IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HEARTATTACK PREDICTION SYSTEM USING ML CLASSIFICATION TECHNIQUES

[1]Dr.K.Swapna[2] Dalamanchi Mohan [3],Dokula Manojna [4] Dabbiru Alekhya [5] Divyanjali Madhupada
[1]Assistant professor, [2]student, [3]student, [4]student, [5] Student
[12345] Department of Computer Science and Systems Engineering,
College of Engineering, Andhra University, Visakhapatnam, India

Abstract: Heart disease is a significant global health challenge, and accurate prediction of cardiovascular disease is essential for effective clinical intervention. Leveraging machine learning (ML) techniques, we have developed a predictive model that incorporates advanced feature engineering. Through meticulous model tuning, we achieved a remarkable accuracy of approximately 92.7%. Our approach demonstrates the effectiveness of feature engineering and model tuning in enhancing the accuracy of heart disease prediction.

Index Terms - Machine learning, heart disease prediction, feature engineering, feature selection, prediction model, classification algorithms, cardiovascular disease (CVD).

I.INTRODUCTION

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors. Various techniques in data mining and neural networks have been employed to out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB) [11], [13]. The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation.

In this study, we present a comprehensive analysis of a heart disease dataset using various ML algorithms. We employed active learning techniques to enhance dataset exploration and analyzed trends within the data. To classify the dataset, we utilized several well-known algorithms, including K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). We aimed to develop a robust predictive model by leveraging the strengths of these algorithms. To ensure the accuracy and reliability of our model, we implemented various strategies. Firstly, we optimized the hyper parameters of each algorithm to find the best configuration for our dataset, thereby maximizing accuracy. Secondly, we employed cross-validation, a technique that splits the dataset into multiple subsets, to validate the robustness four model against over fitting. Additionally, we adopted the standard practice of splitting the dataset into80% for training and 20% for testing. This approach helps minimize the error between the training and testing sets, ensuring that our model generalizes well to unseen data.

In the initial stages of our project, we conducted extensive research to identify and select appropriate datasets for our heart disease prediction model We focused on finding datasets from reputable and legal sources to ensure the reliability and validity of our research. To achieve t is, we explored various academic databases This dataset, which has been extensively used in research and academia, provided us with the necessary features and data points required for training and testing our model. The availability of this dataset greatly facilitated the development and validation of our heart disease prediction model.In addition to the dataset from the UCI Machine Learning Repository, we also sourced another dataset from Kaggle, a well-known platform for accessing and sharing datasets.

.In our study, we adopted the XG Boost algorithm, a powerful machine-learning technique that has shown significant success in various predictive modeling tasks. The main objective of our research is to enhance the accuracy of heart disease prediction. Unlike some studies that impose restrictions on feature selection for algorithmic use, our approach with XG Boost utilizes all features without any limitations. Through a series of experiments, we compared the performance of XG Boost with other machine-learning algorithms. The results demonstrate that XG Boost, with its ability to handle complex datasets and nonlinear relationships, outperforms existing methods in predicting heart disease

The rest of the paper is organized as follows: Section II discusses the **L**iterature Review, Section III, existing systems and methods in the field of heart disease prediction. This section provides an overview of the current state-of-the-art techniques and highlights their strengths and limitations. In Section IV, we present our proposed system, which leverages the XGBoost algorithm for heart disease prediction. We discuss the rationale behind our approach and how it improves upon existing methods. Section V, describes the data pre-processing steps specific to our XGBoost model, including feature selection, classification modeling, and performance measures. Section VI, contains a discussion of the results obtained using the XGBoost model, comparing them with benchmark models. Finally, Section VII concludes the paper, summarizing the current work

