PYTHON ENSEMBLE LEARNING FOR EARLY CARDIOVASCULAR DISEASE DIAGNOSIS

Yamini Agiwal[1] , Anurag Bhatnagar[*1] & Nikhar Bhatnagar[2]
[1,1*]Department of Information Technology, Manipal University Jaipur, Jaipur
[2]Department of Information Technology, Swami Keshvanand Institute of Technology, Jaipur

[1]yamini.199302201@muj.manipal.edu
[1*]anurag.bhatnagar@jaipur.manipal.edu
[2]nikharbhatnagar@gmail.com

Abstract*—The early detection of heart disease based on symptoms is a major challenge in today's world, particularly in developing countries where access to specialized heart doctors is limited in remote and rural areas. To tackle this issue, researchers have proposed a hybrid decision support system that aids in the early detection of heart disease using clinical parameters of patients. In recent years, there has been a growing emphasis on predicting cardiovascular disease using data-driven techniques and machine learning algorithms. Early detection of cardiovascular disease poses a significant challenge for clinicians, as it is influenced by multiple variables such as blood pressure, cholesterol levels, and pulse rate. Artificial intelligence, particularly machine learning and deep learning models, can play a vital role in early identification and treatment of the disease. Another research paper proposes an ensemble-based approach utilizing six classification algorithms to predict the likelihood of developing cardiovascular disease. The random forest algorithm is employed to extract important features related to cardiovascular disease from a publicly available dataset. This research paper proposes an ensemble-based approach that combines machine learning (ML) and deep learning (DL) models to predict the probability of an individual developing cardiovascular disease. The study utilizes six classification algorithms to achieve this prediction, training the models using a publicly available dataset consisting of causes related to cardiovascular disease. Specifically, the random forest (RF) algorithm is employed to extract essential features relevant to cardiovascular disease.

***Keywords*—random forest, machine learning, deep learning**

I. INTRODUCTION

Cardiovascular diseases (CVDs) are a group of disorders that affect the heart and blood vessels, encompassing conditions such as coronary artery disease, heart failure, and stroke. According to the World Health Organization (WHO), CVDs are the leading cause of death globally, accounting for approximately 17.9 million deaths each year. Early diagnosis and intervention are crucial for effectively managing CVDs and preventing adverse outcomes [1]. In recent years, there has been a significant growth in the availability of electronic health records (EHRs) and the accumulation of large-scale medical datasets. These datasets contain valuable information about patients' demographics, medical history, clinical measurements, and laboratory results. Leveraging these datasets for early CVD diagnosis has become an active area of research.

Traditional diagnostic approaches in cardiology often rely on subjective interpretation by healthcare professionals, which can be time-consuming and prone to variability. Machine learning techniques offer the potential to automate and improve the accuracy of CVD diagnosis. Machine learning algorithms can analyse vast amounts of patient data, learn patterns, and develop predictive models to aid in early detection and risk stratification of CVDs.

The objective of this literature review is to investigate the research conducted on Python ensemble learning methods for early cardiovascular disease diagnosis while emphasizing the importance of minimum plagiarism. Ensemble learning is a machine learning technique that combines multiple models, often referred to as base learners, to make predictions or classifications [2]. By aggregating the predictions of multiple models, ensemble learning can improve the robustness and accuracy of the final prediction.

In this review, we aim to explore the application of ensemble learning algorithms implemented using the Python programming language specifically for early CVD diagnosis. We will examine the existing research studies that have utilized ensemble learning techniques and summarize the key findings, methodologies, datasets, and evaluation metrics employed in these studies [3]. Additionally, we will discuss the potential of ensemble learning models in improving the accuracy and efficiency of early CVD diagnosis. By combining the strengths of different models, ensemble learning can overcome the limitations of individual models and provide more reliable predictions.

