IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 12, No. 4, December 2023, pp. 1687~1694
ISSN: 2252-8938, DOI: 10.11591/ijai.v12.i4.pp1687-1694

Machine learning approach for predicting heart and diabetes diseases using data-driven analysis

Usha Sekar, Kanchana Selvarajan
Department of Computer Science, Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur, India

Article Info

Article history:

Received Jun 23, 2022
Revised Jan 28, 2023
Accepted Mar 10, 2023

Keywords:

Adaptive boost
Chi-square
Gradient boost
Prediction
Principal component analysis

ABSTRACT

Environmental changes and food habits affect people's health with numerous diseases in today's life. Machine learning is a technique that plays a vital role in predicting diseases from collected data. The health sector has plenty of electronic medical data, which helps this technique to diagnose various diseases quickly and accurately. There has been an improvement in accuracy in medical data analysis as data continues to grow in the medical field. Doctors may have a hard time predicting symptoms accurately. This proposed work utilized Kaggle data to predict and diagnose heart and diabetic diseases. The diseases heart and diabetes are the foremost cause of higher death rates for people. The dataset contains target features for the diagnosis of heart disease. This work finds the target variable for diabetic disease by comparing the patient's blood sugars to normal levels. Blood pressure, body mass index (BMI), and other factors diagnose these diseases and disorders. This work justifies the filter method and principal component analysis for selecting and extracting the feature. The main aim of this work is to highlight the implementation of three ensemble techniques-Adaptive boost, Extreme Gradient boosting, and Gradient boosting-as well as the emphasis placed on the accuracy of the results.

Corresponding Author:

Usha Sekar
Department of Computer Science, Faculty of Science and Humanities
SRM Institute of Science and Technology
Kattankulathur, Chengalpattu Dt., TamilNadu, India
Email: us3648@srmist.edu.in

1. INTRODUCTION

Healthy living is crucial to a good quality of life. A healthcare professional prevents, treats, and inspects diseases to improve health. Because of the inexactness of the information provided by the patient, it can be challenging to determine a specific disease based on their symptoms [1]. Globally predicting diseases is a crucial challenge in fundamental problems [2]. Many diseases are associated with particular symptoms and signs. It can be inherited, caused by infection, or triggered by stress [3]. Due to the residents' modern lifestyle, there is a risk of mortality and morbidity from diseases like heart disease, chronic respiratory disease, and diabetes [4]. Nowadays, millions of people worldwide suffer and die from many diseases [5]. The majority of people with multiple disorders are also distressed by numerous infections.

In today's society, predicting disease based on early-stage symptoms is a very tough challenge for physicians in the medical field. The field of medical informatics and disease prediction has become increasingly relevant to the community of data scientists in recent years. Data repute, multi-attribution, incompleteness, and a close correlation will occur when manually collecting medical data, making it difficult to identify disease symptoms. The extensive use of computer-based technologies in the health sector has resulted in the availability of colossal health databases for researchers. Many surgical research studies are using these electronic records [6].

Nowadays, all hospitals maintain electronic health data for patients to find the symptoms and diagnose the disease. A health care system can be revolutionized by analyzing and interpreting the information recorded in electronic health records, providing feedback, and implementing changes based on collected data [7].

Machine learning techniques have become increasingly significant in various fields in the past decade, including the health care system and biomedical research [8]. In addition, correctly medicating a patient with a large amount of data is an enormous task. Since the advent of the digital era and technological innovations, several multidimensional patient data sets have been developed, including clinical data, hospital resource information, and patient disease diagnosis information. A complex data set must be analyzed to extract valuable insights [9]. The proposed work aims to develop heart and diabetic disease prediction models from a single dataset incorporating machine learning algorithms, specifically supervised learning methods that employ the ensemble method for more than one disease prediction in a single dataset.

Heart disease has been the leading cause of death worldwide in recent decades [10]. Since the heart is a significant part of the human body, various factors cause heart disease, and people exhibit different symptoms [11]. People consider the disease diabetes is a high sub challenge with deadly chronic disease [12]. Due to the increase in sugar in blood and fat, people affect their daily lives a lot.

