CardioVisionary: Boosting Based Cardiac Disease Prediction using ML Techniques

Abstract- **Cardiovascular disease continues to be one of the world's leading cause of mortality, emphasizing the importance of accurate and precise diagnosis and intervention. This study utilizes UCI Heart Disease dataset, consisting of 920 samples and 16 attributes, to investigate potential of various machine learning algorithms in cardiovascular disease prediction. Using several machine learning algorithms including Random Forest, Logistic Regression, Support Vector Machine (SVM) and Gradient Boosting, the severity of cardiovascular disease is predicted on a scale of 0 to 4 after the data is pre-processed which involved filling the missing values and eliminating duplicates. Gradient Boosting was made the optimal choice as it achieved the highest accuracy of 98.37%. The model was dumped into a pickle file, which is incorporated into a user-friendly web-platform using HTML, CSS and JavaScript, allowing for real-time cardiovascular disease diagnosis based on the user inputs. The system not only provides information on predicted severity of cardiovascular disease but also provides the corresponding measures and precautions to be taken by the respective patient. This research demonstrates how machine learning can help medical healthcare professionals diagnose cardiovascular problems early on more accurately.**

Keywords- **Cardiovascular Disease Severity, Cardiovascular Insights, Data-Driven Healthcare, UCI Heart Disease dataset, Logistic Regression, SVM, Gradient Boosting, Random Forest, Web-platform, Feature Scaling, Smart Medical diagnostics, StandardScaler.**

I. INTRODUCTION

Cardiovascular diseases, particularly heart disease, is one of the main causes of death globally. According to the World Health Organization (WHO) [1], cardiovascular diseases accounts for the approximately 32% of deaths globally. The reduction of mortality rates and improvement of patient outcomes depends heavily on the early detection and efficient management of the cardiovascular problems. However, developing more effective and dependable instruments is imperative because the existing techniques of diagnosing cardiovascular diseases are rather traditional which are prone to be time-intensive, expensive, and subject to human error.

Recent advancements in Machine Learning have shown great promise in the field of medical diagnostics, which have made it possible for automated systems to evaluate massive datasets and provide precise and accurate predictions. By improvising the use of ML algorithms, healthcare systems can be improved to discovers the trends and patterns which many are not evident through conventional analysis. This may result in earlier diagnosis, allowing medical healthcare professionals to take pre-emptive measures beforehand.

This paper focuses on the development of a web-integrated cardiovascular disease prediction system, using machine learning techniques. The system is constructed on the foundation of UCI Heart Disease dataset [2] which comprises of 920 entries and 16 parameters including age, blood pressure, cholesterol levels, fasting blood sugar, etc. To forecast the severity of cardiovascular disease on a scale of 0 to 4, various algorithms including Logistic Regression, Support Vector Machine (SVM), Gradient Boosting and Random Forest are implemented, in which Gradient Boosting achieved the highest accuracy of 98.37%.
In addition to making the predictions, the model is integrated into a web platform that enables the users to enter their personal details and receive real-time predictions, along with the personalized advice and recommendations for managing their conditions. This system aims to assist the patients with improved self-management tools and help medical healthcare professionals to make decisions more quickly effectively and intelligently.

II. LITERATURE REVIEW

There has been quite significant academic research on the application of machine learning (ML) in healthcare, particularly in the diagnosis of cardiovascular disease. The UCI Heart Disease dataset was first presented as a benchmark for cardiovascular disease diagnosis [3]. The inclusion of wide variety of attributes such as age, sex, cholesterol levels, fasting blood sugar, etc. makes it ideal for accessing various prediction models. One of the early studies was conducted by Detrano et al. [3], who used logistic regression to demonstrate it predictive capabilities. However, it easily fails while dealing with complex and non-linear relationships in the data.
While predicting cardiovascular risk, traditional methods like SVM and Logistic Regression are commonly used due to its ease of interpretation and simplicity despite its ineffectiveness in capturing complex and non-linear patterns. Conversely, due to the superior predictive performance, ensemble learning techniques such as Random Forest and Gradient Boosting have gained significant attention due to their higher performance in prediction. Research conducted by Pandey et al. [4] and Sharma et al. [5] has demonstrated that the ensemble models achieve higher accuracy levels exceeding 90% to predict cardiac disease. Although, they can perform better, they require large number of computational resources and careful hyperparameter optimisation. Gradient Boosting excels in managing the imbalanced datasets, that are more common in medical diagnostics by sequentially minimizing prediction errors.
SVMs are widely used are they are resilient when handling high-dimensional data. Patel and Shah [6] demonstrated the effectiveness of SVMs for classifying cardiac diseases than other conventional algorithms. Despite this, SVM models are needed

to require meticulous tuning of the parameters for the best performance and are quite sensitive to kernel selection as well.

Preprocessing methods like scaling data and handling missing data are crucial for improving the performance of the machine learning model. Zubair et al. [7] emphasized the importance of preprocessing techniques, demonstrating properly scaled and cleaned data significantly improves the prediction accuracy. Similarly, preprocessing techniques such as duplication removal, handling missing values and feature scaling with StandardScaler are incorporated in this research.

