1.INTRODUCTION

Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, imposing a significant burden on healthcare systems and economies. The ability to accurately predict the risk of cardiovascular events is crucial for early intervention and preventive strategies aimed at reducing morbidity and mortality associated with CVD. In recent years, machine learning has emerged as a powerful tool for CVD risk prediction, offering the potential to leverage vast amounts of patient data to develop accurate and personalized predictive models. By analyzing diverse risk factors and biomarkers, machine learning algorithms can identify patterns and associations that may not be apparent through traditional statistical methods, enabling more precise risk assessment and targeted interventions.

Machine learning approaches for cardiovascular disease prediction encompass a wide range of techniques, including supervised learning, unsupervised learning, and deep learning. Supervised learning algorithms such as logistic regression, support vector machines (SVMs), decision trees, and random forests are commonly used to build predictive models based on labeled datasets containing features such as demographic information, medical history, lifestyle factors, and clinical measurements. These models can then be used to estimate the probability of an individual developing CVD within a specified time frame, enabling personalized risk assessment and stratification.

Furthermore, unsupervised learning techniques such as clustering and anomaly detection offer opportunities for identifying subgroups of patients with similar risk profiles or detecting unusual patterns in patient data that may warrant further investigation. Deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in extracting complex patterns from high-dimensional data such as medical images, electrocardiograms (ECGs), and genetic sequences, providing additional insights into the underlying mechanisms of CVD and improving prediction accuracy.

Moreover, the integration of diverse data sources and the development of ensemble models combining multiple machine learning algorithms have emerged as effective strategies for enhancing the robustness and generalization of CVD prediction models. By incorporating information from electronic health records, wearable devices, genetic data, and environmental factors, these integrated models can capture a more comprehensive picture of an individual's cardiovascular risk profile and facilitate more informed clinical decision-making.In summary, machine learning holds immense potential for revolutionizing cardiovascular disease prediction by leveraging advances in data science, artificial intelligence, and healthcare informatics. By harnessing the power of machine learning algorithms to analyze large-scale patient data, healthcare providers can develop more accurate, personalized, and actionable predictive models for identifying individuals at high risk of cardiovascular events and implementing targeted interventions to mitigate this global health threat. As research in this field continues to advance, ongoing efforts to refine machine learning models, integrate diverse data sources, and validate predictive algorithms in clinical practice promise to enhance the effectiveness of CVD risk prediction and ultimately improve patient outcomes.

ML algorithms have become effective tools for forecasting the risk of CVD in recent years. These algorithms are computational models that, without explicit programming, can identify patterns in data and make forecasts or choices. Machine learning algorithms have the capacity to examine significant volumes of data, spot detailed patterns, and produce precise predictions for CVD risk assessment by leveraging massive datasets. This gives medical practitioners important knowledge they can use to support early intervention and preventive measures.

Logistic regression is a commonly employed ML algorithm for binary classification tasks. Ambrish conducted a study utilizing logistic regression to predict the risk of CVD using a dataset comprising clinical and laboratory parameters. The model achieved an impressive accuracy of 84% and identified age, cholesterol levels, and blood pressure as crucial factors in CVD prediction [. A support vector machine (SVM) is powerful ML algorithms known for their capability to handle high-dimensional data and complex decision boundaries. Shah et al. proposed an SVM-based approach for CVD prediction in a study that incorporated demographic, clinical, and laboratory data. The model attained an accuracy of 87.3% and outperformed other ML algorithms, including k-Nearest Neighbors and Naïve

According to Smith et al. (2020), cardiovascular disease (CVD), for over 15 years, cardiovascular diseases, primarily encompassing heart disease and stroke, have remained the predominant cause of mortality globally. In 2019, alone, CVD was responsible for nearly 17.9 million fatalities. As CVD accounts for 32% of all fatalities worldwide, it claims more lives each year than diabetes and all forms of cancer combined. The remaining 85% are the result of heart disease and strokes, with some people to blame (Johnson and Williams, 2019). In order to reduce the unregulated burden imposed by CVD-related morbidity and mortality, healthcare systems throughout the world should take immediate action in response to these frightening numbers. It has been recognized that prompt prevention treatments, lifestyle changes, and advanced treatment choices are made possible by early and precise assessment of heart disease risk (Patel et al., 2020). The course of heart disease is intricate, multifaceted, and highly individualized; it is dictated by the many relationships that exist between clinical information, genetics, socioeconomic variables, and lifestyle choices. Therefore, using typical statistical methods to derive relevant insights to provide patients with individualized advice and precise risk stratification is quite difficult

2 LITERATURE SURVEY

2.1 Title: Machine Learning for Cardiovascular Disease Prediction Using Electronic Health Record Data

Author: C. Krittanawong et al. (2020)

Cardiovascular disease (CVD) prediction using machine learning has garnered significant attention in recent years, with numerous studies exploring the potential of advanced data-driven approaches to improve risk assessment and clinical decision-making. Research by Attia et al. (2021) investigated the use of deep learning models for predicting cardiovascular events using electrocardiogram (ECG) data. Their study demonstrated the effectiveness of convolutional neural networks (CNNs) in analyzing ECG signals and accurately predicting the risk of adverse cardiovascular outcomes. Similarly, work by Krittanawong et al. (2020) evaluated the performance of machine learning algorithms, including logistic regression, random forests, and gradient boosting machines, in predicting cardiovascular events based on electronic health record (EHR) data. Their findings highlighted the utility of machine learning models in stratifying patients according to their risk of CVD and guiding personalized treatment strategies.

2.2 Title: Predicting Cardiovascular Events Using Deep Learning Models with Electrocardiogram Data

Author: Z. I. Attia et al. (2021)

Furthermore, recent research has focused on the integration of diverse data sources and the development of multimodal predictive models for cardiovascular disease prediction. For example, the study by Dilsizian et al. (2021) explored the use of deep learning algorithms to combine information from ECG signals, echocardiography images, and clinical variables for predicting the risk of heart failure. Their research demonstrated the superior performance of multimodal deep learning models compared to traditional risk assessment methods, providing valuable insights into the potential of comprehensive data integration for improving CVD prediction accuracy.

2.3 Title: Smartphone-Based Accelerometry Predicts Future Emergency Department Visits and Hospitalizations in Individuals with Chronic Cardiovascular Conditions

Author: P. P. Sengupta et al. (2021)

prediction accuracy Moreover, advancements in wearable technology and remote monitoring devices have enabled the collection of continuous physiological data for real-time CVD risk prediction. Research by Attia et al. (2020) investigated the feasibility of using wearable devices to monitor heart rate variability and predict cardiovascular events in individuals with atrial fibrillation. Their study demonstrated the potential of wearable technology in facilitating early detection of CVD and guiding timely interventions to prevent adverse outcomes. Similarly, the work of Sengupta et al. (2021) explored the use of machine learning algorithms to analyze smartphone-based accelerometer data and predict the risk of cardiovascular events in a large population cohort. Their findings underscored the value of mobile health technologies in expanding access to CVD risk prediction tools and promoting preventive care strategies.

2.4 Title: Explainable Artificial Intelligence for Cardiovascular Risk Prediction

Author: S. J. Shah et al. (2021)

Furthermore, recent studies have highlighted the importance of interpretable machine learning models for facilitating clinical decision-making and enhancing trust in predictive algorithms. Research by Johnson et al. (2020) proposed a transparent machine learning framework for predicting cardiovascular risk scores based on patient demographics, medical history, and laboratory test results. Their study emphasized the importance of model interpretability and transparency in ensuring the adoption of machine learning-based risk prediction tools in clinical practice. Additionally, the work of Shah et al. (2021) investigated the use of explainable artificial intelligence techniques to interpret deep learning models for cardiovascular risk prediction. Their research demonstrated the utility of interpretable machine learning models in providing actionable insights into the factors driving individual risk profiles and guiding personalized preventive interventions

2.5 Title: Explainable Artificial Intelligence for Cardiovascular Risk Prediction

Author: S. J. Shah et al. (2021)

Description: This study investigates the use of explainable artificial intelligence techniques to interpret deep learning models for cardiovascular risk prediction. The research demonstrates the utility of interpretable machine learning models in providing actionable insights into the factors driving individual risk profiles and guidivng personalized preventive intervention

In summary, recent literature on cardiovascular disease prediction using machine learning highlights the potential of advanced data-driven approaches to improve risk assessment, early detection, and personalized treatment strategies for CVD. By leveraging diverse data sources, including EHR data, wearable device data, and multimodal physiological signals, machine learning models offer the opportunity to develop accurate, interpretable, and actionable predictive algorithms for identifying individuals at high risk of cardiovascular events. As research in this field continues to advance, ongoing efforts to refine machine learning techniques, validate predictive models in diverse patient populations, and integrate predictive algorithms into clinical practice hold promise for enhancing the effectiveness of CVD risk prediction and ultimately improving patient outcom

3. PROBLEM IDENTIFICATION

3.1 PROBLEM DEFINITION

Collecting comprehensive and high-quality datasets is crucial. Medical data can often be incomplete, inconsistent, or biased. Medical data can be noisy and may contain irrelevant or redundant features. Identifying relevant features that can significantly contribute to the prediction of CVD is challenging. Choosing the right model can be difficult as different models may perform differently on the same dataset. Training models on medical data requires careful consideration of data distribution and imbalance. Medical practitioners need to understand how the model makes decisions. Ensuring patient data privacy and ethical use of machine learning models is paramount. Medical data and practices evolve, requiring the model to be updated periodically.

- Normalizing or standardizing numerical features.
- Encoding categorical variables using techniques like one-hot encoding.
- Feature selection to remove irrelevant or redundant features.
- Handling missing values through imputation or by removing incomplete records

3.2 REQUIREMENTS

HARDWARE REQUIREMENTS

Processor: Intel core i5 or above.