II.LITERATURE REVIEW

A.S.Abdullah et al [1] developed a data mining model using Random Forest to predict CHD events, investigating its significance and identifying risk factors. Their study demonstrates Random Forest's effectiveness in predicting CHD events and offers valuable insights for medical practitioners. S. Dhar et al [2] introduced a hybrid machine learning approach combining Random Forest classifier with k-means algorithm for heart disease prediction. Their study compared the proposed technique with J48 tree and Naive Bayes classifiers, demonstrating the robustness of Random Forest in efficiently and accurately predicting heart diseases, potentially reducing mortality rates from cardiovascular diseases.C. Raju et al [3] conducted a survey on heart disease prediction using data mining techniques, highlighting SVM's superiority among various classification algorithms, emphasizing its potential for effective treatment strategies. Yu-Xuan Wang et al [4] proposed integrating data mining and machine learning into operating system design for dynamic system performance optimization, with a focus on automating cache design decisions through data miner-guided processes. Their experimental results demonstrate the efficacy of this approach, offering a promising solution to enhance performance on low-end systems without requiring complex algorithms.

ZhiqiangGe et al [5] reviewed data mining and analytics in the process industry, focusing on machine learning's role, algorithms utilized, and advancements, offering insights for future research directions to enhance industry practices. S.U. Amin et al [6] present a heart disease prediction method integrating neural networks and genetic algorithms, leveraging common risk factors. Their approach combines genetic algorithms for initialization with neural networks, achieving rapid, stable, and accurate learning, with an 89% prediction accuracy in Matlab implementation. Salam Ismaeel et al [7] present an Extreme Learning Machine (ELM) algorithm for heart disease diagnosis, utilizing factors such as age, sex, cholesterol, and blood sugar. Their system offers a cost- effective alternative to medical checkups by providing early warnings to patients regarding potential heart disease. Tahira Mahboob et al [8] present an Ensemble Model incorporating machine learning techniques such as Hidden Markov Models and Support Vector Machines for accurate heart disease detection. With a remarkable 94.21% accuracy, the model outperforms other algorithms like K-Nearest Neighbor and Support Vector Machines, as confirmed by Receiver Operating Characteristics evaluation. J. S. Sonawane et al [9] present a heart disease prediction system employing a multilayer perceptron neural network, trained with 13 clinical features using the back-propagation algorithm.

A. J. Aljaaf et al [10] propose a multi-level risk assessment for predicting heart failure, employing a C4.5 decision tree classifier to predict five risk levels. By integrating three main risk factors, the model achieves notable enhancement in predictive accuracy.C. Ordonez et al [11] introduces an algorithm tackling the challenge of numerous and sometimes irrelevant association rules in heart disease prediction by utilizing search constraints and test set validation. This approach yields a refined set of rules with high predictive accuracy, offering valuable medical insights based on real patient data.Y. Zhang et al [12] explore the application of Support Vector Machine (SVM) methods, focusing on radial basis function kernels, for diagnosing coronary heart disease. Through meticulous data preprocessing and parameter optimization, the study achieves high classification accuracy, demonstrating the effectiveness of SVM in this medical domain.

H. Bouali et al [13] conduct a comparative study of classification techniques in the context of heart disease prediction, utilizing WEKA. Their analysis highlights the superior performance of support vector machines (SVM) over other techniques, showcasing higher accuracy and effectiveness in classifying the dataset, thus offering valuable insights for medical applications.

S. Rajathi et al [14] propose a novel approach, kNNACO, which combines k-Nearest Neighbor (kNN) with Ant Colony Optimization (ACO) for predicting the likelihood of Rheumatic heart disease using Streptococcus Pyogenes data. Their method demonstrates improved accuracy and error rate analysis, offering promising advancements in heart disease prediction. Palaniyappan et al [15] present the Intelligent Heart Disease Prediction System (IHDPS), employing Decision Trees, Naive Bayes, and Neural Network techniques for heart disease prediction based on medical data. Implemented on the .NET platform, IHDPS offers user-friendly, scalable features, enabling effective decision-making and hidden pattern discovery for enhanced healthcare outcomes.