While much of the existing research focuses on selecting optimal algorithms, only few studies address the practical application of these models in real-world, user-friendly platforms. Most studies focus on algorithm accuracy but do not incorporate the model into patients or healthcare professionals accessible platforms. This study seeks to close that gap, by integrating the highest-performing Gradient Boosting model into a web-based platform for real-time cardiac disease prediction and offering actionable patient advice and guidance, drawing insights from Reddy et al.'s work on AI-enabled healthcare delivery [8].

III. Methodology

This project uses machine learning techniques to predict the severity of cardiac disease using the UCI Heart Disease dataset. The technique as shown in Figure 1 is divided into multiple phases which includes data collection, data preprocessing, feature scaling, model development, model evaluation and deployment via a web interface.

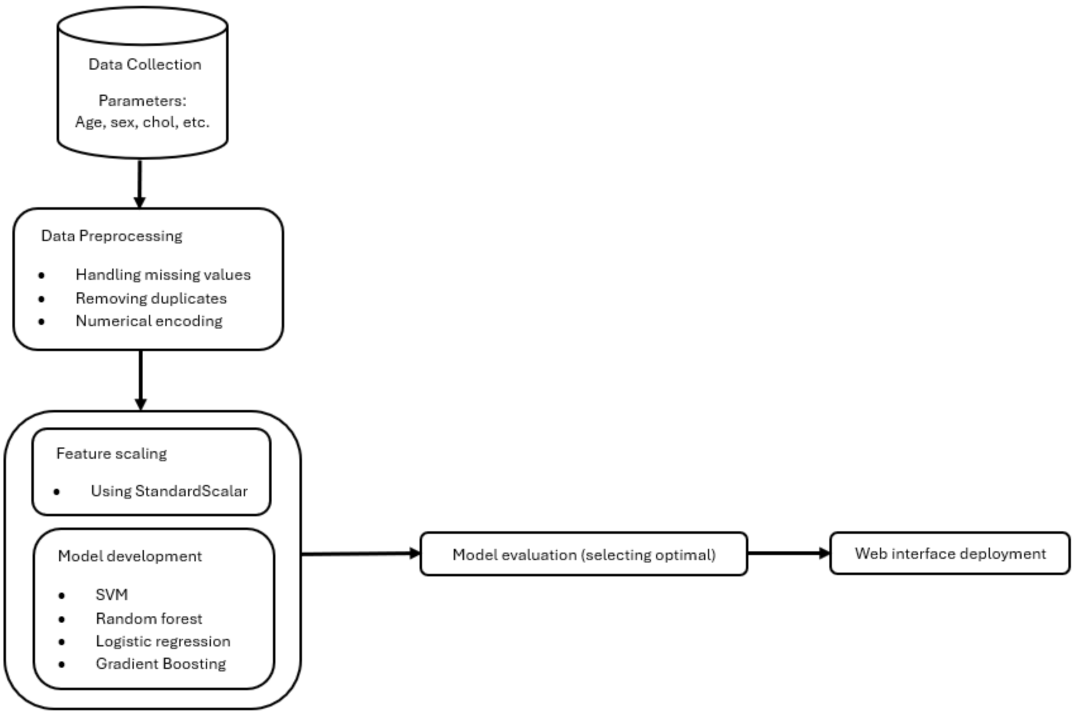

Figure 1 – System model of proposed cardiovascular disease prediction system

A. Dataset and Preprocessing

The UCI Heart Disease dataset consisting of 920 records and 16 variables including age, sex, kind of chest pain, resting blood pressure, fasting blood sugar, cholesterol level and a target variable (num) that represents the severity of cardiovascular disease on the scale of 0-4. The following bar graph illustrates the distribution of the target (num) values (0 – 4) representing the severity of cardiovascular disease.

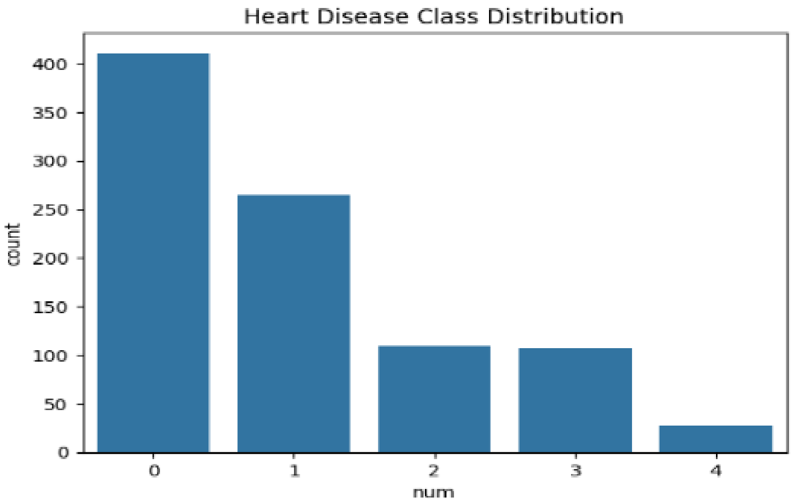

Figure 2 - Distribution of target (num) values in the dataset

The graph suggests that the dataset has maximum records with target variable of no cardiovascular disease (level 0) exceeding 400 records out of 920 records while the target variable of extreme cardiovascular disease (level 4) has bare minimum less than 50 records out of 920 records.