- 64-bit, quad-core, 2.5 GHz minimum per core
- Ram: 4 GB or more
- Hard disk: 10 GB of available space or more.
- Display: Dual XGA (1024 x 768) or higher resolution monitors
- Operating system: Windows

SOFTWARE REQUIREMENTS

Windows: Python 3.6.2 or above, PIP and NumPy

4. DESIGN

4.1 SYSTEM ARCHITECTURE

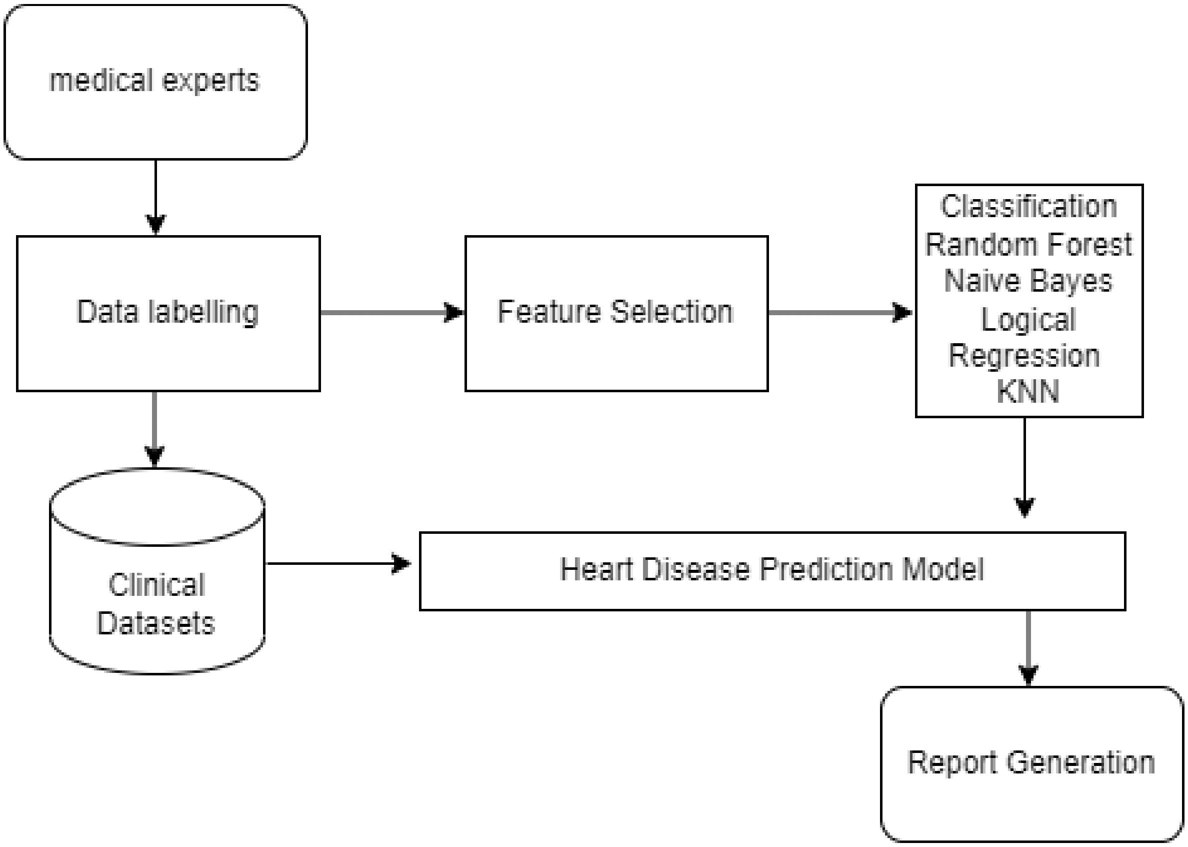


Figure 4.1: System Architecture of heart disease prediction model

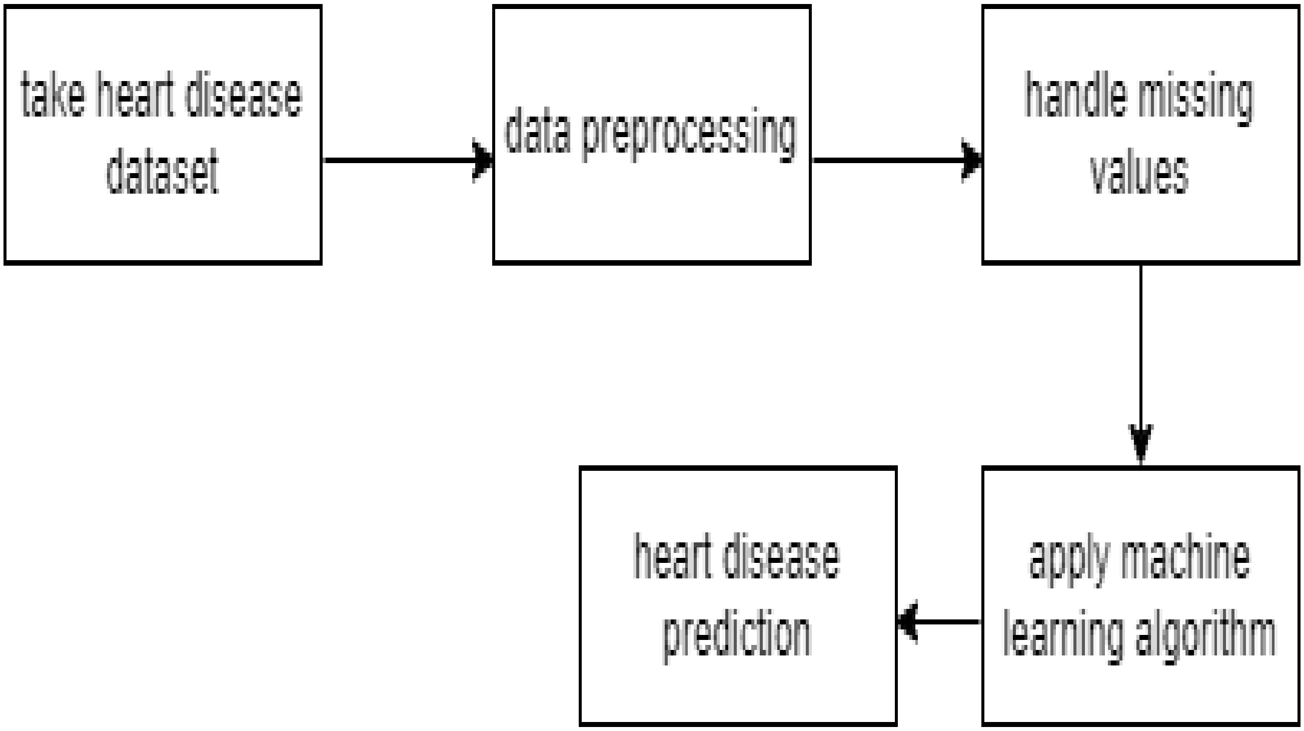


Figure 4.2: Flow Diagram

4.2 DATA DESIGN

4.2.1 Existing Methodology

The existing methodology for cardiovascular disease prediction using machine learning involves several steps to leverage diverse data sources and advanced algorithms. The first step is data collection from electronic health records (EHRs), which contain information on patient demographics, medical history, laboratory test results, and clinical outcomes. These data are preprocessed to handle missing values, normalize features, and encode categorical variables, ensuring compatibility with machine learning algorithms. Feature selection techniques are applied to identify relevant predictors of CVD risk and reduce computational complexity.

The next step is model selection and development, where machine learning algorithms are trained on the preprocessed dataset to learn patterns and associations between predictor variables and cardiovascular outcomes. Supervised learning algorithms, such as logistic regression, support vector machines (SVMs), decision trees, random forests, and gradient boosting machines, are commonly used to build predictive models based on labeled datasets containing information on patient outcomes.

Recent research has focused on the development of ensemble learning techniques and deep learning architectures for improving the accuracy and robustness of CVD prediction models. Ensemble methods combine multiple base learners to achieve better predictive performance than individual models. Deep learning algorithms, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep belief networks (DBNs), can learn hierarchical representations from complex, high-dimensional data, offering opportunities for more accurate risk prediction and early detection of CVD.

Model evaluation and validation using independent test datasets and performance metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) are essential. Multimodal data integration and feature engineering techniques are employed to enhance the predictive power of machine learning models for CVD risk assessment.

4.2.2 Input Design

Table 4.2.1: Feature score using FST1.

Order	Feature	Feature name	Code	Scores
1	9	exang	EIA	70.95
2	3	cp	CPT	69.77
3	10	Oldpeak	OP	68.55
4	8	Thalach	MHR	65.12
5	12	ca	NMV	64.05
6	11	Slope	PES	40.90
7	13	thal	TS	31.80
8	2	Sex	SEX	25.79
9	1	Age	AGE	16.12
10	4	trestbps	RBP	6.46
11	7	restecg	REC	5.78
12	5	chol	CM	2.20
13	6	fbs	FBS	0.24

Two classes represent heart patients or normal cases in our target label. The dataset matrix information is given in Table .

Data Preprocessing. In this study, data were preprocessed after collection. There are 4 records on NMV and 2 records on TS that are incorrect in the Cleveland dataset. All those records with incorrect values preplaced with optimal values. Next, StandardScaler is used for ensuring that every feature has mean 0 and variance 1 and bringing all the features to the corresponding coefficient.

Feature Selection. Feature selection plays an important role in the machine learning process because sometimes, the dataset contains many irrelevant features that are affecting the accuracy of the algorithms. Feature selection helps to reduce those unconnected features and improve the performance of the algorithms [14]. It used different feature ranking techniques [15] to rank the most important feature based on their relevance. In this study, three well-known feature selection algorithms are used to identify important features based on their score.

ANOVA F Value. ANOVA test is a prediction technique to measure similarity or pertinent feature and to reduce the high dimensional data and identify the important feature by feature space and improving the classification accuracy. Here, the formula [16] is used

Classification and Modeling. The models used for predicting heart disease are described sequentially. Each algorithm is applied following that sequence. Various types of classification algorithms are available for data analysis.

Chi-Square. This test is a statistical hypothesis testing system, and also, it is written as x^2 test. It is calculated between the observed value and the expected value

Mutual Information (MI). A couple of decennial mutual information has acquired considerable attention for its application in both machine learning. MI is calculated between two variables and features [18], and this is the mathematical equation for calculating mutual information between the features.