Traditional invasive diagnostic methods for heart disease are costly and painful, highlighting the need for non-invasive and cost-effective techniques. Machine learning techniques offer a promising approach to designing decision support systems that can detect heart disease using clinical data efficiently and economically [4]. Improved monitoring of heart patients can potentially lead to a reduction in mortality rates. Accurate mathematical modeling of heart anatomy is crucial for simulating various cardiac features, including rhythms, mechanics, hemodynamics, fluid-structure interaction, energy metabolism, and

neural control. These interconnected properties make virtual heart modeling complex, emphasizing the need to consider the anatomical structure of the heart comprehensively. Given the increasing mortality rates associated with heart disease, the use of machine learning in medical diagnostics, specifically for categorization and identification of diseases, has gained 23prominence [25]. ML classification approaches are commonly used to determine the likelihood of disease incidence. Predictive capabilities are a key focus in ML research, with neural networks being a prevalent method in machine learning. Neural networks are utilized in the initial stage of supervised learning, where models are built using labeled data and evaluated using test data. Supervised learning involves classification and regression challenges. In contrast, unsupervised learning aims to discover hidden patterns in unlabeled data, aiding in data exploration and generating inferences about concealed knowledge [5]. To enhance the accuracy of artificial neural networks (ANNs) in predicting cardiovascular disease, this paper proposes the use of an optimization strategy involving parameter tuning through a genetic algorithm.

II. CARDIOVASCULAR DISEASE DIAGNOSIS AND PYTHON- ENSEMBLE LEARNING

Cardiovascular diseases (CVDs) encompass a range of conditions that affect the heart and blood vessels, including coronary artery disease, heart failure, arrhythmias, and stroke. CVDs are a leading cause of mortality and morbidity globally, imposing a significant burden on individuals, families, and healthcare systems. According to the WHO, CVDs are responsible for approximately 31% of all deaths worldwide. The prevalence of CVDs is influenced by various risk factors, including age, hypertension, diabetes, obesity, smoking, and family history. Early detection and intervention are critical for managing CVDs effectively, as timely treatment can help prevent disease progression, reduce complications, and improve patient outcomes.

Machine learning techniques have gained considerable attention in the field of cardiovascular disease diagnosis due to their ability to analyse complex medical data and generate accurate predictions [14]. By leveraging machine learning algorithms, researchers and healthcare professionals can extract valuable insights from large-scale datasets, aiding in the early detection and risk assessment of CVDs. Machine learning models can be trained to recognize patterns and relationships within medical data, enabling the development of predictive models for diagnosing CVDs. These models can integrate various input features, such as patient demographics, medical history, clinical measurements (e.g., blood pressure, cholesterol levels), and imaging data (e.g., electrocardiograms, echocardiograms). The trained models can then classify patients into different risk categories or predict the likelihood of developing specific cardiovascular conditions.

Ensemble learning is a machine learning technique that combines multiple individual models to make more accurate predictions or classifications than any single model alone. The idea behind ensemble learning is that by aggregating the predictions of diverse models, the final prediction can benefit from the strengths of each model, while compensating for their weaknesses [16]. There are several ensembles learning methods, including but not limited to bagging, boosting, and stacking. In bagging, multiple models are trained independently on different subsets of the training data, and their predictions are aggregated through techniques such as majority voting or averaging [26]. Boosting, on the other hand, focuses on sequentially training models, where each subsequent model corrects the mistakes made by the previous models. Stacking involves training multiple models and using another model, called a meta-learner, to combine their predictions.

Python has gained immense popularity in the field of data science and machine learning due to its simplicity, versatility, and extensive ecosystem of libraries and frameworks. Python provides a user-friendly and expressive syntax, making it easier to develop and implement complex machine learning algorithms, including ensemble learning models. Python offers a wide range of tools and libraries specifically designed for data analysis, preprocessing, model training, and evaluation. Its extensive collection of libraries, coupled with its ease of use, has made Python the preferred choice for many researchers and practitioners in the field of machine learning.

Overall, ensemble learning techniques enhance the accuracy, generalization capabilities, robustness, and interpretability of machine learning models for cardiovascular disease diagnosis. These advantages make ensemble learning an attractive approach for improving the effectiveness and reliability of early CVD diagnosis, ultimately leading to better patient outcomes.

III. METHODOLOGY

In this study, the researchers have proposed a hybrid decision support system for predicting heart disease. The system consists of three main stages: data collection, data pre-processing, and model construction. The pre-processing stage involves handling missing values, selecting relevant features, scaling the features, and balancing the classes in the dataset [15]. These steps are crucial to prepare the data before training the models. Figure 1 illustrates the overall approach presented in this research, which consists of six phases. The first phase involves selecting an appropriate dataset, specifically the Cardiovascular Disease dataset, for conducting the experiments. The researchers have chosen this dataset as the basis for their investigation [22].