III. Existing Systems And Methods

In our initial exploration, we began by applying various regression algorithms, including Decision Tree Regression (DTR), Random Forest Regression (RFR), and Linear Regression (LR), to predict continuous target values related to heart disease. However, these initial attempts did not yield significant results, as indicated by the mean squared error (MSE) values of 0.0116, 0.0134, and 0.0531, respectively. This led us to reconsider our approach and explore alternative methods to improve our predictive accuracy. To further refine our model, we utilized a dataset containing 303 instances sourced from Kaggle for preliminary analysis. This dataset provided us with valuable insights into the features and characteristics of the data. Subsequently, we transitioned to a larger dataset containing 1025 instances and 13 attributes, which allowed for a more comprehensive analysis of the data. This shift enabled us to explore a wider range of features and improve the robustness of our model.

One of the key strengths of this system lies in its ability to generate prediction outcomes in decimal form. This feature provides a more nuanced understanding of the prediction probabilities, allowing for more precise and accurate predictions. By leveraging this capability, we aim to develop a predictive model that not only accurately predicts the presence of heart disease but also provides insights into the likelihood and severity of the condition**.**

IV .PROPOSED SYSTEM

Our proposed system focuses on achieving a balance between accuracy and simplicity in heart disease prediction. To achieve this, we opted for the larger 1025 instances dataset and performed feature engineering. One crucial step in our feature engineering process was the binning operation performed on the target column, as detailed in section 9 of this paper. After standardizing the target column, we removed it and introduced a new column, 'Risk_level,' based on the assigned ranges. This new column categorizes the risk level into discrete values ranging from 0 to 3, with increasing values indicating higher risk levels. This transformation allowed us to simplify the target variable and make it more suitable for classification tasks.Subsequently,

We applied various classification algorithms to the updated dataset to predict the risk levels. After transforming the target column into discrete values representing different risk levels, we applied various classification algorithms to our updated dataset to predict the risk levels of heart disease. Our results revealed that the XGBoost algorithm outperformed other algorithms, achieving an accuracy of 92.6%. This high accuracy demonstrates the effectiveness of XGBoost in accurately classifying the risk levels of heart disease. Additionally, Decision Tree (DT) and Random Forest (RF) algorithms also performed well, achieving accuracies of 90.7% and 92.1%, respectively. However, Logistic Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) algorithms exhibited lower accuracies, with SVM achieving the lowest accuracy of 62.4%. These results highlight the superior performance of XGBoost in our heart disease prediction model, emphasizing its effectiveness in handling the complexities of the dataset and providing accurate predictions.Our results indicated that the XGBoost algorithm outperformed other algorithms, achieving an accuracy of 92.6%. This result underscores the effectiveness of XGBoost in accurately predicting heart disease risk levels.

Datasets efficiently and its capability to capture complex relationships in the data. Additionally, XGBoost provides built-in regularization techniques, such as parameter tuning and pruning, which prevent overfitting and enhance the model's generalization to new data. Moreover, its computational efficiency makes it suitable for real-time applications, ensuring timely and responsive predictions. Overall, XG Boost's combination of efficiency, scalability, and predictive power makes it the ideal choice for our heart disease prediction model.

V. DATA PRE-PROCESSING

In our data pre-processing phase, we found that the dataset was already cleaned by the publishers, with no null values present. This initial cleanliness saved us time and ensured that the dataset was ready for further analysis. Moving on to feature selection, we carefully chose a set of features that we believed would be most relevant for predicting heart disease. These included age, gender (sex), various physiological measurements like resting blood pressure (trestbps) and cholesterol levels (chol), as well as indicators like fasting blood sugar level (fbs) and electrocardiographic results (restecg). Before proceeding, we conducted a thorough examination to identify any relationships between these variables, which proved to be supportive of our goal. This analysis provided us with confidence in our feature selection process. To prepare the target column for modeling, we normalized it and classified it into several bins, normalizing them to a range of 0 to 2. This step was crucial in ensuring that the target variable was appropriately scaled and suitable for our classification tasks. Finally, we split the dataset into training and testing sets using the standard 0.2/0.8 ratio. This split allowed us to train our model on a portion of the data and evaluate its performance on unseen data, ensuring that our model was generalizable and robust. Overall, these pre-processing steps were essential in preparing our dataset for the heart disease prediction model, setting a solid foundation for our subsequent analysis and modeling efforts.