In health care systems, machine learning techniques can support medical practitioners in promptly and cost-effectively diagnosing various diseases from medical data [13]. Specifying probable disorders could help patients conduct medical tests on targeted medicine. The patient might skip extensive medical tests due to a lack of medical information, leading to severe health problems. In most cases, machine learning identifies the patterns in massive datasets, which can involve human intelligence. Machine learning (ML) approaches can aid in building prediction models that can handle and analyze vast volumes of complex medical data and efficiently find the presence or absence of disease in a patient, which can help address this difficulty [14].

The main aim of recursion enhanced random forest with an improved linear model (RFRF-ILM) is to find the key features. The prediction model produces better performance by combining the classification model. This work compares the essential variables that suggest that coronary artery disease develops more frequently as people age [15]. This paper describes disease progression and predicts disease outcomes. A proposed novel approach [16] uses a model with various features and known classifier techniques to recognize the relevant factors through a machine learning algorithm, leading to better predicting accuracy of cardiovascular disease. Based on the prediction model's hybrid random forests with the linear model (HRFLM), it produced 88.7% of the accuracy value. According to [17], it is tough to identify diabetic disease. A rigorous framework has been developed by rejecting outliers and eliminating missing values. After selecting features, various machine classifiers standardize the data. By estimating the area under receiver operator characteristic curve, the method improved the outcome by weighting the classifier model and producing a better prediction.

The work [18] has proposed a hybrid technique by applying different machine learning classifiers to diagnose cardiovascular disease-the various classifier helps this study to evaluate the performance metrics using weka and keel tools. The primary intent of this work is to choose the best classifier by comparing each classifier's accuracy value. This system [19] used a python tool to perform preprocessing with neighborhood cleaning rule and feature engineering. AutoML, advanced extended gradient boost and advanced ensemble bagging models are applied. Specialists perform this work to identify whether or not someone has cardiovascular disease and diagnose the patient's condition. Studies suggested in [20] used four ML methods to estimate diabetes risk, where bagging and boosting techniques were used to enhance robustness. Among the existing algorithms, the Random Forest algorithm provides the most accurate results. The study employs [21] the AdaBoost and bagging ensemble techniques using the J48 (c4.5) decision tree as a base learner and standalone data mining methodology. The method applied was to classify the patients with diabetes using diabetes risk indicators. In the study, the Adaboost ensemble method outperformed bagging and a standalone J48 decision tree in terms of overall performance.

The [22] work aims to develop a model to predict diabetics. K-nearest neighbor (KNN) helps reduce the processing time, and support vector machine (SVM) allocates a class for all the sample datasets. Selecting features in this work helps build four classifiers. In addition, the researchers used four algorithms in this study to determine the efficacy and accuracy of predicting whether or not people will have diabetes. According to the study [23], A hierarchical ensemble model combines a decision tree and logistic regression classifiers trained independently. The neural network joined with the previous model at the next level provides overall better accuracy.

This work mainly focuses on diagnosing the risk of heart and diabetes diseases and encourages people to have good health. The proposed study reveals that two chronic ailments, such as diabetes and heart disease, can be predicted using the filter method chi-square and principal component analysis (PCA). Creating classification techniques in diagnostics can help to avoid human error. The model utilized ensemble boosting strategies such as Adaboost, Gradient boost, and Extreme Gradient boost to improve prediction accuracy.

Accordingly, the rest of the paper follows section 2 as a method. Section 3 presents a result and discussion. Finally, section 4 covers the conclusion with future work.

2. METHOD

This section describes datasets, feature selection, and the ensemble, such as Adaptive boost, Gradient boost, and Extreme Gradient boost classifier. Figure 1 depicts the pipeline of disease prediction—the proposed system structured into different phases. The phases contained in this work are data collection, data preprocessing and selecting features, feature extraction, splitting the data, classifier models, evaluation metrics, and comparison of ensemble classifier models.

