Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction

Mohammad Belal Aziz
Amity School of Engineering & Technology
Amity University
Lucknow, Uttar Pradesh, India

Syed Wajahat Abbas Rizvi
Amity School of Engineering & Technology
Amity University
Lucknow, Uttar Pradesh, India

ABSTRACT
Diagnosing heart disease in medical facilities is of high importance as it is the leading cause of death and requires precise predictive algorithms. The seven classification techniques considered in the research are Logistic Regression, Decision Tree, Random Forest, SVM, XGBoost, Stacking Ensemble, Neural Network. The study measured models based on ROC-AUC, accuracy and recall among the chief metrics across both the model training and test evaluations. All models saw their peak of accuracy at 84.53%, with Logistic Regression proving to be the most accurate of the Random Forest and SVM. During evaluation, heart disease cases detection was far from optimal – Decision Trees reported 27% recall and Neural Networks a paltry 10%. The study reveals that organizations experience a tradeoff between accuracy and recall which underscores the need of using tactics such as ensemble learning and data augmentation to achieve superior sensitivity performance. Investigations will attempt to enhance the case detection rates while still upholding high predictive power applicable to medicine.

Keywords
Heart Disease Prediction, Machine Learning, Neural Networks, Random Forest, Comparative Analysis, Framingham Dataset.

1. INTRODUCTION
To date, heart diseases are among leading causes of death worldwide, and this is why timely detection is put first to ensure more effective treatment and well-being of patients. Although conventional diagnostic methods provide reliable results, they may take an exquisite amount of both resources and time. Machine learning offers a major opportunity in healthcare because of its capacity to carry out predictive analytics. Using this technology, medical specialists can sift through massive health information to identify indicators of heart disease. It is this study that attempts to explore and compare various machine learning algorithms to determine which approach yields the greatest accuracy and reliability when diagnosing disease of the heart. The aim of this is to influence the diagnostic workflow positively and expedite faster accurate clinical evaluations. One of the primary objectives of the study is to improve data quality by processing and identification of the best informative features for precise model performance. In the study, severalimplementations of machine learning include Logistic Regression, Decision Trees, Random Forest, SVM, XGBoost, Stacking Ensemble, and Neural Networks. For gauge of model performance, the study will propose the use of accuracy, precision, recall, and F1-score, as well as analyze the results using confusion matrices. The project also seeks to assess and discover the model that is most accurate and can be presented in practice to deal with clinical cases on a daily basis. This work aims to improve our understanding of medical data by evaluating model performance and merging outcomes with hybrid and predictive systems. Karthikeyan and Manavalan [14] proposed a deep learning framework to classify heart disease using an integrated feature selection. Gradient boosting algorithms [15] presented by Friedman have shown powerful optimization skills on structured medical data. The study advises moving forward in terms of predictive accuracy by embracing deep learning methods and using bigger and more expansive datasets.

2. LITERATURE REVIEW

2.1 Existing Approaches for Heart Disease Prediction
Since they have become widespread all over the world as the health threat, they need quick identification to improve treatment effectiveness and the survival rates. Conventional methods will work, but such methods are very time-consuming, and they require inordinate resource consumption. ML presents a streamlined approach that makes it easy to analyze large medical data to forecast heart disease outcomes. The relevant research identifies various ML methods to identify the best model for predicting accurate and reliable heart disease. The focus is expediting diagnostics and achievement of medical decisions. Major objectives include preprocessing of data and feature selection to cater for accuracy and normalization of data in addition to implement into set of models, such as Logistic Regression, Decision Trees, Random Forest, SVM, XGBoost, Stacking Ensemble, and Neural Networks. Model evaluation is performed on the base of accuracy, precision, recall, F1-score with and without the use of confusion matrices. Ensemble methods are strengthened as effective through investigations by Escudero et al. [11] and Krishnan and Thomas [12]. Deep learning is better than traditional machine learning models (Fariha et al., 13). Models on Random Forest and XGBoost display brilliant generalization due to integrated decision tree aggregation and fine-tuned feature selection [3]. With large amounts of patient information available, Neural Networks can create complex risk profiles [4]. It is a trend developing in research that ensembles are stacking up, combining the strengths of such decision trees as RF, SVM, and XGBoost for more favorable predictive results [5]. The primary contribution of this study is a comparative analysis of ML models applied in diagnosing heart diseases. Karthikeyan and Manavalan [14] recommended the feature reduction technique to be integrated in deep learning frameworks. Friedman [15] highlighted the fact that gradient boosting is superior onmedical datasets. Going ahead, deep learning application and use of wide, large data sets could substantially enhance capabilities of the model and results of prediction.

