ADVANCED MACHINE AND DEEP LEARNING TECHNIQUES FOR CARDIOVASCULAR DISEASE PRDEICTION

Prof. Anuradha S. Deokar 1, Associate Prof. Dr.M A Pradhan 2,Tanmay S Bhagwat 3, Aditya A Bodake 4, Rudraksha R Kokane 5, Devraj M Malwe 6

1,2,3,4,5,6 Department of Computer Engineering All India Shree Shivaji Memorial Society's College of Engineering, Pune

ABSTRACT

The research focuses the prediction of cardiovascular disease using machine learning and data mining techniques. It highlights the importance of cardiovascular disease prediction due to its high prevalence and impact on global health. The research explains the process of data collection and preprocessing, including steps such as data cleaning, transformation, reduction, and feature scaling. It also focuses the encoding of categorical variables and the importance of feature engineering and selection in improving prediction accuracy. The research then introduces various machine learning techniques, including Random Forest, Support Vector Machine, Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, and Extreme Gradient Boosting, and explains their applications in cardiovascular disease prediction. Model evaluation measures such as accuracy, precision, recall, F1-score, and ROC curves are discussed. The research also touches upon hyperparameter tuning techniques, specifically Grid Search Cross Validation. Overall, the research provides a comprehensive overview of the process and techniques involved in cardiovascular disease prediction using machine learning and data mining.

INDEX TERMS

Machine learning, Deep Learning, Cross Validation, Accuracy, Precision, Recall

I.INTRODUCTION

Cardiovascular disease describes a range of conditions that affect the heart. Cardiovascular disease include: Blood vessel disease, such as coronary artery disease, Irregular heartbeats (arrhythmias), Heart problems you're born with (congenital heart defects), Disease of the heart muscle, heart valve diseases. According to the World Health Organization (WHO), an estimated 17.9 million people died from cardiovascular diseases (CVDs) in 2019, representing 32% of all global deaths. The World Health Organization (WHO) reports that in 2020, approximately 19.1 million deaths were attributed to CVDs globally.

In 2016, India reported 63% of total deaths due to NCDs, of which 27% were attributed to CVDs. CVDs also account for 45% of deaths in the 40–69 years age group. According to Statista, as per the results of a large-scale survey conducted across India, a majority of the people with heart problems in India in 2020 were aged between 45-54 years. Symptoms that may suggest a heart or blood vessel problem are shortness of breath, chest pain, chest pressure, heart palpitations, dizziness, sweating, numbness and weakness. Machine Learning algorithms play an important role for prediction of vast variety of things from historical data and its

analysis. With upcoming developments in data analytics and machine learning techniques, the historical records of patients can be used to predict if person may or may not be suffering from heart condition.

Data Collection: Data collection is a crucial phase in the process of gathering information necessary for analysis, decision-making, or research purposes. It involves systematically acquiring data from various sources, such as databases, surveys, sensors, APIs, or files, to build datasets relevant to the problem or research question at hand. The quality and quantity of collected data directly impact the validity and reliability of subsequent analyses or models. Depending on the nature of the project, data collection methods may vary, ranging from manual data entry to automated data retrieval processes. It's essential to ensure that the collected data aligns with the objectives of the study and adheres to ethical and legal standards, including privacy and consent considerations. Data collection may involve data preprocessing steps, such as cleaning, transforming, or encoding data, to ensure its suitability for analysis. Additionally, data collection often requires careful planning, including defining data collection procedures, selecting appropriate sampling techniques, and designing data collection instruments or protocols. Throughout the data collection process, it's essential to maintain data integrity, accuracy, and security to mitigate potential biases or errors that could affect the validity of the findings or conclusions drawn from the data analysis. But, feeding dataset directly to Machine Learning Algorithms does not provide expected results. For achieving expected result with higher accuracy, the dataset should be processed before feeding them to some algorithm. It is called preprocessing in Machine Learning. Preprocessing data is a fundamental stage in data mining to improve data efficiency. The steps involved in preprocessing datasets are:

Data cleaning: Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in data. This process involves several steps such as removing duplicates, filling in missing values, correcting errors, and dealing with outliers. By cleaning your data, you can ensure that your analysis is based on accurate and reliable information.

Data transformation: This involves transforming the raw dataset into an understandable format. It involves cleaning, filtering, and organizing data so that it can be analyzed effectively.

Data reduction: It is the process of reducing the size of a dataset by removing redundant features or instances. It involves removing data that is no longer useful or relevant to your analysis. There are several techniques for data reduction such as principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE). These techniques can help to reduce the dimensionality of a dataset while preserving the most important information and maintaining variance in data.

Feature scaling: Feature scaling is a technique used to standardize the range of independent variables or features of data. It is a method used to normalize the range of independent variables or features of data. In layman’s terms, feature scaling is like converting all the data into a common unit so that it can be compared on the same scale

II.RESEARCH METHODOLOGY

A.DATA OVERVIEW DATA PREPROCESSING

The heart.csv dataset is used for study of cardiovascular prediction. This dataset is publicly available and taken from famous repository Kaggle. The dataset contains total 12 features. It contains 918 patient records in 918 records it have 725 men samples and 193 female samples.