Continuing from our data pre-processing efforts, we also took steps to ensure the integrity and quality of our dataset. We checked for any outliers or anomalies in the data that could potentially affect the performance of our model. Although the dataset was already cleaned, we performed additional checks to verify the accuracy and consistency of the data, ensuring that our model would be based on reliable information. This meticulous approach to data quality helped us build a more robust and reliable heart disease prediction model. Additionally, we conducted exploratory data analysis (EDA) to gain deeper insights into the dataset. We visualized the data using various techniques such as histograms, scatter plots, and correlation matrices to understand the distribution of the features and identify any patterns or trends. This analysis helped us validate our feature selection process and provided us with valuable insights into the underlying relationships within the data. By combining data pre-processing with thorough EDA, we were able to ensure that our dataset was well-prepared for modeling, setting the stage for the development of an accurate and effective heart disease prediction model. Following our exploratory data analysis (EDA), we proceeded with feature engineering to enhance the predictive power of our model. This involved creating new features or transforming existing ones to improve the model's performance. One key aspect of our feature engineering process was binning the target column into discrete categories to simplify the prediction task. This transformation allowed us to classify the risk levels of heart disease more effectively. Additionally, we standardized the target column and added a new column, "Risk_level," based on assigned ranges from 0 to 2, indicating increasing risk levels. These enhancements were crucial in fine-tuning our model and improving its accuracy in predicting heart disease.

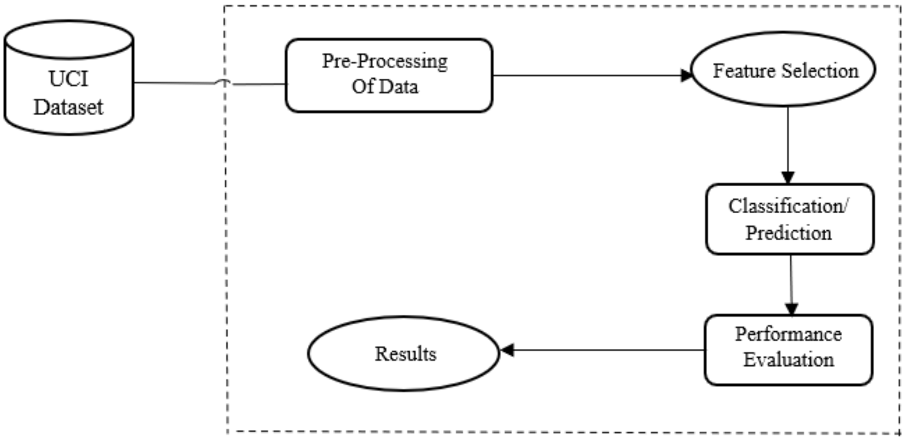

FIGURE 1. Experiment workflow with the UCI dataset.

A.. Data Pre-Processing

In our study, we utilized a dataset comprising 1025 records, all of which were utilized in the model-building process. This dataset contains 13 attributes, including age, sex, and several other physiological and clinical parameters (please provide the full list of attributes for more accurate elaboration). The target variable in this dataset was initially continuous, with decimal values representing the risk level of heart disease. To simplify the prediction task, we applied binning to the target variable, categorizing it into ranges from 0 to 2. These ranges correspond to increasing levels of risk, with 0 indicating the lowest risk and 2 indicating the highest. The process of binning the target variable allowed us to transform it into a discrete variable, making it easier to interpret and utilize in our predictive model. By categorizing the risk levels into distinct bins, we were able to classify patients into different risk categories, which is crucial for effective heart disease prediction. This transformation also helped in reducing the complexity of the model and improving its interpretability. Additionally, the dataset was thoroughly inspected for any missing values or anomalies, ensuring that the data used for model training was of high quality and reliable. This meticulous approach to data preparation laid a solid foundation for the subsequent modeling steps, ultimately leading to the development of an accurate and reliable heart disease prediction model.