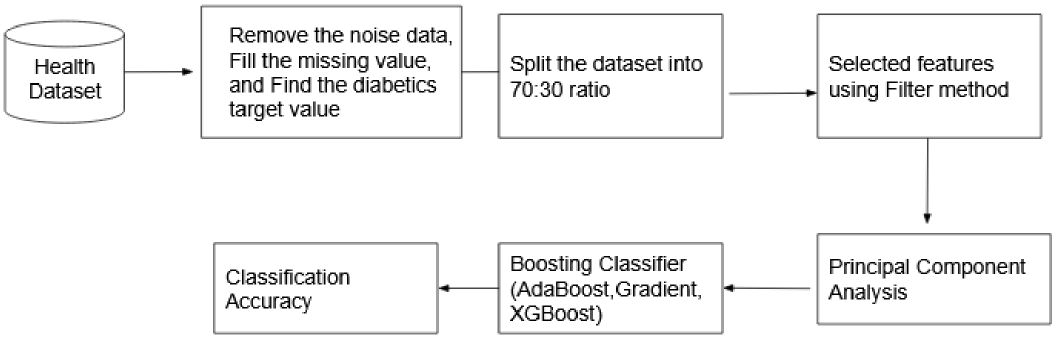
Health Dataset
Remove the noise data, Fill the missing value, and Find the diabetics target value
Split the dataset into 70:30 ratio
Selected features using Filter method
Principal Component Analysis
Boosting Classifier (AdaBoost,Gradient, XGBoost)
Classification Accuracy

Figure 1. Disease prediction pipeline

2.1. Dataset collection

The detection of disease using machine learning is a challenging task. Rather than model complexity, interpretability, or computational burden, the doctor is most concerned with whether the model is reliable and effective in predicting illness. Phase one of the work proposed was a collection of data from the University of California Irvine (UCI) repository. The data collection has 12 attributes and one target attribute. The dataset has continuous or categorical data types.

2.2. Data preprocessing

Data preprocessing is one of the processes considered the most crucial step in classification. The process is to remove the inappropriate inexplicable, and continual features. The feature contains noisy data, a format that cannot be used in the model and fills the missing values using the KNN Impute method. The range of glucose values helps to determine the independent variable for diabetes.

2.3. Feature selection and extraction

In machine learning, feature selection techniques play a pivotal role in selecting the features [24]. The selection method reduced the original feature set into several sub-features to reduce model complexity, improve computational efficiency, and reduce generalization errors caused by irrelevant features. All the features are ranked based on the chi-square method's score. The chi-square approach is a statistical procedure that shows how well the observed frequency data values match the predicted frequency data values for independent variables. Extracting only the best features is essential to maximize a machine learning classifier's performance since irrelevant features can negatively affect performance. This phase involves identifying imperative features within the dataset using principal component analysis.

2.4. Boosting classifier

In a boosting algorithm, the classifiers generate sequentially. Boosting [25] is designed to train a set of classifiers consecutively and then combine them for prediction, where the later classifiers correct mistakes made by the earlier ones. A boosting classifier turns weak classifiers into more robust models to enhance accuracy. This work trained different boosting classifier techniques, i.e., Adaptive boosting, Extreme Gradient boosting machine, and Gradient boosting machine, to predict the heart and diabetics diseases.

2.4.1. Adaptive boosting

The adaptive boosting algorithm has been widely used in classification [26]. In 1997, Freund and Schapire proposed the adaptive boosting (AdaBoost) algorithm. This boosting technique helps weak learners perform better using an ensemble approach. It improves the performance of the classifier when used alongside different algorithm types. Adaptative boosting is exceptionally robust to noise and outliers in data.

2.4.2. Gradient boosting

Gradient boosting (GB) has typically solved the regression and classification problems. The prediction model constructs through a set of decision trees constructed stage-by-stage. Decision trees are frequently used to accomplish Gradient boosting. The primary benefit of Gradient boosting is that it reduces the remaining preceding time in each calculation. In terms of generalization, the GB performs well as an ensemble classifier. The GB implements a regularisation term, regulating the model's complexity and preventing overfitting [27].