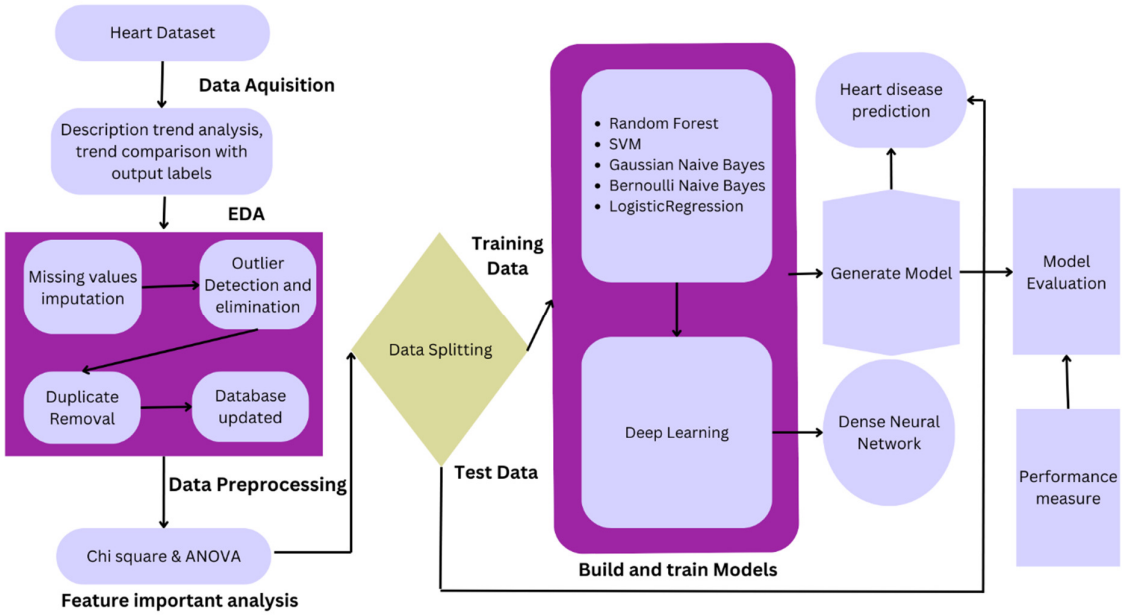
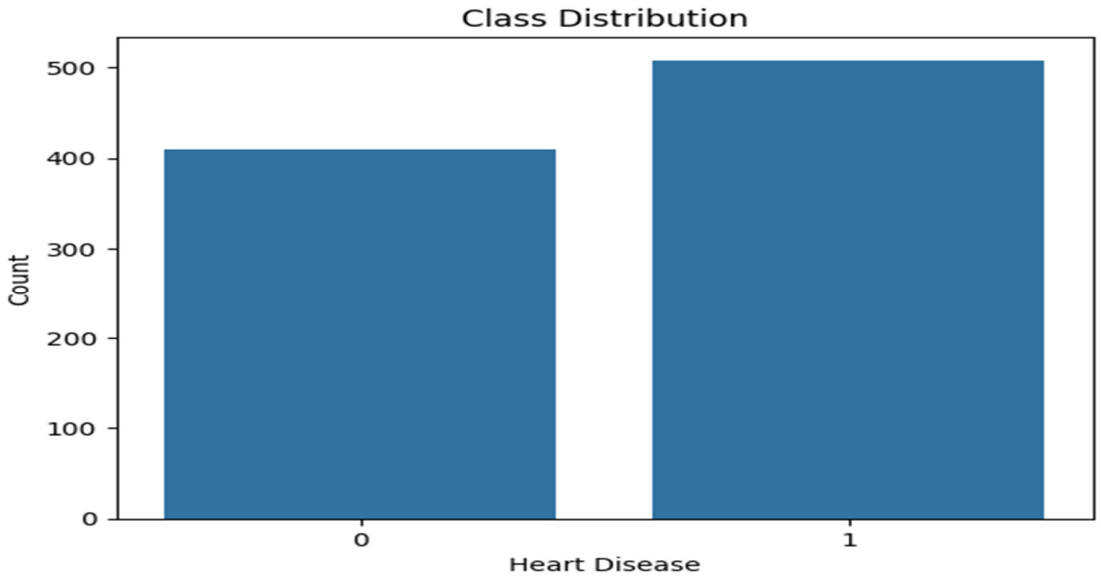


Fig.1 Architecture representation

The dataset's datatype check reveals that 6 attributes have int datatype, while one attribute has float datatype. Analysis of null values indicates no presence of any null values.