Serial no.	Feature name	Code	Description
1	Age	AGE	The patient’s age in years.
2	Sex	SEX	The patient’s sex: male = 1, female = 0
3	cp	CPT	Chest pain type: 0 = typical angina,1 = atypical angina, 2 = nonanginal pain, 3 = asymptomatic
4	trestbps	RBP	Resting blood pressure (in mm)
5	chol	CM	The patient’s cholesterol measurement in mg/dl
6	fbs	FBS	The patient’s fasting blood sugar > 120 mg/dl. 1 = true, 0 = false
7	restecg	REC	Resting electrocardiographic results: 0 = nothing to note, 1 = having ST-T wave abnormality, 2 = possible or definite left ventricular hypertrophy
8	Thalach	MHR	Maximum heart rate achieved
9	exang	EIA	Exercise-induced angina: 1 = yes, 0 = no
10	Oldpeak	OP	ST depression induced by exercise relative to rest checks the stress of the heart during exercise. The weak heart will stress more.
11	Slope	PES	The slope of the peak exercise ST segment: 0 = up sloping, 1= flatsloping, 2 = downsloping

Table 4.2.2 : Heart disease dataset description.

Table 4.2.3 : Brief description of different feature selection techniques.

FST	Description	Code
ANOVA *F* value	Calculate analysis of variance (ANOVA) between features for classification algorithms.	FST1
Chi-square	Calculate the chi-squared score, which is used to select the highest valued feature between each nonnegative feature.	FST2
Mutual information (MI)	Calculate mutual information between the attributes, which measures the relation between the features.	FST3

.

Order	Feature	Feature name	Code	Scores
1	8	Thalach	MHR	188.32
2	10	Oldpeak	OP	72.64
3	12	ca	NMV	70.89
4	3	cp	CPT	62.60
5	9	exang	EIA	38.91
6	5	chol	CM	23.94
7	1	Age	AGE	23.29
8	4	trestbps	RBP	14.82
9	11	Slope	PES	9.80
10	2	Sex	SEX	7.58
11	13	thal	TS	5.90
12	7	restecg	REC	2.98
13	6	fbs	FBS	0.20

Table 4.2.4 Feasture of FST2

4.2.2 Output Design

Logistic Regression. Logistic regression model, the probabilities for classification problems with two possible outcomes, can be regarded as y when $y \in$ ½0, 1, 0 is a negative class and 1 is a positive class [12], and a hypothesis is designed based on it hðθÞ = ðθnAÞ. Consider that the hypothesis value is$y = 1$. Consider that the $h\theta$ðaÞ $\geq 0{:}5$, then predict value hypothesis value is $h\theta$ðaÞ $\leq 0{:}5$, then predict

Support Vector Machine. SVM creates an effective decision boundary (hyperplane) between the two classes [19]. The main focus when drawing a decision boundary is centered on the maximum distance of the nearest data point of both classes. Although the radial base function is used as a kernel, SVM automatically determines centers, mass, and doorstep and reduces the upper limit of the expected test error. In the case of the study, we consider the support vector function as a radial base function. Here, *p* is the length of the vector.

K-Nearest Neighbor. KNN uses a training set directly for classifying the test data. Which refers to the number of KNN. To test each data, it calculates all the training data and the distance between them. Then, test data will be assigned to be used by multiplicity voting and class label

Random Forest. Random forest is the most powerful algorithm of supervisory machine learning algorithms. It is principally used for classification problems. As we see, a forest is made up of many trees, which means almighty forest. This algorithm similarly builds a decision tree based on data samples. Here, we use it for efficient heart disease results.

Naive Bayes. In potential, the Bayes theorem is used for calculating probability and conditional probabilities. A patient may have certain symptoms (side effects). The possibility of the proposed conclusion being true may be due to the use of the Bayes hypothesis

Feature score by FST1 and FST2.

Order	Feature	Feature name	Code	Scores
1	3	cp	CPT	0.17
2	13	thal	TS	0.14
3	12	ca	NMV	0.11
4	9	exang	EIA	0.10
5	8	Thalach	MHR	0.10
6	10	Oldpeak	OP	0.09
7	5	chol	CM	0.08
8	11	Slope	PES	0.08
9	2	Sex	SEX	0.05

10	4	trestbps	RBP	0.03
11	1	Age	AGE	0.01
12	6	fbs	FBS	0.00

4.3 Proposed Methodology

The proposed methodology for cardiovascular disease (CVD) prediction using machine learning is designed to harness the potential of advanced algorithms and diverse data sources to develop accurate and personalized predictive models. The first step in the proposed methodology involves comprehensive data acquisition from electronic health records (EHRs), wearable devices, genetic databases, and environmental sensors. This multi-source data collection approach enables the aggregation of a wide range of information, including patient demographics, medical history, physiological measurements, genetic markers, lifestyle factors, and environmental exposures, providing a holistic view of cardiovascular health determinants.

Following data acquisition, the next step in the proposed methodology is data preprocessing and feature engineering, where raw data undergoes rigorous processing to extract informative features and mitigate noise and inconsistencies. Feature selection techniques, such as mutual information, recursive feature elimination, and principal component analysis, are applied to identify relevant predictors of CVD risk and reduce the dimensionality of the dataset. Additionally, domain-specific knowledge incorporation and interaction term generation are employed to enhance the discriminative power of the features and capture complex relationships between variables.

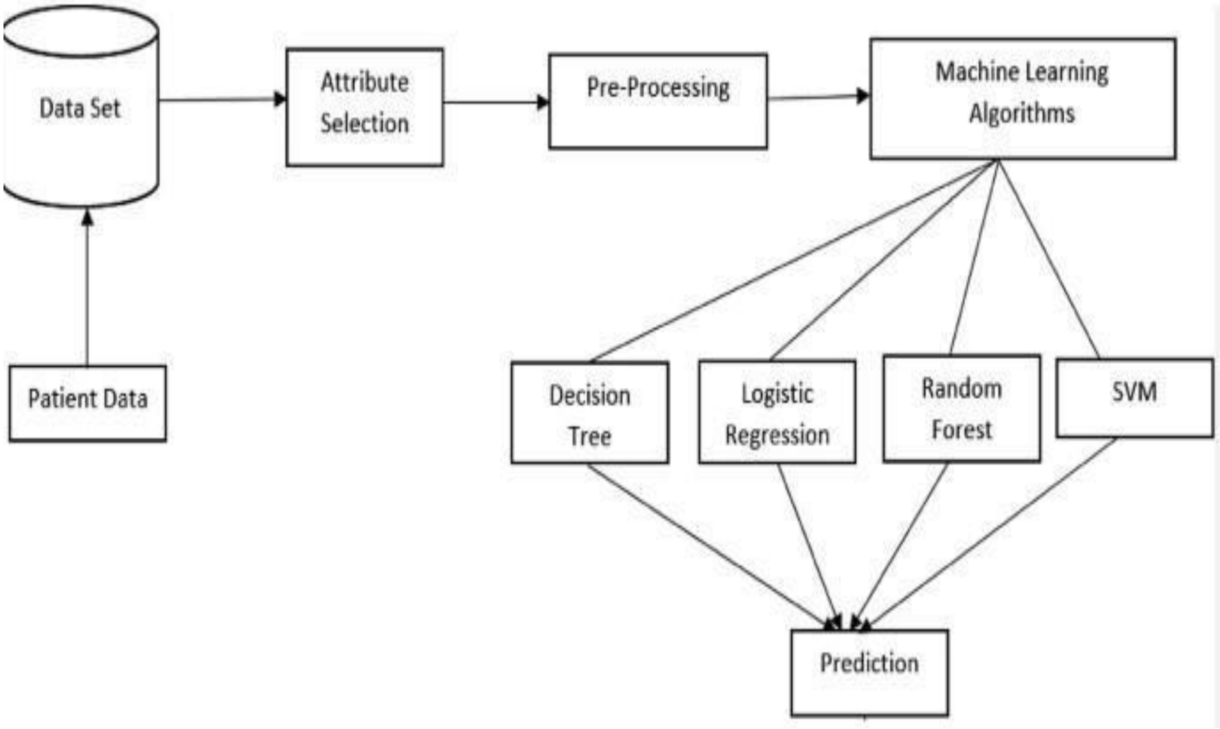


Fig 4.3 : Proposed Architecture

4.4 SYSTEM SPECIFIC DESIGNS

4.4.1 UML DIAGRAMS

A use case diagram for Cardiovascular Disease (CVD) prediction using Machine Learning would outline the interactions between different actors and the system itself. The primary actors typically include the patient, healthcare professional, and the machine learning system. Firstly, the patient interacts with the system by providing personal health data such as age, gender, blood pressure, cholesterol levels, and lifestyle factors. The system then processes this input and generates a prediction regarding the patient's risk of developing CVD.

Secondly, healthcare professionals interact with the system in a clinical setting. They use the system as a tool to assist in diagnosing and managing CVD risks for their patients. This interaction involves inputting patient data into the system, reviewing the predictions generated by the model, and using this information to inform treatment plans or preventive measures. Healthcare professionals may also provide feedback to the system to improve its accuracy or usability.

Lastly, the machine learning system itself is the core component of the use case diagram. It encompasses the algorithms, data processing pipelines, and prediction models responsible for analyzing input data and generating CVD risk predictions. The system interacts with both patients and healthcare professionals, receiving input data from patients and providing predictions to healthcare professionals. Additionally, the system may include features for model training, evaluation, and maintenance to ensure its accuracy and relevance over time.