2.4.3. Extreme gradient boosting

The XGBoost algorithm is said to be a robust boosting algorithm. It's a more advanced variant of the Gradient boosting method that was introduced to predict the errors or residuals from prior models, and then the new model is blended with the old. This model controls overfitting, eliminating interference with outliers and making the model more accurate and stable [28].

2.5. Classification accuracy

The proposed system used a boosting classifier such as Adaboost, XGBoost, and Gradient boost to predict the disease. One of the essential performance indicators for categorization is accuracy. It says that the percentage of the total sample is correctly classified. A classification report performs all the boosting classifiers to produce accurate results. These classifiers combine to assess their performance based on classification accuracy [29]. The accuracy measure assesses the ability of a model to predict the future. The given (1) represents the formula for accurate classification.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where,

- TP: The classifier predicted TRUE, which was the correct class in the case of true positive.
- TN: In the case of real negatives, the classifier predicted FALSE, and it was the suitable class.
- FP: When there are false positives (FP), the classifier predicts TRUE, and the correct class is FALSE.
- FN: Models predict false when they have diseases in the case of false negatives.

3. RESULTS AND DISCUSSION

The proposed system utilizes the same dataset to diagnose heart and diabetic ailments. The dataset contains 12 independent and one dependent feature. The features included in this dataset are id, age, sex, weight, height, gender, blood pressure (both systolic and diastolic), cholesterol, glucose, smoking, alcohol, and physical activity. It has come to be known that the given dataset can also predict diabetic disease by finding separate target variables. Algorithm 1 shows that the target feature of a person with diabetes has been evaluated based on the range of glucose values. In the target attribute, the value is 1 for patients with diabetes, whereas 0 is the value for those who do not have diabetes. The given dataset contains two target attributes: diabetes and heart target. The proposed work develops a prediction model that takes the symptoms from the user and predicts the heart and diabetics diseases.

Algorithm 1. Finding target features for diabetic

```
Require: Input: Health Care dataset.
for ∀ glucose feature do
if data(value)>range
dia_target=1
else
dia_target=0
end if
end for
```

The proposed work uses the Correlation and Chi-square selection method to select features after data preprocessing. A heat map represents the correlation between the target and other features and shows the relationship between the features. Figure 2 and Figure 3 uses a heat map to highlight the correlation between the dataset's attributes in both predictions. This heat map represents values as colors in a two-dimensional representation. In one glance, it provides a quick visual summary of data. The viewer can easily comprehend complex datasets using more elaborate heat maps. The feature method increases the classification accuracy.

According to the principle of feature importance, all the features have a score value that determines the extent. Figure 4 describes the most significant feature for prediction based on the feature importance generated by the filter method. The subset of features used to predict both diseases is different in this work. This works estimated the most significant features as sysbp, glucose, age, chol, ciger for heart disease and api_hi, weight, api, age, and cholesterol for diabetic disease. The highest rank helps to select the features to predict heart and diabetes disease based on the importance score.