The following preprocessing actions were carried out:

1. Handling missing values: Missing data were filled using mean or mode imputation based on the type of the attribute.
2. Eliminating duplicates: To ensure data integrity, duplicate records were eliminated.
3. Numerical Encoding: Numerical format was used to translate the categorical attributes such as sex (Male = 1, Female = 0), fasting blood sugar (True = 1, False = 0).

B. Exploratory Data Analysis

To have the better understanding of the features and patterns of the dataset, exploratory data analysis was conducted which included:

1) Correlation analysis:

The Fig. 3 is the correlation heatmap of the dataset. Some of the major insights obtained from this graph are as follows:

1. It appears that characteristics like thalch, ca and oldpeak are strong indicators for cardiovascular disease.
2. Age and maximum heart rate achieved should also be considered as major attribute for cardiovascular disease are more common in older people with lower heart rates.
3. Cholesterol (chol) surprisingly exhibits a lower correlation, indicating that it might not be a strong factor to predict the cardiovascular disease in this specific dataset.

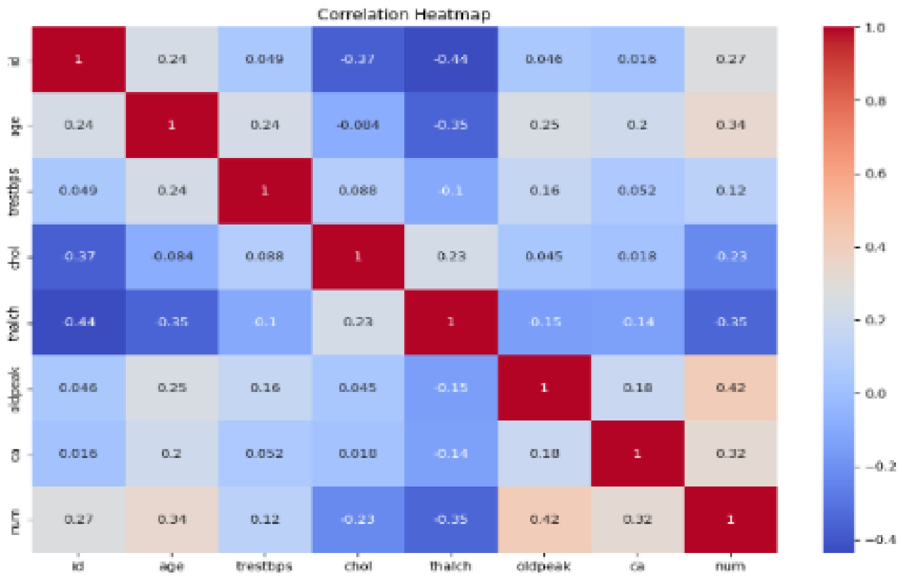

Figure 3 – Correlation heatmap of the characteristics of the dataset

2) Distribution analysis:

The distribution of the strong characteristics for indication of cardiovascular disease were visualized, to identify the generalized patterns in the dataset. The attributes below used to measure the distribution are both clinically significant and highly correlates with the cardiovascular disease risk, providing a strong and solid foundation for model performance and data analysis.

1. Age (age):

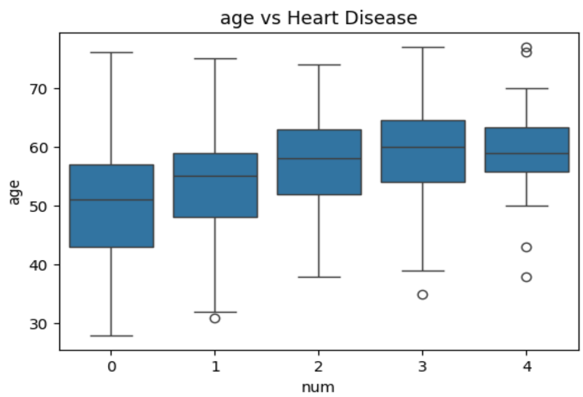

Figure 4 – Distribution of age against level of severity in the dataset

The age distribution showed a notable concentration of patients in the middle-aged range (40-60 years), suggesting that the risk of cardiac disease rises with the age. The analysis demonstrated the significance of age as a predicting factor as older people tend to have higher risk for severe cardiovascular disease.