Graph.1 The database distribution analysis on target values

B. EXPLORATORY DATA ANALYSIS

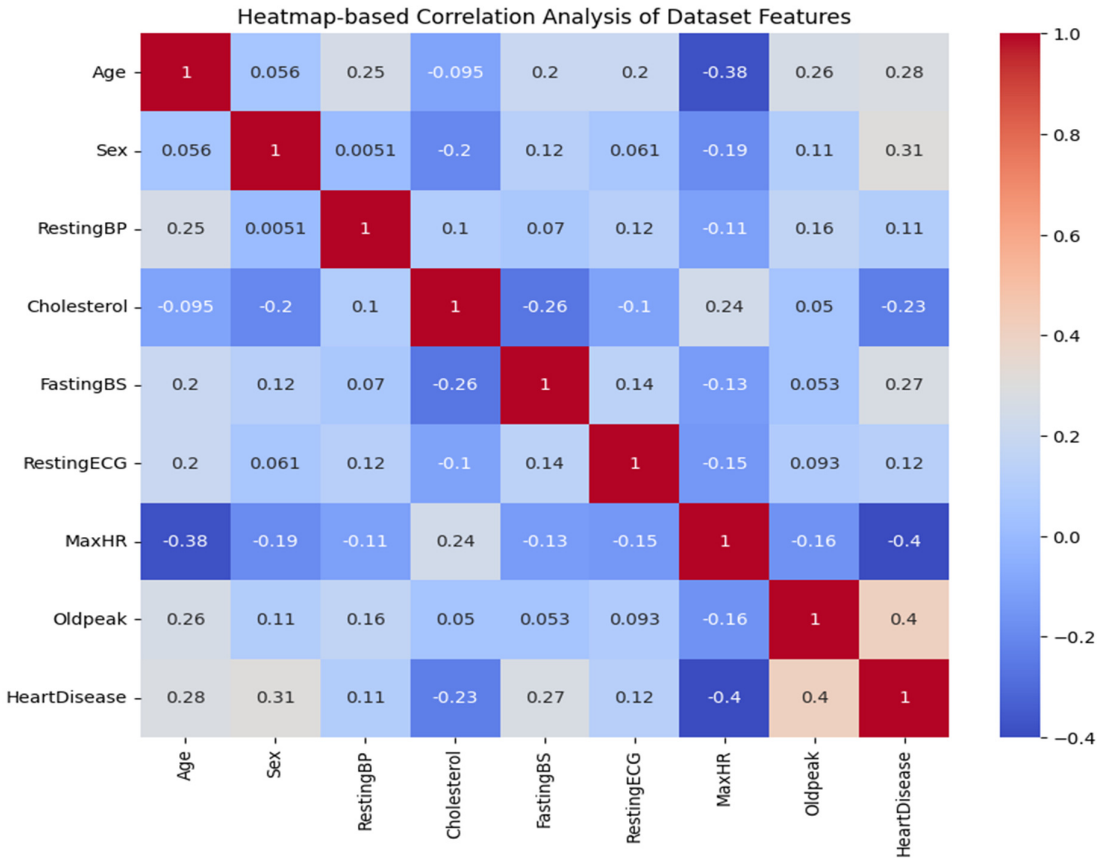
In exploratory data analysis (EDA) delves into the unique characteristics of our dataset, employing charts and heatmap graphs to unveil intricate patterns. This process provides invaluable insights crucial for informed decision-making. By scrutinizing correlations and visualizing dataset-related trends, we uncover nuanced associations, paving the way for deeper analytical exploration.

In the context of building machine learning models for predicting cardiovascular disease, it's crucial to understand the distribution of the target variable, which indicates whether a patient is healthy or has heart failure disease. A balanced dataset ensures that the model is trained on sufficient examples of both classes, leading to better generalization and performance.

In heart dataset, we have observed the following distribution:

- Number of healthy patients: 410
- Number of patients with heart failure disease: 508
- Number of males with heart failure disease: 458
- Number of females with heart failure disease: 50

In graph 2, examination of the correlation among the features in the heart failure dataset reveals a significant correlation across most features, suggesting a strong interrelation within the dataset.



Graph.2 The heatmap-based correlation analysis of dataset features

C.UNIVARIATE FEATURE SELECTION

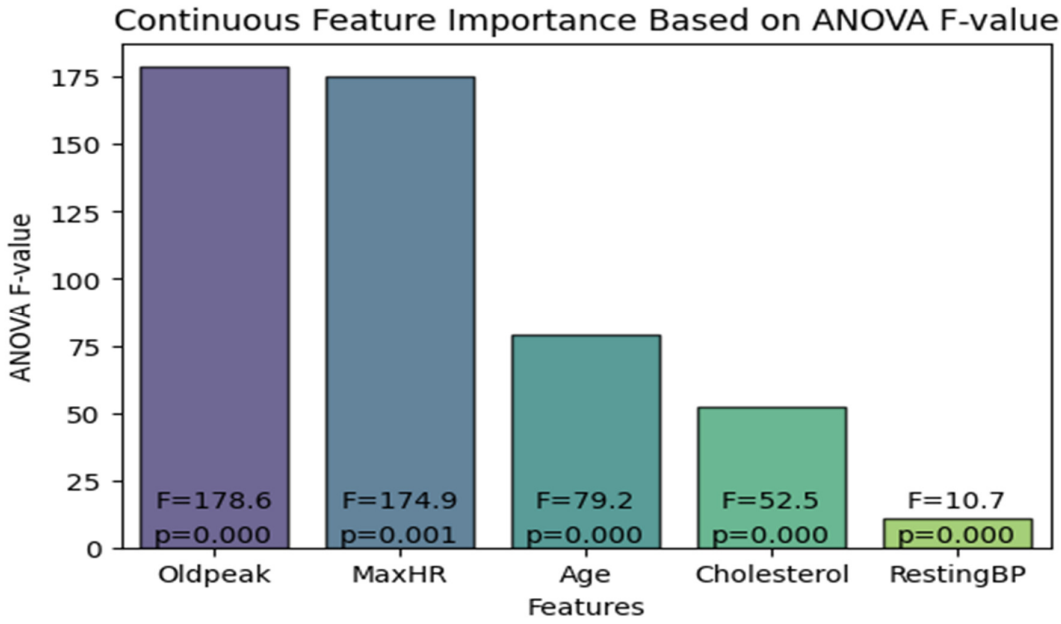
This method selects features based on univariate statistical tests like ANOVA or chi-square tests. It considers each feature independently of the others and selects those with the highest test scores.

