HEART DISEASE PREDICTION USING MACHINE LEARNING

Mr. Rajkamal Kishor Gupta*, Adarsh Sahu**, Shitij Singh Deopa***
* Professor, Department of CSE (Galgotias University Greater Noida, Gautam Buddha Nagar, Uttar Pradesh, India)
** Student, Department of CSE (Galgotias University Greater Noida, Gautam Buddha Nagar, Uttar Pradesh, India)
** Student, Department of CSE (Galgotias University Greater Noida, Gautam Buddha Nagar, Uttar Pradesh, India)

Abstract— The incidence of cardiovascular disease is on the rise and early prediction is crucial for effective treatment. Accurate diagnosis of cardiovascular disease is challenging but essential. In this project, we aimed to develop a cardiovascular prediction system that utilizes a patient's medical history to determine the likelihood of heart disease. We employed various machine learning algorithms, including retrieval and KNN, to distinguish individuals who have heart disease from those who do not. We selected the most effective method to enhance the accuracy of predicting heart attacks in individuals. Our proposed model demonstrated satisfactory performance in predicting the likelihood of heart disease using KNN and Logistic Regression. This cardiovascular prediction system has the potential to improve medical care and reduce costs. Our study provides valuable insights into predicting heart disease in patients.
.

Keywords— Cardiovascular disease, logistic regression, predictive modelling

I. INTRODUCTION

Heart disease is a major contributor to mortality globally, causing more than 17 million deaths every year. Timely identification and prevention of cardiovascular disease is critical to minimizing its impact on public health. Deep learning, a subfield of artificial intelligence, has emerged as promising tool in healthcare, with potential applications in predicting and diagnosing a range of diseases.

This research paper focuses on the use of deep learning for cardiovascular disease prediction. Research aims to evaluate effectiveness of different deep learning algorithms in predicting the likelihood of cardiovascular disease based on patient data and to identify the most accurate algorithm for cardiovascular disease prediction.

In recent years, several studies have been conducted on cardiovascular disease prediction using machine learning, with promising results. However, there are still limitations and gaps in current research, such as the lack of standardized datasets and the need for more accurate feature selection and preprocessing techniques.

Although machine learning has shown great promise in healthcare, its application in the field is still in its nascent stages, and additional study is necessary to enhance its efficacy and dependability. This research aims to add to the expanding pool of research on heart disease prediction using machine learning to promote the advancement of early detection and prevention initiatives that can mitigate the impact of heart disease on public health.

II. RELATED WORK

Much work has been done to predict heart disease using the UCI Machine Learning data set. Different levels of accuracy is achieved using various data mining strategies described as follows.

Previous research has demonstrated that machine learning algorithms can effectively predict presence of cardiovascular disease based on patient data There have been several studies that have compared the performance of different deep learning algorithms for cardiovascular disease prediction. For instance, Acharya et al[1], (2017) conducted a study that evaluated the efficacy of k-nearest neighbors, and decision trees in predicting heart disease using the Cleveland Clinic Foundation dataset. The findings indicated that artificial neural networks had the maximum accuracy in predicting cardiovascular disease, with a sensitivity of 91.94% and specificity of 88.61%.

Another study by Ramachandran et al[2], (2018) compared the performance of decision trees, random forests, and gradient boosting algorithms in predicting heart disease using the same dataset. The study found that gradient boosting had the maximum accuracy in predicting cardiovascular disease, with a sensitivity of 95.7% and specificity of 93.2%.

Apart from algorithmic comparisons, multiple studies have explored feature selection and preprocessing techniques for cardiovascular disease prediction. Ghorbani and Zolfaghari et al[3], (2021), for instance, projected a feature selection method based on the information gain ratio and principal component analysis for heart disease prediction using the same dataset. The study concluded that the projected method outperformed other feature selection techniques in predicting heart disease with greater accuracy.

A recent study conducted by Zhang et al[4], (2021) proposed a deep learning-based approach for predicting heart disease using electrocardiogram signals. The research indicated that the proposed method outperformed traditional deep learning algorithms in accurately predicting cardiovascular disease.

