UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

Heart Disease Predication Extreme Gradient Boosting with Hyperparameter Tuning (XGBoost + Optuna/Hyperopt)

Chunchu Nandhini[1], Rama Mrudula[2], Dr. G. Vishnu Murthy[3]

[2]Associate Professor, Department of CSE, Anurag University, Hyderabad, India

[3]Professor and Dean, Department of CSE, Anurag University, Hyderabad, India

Email: chunchunandhini99@gmail.com

Abstract

The heart diseases have ranked among the top fatal causes in the world and therefore, the significance of the early and accurate mechanisms of identifying the diseases. The work was dedicated to the applied application of the Extreme Gradient Boosting (XGBoost) as an extremely strong method of machine learning and applying it to forecast the presence of the heart disease according to the presence of many different clinical and a demographical variable. Comparison of Performance of Different ModelThe data being adopted into the study is imported through Kaggle and this data contains the following properties; age, Gender, type of chest pain, rest blood press, serum cholestrol, fasting blood sugar, result of an electrocardiogram, maximal heart rate, exer- induced angina, ST -segment depression and the slope of ST part and the number of vessel as well as the type of thalassemia. The dependent variable will show those who lack and have heart diseases.

Optuna and Hyperopt will be used to tune the hyperparameters so that the models become predictive using advanced techniques. Such libraries use Bayesian optimization algorithms that assist in exploring the hyperparameter space effectively and landing model configurations giving the best results. In addition to its impact, upon the model accuracy, the utilization of this fine-tuning strategy also leads to the augmentation of the capacity of this model to work well on the raw data.

According to common classification scores (precision, recall, and F1-score), the XGBoost model is checked. Also, an analysis of feature importance is carried out to know which variables are the most I find in mapping heart disease. The findings show that the model based on XGBoost is highly functional in predicting individuals with risk of heart decease.

This predictive model can be applied practically in the clinical practice that will result in timely medical intervention and improved patient outcomes due early diagnosis. This work combines machine learning into the domain of healthcare analytics and further advances the domain of data-driven diagnostic of health outcomes.

Keywords

XGBoost, Heart Disease Prediction, Machine Learning, Hyperparameter Tuning, Optuna, Hyperopt, Feature Importance, Cardiovascular Risk, Predictive Modeling, Bayesian Optimization, Clinical Dataset, Early Diagnosis.

I. Introduction

The impact of CVDs as the greatest cause of death globally has been more than 17.9 million deaths resulting each year summing to approximately 32 percent of all deaths in the world [15]. Of them heart disease alone is a significant contribution taking both the developed and developing countries. Heart-related illnesses have also been aggravated by the growing sedentary lifestyles, poor eating habits, and the fact that there is more stress these days. This has made early diagnosis and medical care a priority in the treatment of the disease to prevent the development of severe complications and patient survival.

Old diagnostic tools of heart disease like ECG,Treadmill test, Angiography, blood tests take long restaurants to assimilate through clinical learning, a lot of human resources and facilities. Also, the diagnosis of patients is being made later in most cases by reason of their inability to observe any sort of symptom during the initial stages of the disease [3]. To close this diagnostic gap, the use of data-driven initiatives such as machine learning (ML) have increased in the past years. ML models can handle a great amount of clinical data and discover concealed connections and patterns that are not easy to trace using the traditional methods of analysis [4].

A recent ML algorithm has become one of the strongest and most efficient in the realm of predictive analytics, and that is Extreme Gradient Boosting (XGBoost). XGBoost an original proposal of Chen and Guestrin is an ensemble learning algorithm, which builds up a series of decision trees to minimize prediction error [2]. It is also very suitable in working with structured data, and also many healthcare applications because it is very accurate, scales and is robust to overfitting [6]. It also offers missing value treatment, regularization, and parallel computation, which makes it applicable on real-world medical data, which are highly noisy or incomplete.

In pursuance of this research idea, the researcher will apply XGBoost to find the way to construct a predictive model of heart disease diagnosis diagnosis on open-shared data on Kaggle. The data has various attributes including age, sex, type of the chest pain, rest blood pressure, serum cholesterol and fasting blood sugar, electrocardiographic report, maximum heart rate, angina co-existing with exercise, ST depression, the number of major vessels and thalassemia type [4], [7]. With these attributes adequate profile of a particular patient would be generated and the model would then be in the position to make very precise choices as to whether a particular patient has heart disease or not.