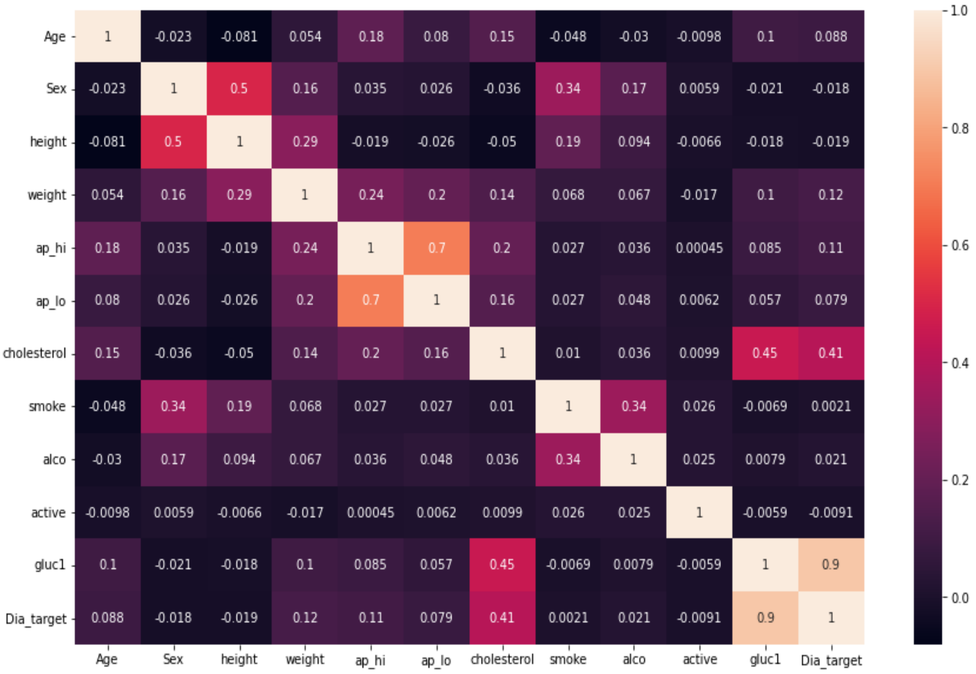

Figure 2. Correlation features of diabetic’s disease

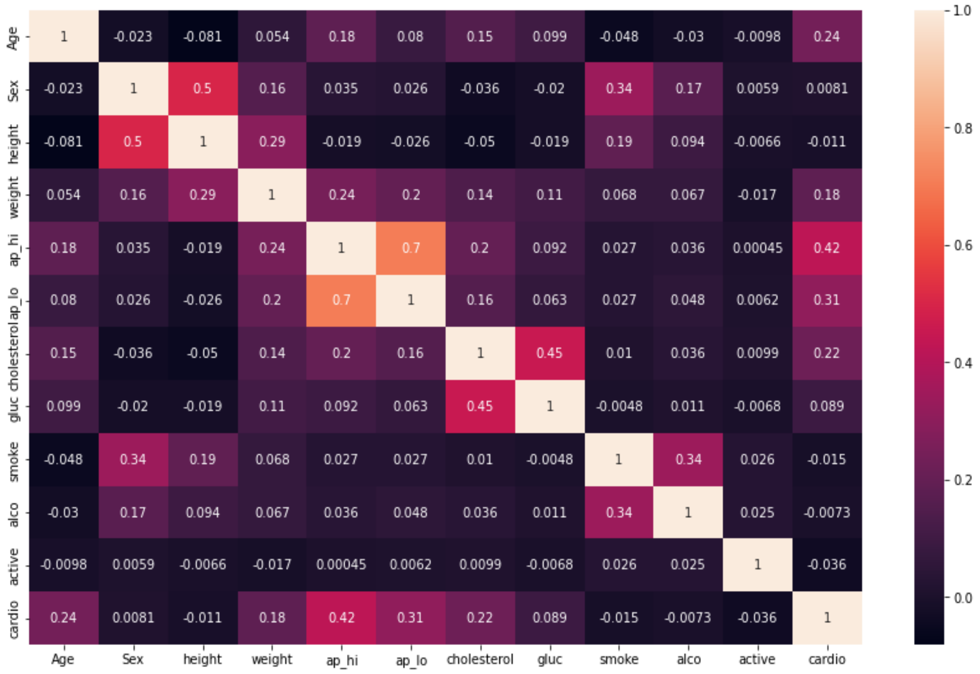

Figure 3. Correlation features of heart disease

In the final phase, the methodology provides a better accuracy by boosting classifiers, including AdaBoosting, Gradient Boosting, and XGBoosting. These results were obtained by combining selected parameters (by using chi-square) with PCA to devise the best classifiers to diagnose the disease. PCA reduces the dimensionality of the input and lowers computation complexity, and speeds up the training process by applying the principle component analysis to the input features. In this study Figure 5 depicts the most significant accuracy outcomes of heart and diabetics diseases.