2. Cholesterol levels (chol):

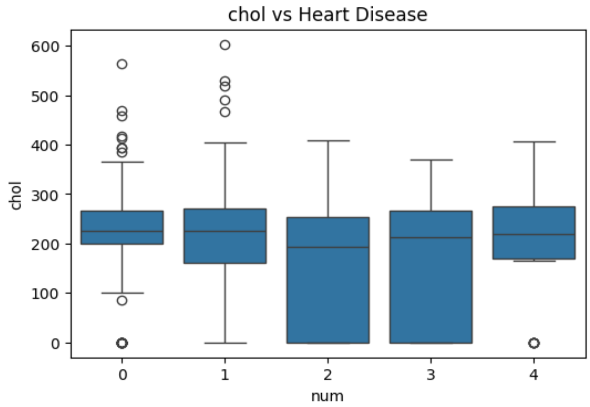

Figure 5– Distribution of cholesterol levels against level of severity in the dataset

A significant proportion of patients had high cholesterol levels (over 240 mg/dl), according to the analysis. Given, that individuals with elevated cholesterol levels were more diagnosed with serious cardiac disease risk, this study points th a clear correlation between cholesterol levels and the likelihood of cardiac disease.

3. Maximum Heart Rate achieved (thalch):

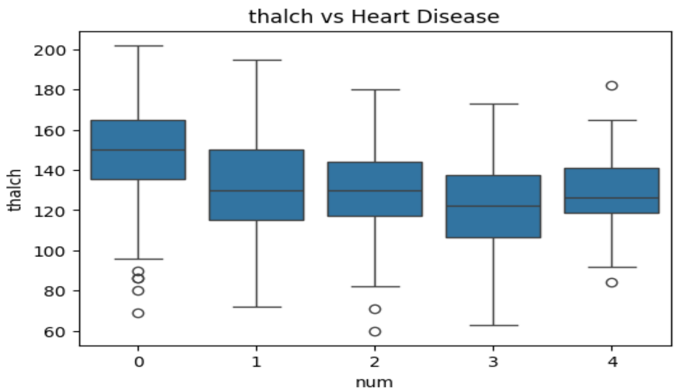

Figure 6– Distribution of maximum heart rate achieved against level of severity in the dataset

The results of the analysis demonstrated the noticeable decline in maximum heart rate achieved attained by the patients with cardiac disease. Those with lower maximum heart rates often had higher severity scores, suggesting that decreased heart rate response to exercise may be a strong indicative of potential cardiac issues.

4. Resting Blood Pressure (trestbps):

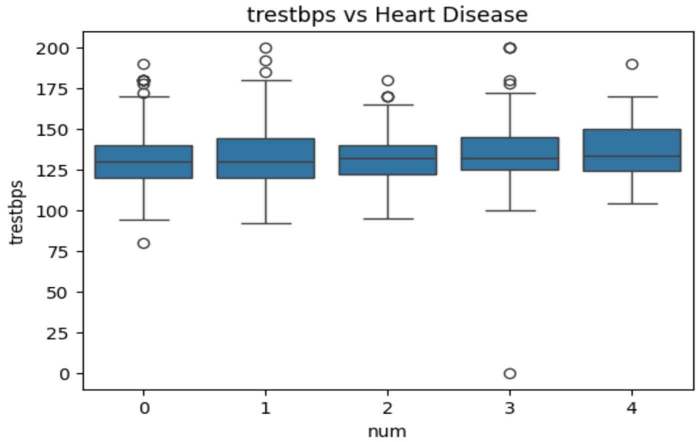

Figure 7 – Distribution of resting blood pressure against level of severity in the dataset

Many individuals were classified as hypertensive (over 140 mmHg of Systolic Blood Pressure) with increased risk of cardiac disease. As per the analysis, higher resting blood pressure was associated with more severe cardiac disease, highlighting its significance in being a predictive factor.

5. Chest Pain type (cp):

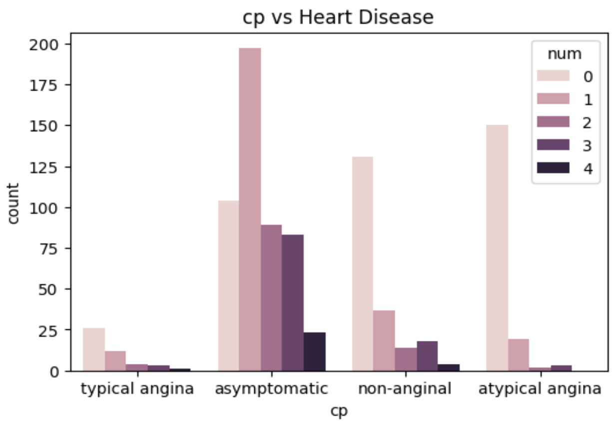

Figure 8 – Distribution of chest pain type against level of severity in the dataset

The analysis of the distribution of chest pain corresponding to severity of cardiac disease suggests that even though asymptomatic individuals doesn't exhibit greater symptoms, they possess higher risk of cardiac disease. On the contrary, atypical anginal pain is primarily associated with either fewer or no cardiac diseases cases in this population.

C. Data Splitting

The dataset was divided into the ratio of 80:20 for training and testing sets respectively. This ensures that the model is trained with a significant proportion of data while preserving 20% for assessment to avoid overfitting and test how well the model generalises.

D. Feature Scaling

A StandardScaler, which fits and modifies the training data and applies the same to the testing data was used to standardise the range of independent variables. This stage makes sure that all the features contribute equally to the model without any dominance by features with bigger scales.