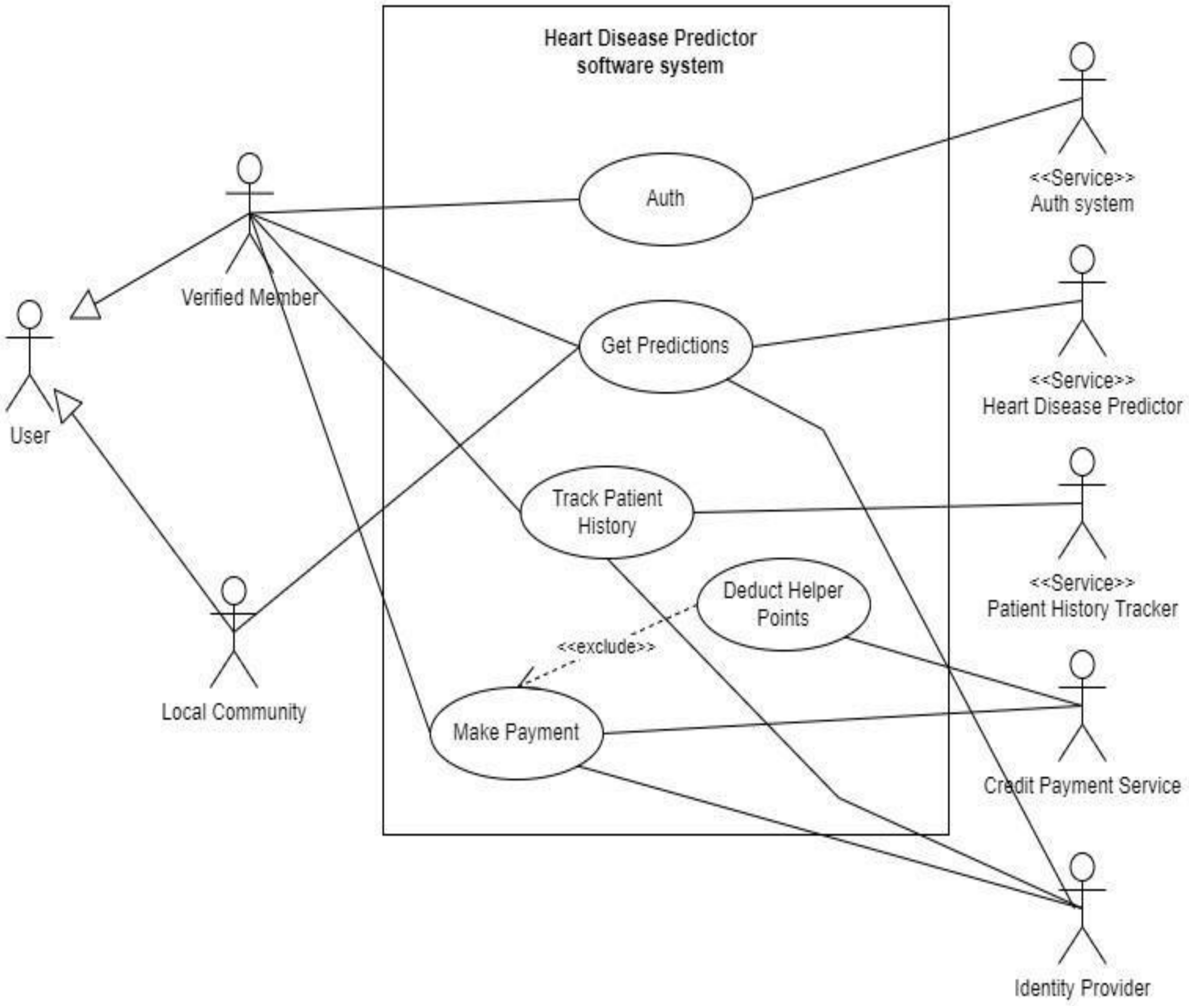


Fig 4.4.1 UML Diagram

4.4.2 ACTIVITY DIAGRAM

An activity diagram for Cardiovascular Disease (CVD) prediction using Machine Learning would illustrate the flow of activities involved in the prediction process. The diagram would begin with the initial activity of data collection, where relevant patient information such as age, gender, blood pressure, cholesterol levels, and lifestyle factors is gathered. This data collection process may involve input from healthcare professionals, electronic health records, or patient self-reports.

Following data collection, the next activity in the diagram would be data preprocessing. This involves cleaning the data to remove any inconsistencies or errors, handling missing values, and normalizing features to ensure uniformity. Once the data is preprocessed, it is ready for use in training the machine learning model. This training activity involves selecting an appropriate algorithm, splitting the data into training and validation sets, and optimizing the model's parameters to achieve the best possible prediction accuracy.

The final activity in the activity diagram would be the prediction process itself. Once the machine learning model is trained, it can be used to predict the risk of CVD for new patients. This involves inputting the patient's data into the model, which then generates a prediction regarding their likelihood of developing CVD. The prediction results can then be presented to healthcare professionals for further analysis and decision-making regarding treatment or preventive measures. Additionally, the prediction process may involve feedback loops where the model's performance is evaluated and refined over time to improve its accuracy and reliability.

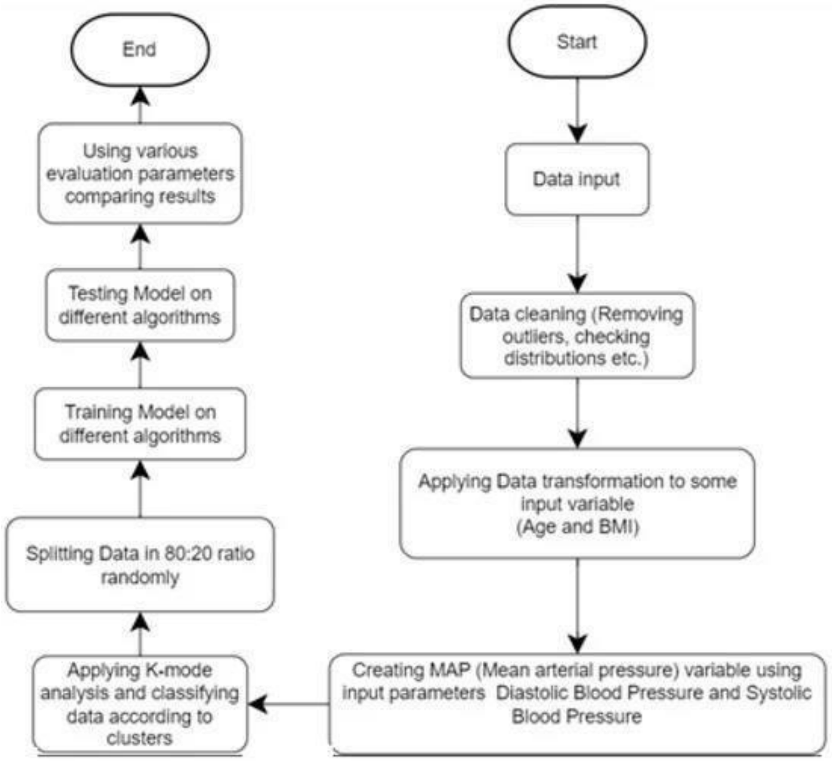
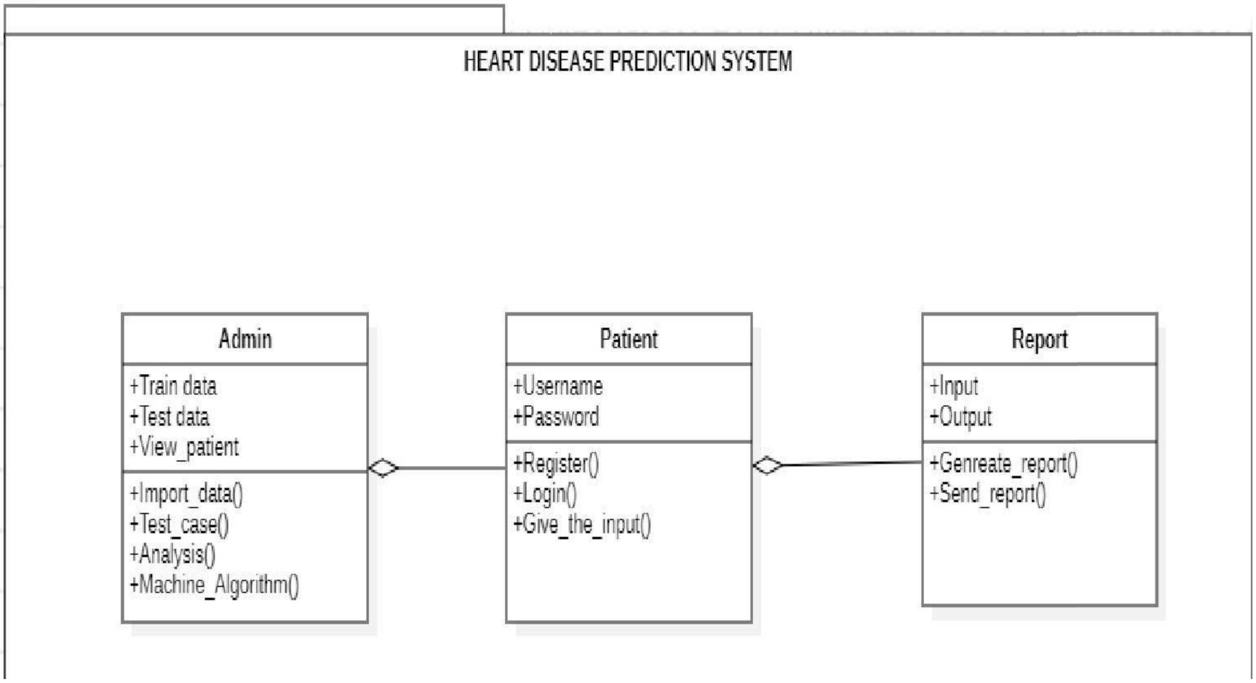

End
Using various evaluation parameters comparing results
Testing Model on different algorithms
Training Model on different algorithms
Splitting Data in 80:20 ratio randomly
Applying K-mode analysis and classifying data according to clusters
Start
Data input
Data cleaning (Removing outliers, checking distributions etc.)
Applying Data transformation to some input variable (Age and BMI)
Creating MAP (Mean arterial pressure) variable using input parameters Diastolic Blood Pressure and Systolic Blood Pressure