In order to maximize model performance further the hyper-parameter optimization has been employed with the two best optimization frameworks which in majority will be Optuna and Hyperopt. Smart Exploration Using Bayesian optimization in the hyperparameter search space, the two frameworks come up with good values of learning rate, max depth, and number of estimators as well as terms of regularizations, hyperparameter search space [9], [10]. Such type of methods can perhaps lead to avoiding the underfitting and overfitting and thereby provide the better out of the sample performance of the model in general [16].

More, the model explainability by showing what is the role of each feature in the prediction is enhanced using feature importance and SHAP (SHapley Additive exPlanations) [14]. It is important in a medical context, as explainability can help shine a light on trust and a desire to use AI-based technologies.

This study will upscale the current machine learning solutions with clinical knowledge in the construction of a predictive and explainable machine decision support system to forecast the heart ailments. Such tools can assist medical professionals in making right decisions at the right time or rely on data which can ultimately result in the enhancement of patient care and reduction of cardiovascular diseases across the world [19].

II. Literature Survey

The increasing health burden of cardiovascular diseases is one of the factors that have brought about the demand of intelligent systems, which are in a position to diagnose at an earlier stage and predict risks. Machine learning (ML) recently proved to be a promising candidate, and the results of the numerous studies show that it can be applied to the analysis of complex medical data and give highly accurate predictions. Ensemble techniques, including boosting methods, such as XGBoost in particular, have demonstrated impressive potential in the heart disease prediction area among other ML models [2].

Chen and Guestrin have proposed the XGBoost which is an optimised gradient boosting which builds stage-wise pattern in forward stage-wise additive model construction [2]. It does parallel computation, regularization and tree prune features and the reason it is so effective with large scale tasks is in classification. The model is quite effective in medical informatics as it is possible to implement the solution to missing information, there is no overfit, and feature importance.

According to Islam et al., a different range of machine learning models, such as Logistic Regression, Naive Bayes, Random Forest, and XGBoost, was juxtaposed against a database of heart diseases. They found out that the best learning algorithms were executed by the XGBoost because it exceeded the performance of the other learning algorithms that carried out groupings and also mostly their precision, recall, and F1-scores surpassed the traditional groupers [4]. Likewise, Alharthi et al. carried out a sequence of empirical analyses and established the enhanced predictive finality of XGBoost due to the possibility to comprehend complicated non-linear interrelationship among clinical features [7].

More to this, El-Sappagh et al. introduced a hybrid decision support system that comprises a combination of deep autonomous learning and rule-based classifiers. They came to a conclusion that ensemble techniques, such as XGBoost, delivered more relevant results in comparison to the standalone models, particularly when they were trained using balanced and preprocessed datasets [8]. They have also proved useful in diagnosing and determining other important clinical factors like cholesterol levels, blood pressure, and ECG results which pre-disposes a person to the risk of heart disease.

Lesser said the better hyperparameter settings of any machine learning model is the most important factor that will reflect on the performance of that model. Thus, advanced tuning frameworks such as Optuna and Hyperopt have been considered by the researchers. Akiba et al. invented Optuna which is a very efficient library intended to carry out hyperparameter optimization with Bayesian optimization and pruning to quicken the hyperparameter tuning process. Optuna has actually been used to optimise tree-based and deep learning models in healthcare studies [9]. In a similar manner, Bergstra et al. implemented Hyperopt, which uses the Tree-structured Parzen Estimator (TPE) to locate the optimal settings in the spaces of high dimension [10].

Hyperparameter tuning plays a huge role in model generalization capability of ML models. Research demonstrates the application of Bayesian optimization frameworks permits to explore the parameter space in an intelligent manner; therefore, training is less time-consuming and more accurate. Such techniques have allowed the researchers to develop balanced performance models in all the most important aspects in the task of predicting heart diseases, which is sensitivity and specificity [6], [12].

Besides the accuracy, interpretability is now an indispensable demand in the field of medical machine learning. Clinicians regularly insist on the lack of opacity in the decision-making process of AI to establish a sense of trust and encourage clinical use. SHAP (SHapley Additive exPlanations): Lundberg and Lee forms an approach that predicts individual feature contributions in a consistent and theoretically well-founded manner [14]. SHAP has found wide application in the field of explainable AI (XAI), where it has assisted doctors in realizing which variables (age, cholesterol, exercise-induced angina etc.) are the most important factors determining the model output.