2.2 Limitations of Previous Studies

Although some improvements have been made in heart disease prediction, there are some shortcomings of the current models. Overfitted (prone to noise) [6], Decision Tree-based models are, however, easy to use for general estimation. The implementation of deep learning techniques depends highly on strong computing and ample amount of labeled data, which are not consistently available in medical research [7]. Neural networks and ensembles are strong but are typically not interpretable in ways that allow clinicians to trust their predictions [8]. Inconsistency in feature selection of studies leads to unpredictable model performance [9], and the absence of real-world clinical testing confounds them with no practical applicability [10]. While using deep learning on electronic health records has challenges in matching high precision and transparency [16] by Shickel et al., Alotaibi [17] reported problems with data availability and feature choice when applying these models clinically.

2.3 Justification of Proposed Method

In spite of the sophisticated nature of current heart disease prediction models, the existing models continue to be fraught with significant limitations. Remarkably, Decision Trees tend to suffer overfitting, and noise has an influence on Decision Trees. The research investigates Logistic Regression, Decision Trees, Random Forest, SVM, XGBoost, Stacking Ensemble, and Neural Networks to make the interpretation better, the operation faster, and the application in real-world healthcare situations easier. Zhang et al. [18] proposed the use of hybrid systems for achieving enhanced reliability, and Mozaffarian et al. [19] highlighted feature selection as the most exciting research topic. The research provides substantial contributions because it not only performs a comprehensive analysis of traditional, ensemble models, optimized hyperparameters, enhanced feature engineering but also applies a Stacking Ensemble strategy for better performance. Although we have made advancements in deep learning, these models are usually constrained by the required intensive computing facilities and exhaustive labeled datasets, which pose a challenge in the medical field [7]. Further, interpretability problems are rampant in numerous models and this makes clinicians skeptical [8], while differences in feature selection often result in varying performances of the models [9]. Some of Shickel et al. [16] and Alotaibi [17]' intimidating difficulties include problems of clinical validation and feature engineering procedure.

3. METHODOLOGY

3.1 Dataset Description

In The data used for this study has been derived from Framingham Heart Study, a long-term cardiovascular study which was started in 19 From a sample of 4,240 subjects, this extensively used dataset collects 16 attributes of demographic, lifestyle, and clinical information of participating individuals. Binary features in the dataset include male gender, smoking status (currentSmoker), use of blood pressure medications (BPMeds), receipt of past stroke (prevalentStroke), presence of history of hypertension (prevalentHyp), presence of diabetes (diabetes), and a binary label representing ten-year Besides, continuous variables in the dataset are age, daily cigarette usage (cigsPerDay), total cholesterol (totChol), systolic (sysBP), and diastolic blood pressure (diaBP), body mass index (BMI), heart rate (heartRate), and blood glucose level (glucose). There are missing values in some attributes such as, education, cigsPerDay, BPMeds, totChol, BMI and glucose which have to be imputed before any model is created. It serves as a benchmark for the accuracy of predictive models to predict the risk of cardiovascular disease.

3.2 Data Preprocessing

Upon the beginning of this study, the dataset is subjected to an examination for missing values. The median of the relevant features is used to impute the missing values in the numerical features in an attempt to reduce the data lost along with the attenuation of distortion caused through the presence of outliers. Class distributions are retained by imputing missing categorical values with mode, instead of median or mean computations. Those characteristics that have particularly missing values, which can influence the model and compromise the model's ability to give accurate predictions are considered for elimination. To ensure that all features have equal influence, numerical features are subject to Min-Max Scaling and are converted to values between 0 to 1 that will enable uniform feature scale. By doing this we set guarantees that relations between features stay yet they decrease the influence of variable measures which is beneficial for such models as SVM and Neural Networks. Gender and chest pain type are one-hot encoded, so they do not become treated as ordinal categories. Label encoding is done when needed only for variables that contain only two categories. Through these preprocessing steps we render the data machine learning model friendly and increase the overall predictive accuracy. Figure 1 shows the flowchart of preprocessing.