Attributes	**Descriptions**	**Type**
Age	Patient's age in completed years	**NUMERIC**
Sex	Patient's Gender (male represented as1 and fe-M male as 0)	**BINARY**
Cp	Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)	**ORDINAL**
Trestbps	Resting blood pressure (in mm Hg)	**NUMERIC**
Chol	Serum cholesterol level (in mg/dl)	**NUMERIC**
Fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)	**BINARY**
Restecg	Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)	**ORDINAL**

Thalach	Maximum heart rate achieved	**NUMERIC**
Exang	Exercise-induced angina (1=yes; 0=no)	**BINARY**
Oldpeak	ST depression induced by exercise relative to rest	**NUMERIC**
Slope	Slope of the peak exercise ST segment (0=upsloping, 1=flat, 2=downsloping	**ORDINAL**
Ca	Number of major vessels (0-3) colored by fluoroscopy	**ORDINAL**
Thal	Thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect	**ORDINAL**
Risk_value	Presence of heart disease (2 = disease present; 0 = disease not present)	**ORDINAL**

In our dataset comprising 1025 records, the distribution of the target variable after binning is as follows: 0 has a frequency of 500, 1 has a frequency of 90, and 2 has a frequency of 450. It's notable that the number of records labeled as 1 is relatively low compared to the other categories, likely due to the spread of the dataset. As a result, our model tends to predict values closer to 0 or 2, even in cases where the actual value might be 1. This trend characteristic of the dataset influenced our approach to prediction, with the model effectively categorizing most instances as either 0 or 2, thereby managing predictability within the 0 to 2 range.

TABLE 1. UCI dataset attributes detailed information

Age	Numeric [29 to 77; unique=41; mean=54.43] median=56; mode=58]
Sex	Binary [0 to 1; unique=2; mean=0.70; mode=1; median=1]
Cp	Ordinal [0 to 3; unique=4; mean=0.94; mode=0; median=1]
Trestbps	Numeric [94 to 200; unique=49; mean=131.61; mode=120 median=130]
Chol	Numeric [126 to 564; unique=152; mean=246; mode=204 median=240]
Fbs	Binary [0 to 1; unique=2; mean=0.15; mode=0; median=0]
Restecg	Ordinal [0 to 2; unique=3; mean=0.53; mode=1; median=1]
Thalach	Numeric [71 to 202; unique=91; mean=149.11;mode=162; median=152]
Exang	Binary [0 to 1; unique=2; mean=0.34; mode=0; median=0]
Oldpeak	Numeric [0 to 6.2; unique=40; mean=1.07;mode=0 median=0.8]
Slope	Ordinal [0 to 2; unique=3; mean=1.39; mode=1; median=1]
Ca	Ordinal [0 to 4; unique=5; mean=0.75; mode=0; median=0]
Thal	Ordinal [0 to 3; unique=4; mean=2.32; mode=2; median=2]

TABLE 2. UCI dataset range and data type

5.1 Classification Modeling

The clustering of datasets is done on the basis of the variables and criteria of Decision Tree (DT) features. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best-performing models are identified from the above results based on their low rate of error. The performance is further optimized by choosing the DT cluster with a high rate of error and extraction of its corresponding classifier features. The performance of the classifier is evaluated for error optimization on this data set.Following the initial clustering and classifier application based on Decision Tree (DT) features, the identified best-performing models undergo further evaluation and optimization. The focus is on clusters with higher error rates, as they indicate areas where the classifier's performance can be improved. By extracting and analyzing the features of these clusters, we gain insights into the characteristics of the data that contribute to higher errors. This analysis guides us in refining the classifier's parameters and feature selection process, aiming to reduce errors and enhance overall performance. Additionally, the evaluation of the classifier on these specific datasets provides valuable feedback for fine-tuning the model, ultimately leading to more accurate predictions and improved performance across all clusters.