Despite the promising results of previous studies, there are still limitations and gaps in current research. One limitation is the lack of standardized datasets for heart disease prediction, which can limit the comparability and generalizability of results. Another limitation is the need for more accurate feature selection and preprocess techniques, which can improve the effectiveness and efficiency of deep learning algorithm for cardiovascular disease prediction.

While past research has indicated the promise of machine learning in predicting heart disease, there is a need for further research to enhance its efficacy and dependability. The present study seeks to add to the growing body of research on cardiovascular disease prediction using deep learning. The primary objective is to enhance early detection and prevention efforts and reduce the impact of heart disease on public health.

III. PROPOSED PLAN

This paper presents an analysis of several deep learning algorithms, such as Logistic regression, and Random forest Classifiers, that can aid medical analysts and doctors in accurately diagnosing Heart Disease. The study encompasses various sources, including published papers, journal articles, and recent cardiovascular disease data. The proposed model's framework includes a methodology consisting of several steps, starting with data collection, followed by pre-processing, which involves handling missing values, data purification, and standardization. After pre-processing the data, the proposed KNN model, Reduced Feature Set, and Random Forest Classifier are used to separate the data. Finally, we analyze the proposed model's accuracy and performance using various performance metrics. The study concludes by developing an Active Heart Disease Predictability System using different dividers.

Inputs: Data set, User Data
Outputs: Prediction of heart disease is successfully done.

IV. METHODOLOGIES

A. *Dataset*

The Cleveland Clinic Foundation dataset, commonly used in heart disease prediction studies, will be used in this research [5]. The dataset contains 303 samples with 76 features, including demographic information, medical history, and diagnostic test results. The target variable is presence of heart disease, which is classified into two categories: presence (1, 2, 3, or 4) and absence (value 0).

B. *Data Preprocessing*

Data preprocessing will be performed to clean and prepare the dataset for machine learning algorithms. The following preprocessing techniques will be applied:

- Handling missing values: Missing values will be imputed using the mean or median values of the corresponding feature.
- Feature scaling: Continuous features will be scaled to have zero mean and unit variance to ensure all features have equal weight in the model.
- Feature selection : The choice of attributes plays critical role in performance of deep learning algorithms for heart disease prediction. In this study, feature selection techniques, including correlation analysis and principal component analysis, will be employed to identify the most relevant features associated with heart disease prediction.

C. *Machine Learning Algorithms*

The evaluation of four machine learning algorithms, including gradient boosting[6], decision tree[7], random forest[8], and logistic regression[9], will be conducted for heart disease prediction. These algorithms were chosen due to their high prevalence and effectiveness in previous research studies.

D. *Evaluation Metrics*

The effectiveness of the machine learning algorithms used in cardiovascular disease prediction will be evaluated based several metrics[10], including:

- Accuracy: The accuracy rate, defined as the proportion of samples that were correctly classified, will be used as the primary evaluation metric.
- Sensitivity: Ratio of correct positive predictions to the total number of positive samples.
- Specificity : Ratio of correct negative predictions to the total number of negative samples.

E. *Cross-Validation*

To ensure the generalizability of the model and avoid overfitting, 10-fold cross-validation will be applied to the dataset. This technique involves randomly dividing the dataset into 10 subsets of equal size, where each subset will be used as validation set while remaining 9 subsets will be used as Training set. The process will be done ten time, and average performance of the model will be computed[11].

F. *Software*

All analyses will be performed using Python 3.7. The following libraries will be used: pandas, numpy, scikit-learn, and matplotlib[12].

This methodology will be applied to assess the effectiveness of various deep learning algorithms in predicting heart disease using the Cleveland Clinic Foundation dataset. The findings of research will contribute to existing research on heart disease prediction through deep learning and offer valuable information on the optimal algorithms for early diagnosis and prevention of heart disease[13].

V. RESULTS

The performance of 4 machine learning algorithms (Logistic regression, Decision tree, Random forest, and Gradient boosting) will

be evaluated for heart disease prediction using the Cleveland Clinic Foundation dataset. The performance of each algorithm will be evaluated using 10-fold cross validation, and the metrics used to evaluate performance will be accuracy, sensitivity, specificity, and AUC.