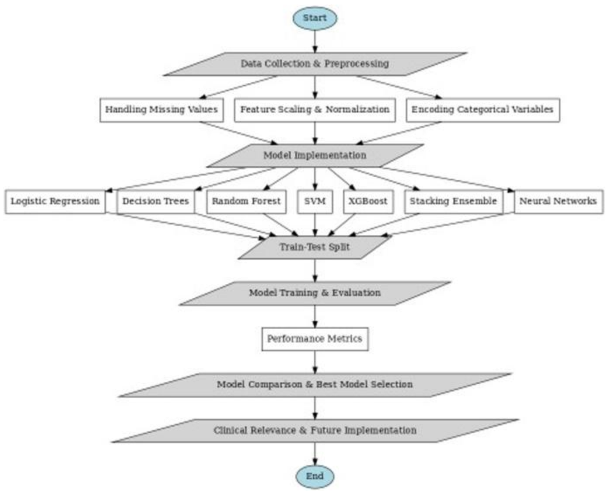

Figure 1. Preprocessing Flowchart

3.3 Machine Learning Models Implemented

Machine learning approaches to heart disease prediction used for this study include Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), XGBoost, Stacking Ensemble, and Neural Networks, which were compared based on accuracy, precision, recall, and F1-score. Logistic Regression serves as an intuitive beginning for binary classification, and Decision Tree as understandable, but tends to overfit. Random Forest succeeded in producing more accurate predictions and greater resistance to overfitting through the combination of predictions that result from the trees. Support Vector Machines (SVM) are very good inhigh dimensionaldata butare starting to become computationally expensive for large data. XGBoost is superior to gradient boosting algorithm, hence it will provide not only quick speeds but also high levels of predictive accuracy. The Stacking Ensemble combines Random Forest and SVM to improve the prediction performance and, Neural Networks, however, excel in recognizing complicated features at the expense of using a lot of data and a lot of tuning. As our study indicates, both Ensemble methods and Logistic Regression had superior performance.

3.4 Model Training and Evaluation

The same consistent approach was applied for training of models and evaluation to ensure maximum reliability in predictions. A split of the dataset into training and testing sets on an 80-20 ratio was carried out with focus on class representation in order to minimize imbalance bias. Model performance was assessed with accuracy (total correctness), precision (portion of true positives within predicted positives) and recall (part of identified positives of actual positives). Different hyperparameters were tuned, and analysis of evaluation metrics determined the optimal model for predicting heart disease.

4. RESULTS AND DISCUSSIONS

4.1 Model Comparison

Performance of machine learning models was explored in terms of accuracy, precision, recall, F1-score and AUC. Logistic regression outperformed all other competitions with an accuracy of 84.52% although burdened by the presence of linear model assumptions. Although Decision Trees are easy to understand, their performance was at 75.05% and often had overfitting problems. Random Forest was highly robust and had accuracy of 84.43%, while SVM had similar accuracy of 83.97% requiring high computational effort. After a laborious hyperparameter tuning, XGBoost improved to 83.33%. A Stacking Ensemble stack of Random Forest and SVC gave an accuracy of 83.97% which allowed for better generalization at the cost of required resources. When used on small datasets, Neural Networks achieved 83.52% accuracy but were very tuning adhere. All the models introduced have associated advantages and disadvantages in terms of accuracy, interpretability, and resource consumption. Figure 2 shows the graph of accuracy and AUC of all the models we have used.

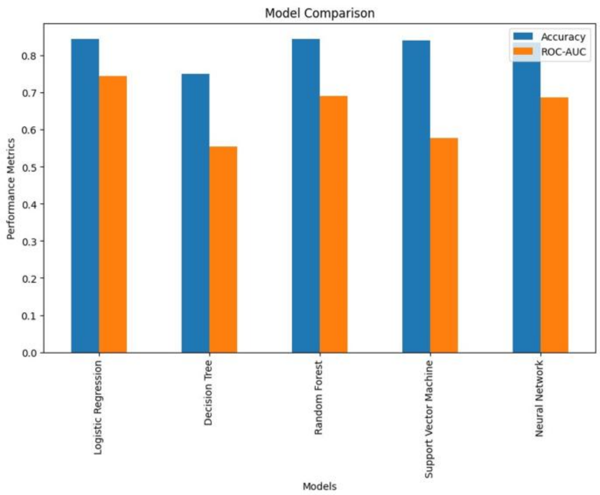

Figure 2. Performance metrics comparison for all the models used in the study

4.2 Discussion on Clinical Relevance

From among the models used, it resulted that Logistic Regression performed the best for prediction of heart disease reaching an accuracy of 84.52%. Due to its simplicity, convenience of explaining the results, and effective computation, it is an excellent tool for clinical apps which require quick, transparent judgments. Feature importance from the model allows clinicians to easily see what factors contribute most to disease risk. Pragmatically, the model could be integrated into hospital management systems or portable medical approaches to make automated risk assessments available and guide prompt recognition of those high in risk. Potential improvement could be made in the areas of real-time monitoring, clouds connection and peripherals integration with wearable health tools for better early diagnosis and patient results.