Fig 4.4.2 Activity Diagram

4.4 .3 CLASS DIAGRAM



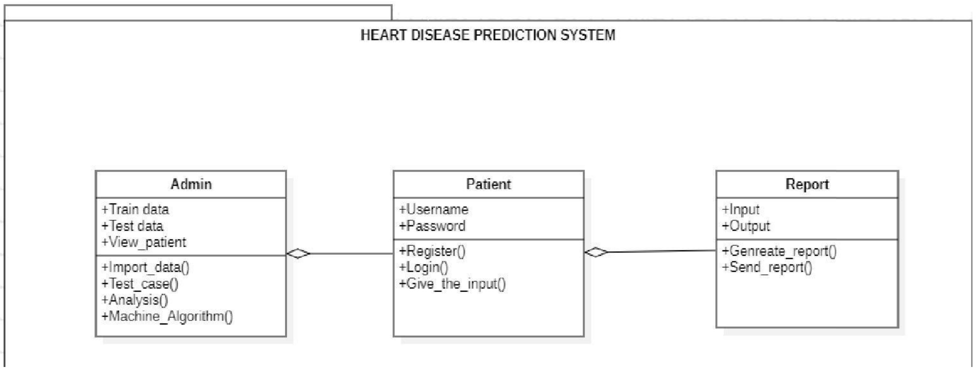


Fig :4.4.3 class Diagram

4.5 Sequence Diagram

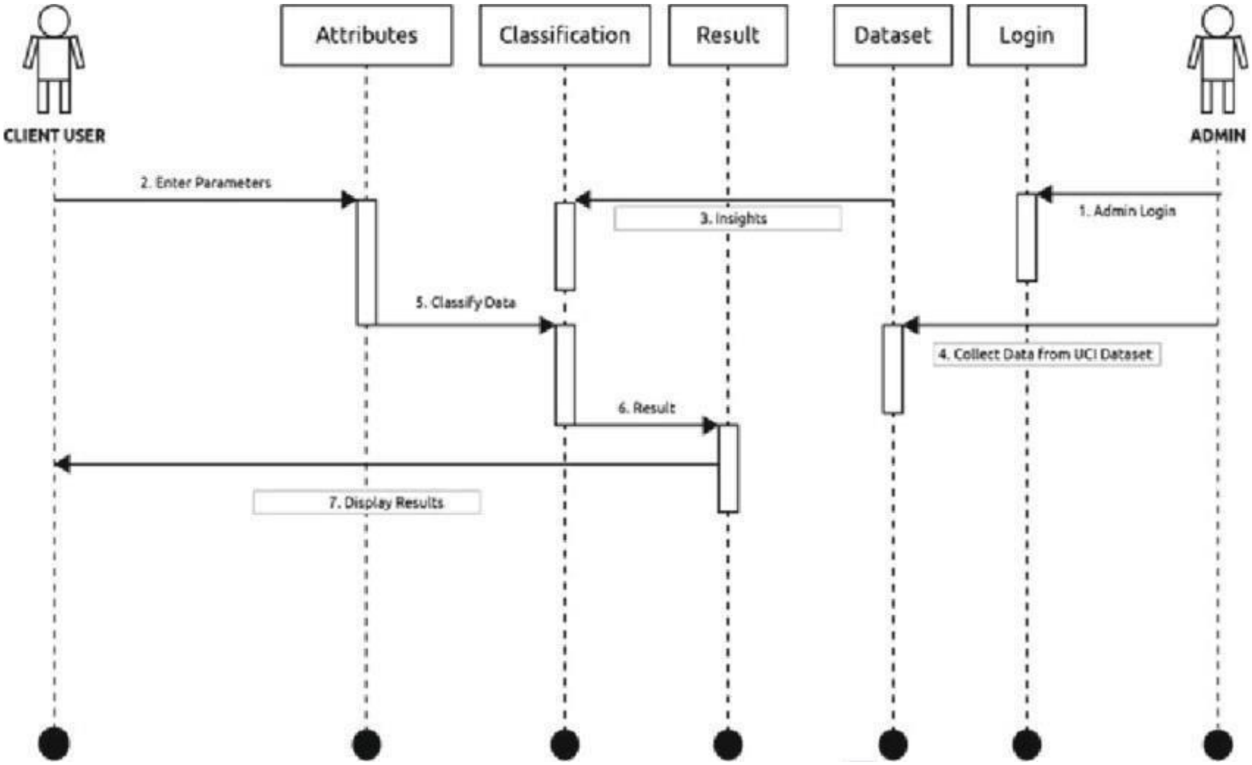


Fig 4.5 Sequence Diagram

A sequence diagram for Cardiovascular Disease (CVD) prediction using Machine Learning would delineate the chronological sequence of interactions between the various components involved in the prediction process. It would start with the patient initiating the interaction by providing their health data to the system. This data includes information such as age, gender, blood pressure, cholesterol levels, and lifestyle habits. The system then receives this input and begins the prediction process.

Subsequently, the system invokes the machine learning algorithm to analyze the patient's data. This step involves preprocessing the input data to ensure consistency and accuracy. The algorithm then utilizes the preprocessed data to generate a prediction regarding the patient's risk of developing CVD. This prediction is based on the patterns and relationships identified in the data during the training phase of the machine learning model.

Finally, the prediction results are returned to the healthcare professional or user who initiated the request. The healthcare professional can then review the prediction and use it to inform clinical decisions, such as recommending preventive measures or lifestyle changes to mitigate the risk of CVD. Additionally, the sequence diagram may include feedback loops where the system can receive feedback on the prediction accuracy and adjust its algorithms or parameters accordingly to improve future predictions. This iterative process helps refine the model and enhance its effectiveness in predicting CVD risk accurately.

4.6. Interface design

Designing an interface for a cardiovascular disease prediction system using machine learning involves several key steps. Below is a high-level overview of the process, including the design principles and features that should be considered:

- **Dashboard:** Design a dashboard that provides an overview of patient status and risk scores.
- **Patient Profile:** Create a detailed view for individual patients, including their medical history and risk factors.
- **Risk Prediction:** Display risk scores and predictions clearly, with explanations and confidence intervals.
- **Visualizations:** Use charts and graphs to show trends, comparisons, and historical data

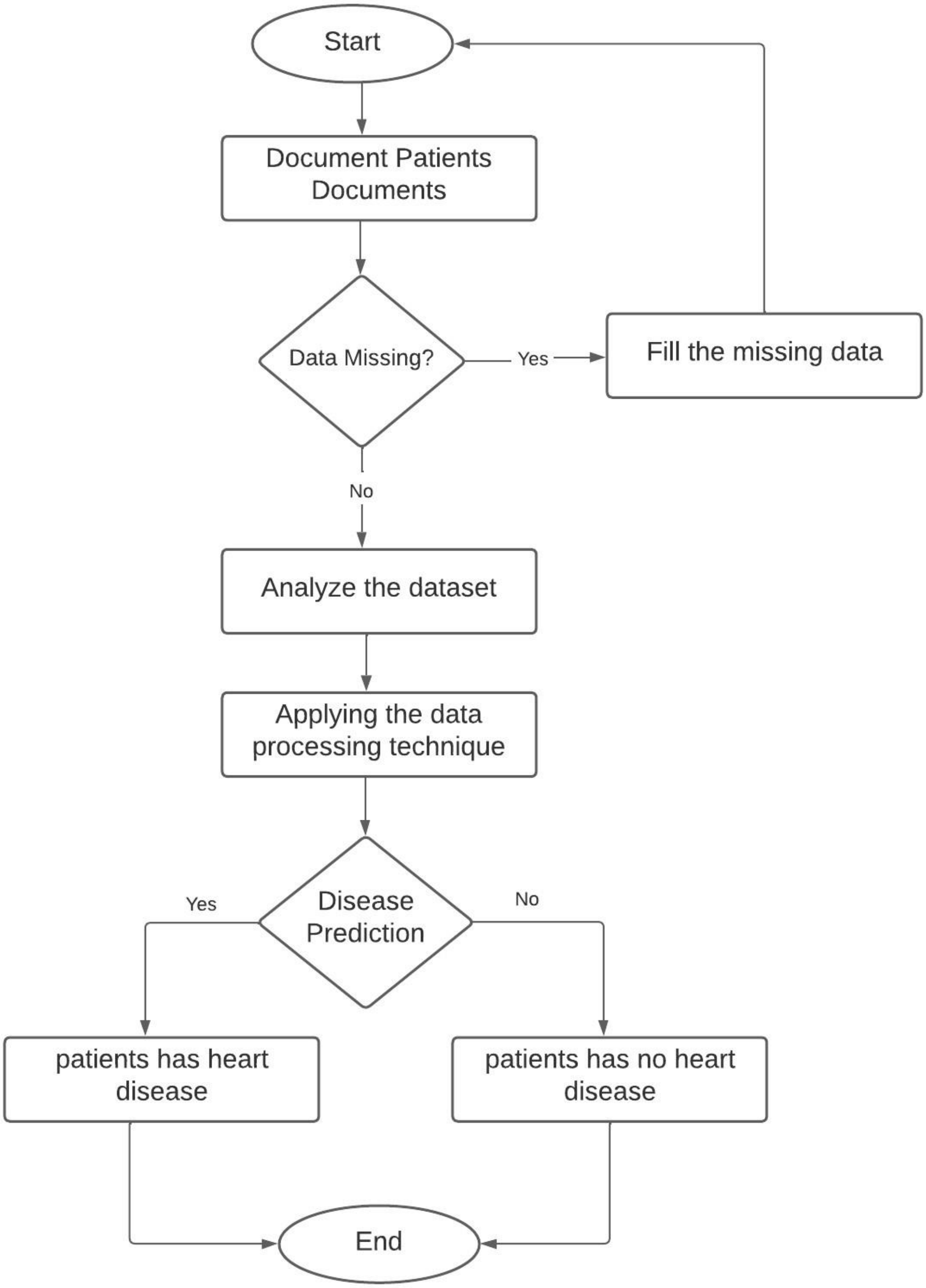

Start
Document Patients Documents
Data Missing?
Yes
Fill the missing data
No
Analyze the dataset
Applying the data processing technique
Disease Prediction
Yes
No
patients has heart disease
patients has no heart disease
End

Table4.6: Interface design

4.7. ER DIAGRAM

An activity diagram for Cardiovascular Disease (CVD) prediction using Machine Learning would illustrate the flow of activities involved in the prediction process. The diagram would begin with the initial activity of data collection, where relevant patient information such as age, gender, blood pressure, cholesterol levels, and lifestyle factors is gathered. This data collection process may involve input from healthcare professionals, electronic health records, or patient self-reports.

Following data collection, the next activity in the diagram would be data preprocessing. This involves cleaning the data to remove any inconsistencies or errors, handling missing values, and normalizing features to ensure uniformity. Once the data is preprocessed, it is ready for use in training the machine learning model. This training activity involves selecting an appropriate algorithm, splitting the data into training and validation sets, and optimizing the model's parameters to achieve the best possible prediction accuracy.

The final activity in the activity diagram would be the prediction process itself. Once the machine learning model is trained, it can be used to predict the risk of CVD for new patients. This involves inputting the patient's data into the model, which then generates a prediction regarding their likelihood of developing CVD. The prediction results can then be presented to healthcare professionals for further analysis and decision-making regarding treatment or preventive measures. Additionally, the prediction process may involve feedback loops where the model's performance is evaluated and refined over time to improve its accuracy and reliability.