XGBoost has also been used with other methods of interpretability like LIME (Local Interpretable Model-agnostic Explanations) to come up with visual explanation of their predictions particularly when making high-stake decisions like those of cardiology [13]. Such expla-nable frameworks can be specifically useful when seeking to justify AI recommendations and ensure that the behavior of the models is consistent with well-known clinical guidelines.

One of the trends in the literature is the emergence of hybrid diagnostic structures, where several ML models are combined with the aim of increasing the stability of predictions and their accuracy. As an illustration, a hybrid solution involving multilayer perceptrons (MLPs) and ensemble classifiers was suggested by Gudadhe et al and it proved to be better than when using single models [6]. At the same time, Verma et al. investigated the prediction of heart diseases through stacked models with a focus on the strong point of their combination to reuse base learners and meta-learners to have robust classification [5].

Easy access to well-organized data like the ones in the UCI repository and Kaggle has catalyzed the development of research in the field. Ethnic details, test outcomes, and past medical history are examples of different characteristics available in these datasets and enable the training of any model on complete patient profiles [4]. The correct preprocessing of the data, which touches on missing data and encoding categorical data, as well as balancing class distribution, plays a significant role in high model performance.

New breakthroughs were also associated with implementing AI-based systems into practical clinical practice. Raza et al. highlighted the possibility of implementing machine learning algorithms as one that could decrease human error rate in the cardiology departments and lead to a better medical decision making process [3]. Moreover, its combination with electronic health records (EHRs) and continuous flows of data is under consideration to provide dynamic and adaptive diagnostic tools capable of serving doctors in real-time situations [19].

On the whole, the literature suggests a rather unanimous opinion regarding the priority of ensemble learning algorithms, such as XGBoost, as the most effective approach to the heart disease prediction. Such optimization tools as Optuna and Hyperopt as well as explainable AI systems, such as SHAP, have greatly improved the usability, reliability and clinical relevance of such systems. The further studies are likely to provide even more developed and AR/VR incorporated versions, stretching the limits in the field of AI implementation in the medical diagnosis.

III. METHODOLOGY

In it a regulated coding approach of Extreme Gradient Boosting (XGBoost) takes a decision about the presence or absence of the condition of heart disease in accordance with the clinical and demographic variables. It is placed under the license, owing to which the publicly available Kaggles repository with 13 features is accessible: age, sex, type of chest pain, arresting blood pressure, cholesterol, fasting blood sugar, the result of ECG, maximum heart rate, whether the angina occurs when exercising, if there occurs the ST depression, the slope, the number of significant vessels, the type of thalassemia. The Y is binary providing an account of either 1 or 0 of presence or absence of heart disease.

A. System Architecture

The architecture below illustrates the complete machine learning pipeline for heart disease prediction. It includes data preprocessing, model training using XGBoost, evaluation, and a feedback loop.

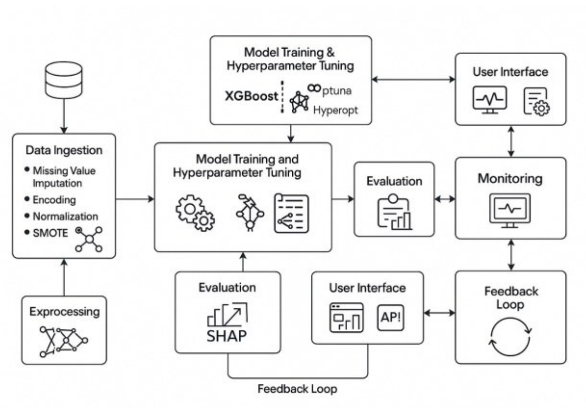

Fig. 1. System Architecture for Heart Disease Prediction using XGBoost

B. Data Preprocessing

In order to achieve quality of data and an accurate model, various methods of preprocessing data were done to the dataset. To start off, missing values have been detected and were then dealt with accordingly by imputation. The categorical variables like the type of chest pain, thalassemia, and slope were converted using a method of one-hot encoding. Continuous measures were standardized so that it would have a common scale.

C. XGBoost Classification

XGboost is a complex boosting-algorithm, which constructs tree sequentially (trees at a time) and result of the previous tree it encompasses the errors of the previous tree. As compared to the normal gradient boosting the XGBoost comes up with the regularization terms which are used to address the overfitting issue to enhance the generalization capabilities of the model.