E. Model Development

In this section, we investigate through the different machine learning algorithms applied to predict the severity of cardiovascular disease (num: 0-4), each tailored to improve and enhance the accuracy and reliability of the prediction.

1) Logistic Regression:

Logistic regression was used for the multiclass categorization to predict the severity of the cardiac disease. The model uses the sigmoid activation function to predict the probability of each class:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$Where,$

- $h_\theta(x)$ = Predicted probability of the cardiac disease severity.
- θ = Model parameters (weights).
- x = Input features such as age, cholesterol, etc.

This approach is straightforward but effective for the problems requiring probability estimates for each class. The model achieved an accuracy of 92.93% accuracy rate in classifying the severity of cardiac disease.

2) Support Vector Machine (SVM):

SVM was used with a Radial Basis Function (RBF) kernel to categorize the severity of the cardiac disease. The decision boundary is based to maximize the difference between the classes:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b$$

$Where,$

- α_i = Lagrange multipliers.
- y_i = Class labels (+1 or -1).
- $K(x_i, x)$ = Kernel function (RBF in this case).
- b = Bias term.

SVM is excellent at handling non-linear connections between features, particularly with the use of RBF kernel. The SVM model was able to achieve the accuracy of 95.65% in classifying the severity of cardiac disease.

3) Random Forest:

Random Forest is an ensemble learning algorithm that build many decision tress and averages their predictions to increase the accuracy and stability. Each tree is trained on a random subset of features:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} Tree_i(x)$$

$Where,$

- n = Number of trees.
- $Tree_i(x)$ = Prediction of the $i - th$ decision tree.

By averaging multiple trees, Random Forest mitigates overfitting, resulting in a robust and reliable model. In this study, it achieved the accuracy of 97.28% in classifying the severity of cardiac disease.

4) Gradient Boosting:

Gradient Boosting algorithm was the best-performing algorithm in this study. It builds the models sequentially, fixing mistakes in each model as it iterates. The objective is to reduce the squared error loss:

$$L\big(y, f(x)\big) = \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2$$

$Where,$

- $L\big(y, f(x)\big)$ = Loss function.
- y_i = True label.
- $f(x_i)$ = Predicted value from the model.

Gradient Boosting is effective and powerful for classification problems because it can iteratively minimize the errors. With the accuracy of 98.37% in classifying the severity of cardiac disease making the model optimal choice for this study.

IV. Results

A. Model Evaluation

Using important performance indicators including accuracy, precision, recall, and F1-score, the performance of the models used in this study are accessed. To determine the optimal performing algorithm for c classifying the severity of cardiac disease. The models are evaluated.

1) Accuracy Comparison:

The accuracy comparison graph below shows the accuracy of each model: Logistic Regression, Gradient Boosting, Random Forest and SVM. Gradient Boosting performs better, achieving the highest accuracy (98.37%), followed by Random Forest (97.28%) across the dataset.

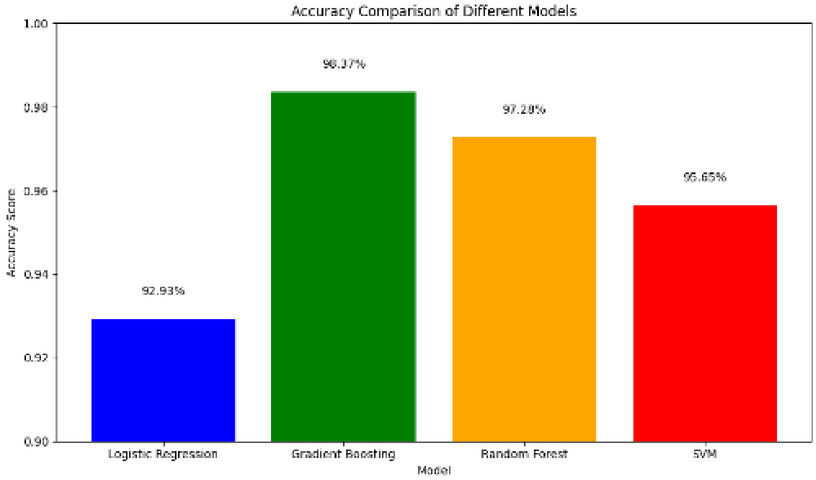

Figure 9 – Accuracy comparison of different models

2) Performance Metrics Comparison:

The graph below demonstrating accuracy, precision, recall and F1-score illustrates how well-balanced and robust Gradient Boosting performs across all metrics, making it the most reliable and consistent model. While other models used in the study exhibit competitive accuracy, their performance fluctuates in precision, recall and F1-score. Gradient Boosting, however, maintains its consistency in all areas, proving its overall reliability and effectiveness.