1) ANOVA

The ANOVA (Analysis of Variance) F-value is a statistical method used to compare means across different groups or categories. In the context of feature selection, the ANOVA F-value measures the difference in variance between the distributions of feature values conditioned on different classes of the target variable. A higher F-value suggests that the means of the feature values vary significantly across different classes of the target variable, making the feature potentially informative for classification tasks. The resulting ANOVA F-values are sorted in descending order, allowing for the prioritization of features based on their importance in discriminating between different classes of the target variable. Additionally, a significance test is conducted by comparing the associated p-values against a chosen threshold (typically 0.05). Features with lower p-values are considered statistically significant and are retained for further analysis or model building. The visualization produced by the ANOVA F-values for each feature, helps to interpret the relative importance of different features in the classification task. Each bar is annotated with its corresponding F-value and p-value, providing additional insight into the significance of each feature. The Formula for ANOVA is:

$$F = \frac{MST}{MSE}$$

Where F is ANOVA coefficient, MST is Mean sum of square due to treatment and MSE is Mean sum of square due to error.



Graph.3 Continuous feature importance based on AVOVA F-value

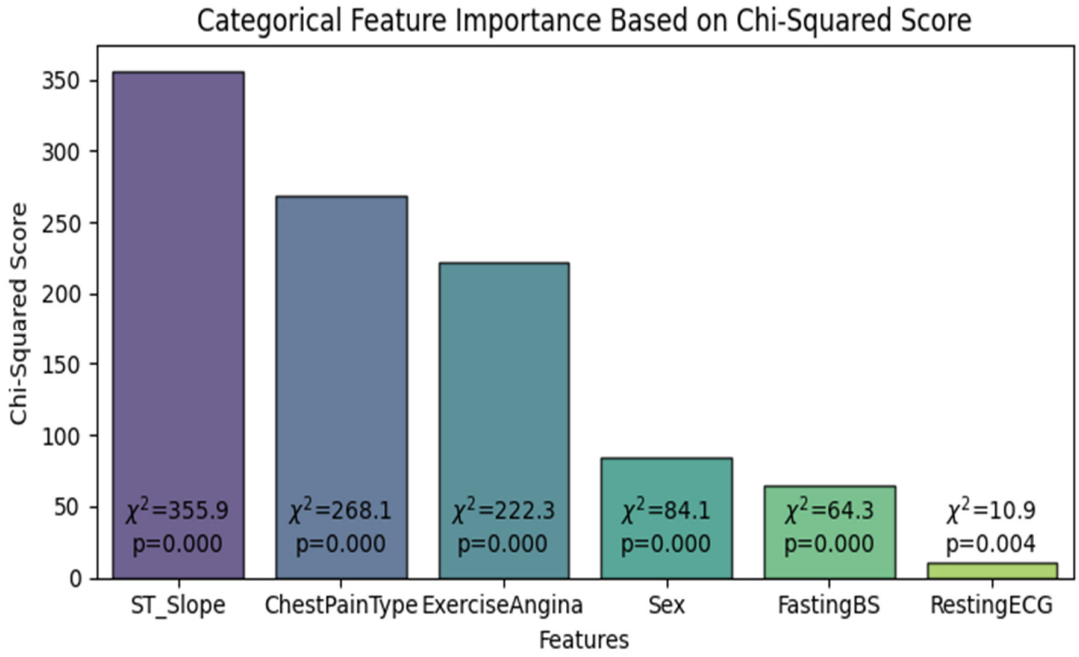
In above graph f value of oldpeak is 178.6 and p value is 0,also f value of MaxHR is 174.9 and p value is 0.001,along with that f value of Age Feature is 79.2 and p value is 0 and lastly f value of RestingBP is 10.7 and p value is 0.

2) Chi-Square

The Chi-squared test is a statistical method used to determine if there is a significant association between two categorical variables. In the context of feature selection, the Chi-squared test measures the independence between a categorical feature and a categorical target variable. This fundamental process underscores the essence of Chi-Square methodology. By prioritizing the most informative attributes, this approach holds the potential to significantly augment performance while streamlining the dataset, thereby mitigating unnecessary complexity. Through its discerning analysis, Chi-Square empowers practitioners to distill datasets to their most essential elements, facilitating more efficient and accurate machine learning models.