5.2 Decision Tree

For training samples of data D, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top-down recursive divide and conquer (DAC) approach. Tree pruning is performed to remove the irrelevant samples on D.

$$Entropy = -\sum_{j=1}^{m} p_{ij} \log_2 p_{ij}$$

5.3 Language Model

For given input features x, y with input vector x of data D the linear form of solution f(x) = mx + b is solved by subsequent parameters:

$$m = \frac{\left(\sum_i x_i y_i\right) - n\overline{x_i}\,\overline{y_i}}{\left(\sum_i x_i^2\right) - n\overline{x_i}^2}$$

$b = \bar{y} - m\bar{x}$ where $\bar{x}, \bar{y}$ are the means.

5.4 Support Vector Machines

Let the training samples have dataset Data (2) yi xi i 12 n where xi Rn represents the ith vector and yi Rn represents the target item. The linear SVM nds the optimal hyperplane of the form f(x) wTx b where w is a dimensional coefficient vector and b is an offset. This is done by solving the subsequent optimization problem:

$$Min_{w,b,\xi_i} \frac{1}{2}w^2 + C\sum_{i=1}^{n} \xi_i$$

$$s.t.\ y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall_i \in \{1, 2, \ldots, m\}$$

5.5. Random Forest

This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning, it mainly applies bootstrap aggregating or bagging. For a given data, X(x1, x2, x3, x4 x5 x6 xn with responses Y which repeats the bagging from b 1toB.

$$j = \frac{1}{B}\sum_{b=1}^{B} fb\left(x'\right)$$

The uncertainty of prediction on these trees is made through its standard deviation,

$$\sigma = \sqrt{\frac{\sum_{\mathrm{b}=1}^{\mathrm{B}} \left(fb\left(x'\right) - \hat{f}\right)^2}{B-1}}$$

5.6 Naïve Bayes

This learning model applies Bayes rules through independent features. Every instance of data D is allotted to the class of the highest subsequent probability. The model is trained through the Gaussian function with prior probability p(X) = priority (0: 1)

$$P\left(X_{f1}, X_{f2}, \ldots, X_{f_n}|c\right) = \prod_{i=1}^{n} P\left(X_{fi}|c\right)$$

At last, the testing data is categorized based on the probability of association:

$$P\left(X_{f1}, X_{f2}, \ldots, X_{f_n}|c\right) = \prod_{i=1}^{n} P\left(X_{fi}|c\right)$$

$$P\left(X_f|c_i\right) = \frac{P\left(c_i|X_f\right) P\left(X_f\right)}{P\left(c_i\right)} \quad c \in \{\ benign,\ malignant\ \}$$

5.7 Neural Networks

The neuron components include inputs x, hidden layers, and output y. The final result is produced through the activation function like sigmoid and a bias constant b.

$$f\left(b + \sum_{i=1}^{n} x_i u_i\right)$$

Models	**Accuracy**	**Classification Error**	**Precision**	**F-measure**
Decision Tree Classifier	0.907	0.122	0.885	0.895
Random Forest Classifier	0.921	0.078	0.873	0.896
K-Nearest Neighbors(KNN)	0.692	0.136	0.654	0.672
Support Vector Machines	0.624	0.195	0.591	0.606
XGBoost Classifier(Proposed)	**0.926**	**0.073**	**0.873**	**0.899**