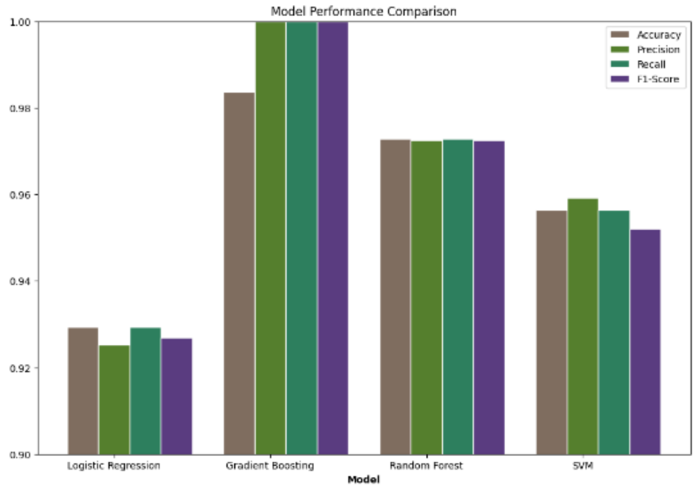

Figure 10 – Model Performance Comparison

3) Confusion Matrix:

The confusion matrix of the Gradient Boosting shows nearly flawless performance, precisely predicting nearly all severity levels (num: 0-4). The model’s exceptional accuracy and resilience in predicting the severity of cardiac disease, is thereby confirmed by the minimal miscalculations.

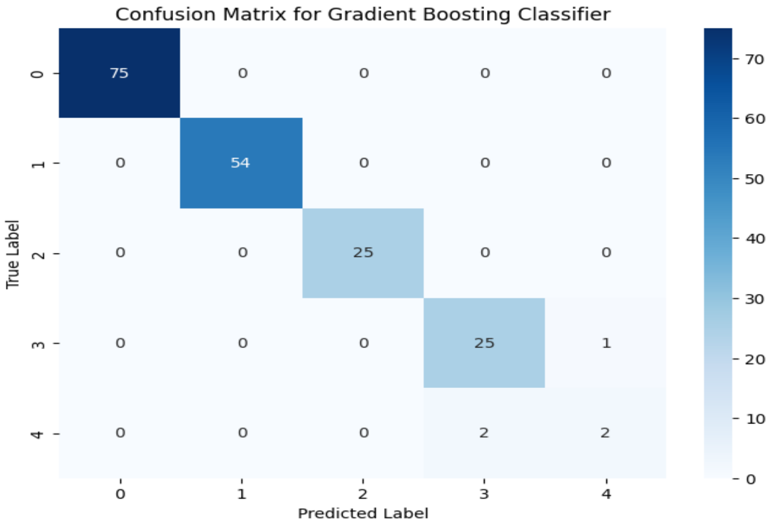

Figure 11 – Confusion Matrix for Gradient Boosting Classifier

B. Best Model

After evaluating and analysing all the models, Gradient Boosting emerges as the optimal choice because of its reliable and consistent results in terms of accuracy, precision, recall and F1-score. The model is highly effective for predicting the severity of cardiovascular disease because of its sequential learning approach, where it lowers mistakes at each iteration. Its suitability for this purpose is further reinforced by its strong confusion matrix having nearly flawless classification performance, with minimal misclassifications across all severity levels (num: 0-4). Thus, being the best model, Gradient Boosting model is stored into a pickle file and exported to use it a web-based platform for real-time predictions.

V. Web Application Development

The web application was built to include the cardiovascular disease prediction model into a user-friendly user interface. The program uses the input of the patient’s medical details to estimate the severity of the cardiac disease (num: 0-4). The key technologies used in this web-interface development are:

1. Frontend: The web-interface was designed using HTML, CSS and JavaScript. The form fields are used to collect user inputs such as age, sex, cholesterol levels, resting blood pressure, etc.
2. Backend: The machine learning model (Gradient Boosting) is handled by Python, which was created using Flask. The prediction is generated after the input data is pre-processed and passed through the model.

3. Model Integration: The Flask server was loaded with the trained Gradient Boosting model, which was saved using Pickle. Following data submission, the backend uses this model to predict the severity of heart disease and provides the user with the result.
4. User Feedback: Based on the predicted severity of cardiac disease, the web application shows the user personalized safety measures and health recommendations.

Screenshots of the web application functionality are incorporated in Figure 12 and Figure 13.