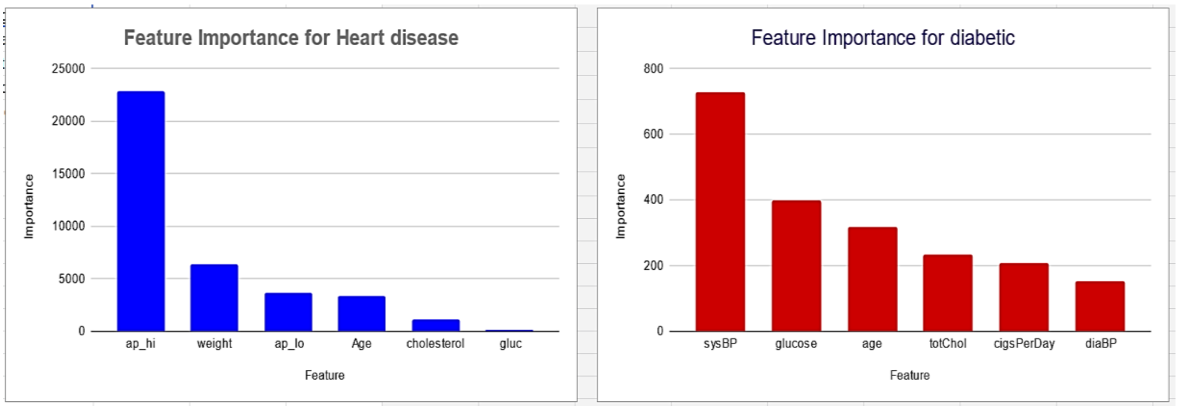

Figure 4. Feature importance of heart and diabetics disease

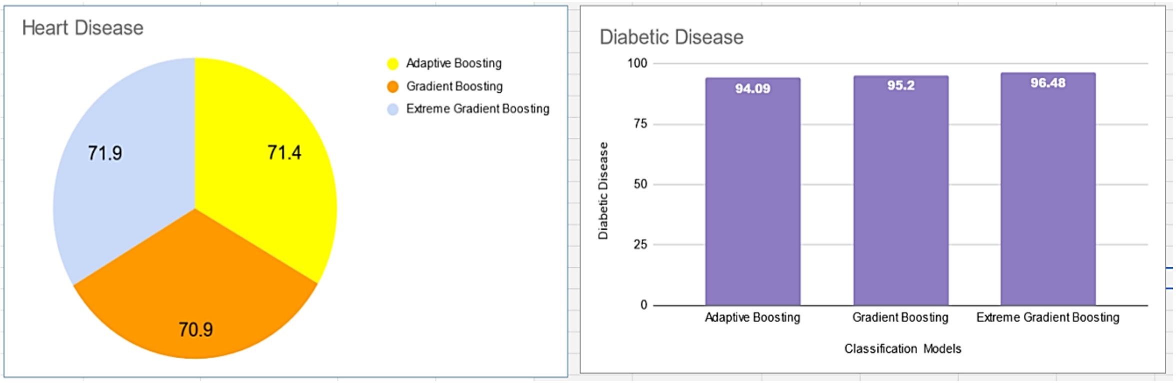

Figure 5. Classification accuracy of heart and diabetic disease

The result section compares all boosting classifiers in both diseases' predictions. The classification results are shown in Table 1. The table shows the classification results for heart disease and people with diabetes using the boosting classifier and concluded that Extreme Gradient boosting (XGBoosting) performed well and produced the highest accuracy value.

Table 1. Classification result

Classifier	Accuracy Score	
	Heart Disease	**Diabetic Disease**
Adaptive Boosting	71.4	94.09
Gradient Boosting	70.9	95.2
Extreme Gradient Boosting	71.9	96.48

4. CONCLUSION

This study aims to develop a dependable and accurate predictive model for heart and diabetic disease. It has used a single dataset for predicting heart and diabetic disease. The given dataset has both target variables for heart and diabetic diseases. Filter method Chi-square selected the feature to diagnose disease. PCA was used to extract the features. The three ensemble boosting classifiers: are Adaboost, Gradient boost, and XGBoost. Results showed that XGBoost provides a higher accuracy value than other boosting algorithms in both disease predictions. Future work needs a better performance metric value by implementing a hybrid model for both diseases.