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where , x^2 represents chi squared , O_i represents observed value , E_i represents expected value



Graph 4. Categorical feature importance based on chi-square score

In above graph chi square value of ST_slope is 355.9,ChestPainType is 268.1, ExerciseAngina is 222.3, Sex is 84.1, FastingBS is 64.3,RestingECG is 10.9

D. DATA SPLITTING

Data splitting is a crucial step in machine learning to prevent models from overfitting and to assess their performance on unseen data. In our study, we divided the heart failure dataset into two parts: a training set and a testing set. The training set, which comprises 80% of the data, is used to train our machine learning classifiers. The remaining 20% of the data is reserved for

testing, allowing us to evaluate how well our trained models perform on new, unseen data. In employed well-established and traditional machine learning classifiers that have been trained thoroughly and have shown excellent accuracy in previous research. This ensures that our models are reliable and capable of making accurate predictions on real-world data.

E. FEATURE SCALING

Feature scaling is a crucial data preprocessing technique used in machine learning to standardize the range of independent variables or features in a dataset. Its primary objective is to ensure that features with different scales or units are on a comparable scale, which is essential for the proper functioning of many machine learning algorithms. There are several common methods of feature scaling, including standardization (Z-score normalization), normalization (Min-Max scaling), robust scaling, and log transformation. Standardization scales features to have a mean of 0 and a standard deviation of 1, while normalization scales features to a specified range, typically between 0 and 1. Robust scaling uses statistics that are robust to outliers, and log transformation is useful for normalizing skewed data distributions. Feature scaling should be applied after preprocessing steps like imputation of missing values and before training machine learning models. It helps improve the convergence and performance of machine learning algorithms, particularly those based on distance metrics or gradient descent optimization, by ensuring that features are on a comparable scale.

F. ENCODING CATEGORICAL VARIABLES

Categorical variables represent qualitative attributes or categories, such as gender, color, or product type, which are not inherently numerical. Common methods of encoding categorical variables include one-hot encoding, label encoding, and ordinal encoding. One-hot encoding creates binary columns for each category, where each column represents a unique category and contains a 1 if the observation belongs to that category and 0 otherwise. Label encoding assigns a unique integer to each category, replacing each category with its corresponding integer value. Ordinal encoding maps categories to integer values based on their ordinal relationship, preserving the order of categories. Encoding categorical variables is necessary because most machine learning algorithms require numerical input. By converting categorical variables into numerical format, it allows algorithms to process and analyze the data effectively. However, the choice of encoding method depends on the nature of the categorical variables and the requirements of the machine learning algorithm being used. Proper encoding of categorical variables ensures that the information contained within them is accurately represented in the dataset, thereby improving the performance and interpretability of machine learning models. It is typically performed as part of the data preprocessing pipeline before training the machine learning model.

III.MACHINE LEARNING TECHNIQUES

A) Random Forest (RF)

Random Forest (RF) stands out as a powerful technique in supervised machine learning, adept at handling both classification and regression tasks. Its strength lies in its utilization of multiple decision trees to address intricate problems efficiently. Essentially, an RF classifier aggregates the outputs of numerous decision trees, each trained on different subsets of the dataset. This ensemble approach significantly enhances predictive accuracy. Although adding more decision trees can boost accuracy, there's a point of diminishing returns where further additions don't

yield substantial improvements. One of Random Forest's notable advantages is its capability to mitigate overfitting—a common pitfall where a model becomes too tailored to the training data and struggles to generalize to new data. Instead, Random Forest achieves a delicate balance between fitting the data and sustaining high performance, rendering it an invaluable asset in the realm of machine learning.

$$\hat{y}_i = \frac{1}{M}\sum_{j=1}^{M} f_j(x_i)$$

B) Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a widely used supervised learning method, suitable for both classification and regression problems. Its primary goal is to establish optimal decision thresholds. SVM achieves this by creating a crucial decision boundary called a hyperplane, which effectively separates different classes within a multidimensional space. What makes this hyperplane particularly valuable is its ability to simplify the task of assigning new data points to their correct categories. SVM selects support vectors at the extreme edges of the data distribution to define this hyperplane. These support vectors are essentially the most critical data points for class separation, which is why the method is named a "support vector machine." This approach ensures that SVM can accurately categorize new data points by considering these pivotal support vectors that define the hyperplane.

$$\vec{w} \cdot \vec{x} + b = 0$$

where $\vec{w}$ is the weight vector, $\vec{x}$ is the input vector, and b is the bias term.