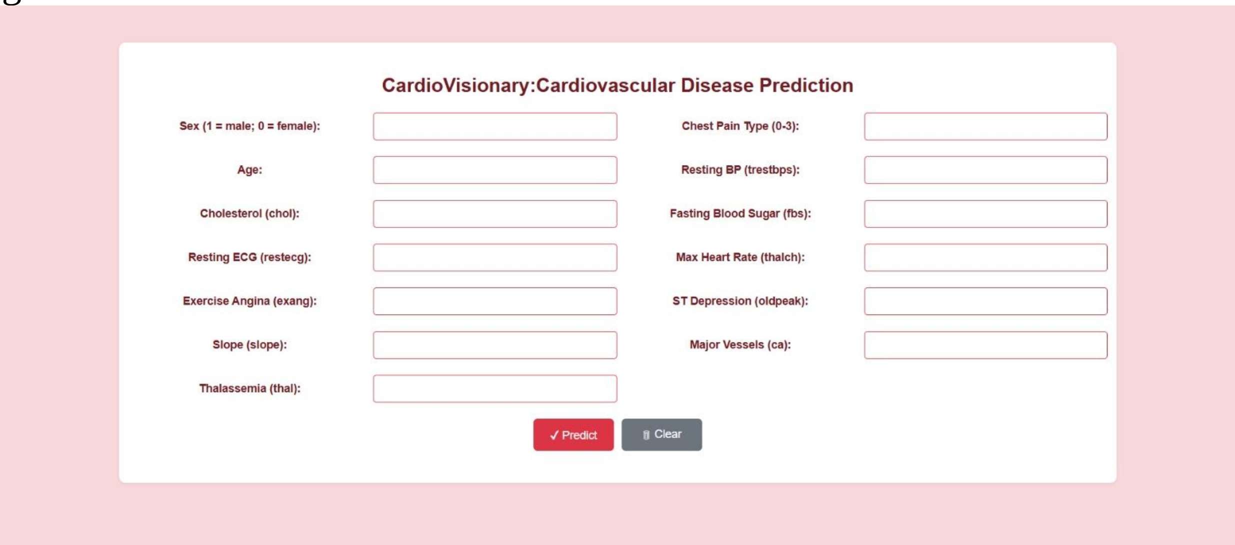

Figure 12 – User Input Section

The image illustrates the form where the users input their medical information, including age, sex, cholesterol, etc. It helps in demonstrating how user-friendly web-interface is crucial for data collection.

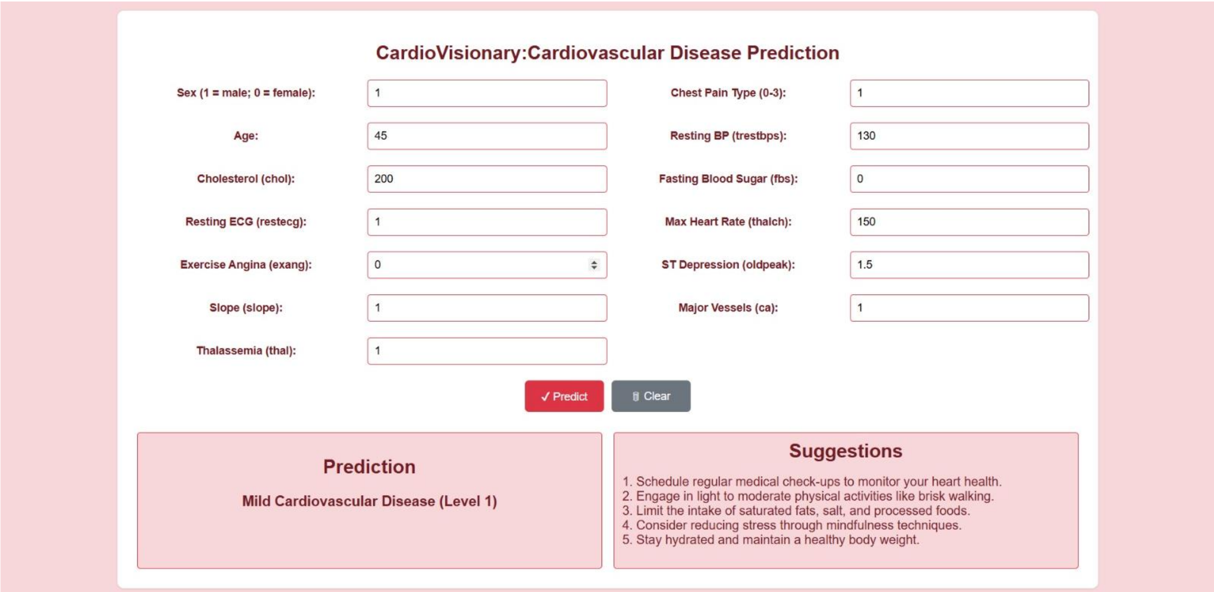

Figure 13 – Predicting Output and Recommendations

The Gradient Boosting model provides a prediction that indicates the severity degree of cardiac disease (from 0 to 4) based on the user's input. Along with the result, the application delivers practical recommendations and precautions tailored to the predicted degree of severity. This demonstrates how this program helps user to understand their health condition and take next steps for betterment.

VI. Conclusion

In this project, machine learning algorithms were successfully utilized to predict the severity of cardiovascular disease, with the Gradient Boosting model emerging as the most accurate, robust and reliable model. By demonstrating its superiority in terms of performance metrics such as accuracy, recall, precision and F1-score as well as showcasing its exceptional nearly flawless confusion matrix, Gradient Boosting proved its effectiveness in handling the complexity of cardiovascular disease prediction. The inclusion of the model with a web-interface ensures accessibility, allowing users to interact and evaluate their health risks based on the key medical parameters.

In addition to automating the diagnostic procedure, this system offers valuable insights that can support in early detection and prevention of cardiovascular diseases. The seamless implementation of the web-interface makes it a viable alternative for real-world applications for enhancing healthcare results by offering patients with an easy-to-use tool to evaluate and potentially improve their cardiovascular health. Future improvements may involve expanding the dataset and adding new risk variables, for more accurate and precise predictions.