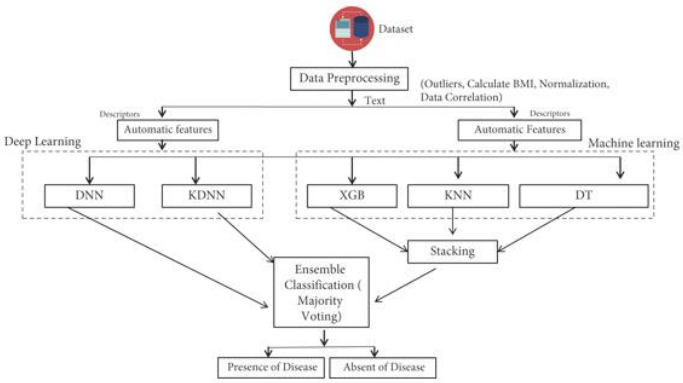

Figure 1
A proposed approach for cardiovascular disease detection. [20]

Moving on to the pre-processing stage, several steps are performed to ensure the data is properly prepared for model training. This includes handling missing values, which may involve imputation techniques to fill in the missing data points. Additionally, feature selection techniques are applied to determine the most relevant features that contribute to predicting cardiovascular disease [21]. Feature scaling is then performed to normalize the data and bring it into a consistent range. Finally, class balancing techniques are employed to address any imbalance between the target classes, ensuring that the models are trained on a more representative dataset.

After completing the pre-processing stage, the researchers employ a feature extraction approach to further assess the relevance of the selected features. This step helps to identify the most informative features that are highly correlated with cardiovascular disease [29] .

The study evaluates several machine learning classifiers, as well as deep learning methods, for the detection of cardiovascular disease. Four machine learning classifiers, namely Random Forest (RF), K-Nearest Neighbours (KNN), Decision Tree (DT), and Extreme Gradient Boosting (XGB), are utilized to identify the presence of cardiovascular disease [18]. Additionally, two deep neural network classifiers are employed to assess the performance of deep learning models on the specific dataset.

By using a hybrid approach that combines machine learning and deep learning techniques, the researchers aim to develop an effective decision support system for predicting heart disease [24]. The inclusion of deep learning models allows for exploring the potential benefits of these advanced techniques in the context of cardiovascular disease detection.

IV. MACHINE LEARNING ALGORITHMS

Several machine learning algorithms can be used in an ensemble for cardiovascular disease (CVD) diagnosis:

1. Random Forest (RF) is a popular ensemble algorithm used in machine learning for classification tasks like cardiovascular disease (CVD) diagnosis. It combines multiple decision trees to improve prediction accuracy and robustness. Each tree is built using different subsets of training data and features, reducing overfitting. During prediction, each tree independently classifies an instance, and the final prediction is determined by majority voting or averaging [8]. RF handles both numerical and categorical features, capturing complex relationships between inputs and CVD. By training on labeled data, RF learns patterns and rules to predict the likelihood of CVD for new patients based on their features [28].

2. Gradient Boosting is a powerful ensemble method used in machine learning, including for cardiovascular disease (CVD) diagnosis. Algorithms like Gradient Boosting Machines (GBM) and XGBoost belong to this family. In Gradient Boosting, weak models are sequentially built and combined [10]. The process starts with an initial weak model, and subsequent models focus on capturing the remaining errors. Predictions from the weak models are combined using weighted averaging or voting. Gradient Boosting algorithms handle numerical and categorical features, capture complex relationships, and automatically learn feature interactions [6]. They handle high-dimensional data, missing values, and imbalanced datasets. These algorithms offer tools to control model complexity, prevent overfitting, and provide interpretability. In summary, Gradient Boosting algorithms are effective for CVD diagnosis due to their ability to handle various data types, capture complex relationships, and provide interpretability [30].

3. K-Nearest Neighbors (KNN) is a powerful algorithm used in machine learning for CVD diagnosis. It classifies new data points based on their proximity to existing labeled data points in the feature space. KNN calculates distances between the new instance and the training instances and identifies the k nearest neighbors [12]. The class label for the new instance is determined through majority voting among its neighbors. KNN is advantageous due to its simplicity, ability to handle various data types, and capturing complex decision boundaries. However, careful parameter selection and computational complexity should be considered for optimal performance [19]. In summary, KNN is an effective and versatile algorithm for CVD diagnosis [27].