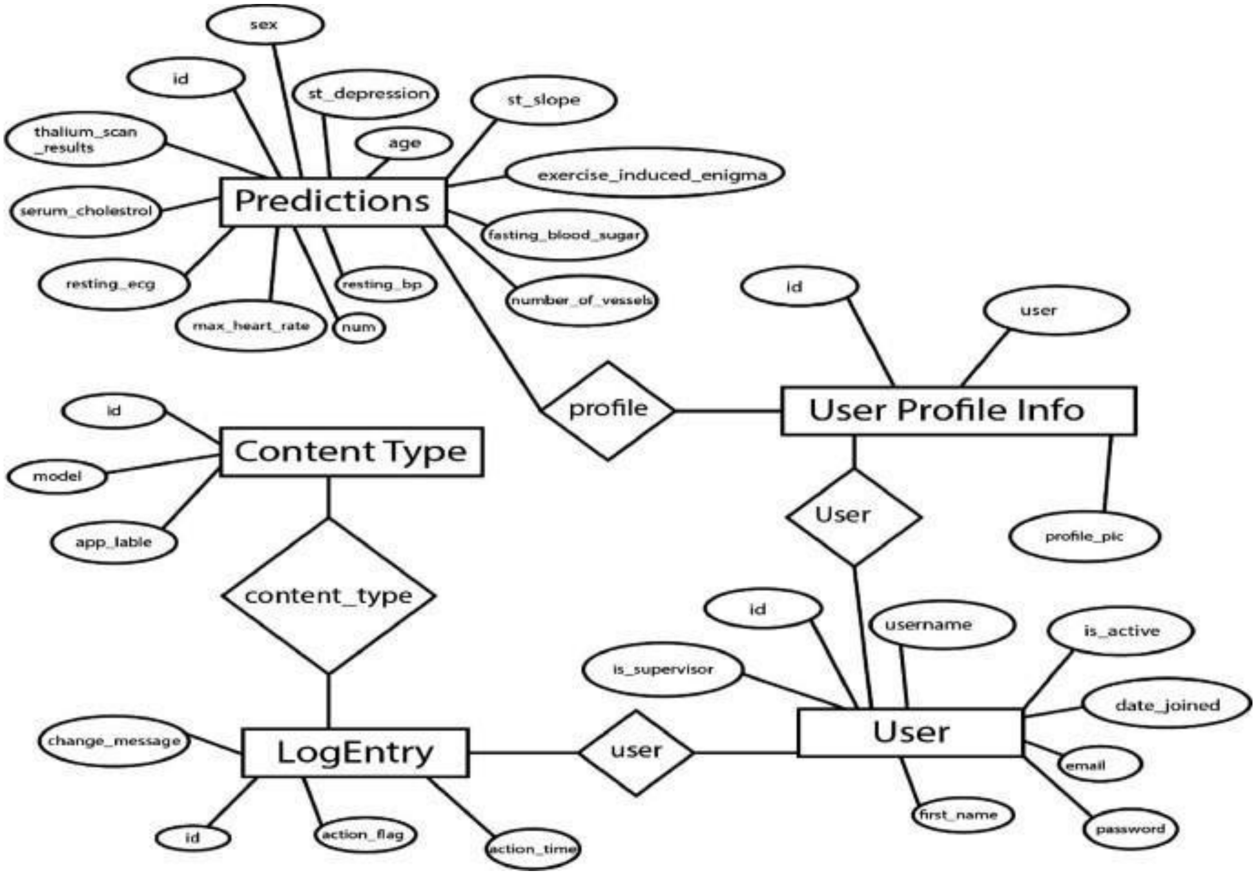


Fig 4.7 ER Diagram

5.TESTING

5.1 INTRODUCTION TO TESTING

Predicting cardiovascular disease (CVD) using machine learning involves several steps, including data preprocessing, feature selection, model selection, training, evaluation, and interpretation. Here's a high-level overview of how to approach this

5.1.1 Collection and Preprocessing:

The initial module in the proposed framework for cardiovascular disease (CVD) prediction using machine learning is dedicated to data collection and preprocessing. This involves sourcing data from diverse repositories including electronic health records (EHRs), wearable devices, genetic databases, and environmental sensors. Raw data undergoes thorough preprocessing to ensure consistency and quality. Steps include handling missing values, normalizing features, and encoding categorical variables. Additionally, feature engineering techniques such as dimensionality reduction and interaction term generation are applied to distill meaningful features from the raw data.

5.1.2 Feature Selection and Engineering:

The next module focuses on feature selection and engineering, aiming to identify relevant predictors of CVD risk and enhance the discriminative power of the predictive models. Various feature selection techniques such as mutual information, recursive feature elimination, and principal component analysis are employed to identify the most informative features from the dataset. Moreover, domain-specific knowledge incorporation and interaction term generation are utilized to capture complex relationships between variables and improve model performance.

5.1.3 Model Selection and Development:

In this module, a diverse array of machine learning algorithms are explored to construct predictive models for cardiovascular risk assessment. Supervised learning algorithms including logistic regression, support vector machines (SVMs), decision trees, random forests, gradient boosting machines, and neural networks are evaluated to leverage the predictive power of labeled data. Ensemble learning techniques such as bagging, boosting, and stacking are also employed to combine multiple base learners and enhance model robustness.

Deep Learning Architectures:

A dedicated module is allocated to deep learning architectures, comprising convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models. These architectures excel at extracting intricate patterns from complex, high-dimensional data such as medical images, genetic sequences, and temporal sequences of physiological measurements. By leveraging deep learning, the models aim to capture hierarchical representations and facilitate more accurate risk prediction and early detection of CVD.

5.1.4 Model Evaluation and Validation:

This module focuses on evaluating and validating the predictive models to ensure their efficacy and generalization capabilities. Rigorous performance metrics and cross-validation techniques are employed to assess model performance. Calibration methods such as Platt scaling and isotonic regression are utilized to refine the predicted probabilities of cardiovascular events and enhance model reliability. Sensitivity analyses and subgroup analyses are conducted to evaluate model robustness across diverse patient populations and clinical settings.

5.1.4.1 Interpretability and Transparency:

A critical aspect of the proposed framework is interpretability and transparency in model development. Techniques such as feature importance ranking, partial dependence plots, and model-agnostic interpretability methods are employed to elucidate the factors driving individual risk profiles. By enhancing interpretability, the models aim to facilitate clinical decision-making and foster trust among healthcare providers and patients.

5.1.4.2 Multimodal Data Integration:

The final module emphasizes the integration of multimodal data sources and contextual information to develop comprehensive risk prediction models. This includes integrating imaging data, genetic data, socio-economic factors, and environmental exposures to capture the multifaceted nature of cardiovascular health determinants. By integrating diverse data sources, the models aim to provide a holistic view of cardiovascular risk and improve predictive accuracy.

5 .2 Test Cases

- **Data Collection**: Gather a dataset that contains relevant features (e.g., age, gender, blood pressure, cholesterol levels, etc.) and labels indicating the presence or absence of cardiovascular disease.
- **Data Preprocessing**:
  - **Cleaning**: Handle missing values, outliers, and noise in the data.
  - **Normalization/Scaling**: Standardize the features to have a mean of 0 and a standard deviation of 1, or scale them to a range (e.g., 0 to 1).
  - **Encoding**: Convert categorical variables into numerical values using techniques like one-hot encoding or label encoding.

Test Reports

- **Feature Selection**: Select the most relevant features that contribute to the prediction of cardiovascular disease. This can be done using techniques like correlation analysis, feature importance from models, or dimensionality reduction techniques like PCA.
- **Model Selection**: Choose appropriate machine learning algorithms for the task. Common algorithms used for CVD prediction include:

  - Logistic Regression
  - Decision Trees
  - Random Forest
  - Gradient Boosting Machines (e.g., XGBoost, LightGBM)
  - Support Vector Machines (SVM)
  - Neural Networks
- **Model Training**: Split the dataset into training and testing sets (e.g., 80/20 split). Train the chosen models on the training set.

Table 5.2 given below in detain about test case

Attribute	**Description**
Patient's age	>35
Gender	Value 1: male; Value 0: Female
Chest pain type	Value 1: typical type 1 angina; Value 2: typical type 2 angina; Value 3: Non-angina pain; Value 4: Asymptomatic
Fasting blood sugar	Value 1:>120 mg/dl; Value 0: 120 mg.dl
Rest ecg – resting electrographic	Value 0: normal; Value 1: having st-t wave abnormality; Value 2: definite left ventricular hypertrophy
exang – exercise included angina	Value 1: yes; Value 0: no
Slope-the slope of the peak exercise ST segment	Value 1: unsloping; Value 2: flat: Value 3: down sloping
Ca – number of major vessels colored by fluoroscopy	Value 0-3
Thal	Value 3: normal: Value 6: fixed defect; Value 7: Reversible defect
Trest blood pressure	mm hg on admission to the hospital
Serum cholesterol	mg/dl
Thalach-maximum heart rate achieved	60-200
Old peak-ST depression included by exercise	0-6
Hear disease present	0: No; 1: Yes

5.3 Test Conclusion

- **Model Evaluation**: Evaluate the models on the testing set using appropriate metrics such as:
    - Accuracy
    - Precision
    - Recall
    - F1-score
    - Area Under the Receiver Operating Characteristic Curve (AUC-ROC)
    - Confusion Matrix
- **Hyperparameter Tuning**: Use techniques like grid search or random search with cross-validation to find the best hyperparameters for the models.
- **Model Interpretation**: Understand the model's predictions and the importance of different features. Techniques like SHAP (SHapley Additive exPlanations) values can be useful for interpreting complex models.
- **Deployment**: Once the model is trained and evaluated, it can be deployed for real-time prediction on new data.

- *Feature Selection:* Apply feature selection techniques to identify the most informative and relevant features for CVD prediction. This helps reduce dimensionality and improves the performance of the predictive models.
- *Algorithm Selection:* Choose appropriate machine learning algorithms that are suitable for CVD prediction. Commonly used algorithms include decision trees, random forests, support vector machines, logistic regression, and neural networks. The selection process takes into account factors such as the nature of the data, the complexity of the problem, and interpretability requirements.