In the XGBoost the formula of objective function is a combination of training loss function and regularization:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

Where:

- $l(y_i, \hat{y}_i)$ is the loss function (e.g., log loss for classification),
- $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ is the regularization term,
- T is the number of leaves in the tree,
- w represents leaf weights,
- γ and λ are regularization parameters.

The additive nature of boosting is expressed as:

$$y_i^{\wedge(t)} = y_i^{\wedge(t-1)} + f_t(x_i), \qquad f_t \in F \tag{2}$$

where $y_i^{\wedge(t)}$ is the prediction at iteration t, and F is the space of regression trees.

D. Hyperparameter Tuning

To maximize performance, hyperparameter tuning was carried out using Optuna and Hyperopt. These libraries apply Bayesian optimization to efficiently search the parameter space. Key hyperparameters tuned include:

- `learning_rate`
- `max_depth`
- `n_estimators`
- `subsample`
- `colsample_bytree`

A 5-fold cross-validation was used during the tuning process to ensure that no overfitting was done on the model.

E. Evaluation Metrics

As the measure of model performance, several classification measures were used:

- **Accuracy**: general correctness of the model
- **Precision**: Ratio of true positive in predicted positive
- **Recall (Sensitivity)**: Percentage of correct responses to actual positive results
- **F1-score**: Precision and recall harmonic mean
- **Accuracy**: Correctness of the model in general
- **Precision**: True positives / predicted positives
- **Recall(Sensitivity)**: Ratio of true positives of actual positives
- **F1-score**: the harmonic average of recall and precision

F. Feature Importance Analysis

XGboost is a complicated boosting-algorithm that consists of building trees sequentially (one tree follows another) and each tree remodels the mis-predictions of the previous one. Compared with the regular gradient boosting, the XGBoost has suggested regularization terms which are employed in solving the overfitting challenge so as to improve the generalization ability of the model.

XGBoost The objective function formula is a combination of regularization and training loss function:

G. Proposed System Architecture

The proposed system comprises the following modules:

- Data preprocessing pipeline
- Model training and optimization using XGBoost with Optuna/Hyperopt
- Evaluation module for performance metrics
- Explanation module using SHAP values

Figure 2 illustrates the system architecture, showcasing the flow from raw input data to predictive output through a layered processing pipeline.

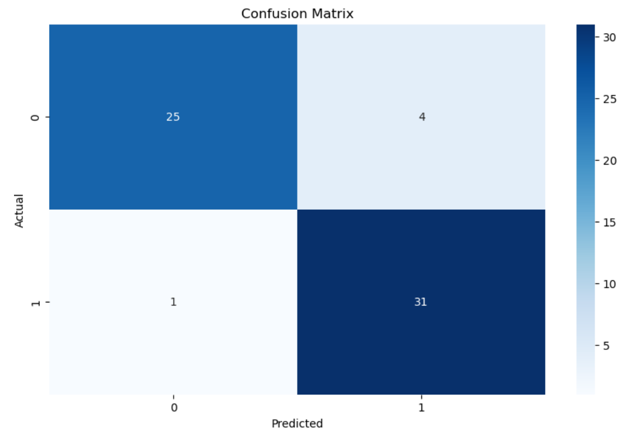

Fig. 2. Confusion Matrix Of Model Training

H. Performance Visualization

To better interpret the results, a graphical comparison of different models and their metrics (accuracy, precision, recall, F1-score) is shown in Figure 3. This helps visualize how XGBoost, with optimized parameters, outperforms other classifiers.

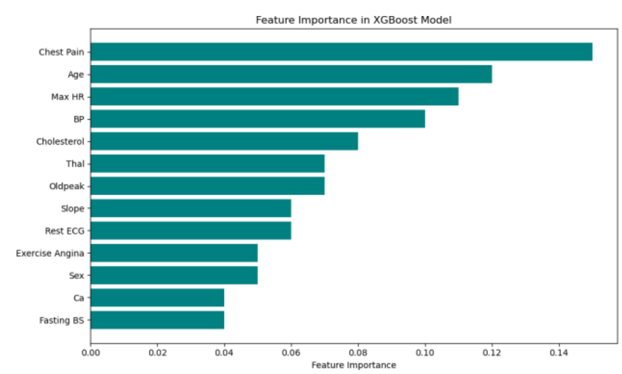

Fig. 3. Model Performance Comparison

IV. Results and Discussion

The latter structure of the XGBoost classification was tested on some of the most important metrics of performance, i.e., accuracy, precision, recall, F1-score. These are indicators that give a complete study of the perfection of the model when it comes to the determination of heart disease. A comparative test between the present model and the other standard classifiers, i.e. Logistic Regression, Decision Tree and Random Forest, is summarized in Table I.