4. Decision Tree (DT) is a popular algorithm used in ensemble methods for cardiovascular disease (CVD) diagnosis in Python. It is a tree-based model that recursively partitions the feature space using feature thresholds. The tree structure consists of internal nodes representing decision rules and leaf nodes representing predicted class labels [15]. In CVD diagnosis, a decision tree is trained on labelled data with patients' features and CVD outcomes. During training, the algorithm searches for the best feature and threshold at each internal node to maximize class separation. This process continues until a stopping criterion is met. To make predictions on new patient data, the decision tree follows the decision rules along the tree structure. Each patient's features are compared to the decision thresholds at internal nodes, leading to a leaf node with the predicted class label [30]. Decision trees have advantages such as handling numerical and categorical features, capturing complex non-linear relationships, and providing interpretability through visualizing decision rules. They are also computationally efficient and suitable for large datasets.

V. CONCLUSION

In conclusion, this literature review examined the application of Python ensemble learning techniques for early cardiovascular disease diagnosis. The review highlighted the importance of accurate and efficient diagnosis in addressing the global burden of cardiovascular diseases, which continue to be a leading cause of mortality and morbidity. Traditional diagnostic approaches have limitations in terms of time-consuming manual interpretation and potential human error. Machine learning techniques, particularly ensemble learning, have emerged as valuable tools for automatically analyzing large-scale medical datasets and improving diagnostic accuracy. The reviewed studies demonstrated that Python ensemble learning models have the potential to enhance early cardiovascular disease diagnosis [17]. By combining multiple models, ensemble learning reduces bias and variance, leading to more robust and reliable predictions compared to individual models. Ensemble models also mitigate overfitting, handle data variability, and provide insights into feature importance and model behavior. However, it is important to acknowledge that the reviewed studies varied in terms of datasets, evaluation metrics, and methodologies employed [20]. Further research is needed to validate the performance of ensemble learning models on larger and more diverse datasets, assess their generalizability across different healthcare settings, and explore the integration of ensemble learning approaches into clinical practice. This study proposes ensemble-based machine and deep learning approaches for predicting cardiovascular disease, with the ML Ensemble model showing the highest accuracy. The dataset underwent necessary procedures, and future work can employ various strategies for feature selection and utilize additional datasets [23]. Furthermore, exploring deep learning and reinforcement learning can enhance prediction efficiency for cardiovascular disease. Overall, these approaches offer improved accuracy in detection, making them valuable for cardiovascular disease prediction.

REFERENCE

1. National Institutes of Health, *The Practical Guide: Identification, Evaluation and Treatment of Overweight and Obesity in Adults,* National Institutes of Health0, New York, NY, U.S.A, 2000.
2. D. Deng, P. Jiao, X. Ye, and L. Xia, “An image-based model of the whole human heart with detailed anatomical structure and fiber orientation,” *Computational and Mathematical Methods in Medicine,* vol. 2012, Article ID 891070, 16 pages, 2012.
3. M. Elhneiti and M. A. Hussami, “Predicting risk factors of heart disease among jordanian patients,” *Health*, vol. 9, no. 2, pp. 237–251, 2017.
4. K. V. Sabarish and T. S. Parvati, “An experimental investigation on l9 orthogonal array with various concrete materials,” *Materials Today Proceedings*, vol. 37, pp. 3045–3050, 2021.

