart disease patients," *ICAITA*, pp. 125-137, 2012.
[41] Uzma Ansari, Dipesh Sharma, Jyoti Soni and Sunita Soni, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers," *International Journal on Computer Science and Engineering (IJCSE*), vol. 3, no. 6, pp. 2385-2392, Jun. 2011.
[42] A. Yassi, M. Yaghoobi, M. Yassi, "Distinguishing and Clustering breast cancer according to hierarchical structures based on chaotic multispecies partical swarm optimization," *Iranian Conference on Intelligent Systems*, pp. 1-6, Feb. 2014.
[43] Lundin M., Lundin J., Burke B.H., Toikkanen S., Pylkkänen L. and Joensuu H., "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer," *Oncology International Journal for Cancer Resaerch and Treatment*, vol. 57, no. 4, pp. 281-286, 1999.
[44] Delen D., Walker G., Kadam A., "Predicting breast cancer survivability: a comparision of three data mining methods," *Artif Intell Med*, vol. 34, no. 2, pp. 113-27, 2005.
[45] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques," *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining*, 2006.
[46] H. Koyuncu and R. Ceylan, "Artificial neural network based on rotation forest for biomedical pattern classification," in *36th International Conference on Telecom-munications and Signal Processing (TSP), IEEE*, pp. 581–585, 2013.
[47] Jamarani S. M. h., Behnam H. and Rezairad G. A., "Multiwavelet Based NeuralNetwork for Breast Cancer Diagnosis," *International Enformatika Conference, IEC'05*, pp. 19-21, 2005.
[48] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Medi-cal data classification using interval type-2 fuzzy logic system and wavelets," *Applied Soft Computing*, vol. 30, pp. 812–822, 2015.
[49] Z. Mahmud and S. Sulong, "Confounding effects of age, marital status and treatment on cervical cancer stages among malaysian women," *Conference: Statistics in Science, Business, and Engineering (ICSSBE)*, 2012.
[50] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2239–2249, 2014.
[51] W. Kuo, R. Chang, D. Chen and C. C. Lee, "Data Mining with Decision Trees for Diagnosis of Breast Tumor in Medical Ultrasonic Images," *Breast Cancer Researchs and Treatment*, vol. 66, no. 51-57, 2001.
[52] Chul-Heui Lee, Seon-Hak Seo, Sang-Chul Choi, "Rule Discovery using hierarchical clasification structure with rough sets," *IFSA World Congress and 20th NAFIPS International Conference*, 2001.

BIOGRAPHIES OF AUTHORS

Ms. G. Saranya received B.Tech degree in Information Technology from Sri Venkateswara College of Engineering, Sriperumbudur, Chennai, India in 2009. M.Tech degree in Information Technology from Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi, Chennai, India in 2011.She is currently pursuing her Ph.D in Computer Science and Engineering at Sathyabama Institute of science and Technology, Chennai, India and works as Assistant Professor in Information Technology Department at SRM Institute of Science and Technology, India. She has 7+ years of teaching experience and 2 years of Industry Experience. Her research interests include Big Data Analytics, Machine learning and Deep learning.

Dr. A. Pravin received the B.E degree in Computer Science & Engineering from Bharath Niketan Engineering College, Madurai Kamaraj University, Madurai, India in 2003 , M.E degree in Computer Science & Engineering from Sathyabama University, Chennai, India in 2005 and Ph.D degree in Computer Science & Engineering at Sathyabama University, Chennai, India in 2014. He works currently as an Associate Professor for the Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology, Chennai and he has 14 Years of teaching experience. He has participated and presented many Research Papers in International and National Conferences and also published many papers in International and National Journals. His area of interests includes Software Engineering, Data mining , Internet of Things and Big data**.**