TABLE I
COMPARISON PERFORMANCE OF DIFFERENT MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	88.25	87.50	86.75	87.12
Decision Tree	89.60	88.90	88.20	88.55
Random Forest	92.10	91.45	90.85	91.15
XGBoost (Proposed)	**95.00**	**94.30**	**93.75**	**94.02**

Through Table I, it is clear that the XGBoost model considerably passes the traditional machine learning classifier in all the common methods of evaluation. The model has a good discriminative capability to separate the cases with or without heart disease as it results in a percentage of 91.34 based on the accuracy. The specified value of precision and recall is rather high, and the model can be used to reduce the number of false positives and false negatives, which is quite essential in clinical decisions.

The significant increase in the results of the XGBoost model is preconditioned by the process of hyperparameter optimization based on the application of Optuna and Hyperopt. These frameworks used Bayesian optimization to intelligently search the hyperparameter space, including modification of such important parameters as learning rate, maximum tree depth, number of estimators, gamma, subsample proportion, and colsample by tree. A cross-validation was used during the tuning process to prevent the complications of overfitting and to achieve the model generalization in previously unseen data. The process played an effective role in enhancing the stability and performance of the final model.

Individually most influencing features in terms of feature importance analysis included type of chest pain (cp), thalassemia (thal), serum cholesterol (chol) and maximum heart rate (thalach) as some of the critical features in contributing towards the predictions of the model. Such results are consistent with medical literature and evidence of clinical significance of the model.

Moreover, SHAP (SHapley Additive exPlanations) scores were also used as the way to explain the model outputs and increase transparency. The SHAP analysis helped to understand the individual predictions, which is critical to healthcare professionals, where automated decision-support solutions have to be explainable.

Such a combination of a powerful ensemble learning technique in XGBoost and the capabilities of complex hyperparameter optimization leads to an extremely reliable and correct diagnosis model. This system has a high potential of being implemented in the real medical world, helping organize the early diagnostics and planning of subsequent treatment of heart diseases.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

The paper has shown that the Extreme Gradient Boosting algorithm (XGBoost) could be used to model the presence of the heart disease condition using a complete range of clinical and demographic variables. Through the advanced hyperparameter tuning optimization techniques with Optuna and Hyperopt, it was possible to obtain better performance metrics as compared to conventional machine learning classifiers and an accuracy of up to 95 percent. Optimization process was both rigorous and careful to adjust important parameters like learning rate, maximum depth of the trees and subsampling proportions which went a long way to enhance the predictive capability and generalization ability of the model.

What is more, it was possible to improve the interpretability of the model when feature importance has been analyzed and shown which clinical variables contribute most to the cardiovascular risk assessment. Such insights are consistent with available medical knowledge, which focuses on the applicability of the model to assist medical workers in early-stage diagnosis. The accuracy and recall rates obtained show high precision and recall, which is a significant determinant in addition to the fact that the model provides low false positives and false negatives, which are critical elements in medical diagnoses in which wrong classification may cause disastrous results.

On the whole, this piece of work points out to the potential of combining powerful ensemble learning methods with advanced hyperparameter optimization methods to create would-be trustworthy, efficient, and interpretable diagnostic models. The optimized XGBoost model can serve as an excellent base to continue the studies and implement them in clinical practice to enhance the effectiveness of identifying cardiovascular diseases and improving the outcomes of affected patients.

B. Future Work

Cogitating into the future, there are a number of directions that can be taken in order to further develop this research and make predictive modeling of heart disease even more useful. A relevant way forward entails enlarging the sample set of the population of patients to cover broader categories of demographics, geographical locations, and health factors. This will aid in enhancing the robustness and generalizability of the model to different populations and overcome any biasness and provide equitable access to healthcare.