ACKNOWLEDGEMENTS

I am pleased to thank my research supervisor Dr. S. Kanchana for her guidance and enthusiastic encouragement of my research work. This article receives no financial support for its research, authorship, and/or publication.

REFERENCES

[1] A. K. Yadav, R. Shukla, and T. R. Singh, "Machine learning in expert systems for disease diagnostics in human healthcare," *Machine Learning, Big Data, and IoT for Medical Informatics*, pp. 179–200, 2021, doi: 10.1016/B978-0-12-821777-1.00022-7.

[2] P. G. Shynu, V. G. Menon, R. L. Kumar, S. Kadry, and Y. Nam, "Blockchain-based secure healthcare application for diabetic-cardio disease prediction in fog computing," *IEEE Access*, vol. 9, pp. 45706–45720, 2021, doi: 10.1109/ACCESS.2021.3065440.

[3] K. Burse, V. P. S. Kirar, A. Burse, and R. Burse, "Various preprocessing methods for neural network based heart disease prediction," *Advances in Intelligent Systems and Computing*, vol. 851, pp. 55–65, 2019, doi: 10.1007/978-981-13-2414-7_6.

[4] P. Priyanga, V. V. Pattankar, and S. Sridevi, "A hybrid recurrent neural network-logistic chaos-based whale optimization framework for heart disease prediction with electronic health records," *Computational Intelligence*, vol. 37, no. 1, pp. 315–343, 2021, doi: 10.1111/coin.12405.

[5] A. Elumalai, P. B. Maruthi, N. Gautam, S. Priyadharshini, and M. Suganthy, "RETRACTED ARTICLE: Optimal prediction of attacks and arterial stiffness effects on heart disease by hybrid machine learning algorithm," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, p. 83, 2022, doi: 10.1007/s12652-020-02706-4.

[6] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, 2019, doi: 10.1186/s12911-019-1004-8.

[7] A. M. Khedr, Z. Al Aghbari, A. Al Ali, and M. Eljamil, "An efficient association rule mining from distributed medical databases for predicting heart diseases," *IEEE Access*, vol. 9, pp. 15320–15333, 2021, doi: 10.1109/ACCESS.2021.3052799.

[8] A. K. Dubey, "Optimized hybrid learning for multi disease prediction enabled by lion with butterfly optimization algorithm," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 46, no. 2, 2021, doi: 10.1007/s12046-021-01574-8.

[9] R. Manne and S. C. Kantheti, "Application of artificial intelligence in healthcare: chances and challenges," *Current Journal of Applied Science and Technology*, pp. 78–89, 2021, doi: 10.9734/cjast/2021/v40i631320.

[10] R. C. Ripan *et al.*, "A data-driven heart disease prediction model through k-means clustering-based anomaly detection," *SN Computer Science*, vol. 2, no. 2, 2021, doi: 10.1007/s42979-021-00518-7.

[11] R. Kumar and P. Rani, "Comparative analysis of decision support system for heart disease," *Advances in Mathematics: Scientific Journal*, vol. 9, no. 6, pp. 3349–3356, 2020, doi: 10.37418/amsj.9.6.15.

[12] U. Ahmed *et al.*, "Prediction of diabetes empowered with fused machine learning," *IEEE Access*, vol. 10, pp. 8529–8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

[13] L. Men, N. Ilk, X. Tang, and Y. Liu, "Multi-disease prediction using LSTM recurrent neural networks," *Expert Systems with Applications*, vol. 177, 2021, doi: 10.1016/j.eswa.2021.114905.

[14] M. N. Uddin and R. K. Halder, "An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach," *Informatics in Medicine Unlocked*, vol. 24, 2021, doi: 10.1016/j.imu.2021.100584.

[15] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han, and J. Yu, "Recursion enhanced random forest with an improved linear model (RERF-ILM) for heart disease detection on the internet of medical things platform," *IEEE Access*, vol. 8, pp. 59247–59256, 2020, doi: 10.1109/ACCESS.2020.2981159.