References

[1]. World Health Organization, "Cardiovascular diseases (CVDs)," *World Health Organization*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2]. UCI Machine Learning Repository. (1988). *Heart Disease Data Set*. Irvine, CA: University of California, School of Information and Computer Science. Retrieved October 24, 2024, from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[3]. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V., "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, no. 5, pp. 304-310, 1989. DOI: 10.1016/0002-9149(89)90524-9.

[4]. Pandey, M., Tripathi, V., & Dwivedi, P. (2020). Heart disease prediction using machine learning and data mining techniques. *International Journal of Engineering Research & Technology (IJERT)*, *9*(5), 672-676.

[5]. Sharma, P., Verma, K., & Malhotra, N. (2021). Application of machine learning algorithms in heart disease prediction. *Journal of Emerging Technologies and Innovative Research, 8*(2), 125-132.

[6]. Patel, J., & Shah, S. (2019). Support vector machine models for heart disease prediction. *International Journal of Computer Applications, 178*(7), 10-13.

[7]. Zubair, M., Hassan, S., & Farooq, F. (2021). Improving heart disease prediction accuracy using data preprocessing and machine learning models. *International Journal of Computer Science and Network Security, 21*(4), 204-211.

[8]. Reddy, S., Fox, J., & Purohit, M. P., "Artificial intelligence-enabled healthcare delivery," *Journal of the Royal Society of Medicine*, vol. 112, no. 1, pp. 22-28, 2019. DOI: 10.1177/0141076818815510.

[9]. Ahir, M., Gandhi, V., & Chaudhary, A. (2019). Predictive analysis of heart disease using machine learning: A case study. International Journal of Computer Applications, 182(18), 25-29.

[10]. Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2020). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Networks and Applications, 25(6), 1706-1714. DOI: 10.1007/s11036-020-01623-w.

[11]. Tan, J. H., Hagiwara, Y., Pang, W., & Lim, I. (2018). Application of machine learning algorithms for the classification of cardiovascular diseases. Journal of the American College of Cardiology, 72(12), A123.

[12]. Le, T. T., Vodopyanov, A., & Islam, M. A. (2020). Predicting heart disease using machine learning algorithms in 2020: A case study of heart disease prediction system. Procedia Computer Science, 168, 456-463.

[13]. Gupta, A., & Mukherjee, S. (2020). A review of heart disease prediction techniques using machine learning. International Journal of Computer Applications, 176(6), 35-42.

[14]. Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology, 2(6), 56-66.

[15]. Huan, D., & Xing, H. (2020). A comprehensive review of feature selection methods for heart disease prediction. Journal of Computer Applications, 182(1), 19-28.

[16]. Shah, B., & Kothari, R. (2021). Heart disease detection using decision tree classifier. International Journal of Advanced Trends in Computer Science and Engineering, 10(2), 456-463.

[17]. Tsai, C. F., & Lee, C. W. (2019). An overview of machine learning in healthcare: Applications in heart disease. International Journal of Computer Applications, 173(5), 21-29.

[18]. Kumar, M., Singh, P., & Kumar, P. (2021). Evaluation of machine learning techniques for heart disease prediction. International Journal of Recent Technology and Engineering, 8(6), 345-352.

[19]. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

[20]. Abdul, A. S., & Sarmad, M. (2020). Analyzing the performance of different machine learning algorithms for heart disease prediction. International Journal of Computer Applications, 177(10), 11-18.

[21]. Cheng, J., & Chou, S. (2019). Machine learning in medical imaging for cardiovascular disease diagnosis. IEEE Access, 7, 153120-153130.

[22]. Malik, K., & Tiwari, S. (2020). Analysis of machine learning algorithms for heart disease prediction. International Journal of Engineering and Technology, 12(3), 98-104.

[23]. Liang, C., & Lee, C. (2020). A study of machine learning techniques for heart disease prediction. International Journal of Computer Science and Engineering, 11(7), 543-549.

[24]. Li, X., & Yuan, J. (2019). Review on machine learning methods for heart disease prediction. Journal of Biomedical Informatics, 100, 103-115.

[25]. Marling, C., & Whitehouse, P. (2001). Case-based reasoning in the care of Alzheimer's patients. Expert Systems with Applications, 21(3), 49-56.

[26]. Priyadarshini, A., & Shah, M. (2020). Feature selection techniques for heart disease prediction using machine learning. International Journal of Emerging Technology and Advanced Engineering, 10(4), 34-45.

[27]. Kaur, A., & Arora, S. (2020). Machine learning models for heart disease prediction: A review. International Journal of Engineering Research & Technology, 9(7), 201-205.

[28]. Mosavi, A., & Varkonyi-Koczy, A. R. (2020). Heart disease prediction using machine learning techniques: A review of recent advancements. International Journal of Engineering and Technology Innovation, 10(2), 152-160.

[29]. Tufail, M., & Chouhan, A. (2020). Heart disease prediction using ensemble machine learning methods. Procedia Computer Science, 175, 47-56.

[30]. Puri, A., & Narang, M. (2020). Comparing the performance of machine learning models for predicting heart disease. International Journal of Engineering Trends and Technology, 68(4), 72-78.