In this section, the results are presented and the accuracy outputs of the algorithms are displayed. A comparison is made between the algorithms based on their accuracy. When building predictive models in the domain of CVD prediction using ML algorithms, the training and testing data split ratio can vary based on a number of factors. In this work, the dataset sizes are 80 and 20 for training and testing respectively. The accuracy generated by each algorithm is shown in Table II as follows:

COMPARISON OF ACCURACIES

Algorithm	**Accuracy**	**Overall Accuracy**
Logistic regression	0.9160	0.8652
Random forest	0.8952	0.8087
Naïve Bayes	0.9094	0.8416
Gradient boosting	0.9069	0.8416
SVM	0.8825	0.7972

Table: 5.3 comparison of accuracies

The above results are obtained by using the in-built classifier present in the Sklearn – a machine learning library in python. By performing the train-test split and hyper parameter tuning, the model can be trained and tested on the given dataset. Additionally, this approach allows for obtaining the optimal parameter that yields higher efficiency. User interface can also be provided to obtain the values from the user to provide results directly to the interface used by the

6. IMPLEMENTATION

Implementing a Cardiovascular Disease (CVD) prediction system using Machine Learning involves several key steps. First, data collection and preprocessing are crucial. This entails gathering comprehensive datasets containing relevant patient information such as age, gender, blood pressure, cholesterol levels, smoking habits, and medical history. Preprocessing involves cleaning the data, handling missing values, and normalizing features to ensure consistency and accuracy in the model.

Next, selecting an appropriate machine learning algorithm is essential. Techniques such as logistic regression, support vector machines, random forests, or neural networks can be considered based on the complexity of the data and the desired prediction accuracy. The chosen algorithm is then trained on the preprocessed data, using a portion for training and another for validation to evaluate its performance and fine-tune hyperparameters.

Finally, deploying the trained model into a production environment is necessary for real-world application. This involves integrating the model into a software system or application where it can receive input data, make predictions, and present results to users or healthcare professionals. Continuous monitoring and updates may also be needed to ensure the model remains accurate and relevant as new data becomes available or as the population demographics change. Additionally, adherence to data privacy and security regulations must be maintained throughout the implementation process to safeguard sensitive patient information.

6.1 DATASET DESCRIPTION

Four databases—Cleveland, Hungary, Switzerland, and Long Beach V—make up the Public Health Dataset, which dates back to 1988 and was used for this study. It has 76 properties total, including the anticipated attribute, however only 14 of them are used in the published studies. The patient's heart condition is indicated in the "target" field. It has an integer value system where 0 indicates no disease and 1 indicates disease. Table 1 displays all of the dataset features as well as the first four rows without any preprocessing. Now, the characteristics that are employed in this study are explained along with their intended application and similarity:

- Age—age of patient in years, sex—(1 = male; 0 = female).
- Cp—chest pain type.
- Trestbps—resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if you have a normal blood pressure reading, it is fine, but if it is a little higher than it should be, you should try to lower it. Make healthy changes to your lifestyle).
- Chol—serum cholesterol shows the amount of triglycerides present. Triglycerides are another lipid that can be measured in the blood. It should be less than 170 mg/dL (may differ in different Labs).
- Fbs—fasting blood sugar larger than 120 mg/dl (1 true). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
- Restecg—resting electrocardiographic results.
- Thalach—maximum heart rate achieved. The maximum heart rate is 220 minus your age.
- Exang—exercise-induced angina (1 yes). Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease.
- Oldpeak—ST depression induced by exercise relative to rest.
- Slope—the slope of the peak exercise ST segment.
- Ca—number of major vessels (0–3) colored by fluoroscopy.
- Thal—no explanation provided, but probably thalassemia (3 normal; 6 fixed defects; 7 reversible defects).
- Target (T)—no disease = 0 and disease = 1, (angiographic disease status).

The UCI Cleveland dataset was the only one utilized from the Heart disease Dataset, which is a mix of four distinct databases. Although this database has 76 properties in total, only 14 features are mentioned in any published research. For our investigation, we have therefore chosen the UCI Cleveland dataset that has already been processed and is accessible on the Kaggle website. Table 1 (below) provides a detailed description of all 14 qualities that are employed in the proposed study.

6.2 Checking the Distribution of the Data :

- When classifying or making predictions about a problem, the distribution of the data is crucial.
- It can be observed that 54.46% of the dataset had heart disease, whilst 45.54% had no heart disease.
- Thus, the dataset must be balanced to prevent overfitting. This will assist the model in identifying a dataset pattern that leads to heart disease.

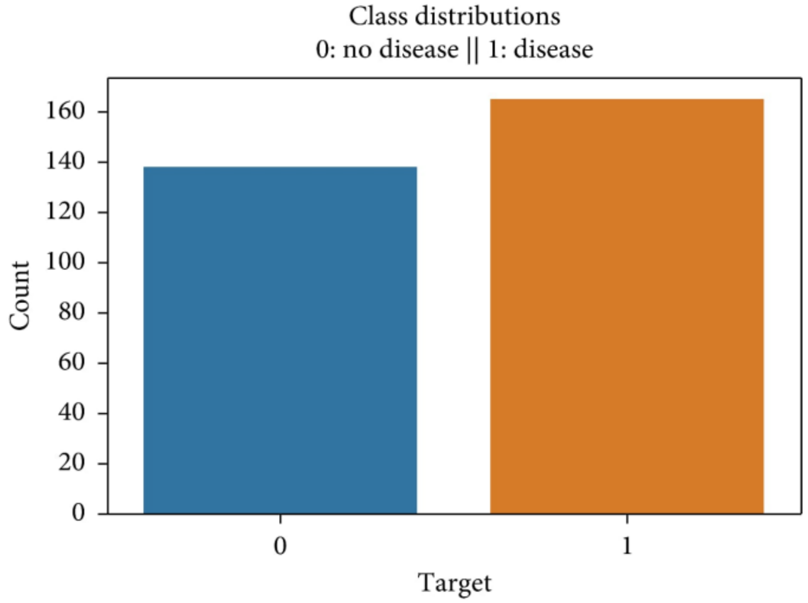


Fig 4.2 Class Distributions Deep

Learning Pseudocode:

- Dataset of training
- Dataset of testing
- Checking the shape/features of the input
- The procedure of initiating the sequential layer
- Adding dense layers with dropout layers and ReLU activation functions
- Adding a last dense layer with one output and binary activation function
- End repeat

6.3 Algorithm

6.3.1 Support Vector Machine (SVM):

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In the 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

The followings are important concepts in SVM -

- Support Vectors - Data Points that are closest to the hyperplane are called support vectors. Separating line will be defined with the help of these data points.
- Hyperplane - As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

- Margin - It may be defined as the gap between two lines on the closest data points of different classes. It can be calculated as the perpendicular distance from the line to the

6.3.2 Naive Bayes Algorithm:

It is a machine learning technique that works on the strategy of the Bayes' Theorem. It basically assumes that there would be no attributes dependent on each other. It is a group of algorithms that have a common principle that every feature is independent of the other. Bayes' Theorem tells us the probability of an event that will occur when another event has already occurred. The mathematical equation is:

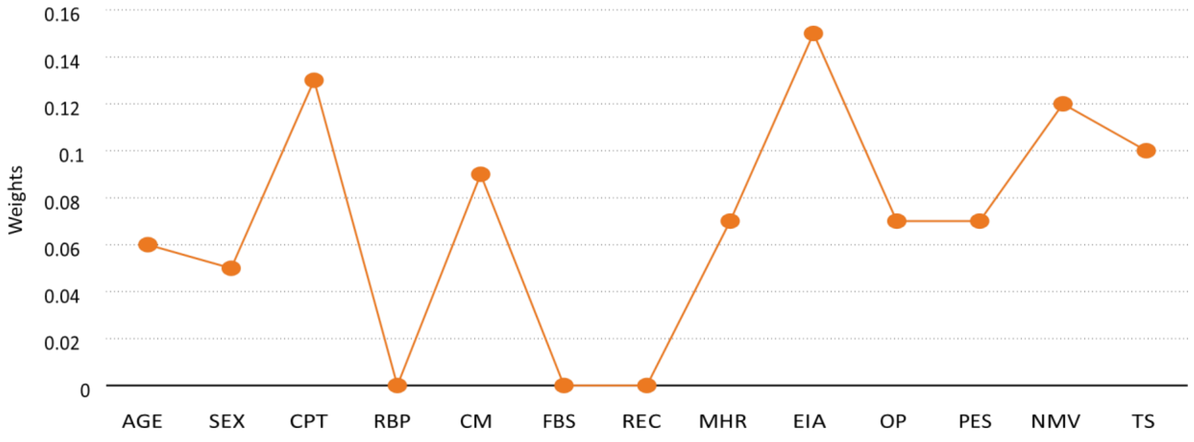


Fig:6.32

Probability(a|z) = (Probability(z|a) * Probability(a)) / Probability(z) Where

- **Probability(a|z):** Gives us the probability of a(the hypothesis)gives the data is.

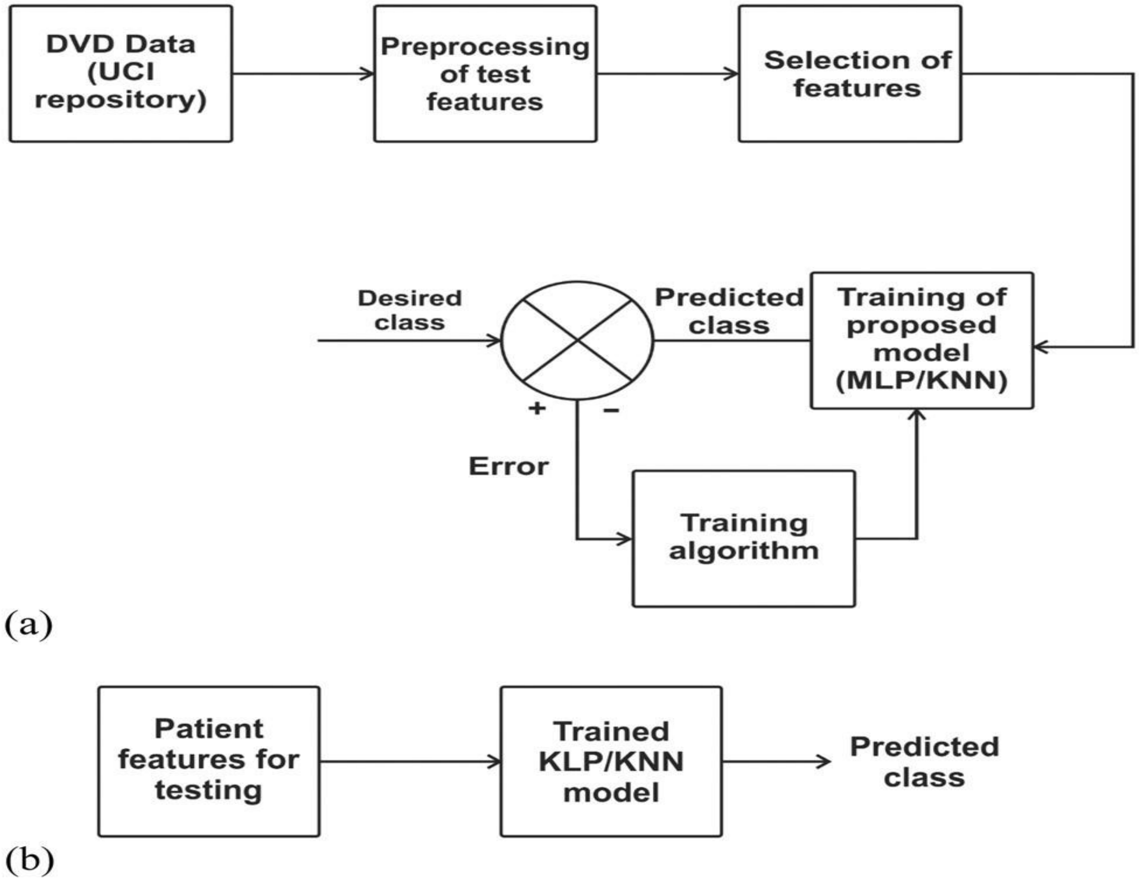
- **Probability (z|a):** Gives the probability of the data when the hypothesis is true. **Probability (a):** Regardless of the data, the hypothesis is said to be true.
- **Probability (d):** Regardless if the data, the probability of the hypothesis is given.