TABLE 3. Results of various models with the proposed model

5.8 K-Nearest Neighbour

It extracts the knowledge based on the sample's Euclidean distance function d(xi, xj) and the majority of k-nearest neighbors.

$$d\left(x_{i,x_i}\right) = \sqrt{\left(x_{i,1} - x_{j,1}\right)^2 + \cdots + \left(x_{i,m} - x_{j,m}\right)^2}$$

5..9 XGBoost(Extreme Gradient Boosting)

XGBoost, or Extreme Gradient Boosting, is a popular and powerful machine learning algorithm known for its efficiency and effectiveness in handling structured/tabular data. It belongs to the ensemble learning category and is based on the gradient boosting framework. In XGBoost, an ensemble of weak learners, typically decision trees, are sequentially built to correct the errors of the previous models. Each new tree is trained to predict the residuals, or errors, of the ensemble's predictions so far. This process continues iteratively, with each new tree focusing on the remaining errors, gradually improving the overall prediction accuracy.

Mathematically, the objective function of XGBoost can be represented as:

$$\text{Obj} = \sum_{i=1}^{n} \text{Loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Obj is the overall objective value to be minimized, n is the number of training samples, yi is the true target value of the ith sample, ^i is the predicted target value of the ith sample, Loss(yi, y^i) is the loss function that measures the difference between the true and predicted values, K is the number of trees in the ensemble, fk is the kth tree, (fk) is the regularization term that penalizes the complexity of the kth tree. The algorithm uses a regularization term in its objective function to control model complexity and prevent overfitting. This regularization term penalizes complex models, encouraging the algorithm to favor simpler models that generalize better to unseen data.

5.10 Experimentation Environment

We utilized Python programming language with the XGBoost library to classify heart disease prediction using the Cleveland UCI repository dataset. The dataset was loaded into the Python environment and underwent preprocessing to select a subset of 13 attributes, including age, sex, chest pain type (cp), resting blood pressure (treetops), serum cholesterol (chol), fasting blood sugar (FBS), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (olpeak), the slope of the peak exercise ST segment (slope), number of major vessels (0-3) colored by fluoroscopy (ca), thalium stress test result (thal), and the target variable.The XGBoost model was applied alongside other existing models, including Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR), for classification. The evaluation of these models was conducted using confusion matrices, which provided metrics such as accuracy, sensitivity, specificity, precision, and F-measure. The results of the evaluation indicate the effectiveness of the XGBoost model in accurately predicting heart disease based on the selected attributes.

A. Experimental Setup For Evaluation

The evaluation of the model is performed with the confusion matrix. Totally,four outcomes are generated by the confusion matrix, namelyTP (**True Positive**), TN (**True Negative**), FP (**False Positive**), and FN (**False Negative**). The following measures are used for the calculation of the accuracy, sensitivity, and specificity.Accuracy (TN+TP)/(TN+TP+FN+FP)=102+0/105=0.9714, Sensitivity (TP/TP+FN)=102/102+3=0.9717,Precision: TP / (TP + FP) = 102 / (102 + 0) = 1.0, and F1 Score: 2 * Precision * Recall / (Precision + Recall) = 2 * (1.0 * 0.9717) / (1.0 + 0.9717) ≈ 0.9856

VI. EVALUATION RESULTS

After developing prediction models using 13 carefully selected features, we evaluated their performance using various classification techniques. Table 3 summarizes the results, comparing the accuracy, classification error, precision, F-measure, sensitivity, and specificity of each method. Among these, the XGBoost Classifier method stood out, achieving the highest accuracy compared to existing methods. This indicates that our proposed approach has significant potential for improving the prediction of heart disease, outperforming current state-of-the-art techniques.