C) Logistic Regression

Logistic regression serves as a method for modeling the probability of a discrete outcome based on input variables. Typically, logistic regression models are employed for binary outcomes, representing scenarios with two possible values like true/false or yes/no. However, multinomial logistic regression extends this capability to scenarios with more than two potential discrete outcomes. This modeling approach is particularly valuable in classification problems, where the objective is to classify new samples into specific categories. Given that many aspects of cybersecurity involve classification tasks, such as detecting attacks, logistic regression emerges as a valuable analytical technique in this domain.

$$\mathrm{F(x)} = \frac{1}{1+e^{-x}}$$

D) Gaussian Navie Bayes (GNB)

Gaussian Naive Bayes represents a subtype of the Naive Bayes method tailored for datasets where attributes are continuous and follow a Gaussian (normal) distribution. Within the Sklearn library framework, Gaussian Naive Bayes is categorized as a classification algorithm, specifically designed for datasets with continuous features exhibiting a Gaussian distribution. To comprehend Gaussian Naive Bayes thoroughly, it's essential to grasp the foundational principles underlying its functionality. One of the key advantages of GNB is its computational

efficiency and scalability, making it suitable for handling large datasets with numerous continuous features. The algorithm requires estimating the mean and variance of each feature for each class, a process that can be efficiently executed even with extensive datasets. Additionally, GNB tends to perform well in high-dimensional spaces, where other classification algorithms may encounter difficulties due to the curse of dimensionality.

$$f(x|u,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(\frac{(x-\mu)^2}{2\sigma^2})}$$

Where , x is the input value

μ is the mean (average) of the distribution

σ is the standard deviation (spread or "width") of the distribution

e is the base of the natural logarithm (approximately equal to 2.71828)

E) Bernoulli Naive Bayes

Bernoulli Naive Bayes stands as a specialized subtype within the Naive Bayes Algorithm family, primarily employed for binary feature classification. In essence, it categorizes features into binary outcomes, such as 'Yes' or 'No', '1' or '0', 'True' or 'False', among others. An essential characteristic of Bernoulli Naive Bayes is its assumption of feature independence, meaning each feature contributes independently to the classification process. This classification method finds practical applications in various domains, with notable uses including spam detection, text classification, sentiment analysis, and determining the presence of specific words within documents. Bernoulli Naive Bayes can yield effective results when applied to binary feature datasets with a considerable degree of independence between features.

$$p(x_i|y) = p(i|y)x_i + (1 - p(i|y))(1 - x_i)$$

Where, p($x_i|y$) is the conditional probability of x_i occurring provided y has occurred

i is the event

x_i holds binary value either zero or 1

F) Extreme Gradient Boosting

XGB, short for Extreme Gradient Boosting, stands out as a supervised machine learning technique widely employed for classification and regression tasks. It falls under the umbrella of ensemble learning methodologies, which involve amalgamating various machine learning algorithms to enhance predictive performance. What sets XGB apart is its adeptness in combining multiple decision trees in a sophisticated manner. Unlike traditional approaches that yield deep trees, XGB adopts a sequential training process, gradually refining the ensemble with each iteration. Each subsequent tree is strategically crafted to rectify the errors introduced by its predecessors, thereby facilitating incremental improvements in the model's predictive capability. Within XGB, model refinement occurs iteratively through the application of the gradient descent algorithm, which systematically adjusts the model's parameters to minimize prediction errors. This iterative optimization process culminates in a model that is both highly accurate and computationally efficient. In the realm of machine learning, XGB emerges as a potent tool, particularly prized for its ability to deliver superior performance and nuanced model tuning.

$$\hat{y}_i = \sum k = 1^k f_k(x_i), f_k \in f$$

The symbol $\hat{y}_i$ denotes the predicted value for the i-th instance, while f_k signifies the k-th weak learner incorporated into the ensemble.

IV. DEEP LEARNING

In this research, a deep learning approach was employed to investigate its effectiveness in predicting heart failure. The research aimed to evaluate the performance of deep learning models in comparison to traditional machine learning classifiers. The primary deep learning model, **dl_model1**, was constructed using a sequential architecture consisting of multiple dense layers with rectified linear unit (ReLU) activation functions. The model was trained using the Adam optimizer and binary cross-entropy loss function to classify instances as either positive or negative for cardiovascular disease. To optimize the model's hyperparameters, a grid search cross-validation technique was employed. This involved exploring different combinations of hyperparameters such as the number of layers, the number of neurons per layer, activation functions, and optimizers. The objective was to identify the optimal configuration that maximized the model's accuracy on the training data.

Subsequently, another deep learning model, **dl_model2**, was built using the best-performing hyperparameters identified during the grid search. This model was trained and evaluated on the heart failure dataset to assess its predictive performance. Additionally, the study included experiments with traditional machine learning classifiers, such as logistic regression, random forest, support vector machine (SVM), Gaussian Naive Bayes (GNB), Bernoulli Naive Bayes (BNB), and XGBoost. These classifiers were trained and evaluated on the same dataset for comparative analysis with the deep learning models.

The performance of each model variant was evaluated using standard evaluation metrics such as accuracy, which measures the proportion of correctly classified instances. The models were trained on a training subset comprising 80% of the data and tested on a separate testing subset comprising the remaining 20%. The research findings aimed to provide insights into the effectiveness of deep learning techniques for heart failure prediction and their comparative performance with traditional machine learning approaches. The results obtained from the experimentation, including the best-performing model configurations and their corresponding performance metrics, were reported to contribute to the existing body of knowledge in the field of predictive modelling for medical diagnosis, particularly in the context of cardiovascular disease detection.