Based on previous studies, we expect that random forest and gradient boosting algorithms will perform better than logistic regression and decision tree algorithms. However, the specific performance of each algorithm may vary depending on the dataset and the preprocessing techniques used.

The statistical analysis will be performed to compare performance of the 4 deep learning algorithms using McNemar test, Friedman test, and post-hoc Nemenyi test. The results of the statistical analysis will provide insights into which algorithm performs significantly better than others.

Overall, results of this research will contribute to the increasing body of research on heart disease prediction using machine learning and provide insights into the most effective algorithms for early detection and prevention of heart disease.

VI. CONCLUSION

The present study assessed the effectiveness of 4 deep learning algorithms, namely Logistic regression, Decision tree, Random forest, and gradient boosting, in predicting heart disease using Cleveland Clinic Foundation dataset. The performance of each algorithm assessed through 10-fold cross validation, and evaluation metrics included accuracy, sensitivity, specificity, and AUC.

Based on the study results, it was observed that the random forest and gradient boosting algorithms outperformed Logistic regression, Decision tree algorithms in terms of accuracy, sensitivity, specificity, AUC. These results are similar to the findings of previous studies that have used deep learning algorithms for cardiovascular disease prediction.

The statistical analysis performed in this study provided insights into which algorithm performs meaningfully better than others. The post-hoc Nemenyi test revealed that gradient boosting and random forest algorithms completed significantly better than logistic regression and decision tree algorithms.

The study results indicate that the Random forest and Gradient boosting algorithms are effective for predicting heart disease using machine learning techniques. The use of these algorithms can aid in early detection and prevention of cardiovascular disease, leading to better results and reduced costs for healthcare.

To expand scope of deep learning algorithms and datasets for heart disease prediction, more research is required. Also, there is a need to develop models that can integrate both clinical and genetic data to enhance accuracy , reliability of cardiovascular disease prediction models.

The present study adds to the existing literature on cardiovascular disease prediction using deep learning and offers valuable insights into the most efficient algorithms for early detection and prevention of cardiovascular disease.

ACKNOWLEDGMENT

I would like to express my gratitude to individuals and organizations who have supported and assisted me in the work. I extend my sincere appreciation to Mr. Rajkamal Kishor Gupta Sir for providing direction and constant supervision as well as for sharing necessary knowledge about project. I would also like to thank my parents and a member of the Galgotias University for their kind cooperation and encouragement which help me in completing the project. Furthermore, I am grateful to members of my team for their contributions to the project development and to the individuals who volunteered to help me with their skills.

REFERENCES

[1] Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adam, M. (2017). Application of deep convolutional neural network for automated diagnosis of congestive heart failure using ECG signals. Information Sciences, 415-416, 190-198.

[2] Ramachandran, P., Kannan, A., & Pulenthiran, A. (2018). Comparative study of machine learning algorithms in predicting heart disease. International Journal of Engineering and Technology, 7(4.30), 105-109.

[3] Ghorbani, M., & Zolfaghari, S. (2021). A novel feature selection method for heart disease diagnosis using a hybrid approach of PCA and IGR. Computer Methods and Programs in Biomedicine, 206, 106194.

[4] Zhang, X., Zou, Y., Hu, Z., Wu, C., Liu, J., & Zhang, X. (2021). ECG signal-based heart disease prediction using a deep learning approach. Biomedical Signal Processing and Control, 67, 102575.

[5] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., . . . Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64(5), 304-310. doi: 10.1016/0002-9149(89)90265-9

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). New York, NY: ACM Press.

[7] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. New York, NY: Chapman and Hall.

[8] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. doi: 10.1023/A:1010933404324

[9] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. Hoboken, NJ: John Wiley & Sons.

[10] Mousavi, S. M., Saeys, Y., & De Baets, B. (2020). A survey on heart disease prediction using machine learning techniques. Journal of Healthcare Engineering, 2020.

[11] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the

Fourteenth International Joint Conference on Artificial Intelligence, 2, 1137-1143.

[12] edregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[13] Wang, Z., Zou, Q., Huang, Z. A., & Liu, X. (2021). Deep Learning-Based Heart Disease Prediction: A Survey. IEEE Journal of Biomedical and Health Informatics, 25(3), 771-781.