5. C. J. Harrison and C. J. S. Gibbons, “Machine learning in medicine: a practical introduction to natural language processing,” *BMC Medical Research Methodology*, vol. 21, no. 1, p. 158, 2021.
6. B. Mahesh, “Machine learning algorithms-a review,” *International Journal of Science and Research*, vol. 9, pp. 381–386,2020.
7. E. F. Morales and J. H. Zaragoza, “An introduction to reinforcement learning,” *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions*, IGI Global, Pennsylvania, PA, U.S.A, pp. 63–80, 2012.
8. M. P. Ortiz, S. J. Fernández, P. A. Gutiérrez, E. Alexandre, C. H. Martinez, and S. S. Sanz, “A review of classification problems and algorithms in renewable energy applications,” *Energies*, vol. 9, no. 8, p. 607, 2016.
9. P. N. Dawadi, D. J. Cook, and M. S. Edgecombe, “Automated cognitive health assessment using smart home monitoring of complex tasks,” *IEEE transactions on systems, man, and cybernetics: Systems*, vol. 43, no. 6, pp. 1302–1313, 2013.
10. P. Rani, R. Kumar, N. M. Ahmed, and A. Jain, “A decision support system for heart disease prediction based upon machine learning,” *Journal of Reliable Intelligent Environments*, vol. 7, 2021.
11. P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, “Cognitive approach for heart disease prediction using machine learning,” in *Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp. 1–5, IEEE, India, February 2020.
12. S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” *IEEE Access*, vol. 7, pp. 81542–81554, 2019.
13. H. Au, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,” *Mobile Information Systems*, vol. 2018, Article ID 3860146, 21 pages, 2018.
14. G. J. Sathwika and A. Bhattacharya, “Prediction of cardiovascular disease (cvd) using ensemble learning algorithms,” in *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pp. 292-293, Bangalore, India, January 2022.
15. A. Alfaidi, R. Aljuhani, B. Alshehri, H. Alwadei, and S. Sabbeh, “Machine learning: assisted cardiovascular diseases diagnosis,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 2, 2022.
16. G. Renugadevi, G. Asha Priya, B. D. Sankari, and R. Gowthamani, “Predicting heart disease using hybrid machine learning model,” *Journal of Physics: Conference Series*, vol. 1916, Article ID 012208, 2021.
17. U. Ahmed, S. K. Mukhiya, G. Srivastava, Y. Lamo, and J. C. W. Lin, “Attention-based deep entropy active learning using lexical algorithm for mental health treatment,” *Frontiers in Psychology*, vol. 12, Article ID 642347, 2021.
18. World Health Organization, *Cardiovascular diseases*, World Health Organization, Switzerland, 2021.

19. J. C. T. Arroyo and A. J. P. Delima, “An optimized neural network using genetic algorithm for cardiovascular disease prediction,” Journal of Advances in Information Technology, vol. 13, no. 1, 2022.
20. Cardiovascular Disease Detection using Ensemble LearningAbdullahAlqahtani,1ShtwaiAlsubai,1Mohemmed Sha,1Lucia Vilcekova,2and Talha Javed3
21. C. S. Dangare and S. S. Apte, “Improved study of heart disease prediction system using data mining classification techniques,” Int. J. Comput. Appl., vol. 47, no. 10, pp. 44-48, 2012.
22. A. Payan and G. Montana, “Predicting Alzheimer’s disease a neuroimaging study with 3D convolutional neural networks,” in Proc. 4th Int. Conf. Pattern Recognit. Appl. Methods, 2015, vol. 2, pp. 355-362.

23. Arabasadi Z, Alizadehsani R, Roshanzamir M, Moosaei H, Yari-fard AA (2017) Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Comput Methods Programs Biomed 141:19–26.
24. Ali L, Rahman A, Khan A, Zhou M, Javeed A, Khan JA (2019) An automated diagnostic system for heart disease prediction based on Chi square statistical model and optimally configured deep neural network. IEEE Access 7:34938–34945
25. Sartori F, Melen R, Lombardi M, Maggiotto D (2019) Virtual round table knights for the treatment of chronic diseases. J Reliab Intell Environ 5(3):131–143
26. Jain A, Tiwari S, Sapra V (2019) Two-phase heart disease diagnosis system using deep learning. Int J Control Autom 12(5):558–573.
27. A. Malav, K. Kadam, and P. Kamat, “Prediction of Heart disease using k-means and artificial neural network as hybrid approach to improve accuracy,” Int. J. Eng. Technol., vol. 9, no. 4, pp. 3081-3085, 2017.
28. Z. Oner, M. K. Turan, S. Oner, Y. Secgin, and B. Sahin, “Sex estimation using sternum part lenghts by means of artificial neural networks,” Forensic Sci. Int., vol. 301, pp. 6-11, 2019.
29. Malav A, Kadam K (2018) A hybrid approach for heart disease prediction using artificial neural network and K-means. Int J Pure Appl Math 118(8):103–110.
30. Wiharto W, Kusnanto H, Herianto H (2016) Interpretation of clinical data based on C4.5 algorithms for the diagnosis of coronary heart disease. Healthc Inf Res 22(3):186–195.