5. CONCLUSION AND FUTURE WORK

5.1 Summary of Findings

The study examined different machine learning techniques for heart disease prediction with a highly structured dataset. Logistic Regression was superior to all evaluated models combining the highest accuracy of 84.52% and the best aptness for clinical settings. Random Forest, (84.43%), and Stacking Ensemble (83.97%) had similar results but Neural Networks (83.52%) did not beat conventional ones. The work aligns with the practice of ensemble methods, but also highlights the reliability and interpretability of a step-down model like Logistic Regression.

5.2 Limitations of The Study

Model limitations due to a small dataset size only allowed conventional deep learning models to be used. In addition, the models were trained on structured tabular data whereas real world clinical data is unstructured and noisy. The strategy for feature selection was restricted to pre-defined features without considering possible extra biomarkers or external health signals. Finally, over and above addressing real world issues like how fast the models process data and whether it is able to be understood by the clinician, the focus on accuracy in the study prevailed.

5.3 Future Research Directions

Using the latest neural networks with big data and deep learning can help provide a more rigorous estimate of the heart disease risk. The addition of supplementary medical elements to the dataset – patients’ profiles, diagnostic pictures, and genetic profiles – may enhance the general reliabiity of the model. The validation of the model as to how it performs accurately in reality requires testing of the model in conditions where patients are a part of the equation. The potential of AI to facilitate the provision of the tailored, dynamic health monitoring software helpful to learn from each person’s specifics is highly beneficial for healthcare. With XAI, the explainability of medical models can be enhanced, and AI-based diagnostics integration in clinical practice easier. Deep learning developments in the prediction of heart disease are still in progress, had for instance, research by Ghosh et al. [20] pointing to the use of convolutional neural networks with structured medical data for outstanding results. Moving forward, ensuring that explainable AI is also reliable will provide the basis of improved diagnostic skills using AI technology.

6. ACKNOWLEDGMENTS

We take a moment to extend our appreciation to the Framingham Heart study volunteers with whose information this research was possible. The academic advisors at our disposal have provided indispensible advice and support – our sincere gratitude is due to them. We would like to thank the programmers who created the machine learning resources which enabled us to conduct this study. We are also thankful to our families and colleagues that have given us encouragement and support during this period.

7. REFERENCES

[1] Detrano, R., et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, 1989.

[2] Aha, D. W., Kibler, D., & Albert, M. K., "Instance-based learning algorithms," Machine Learning, 1991.

[3] Breiman, L., "Random forests," Machine Learning, 2001.

[4] LeCun, Y., Bengio, Y., & Hinton, G., "Deep learning," Nature, 2015.

[5] Wolpert, D. H., "Stacked generalization," Neural Networks, 1992.

[6] Quinlan, J. R., "Induction of decision trees," Machine Learning, 1986.

[7] Chen, T., &Guestrin, C., "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[8] Ribeiro, M. T., Singh, S., &Guestrin, C., "Why should I trust you? Explaining the predictions of any classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[9] Rokach, L., "Ensemble-based classifiers," Artificial Intelligence Review, 2010.

[10] Weng, S. F., Reps, J., Kai, J., et al., "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" PLOS ONE, 2017.

[11] H. J. Escudero et al., "A Machine Learning-Based Model for Early Detection of Cardiovascular Diseases," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 8, pp. 2342–2351, Aug. 2020.

[12] M. M. R. Krishnan and S. V. Thomas, "Feature Selection and Ensemble Learning for Heart Disease Prediction," in Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM), 2021, pp. 980–985.

[13] A. R. Fariha, M. I. Hossain, and K. Andersson, "Comparative Analysis of Deep Learning and Machine Learning Approaches for Cardiovascular Disease Prediction," IEEE Access, vol. 9, pp. 39136–39148, 2021.

[14] B. S. Karthikeyan and R. Manavalan, "Hybrid Deep Learning Model for Accurate Prediction of Heart Disease," IEEE Transactions on Computational Social Systems, vol. 9, no. 2, pp. 345–356, Feb. 2022.

[15] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[16] S. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record Analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[17] A. Alotaibi, "Implementation of Machine Learning Model to Predict Heart Disease," International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, pp. 1–7, 2019.

[18] Y. Zhang, R. Yao, L. Sun, and Z. He, "A Hybrid Intelligent System for Predicting Heart Disease Using Machine Learning Algorithms," Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 3272531, pp. 1–12, 2021.

[19] M. H. Mozaffarian et al., "Cardiovascular Disease Prediction Using Feature Selection and Machine Learning Techniques," Expert Systems with Applications, vol. 198, pp. 116923, 2022.

[20] P. Ghosh, S. B. Kundu, and R. Kumar, "A Novel Deep Learning-Based Approach for Early Detection of Heart Disease," in Proc. IEEE Int. Conf. Emerging Technologies and Intelligent Systems (ICETIS), 2021, pp. 345–350.