V. MODEL EVALUATION

1. Accuracy: Accuracy measures the proportion of correctly classified instances out of the total number of instances. It provides an overall assessment of the model's correctness but may not be suitable for imbalanced datasets.

$$Accuracy = \frac{Number\ of\ Correct\ predicions}{Total\ number\ of\ predictions}$$

2. Precision: Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It indicates the model's ability to correctly identify positive instances without misclassifying negative instances as positive.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

3. Recall (Sensitivity): Recall measures the proportion of true positive predictions out of all actual positive instances in the dataset. It indicates the model's ability to capture all positive instances without missing any.

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

4. F1-score: The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance, considering both precision and recall. F1-score is particularly useful when the dataset is imbalanced.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. ROC curves (Receiver Operating Characteristic curves): ROC curves plot the true positive rate (TPR), also known as recall or sensitivity, against the false positive rate (FPR), also known as the fall-out, at various threshold settings. ROC curves visualize the trade-off between sensitivity and specificity (1 - FPR) for different classification thresholds. AUC-ROC (Area Under the ROC Curve) summarizes the overall performance of the model, with higher AUC values indicating better discrimination between positive and negative classes.

$$True\ Positive\ Rate(TPR) = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$False\ Positive\ Rate(FPR) = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

Algorithm : Cardiovascular Disease Prediction Algorithm

Input: Heart Disease Dataset

For each type of dataset (unbalanced and balanced):

Conduct exploratory data analysis (EDA):

Examine data distribution, structure, and attributes.

Identify potential correlations and trends.

Compare feature distributions across different categories.

Preprocess the data:

Handle missing values using imputation techniques:

Replace missing values with appropriate estimates (mean, median, mode).

Employ advanced imputation methods.

Perform feature scaling or normalization if necessary:

Standardize or scale features to a similar range.

Normalize features to a common scale.

Address outliers:

Detect and handle outliers using statistical methods.

Perform feature importance analysis:

Determine feature importance using techniques like Random Forest or feature selection algorithms:

Assess the impact of each feature on the target variable.

Rank features based on their predictive power.

Split the dataset into training and testing sets:

Divide the dataset into subsets for model training and evaluation.

For each imputation technique:

Preprocess the data:

Handle missing values using the selected imputation method.

Perform feature scaling or normalization.

Address outliers if necessary.

Build and train models:

Train machine learning models and deep learning models.

Validate models on the testing set.

Evaluate model performance using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score).

Select top-performing models:

Identify models with the best performance across different imputation techniques.

Combine top-performing models using an ensemble approach:

Aggregate predictions from multiple models to improve overall performance.

Final model prediction and evaluation:

Generate final predictions using the ensemble model.

Evaluate the final predictions using appropriate evaluation metrics.

Output: Final prediction results.

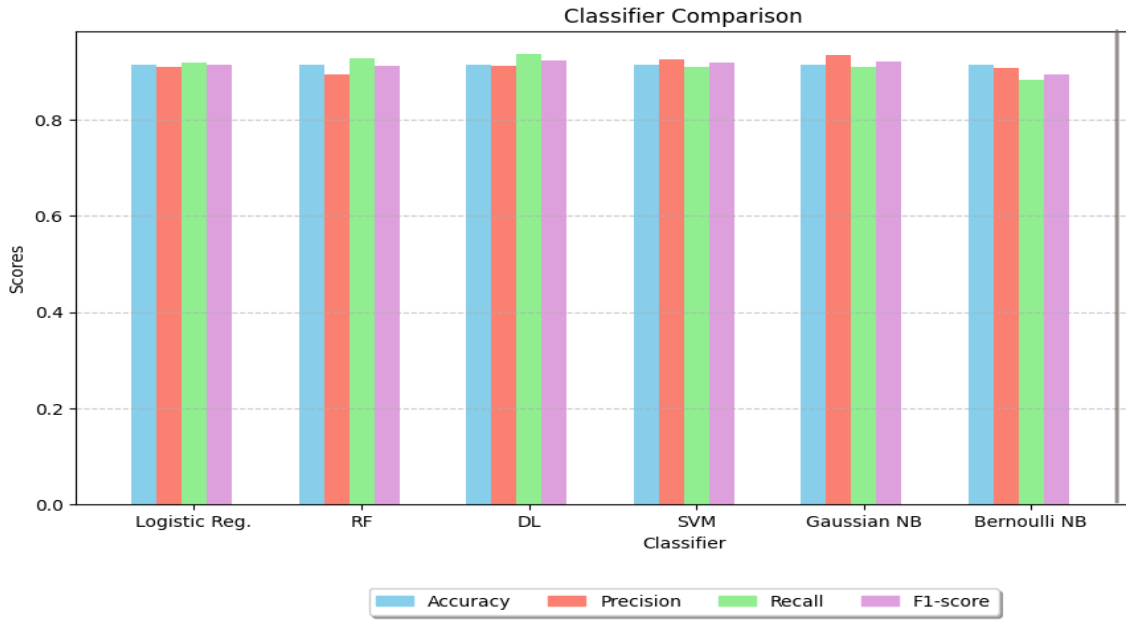
Table 1: Algorithm

VI.RESULT

Model name	accuracy	precision	recall	f1	specificity
DL model using median imputation	0.923913	0.921053	0.954545	0.937500	0.87837
DL model using mode imputation	0.923913	0.928571	0.945455	0.936937	0.891892

Table2:Represents accuracy,precision,recall,f1 score and specificity

Among the deep learning models, the one utilizing median imputation achieves the highest accuracy of 92.39%. Additionally, it demonstrates strong precision, recall, and F1 score metrics, with precision at 92.10%, recall at 95.45%, and F1 score at 93.75%. Moreover, it maintains a commendable specificity of 87.78%.