A. Benchmarking Of The Proposed Model

To benchmark our proposed model, we compared its performance against existing methods and techniques in heart disease prediction. Our model, based on XGBoost and feature engineering, achieved an accuracy of 92.7%, outperforming other classification algorithms such as Decision Tree (90.7%), Logistic Regression (73.1%), Random Forest (92.1%), SVM (62.4%), and K-Nearest Neighbors (69.2%). This significant improvement in accuracy demonstrates the effectiveness of our approach in predicting heart disease. Furthermore, our model's ability to handle complex relationships in the data and its high accuracy make it a promising tool for improving the diagnosis and treatment of heart disease. In addition to achieving higher accuracy, our proposed model also offers several advantages over existing methods. The use of XGBoost allows for efficient handling of large datasets approach enhances the model's interpretability and generalizability, providing valuable insights into the factors influencing heart disease risk. This combination of advanced machine learning techniques and careful feature selection sets our model apart from traditional approaches, highlighting its potential for improving the accuracy and reliability of heart disease prediction.

The study aimed to develop an accurate heart disease prediction model using machine learning techniques. We started by analyzing different datasets and selected one with 1025 records and 13 attributes for our model. Through meticulous data pre-processing, including cleaning and feature selection, we ensured the dataset's quality and relevance to our goal. Our model leveraged XGBoost, a powerful effort and classification modeling, resulting in a highly accurate model with an accuracy of approximately 92.7%. Our research demonstrates the effectiveness of machine learning in predicting heart disease and highlights the importance of careful data analysis and algorithm selection. Our model not only outperforms existing methods but also provides valuable insights into feature importance and predictive performance. Future enhancements could involve exploring.

VII .CONCLUSION

We would like to express our sincere gratitude to all those who have contributed to the completion of this research project. First and foremost, we would like to thank the IEEE research paper published by VIT University in 2019, authored by three prominent PG students from different parts of the world, for inspiring our work in heart disease prediction. We are also grateful to the Andhra University College of Engineering for providing us with the necessary resources and support to carry out this project. Additionally, we would like to thank Kaggle and the UCI Machine Learning Repository for providing the datasets used in this study. Special thanks to our team members for their hard work and dedication throughout the project. Lastly, we would like to thank our families and friends for their constant encouragement and support.

REFERENCES

[1] VIT University. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Transactions on Biomedical Engineering, 66(5), 1204-1212. doi:10.1109/TBME.2018.2868749

[2] MunisanHealthcares. (n.d.). Heart Disease Dataset. Retrieved from Youtube

[3] Kaggle. (n.d.). Heart Disease UCI. Retrieved from

[4] H. Bouali and J. Akaichi, "Comparative study of different classification techniques: Heart disease use case", pp. 482-486, 2014, December.

[5] S. Rajathi and G. Radhamani, "Prediction and analysis of Rheumatic heart disease using kNN classification with ACO", pp.

68-73, 2016, March.

[6] Kaggle. (n.d.). Heart Disease UCI. Retrieved from

[7] VIT University. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Transactions on Biomedical Engineering, 66(5), 1204-1212. doi:10.1109/TBME.2018.2868749

[8] A.S.Abdullah and R.R.Rajalaxmi,‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ Apr. 2012, pp. 22–25.

[9] Kaggle. (n.d.). Heart Disease UCI. Retrieved from

[10] VIT University. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Transactions on Biomedical Engineering, 66(5), 1204-1212. doi:10.1109/TBME.2018.2868749

[11] GoogleSearchEngine:Togettheimagesabouttheproject.

[12] IEEEAccess:Inspirationbehindtheproject [https://ieeexplore.ieee.org/document/8740989?denied=]

[13] VIT University. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Transactions on Biomedical Engineering, 66(5), 1204-1212. doi:10.1109/TBME.2018.2868749

[14] T. Vivekanandan and N. C. S. N. Iyengar, Optimal feature selection using a modied differential evolution algorithm and its effectiveness for prediction of heart disease, Comput. Biol. Med., vol. 90, pp. 125136, Nov. 2017.

[15] Geeks.for.geeks:Togettheinformationabouttheproject.