[16] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[17] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[18] F. Z. Abdeldjouad, M. Brahami, and N. Matta, "A hybrid approach for heart disease diagnosis and prediction using machine learning techniques," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12157 LNCS, pp. 299–306, 2020, doi: 10.1007/978-3-030-51517-1_26.

[19] M. Fayez and S. Kurnaz, "RETRACTED ARTICLE: Novel method for diagnosis diseases using advanced high-performance machine learning system (Applied Nanoscience, (2023), 13)," *Applied Nanoscience (Switzerland)*, vol. 13, no. 3, p. 1787, 2023, doi: 10.1007/s13204-021-01990-6.

[20] N. Nai-Arun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," *Procedia Computer Science*, vol. 69, pp. 132–142, 2015, doi: 10.1016/j.procs.2015.10.014.

[21] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Procedia Computer Science*, vol. 82, pp. 115–121, 2016, doi: 10.1016/j.procs.2016.04.016.

[22] M. Panda, D. P. Mishra, S. M. Patro, and S. R. Salkuti, "Prediction of diabetes disease using machine learning algorithms," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 284–290, 2022, doi: 10.11591/ijai.v11.i1.pp284-290.

[23] M. Abedini, A. Bijari, and T. Banirostam, "Classification of Pima Indian diabetes dataset using ensemble of decision tree, logistic regression and neural network," *Ijarcce*, vol. 9, no. 7, pp. 1–4, 2020, doi: 10.17148/ijarcce.2020.9701.

[24] E. Nasarian *et al.*, "Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach," *Pattern Recognition Letters*, vol. 133, pp. 33–40, 2020, doi: 10.1016/j.patrec.2020.02.010.

[25] B. A. Tama and K. H. Rhee, "Tree-based classifier ensembles for early detection method of diabetes: an exploratory study," *Artificial Intelligence Review*, vol. 51, no. 3, pp. 355–370, 2019, doi: 10.1007/s10462-017-9565-3.

[26] Y. Wang and L. Feng, "An adaptive boosting algorithm based on weighted feature selection and category classification confidence," *Applied Intelligence*, vol. 51, no. 10, pp. 6837–6858, 2021, doi: 10.1007/s10489-020-02184-3.

[27] P. Theerthagiri and J. Vidya, “Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques,” *Expert Systems*, vol. 39, no. 9, 2022, doi: 10.1111/exsy.13064.

[28] H. Jiang *et al.*, “Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease,” *International Journal of Medical Informatics*, vol. 145, 2021, doi: 10.1016/j.ijmedinf.2020.104326.

[29] D. Ananey-Obiri and E. Sarku, “Predicting the presence of heart diseases using comparative data mining and machine learning algorithms,” *International Journal of Computer Applications*, vol. 176, no. 11, pp. 17–21, 2020, doi: 10.5120/ijca2020920034.

BIOGRAPHIES OF AUTHORS

Usha Sekar received the B.Sc. & MCA. degree, respectively, from Madurai Kamaraj Univeristy. She has worked as an Assistant Professor for 12 yrs in SRM Institute of Science & Technology. Now, currently she is pursuing Ph.D as Full Time Research Scholar in Department of Computer Science, SRM Institute of Science & Technology, Kattankulathur, Chennai, India. Her research area includes Image Processing, Data Mining, Cloud Computing, Machine Learning, and Deep Learning. She has published a paper in international journal and presented paper in national and international conference. She can be contacted at email: us3648@srmist.edu.in

Dr. Kanchana Selvarajan Working as an Assistant Professor in the Department of Computer Science at SRM Institute of Science and Technology, Chennai. She obtained her Ph.D degree from Bharathiar University. She has published more than 15 research papers in National and International Journals and presented paper in a Conferences. She has received Best poster Presentation Award in ISCA-2015. Her research interest includes Data Mining, Machine Learning, IOT, and Cloud Computing. She can be contacted at email: kanchans@srmist.edu.in