Graph .6 Classifier Comparisons

VIII.CONCLUSION AND FUTURE WORK

Deep learning emerges as a powerful tool for enhancing the precision and efficacy of cardiovascular disease diagnosis and prognostication. This model has demonstrated remarkable improvements in accuracy, sensitivity, and specificity compared to existing methodologies. Looking ahead, aim to fortify this approach by integrating image data from patients with cardiovascular disease. These images will be sourced from comprehensive laboratory examinations and imaging procedures. To leverage this visual data effectively, plan to employ Convolutional Neural Networks (CNNs), renowned for their ability to discern intricate patterns within images. By applying CNNs to these images, anticipate achieving unprecedented levels of diagnostic accuracy. One of the key advantages of CNNs lies in their capacity to automatically identify salient features crucial for accurate diagnosis.

IX.REFERENCES

[1] A. M. Qadri, A. Raza, K. Munir, and M. Almutairi, "Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning," vol. 11, p. 56214, Jan. 2023, DoI: 10.1109/access.2023.3281484.

[2] R. Kapila, T. Ragunathan, S. Saleti, T. J. Lakshmi, and M. W. Ahmad, "Heart Disease Prediction using Novel Quine McCluskey Binary Classifier (QMBC)," vol. 11, p. 64324, Jan. 2023, DoI: 10.1109/access.2023.3289584.

[3] Abdellatif, Abdallah, et al. "An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods." ieee access 10 (2022), DoI: 10.1109/ACCESS.2022.3191669.

[4] Aishwarya D., Pratiksha K. et al "Interpreting Cardiovascular Disease using Random Forest and Explainable AI" IJRASET Volume 11, Issue V, May 2023 DoI: 10.22214/ijraset.2023.52922.

[5] K. Vayadande et al., "Heart Disease Prediction using Machine Learning and Deep Learning Algorithms," May 2022, DoI: 10.1109/cises54857.2022.9844406.

[6] A. A. Almazroi, E. Aldhahri, S. Bashir, and S. Ashfaq, "A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning," vol. 11, p. 61646, Jan. 2023, DoI: 10.1109/access.2023.3285247.

[7] Aishwarya Dabir et al "Analysis of Cardiovascular Disease using Machine Learning Techniques", Volume 11 Issue V May 2023, DoI: 10.22214/ijraset.2023.52789.

[8] Prof. Guruprasad G1, Mr. Sarthik Poojary2, Ms. Simran Banu3, Ms. Azmiya Alam4, Mr.

Harshith K R, —Emotion recognition from audio using librosa and mlp classifier/IRJET 2021.

[9] S. Khan and S. T. Rasool, ''Current use of cardiac biomarkers in various heart conditions,'' *Endocrine, Metabolic Immune Disorders-Drug Targets*, vol. 21, no. 6, pp. 980–993, Jun. 2021

[10] P. Dileep, K. N. Rao, P. Bodapati, S.Gokuruboyina, R. Peddi, A. Grover, and A. Sheetal, ''An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm,'' Neural Comput. Appl., vol. 35, no. 10, pp. 7253–7266, Apr. 2023.

[11] A. Kumar, P. Kumar, A. Srivastava, V. D. A. Kumar, K. Vengatesan, and A. Singhal, ''Comparative analysis of data mining techniques to predict heart disease for diabetic patients,'' in *Proc. Int. Conf. Adv. Comput. Data Sci.* Singapore: Springer, 2020, pp. 507–518.

[12] M. S. Oh and M. H. Jeong, ''Sex differences in cardiovascular disease risk factors among Korean adults,'' *Korean J. Med.*, vol. 95, no. 4, pp. 266–275, Aug. 2020.

[13] R. Katarya and S. K. Meena, ''Machine learning techniques for heart disease prediction: A comparative study and analysis,'' *Health Technol.*, vol. 11, no. 1, pp. 87–97, Jan. 2021.

[14] T. Amarbayasgalan, V. Pham, N. Theera-Umpon, Y. Piao, and K. H. Ryu, ''An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets,'' *IEEE Access*, vol. 9, pp. 135210–135223, 2021.

[15] B. A. Tama, S. Im, and S. Lee, ''Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble,'' *BioMed Res. Int.*, vol. 2020, pp. 1–10, Apr. 2020.

[16] K. Shah, H. Patel, D. Sanghvi, and M. Shah, ''A comparative analysis of logistic regression, random forest and KNN models for the text classification,'' *Augmented Human Res.*, vol. 5, no. 1, pp. 1–16, Dec. 2020.

[17] D. K. Chohan and D. C. Dobhal, ''A comparison based study of supervised machine learning algorithms for prediction of heart disease,'' in *Proc. Int. Conf. Comput. Intell. Sustain. Eng. Solutions (CISES)*, May 2022, pp. 372–375.