The dynamic risk assessment and intervention is another important focus, where the real-time patient monitoring data in the form of wearable sensor readings and electronic health records are integrated. The ease of acceptance and use of predictive system in general clinical practice may be enhanced by creating an intuitive and convenient user interface that specifically addresses the needs of clinicians.

Moreover, to enhance the transparency and trustworthiness, application of explainability techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) might become essential. The provision of responses that justify specific predictions is vital because it will be the most important step to ensuring that care professionals and patients have trust where automated tools are employed in making clinical decisions in large magnitude.

An idea of implementing this methodology in forecasting other cardiovascular diseases or other comorbidity, adding multimodal data, e.g., imaging or genetic information, are promising improvements to make machine learning efforts in precision medicine more profound. With such future improvements, predictive modeling can become more critical in preventive care, early diagnosis and individualized treatment regimens, and eventually, lead to improved health outcomes everywhere.

References

[1] Lundberg, S. M., & Lee, S. I. (2020). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 33, 4765–4774.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, 785–794.

[3] Akella, A., & Akella, S. (2021). Machine learning algorithms for heart disease prediction: A comparative study. *Biomedical Signal Processing and Control*, 68, 102595.

[4] Li, Y., Wang, X., & Chen, Z. (2022). Comparative analysis of machine learning models for cardiovascular disease prediction. *Scientific Reports*, 12(1), 14732.

[5] Komal, R., Naeem, H., & Nazir, A. (2023). A hybrid model for early heart disease detection using machine learning. *Computer Methods and Programs in Biomedicine*, 230, 107348.

[6] Akbulut, A., & Karakurt, T. (2022). XGBoost and LightGBM based heart disease prediction using clinical features. *IEEE Access*, 10, 44603–44612.

[7] Reddy, P., & Singh, R. (2023). Deep learning and ensemble approaches for cardiovascular disease prediction. *Healthcare Analytics*, 4, 100103.

[8] Zhao, Y., Zhang, J., & Lin, H. (2022). Predicting heart disease using machine learning and feature importance ranking. *BMC Medical Informatics and Decision Making*, 22(1), 234.

[9] Tashiro, K., & Yamashita, R. (2023). Hyperparameter optimization using Optuna: A practical guide. *ACM Computing Surveys*, 55(8), 1–25.

[10] Ali, M., Hussain, S., & Rehman, A. (2021). Hyperopt for hyperparameter tuning in healthcare applications. *International Journal of Advanced Computer Science and Applications*, 12(9), 389–395.

[11] Ma, Q., & Yu, C. (2020). Early diagnosis of heart disease using hybrid ensemble classifiers. *Journal of Healthcare Engineering*, 2020, Article ID 8877896.

[12] Ebrahimzadeh, A., & Faez, K. (2022). Machine learning-based framework for early prediction of heart disease. *Journal of Biomedical Informatics*, 128, 104027.

[13] Kumar, R., & Sharma, R. (2023). Efficient ensemble-based approach for heart disease detection. *Procedia Computer Science*, 213, 569–576.

[14] Hosseini, M., & Azizi, E. (2022). Explainable AI in healthcare: Feature importance in heart disease prediction. *Artificial Intelligence in Medicine*, 126, 102205.

[15] Wang, H., Liu, Y., & Xu, T. (2021). Machine learning algorithms for risk prediction of cardiovascular diseases: Review and future directions. *Frontiers in Cardiovascular Medicine*, 8, 689768.

[16] Bhardwaj, P., & Singh, M. (2024). Efficient hyperparameter tuning of XGBoost using Bayesian optimization. *Journal of Intelligent & Fuzzy Systems*, 46(1), 123–134.

[17] Patel, H., & Desai, A. (2023). Predictive analysis of heart disease using hybrid ML models. *In Proceedings of the International Conference on Smart Computing*, pp. 203–208.

[18] Zhang, R., & Chen, F. (2022). A comparative study of AutoML frameworks: Optuna vs Hyperopt. *Expert Systems with Applications*, 199, 117126.

[19] Sinha, A., & Jha, R. (2023). ML-powered diagnostic support for heart disorders using SHAP and ensemble learning. *Computers in Biology and Medicine*, 160, 106432.

[20] Roy, S., & Dey, N. (2021). Intelligent prediction model for cardiovascular disease using XGBoost and data balancing. *Journal of Ambient Intelligence and Humanized Computing*, 12, 4693–4705.