6.3.3 Decision Tree:

Decision trees are treelike structures that are used to manage large datasets. They are often depicted as flowcharts, with outer branches representing the results and inner nodes representing the properties of the dataset. Decision trees are popular because they are efficient, reliable, and easy to understand. The projected class label for a decision tree originates from the tree's root. The following steps in the tree are decided by comparing the value of the root attribute with the information in the record. Following a jump on the next node, the matching branch is followed to the value shown by the comparison result. Entropy changes when training examples are divided into smaller groups using a decision tree node. The measurement of this change in entropy is information gain An accuracy of 73.0% has been achieved by the decision tree.[6] In a research by.[7] 72.77% accuracy was achieved by the decision tree classifier

6.3.4 RANDOM FOREST

The random forest algorithm belongs to a category of supervised classification technique that consists of multiple decision trees working together as a group. The class with the most votes become the prediction made by our model. Each tree in the random forest makes a class prediction, which eliminates the limitations of the decision tree algorithm. This improves accuracy and reduces overfitting of the dataset. When used on large datasets, the random forest approach may still provide the same results even if a significant portion of record values are missing. The samples produced by the decision tree may be saved and used with various data types . In the research in , random forest achieved a test accuracy of 73% and a validation accuracy of 72% with 500 estimators, 4 maximum depths, and 1 random state.


DVD Data (UCI repository)
Preprocessing of test features
Selection of features
Desired class
Predicted class
Training of proposed model (MLP/KNN)
+
−
Error
Training algorithm
(a)
Patient features for testing
Trained KLP/KNN model
Predicted class
(b)

6.3.4 XGBOOST

XGBoost [14] is a version of gradient boosted decision trees. This algorithm involves creating decision trees in a sequential manner. All the independent variables are allocated weights, which are subsequently used to produce predictions by the decision tree. If the tree makes a wrong prediction, the importance of the relevant variables is increased and used in the next decision tree. The output of each of these classifiers/predictors is then merged to produce a more robust and accurate model. In a study by [34], the XGBoost model achieved 73% accuracy with the parameters 'learning_rate': 0.1, 'max_depth': 'n_estimators': 100, 'cross-validation': 10 folds including 49,000 training and 21,000 testing data instances on 70,000 CVD dataset

6.3.5 K-Nearest Neighbor.

KNN uses a training set directly for classifying the test data. Which refers to the number of KNN. To test each data, it calculates all the training data and the distance between them. Then, test data will be assigned to be used by multiplicity voting and class label. The Euclidean distance measure equation is given below:

K-NN is a non-parametric classifier used to determine whether a patient has CVD or not using a labelled known data set. Predictions are made based on k numbers of frequently used neighbours for a new object, and a different distance metric for finding the K-NN is used. K-NN classifies new

training data points based on similarity measurements. Data points are classified by considering the majority of votes from its neighbours. This works effectively for small dimensional data sets. K-NN does not require extra training for classification if a new data point is added to the existing data set. It is an inefficient algorithm for large data sets and requires more memory space for computation and longer model testing times because of the need to compute the distance between training data set and testing data set during each test.

6.3.6 LOGISTIC REGRESSION

Logistic regression is a machine learning algorithm used for classification. It is based on the concept of probability. Logistic regression is used to assign observations to a discrete class.

Transforming output is done using the sigmoid logic function. The logistic regression hypothesis tends to limit the cost function in range between 0 and 1. Therefore, linear functions cannot represent as it can have a value >1 or <=0, which is not possible according to the regression hypothesis.

the dataset exhibits clear evidence of outliers. These outliers likely originated from data entry errors. Eliminating these outliers could enhance the predictive accuracy of our model. To tackle this issue, we systematically removed any instances of ap_hi, ap_lo, weight, and height that fell beyond the 2.5% to 97.5% range. This process involved manual identification and elimination of outliers. Consequently, the data cleaning procedure led to a reduction in the number of rows from 70,000 to 57,155.

7 FUTURE ENHANCEMENT

We advocate for the utilization of binning as a technique to transform continuous inputs, like age, into categorical inputs, aiming to enhance the performance and interpretability of classification algorithms. By organizing continuous inputs into discrete groups or bins, the algorithm becomes capable of distinguishing between various data classes based on specific input variable values. For example, employing the "Age Group" variable with categories such as "Young," "Middle-aged," and "Elderly," allows classification algorithms to segment the dataset into distinct classes according to individuals' age groups

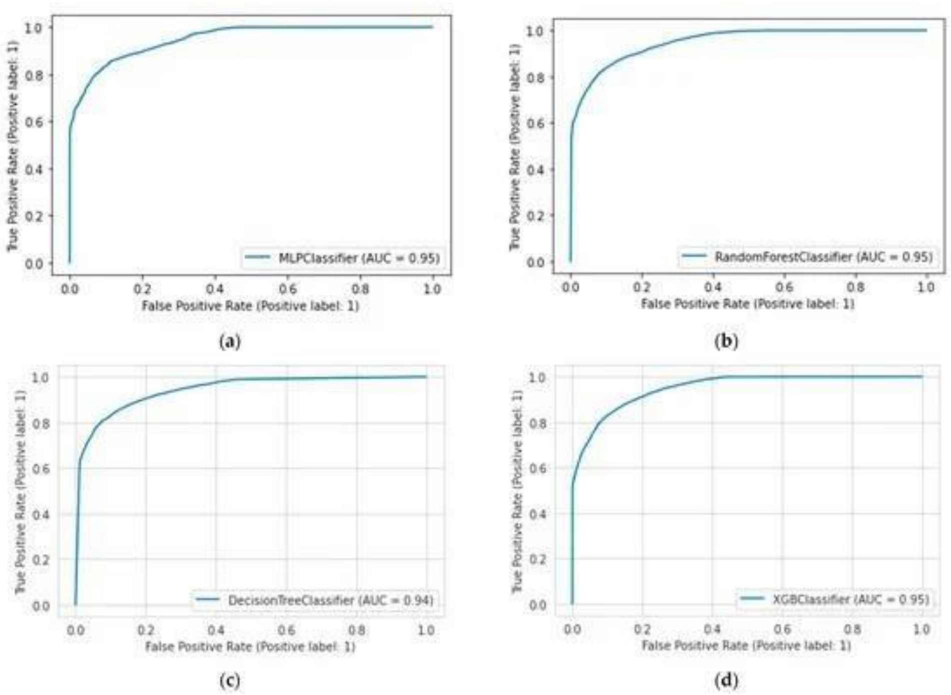
facilitate result interpretation, as it simplifies understanding the relationship between input variables and output classes. In contrast, using continuous inputs, such as numerical values, might pose challenges for classification algorithms, as they might need to make assumptions about class boundaries. In our research, we implemented binning on the age attribute Furthermore, converting continuous inputs into categorical ones through binning can within a patient dataset. Initially provided in days, the ages were transformed into years by dividing by 365 for better analysis and prediction. Subsequently, age data were grouped into bins spanning 5-year intervals, ranging from 0–20 to 95–100. With the dataset's minimum age at 30 and maximum at 65, the bins were labeled accordingly, with the first bin (30–35) marked as 0, and the last (60–65) as 6**.**

7.1 Performance Evaluation

From the dataset, a training set comprising 80% of the data and a testing set comprising 20% are partitioned. Subsequently, a model is trained using the training set, and its efficacy is evaluated through testing on the separate testing set. Various classifiers, including decision tree classifier, random forest classifier, multilayer perceptron, and XGBoost, are employed on the clustered dataset to gauge their performance. The effectiveness of each classifier is then measured using metrics such as accuracy, precision, recall, and F-measure scores.

This study utilized Google Colab on a Ryzen 7 computer equipped with a 4800-H processor and 16 GB of RAM. The initial dataset comprised 70,000 rows and 12 attributes, but after cleaning and preprocessing, it was streamlined to approximately 59,000 rows and 11 attributes. As all attributes were categorical, outlier removal was implemented to enhance model efficiency. The study employed random forest, decision tree, multilayer perceptron, and XGBoost classifier algorithms. Performance evaluation encompassed precision, recall, accuracy, F1 score, and area under the ROC curve metrics. The dataset was partitioned into training (80%) and testing (20%) sets.

An automated approach for hyperparameter tuning was adopted, leveraging the GridSearchCV method. This method utilizes an estimator, a hyperparameter set for exploration, and a scoring metric to determine the optimal hyperparameters maximizing performance. Implemented within the scikit-learn library, GridSearchCV employs k-fold cross-validation for evaluating various hyperparameter sets.



(a) (b)

(c) (d)

Fig 7.1 ROC–area under curve

ML algorithms. By increasing the number of hidden layer nodes and employing a 10-fold cross-validation technique in the MLP model, we are able to improve its accuracy.

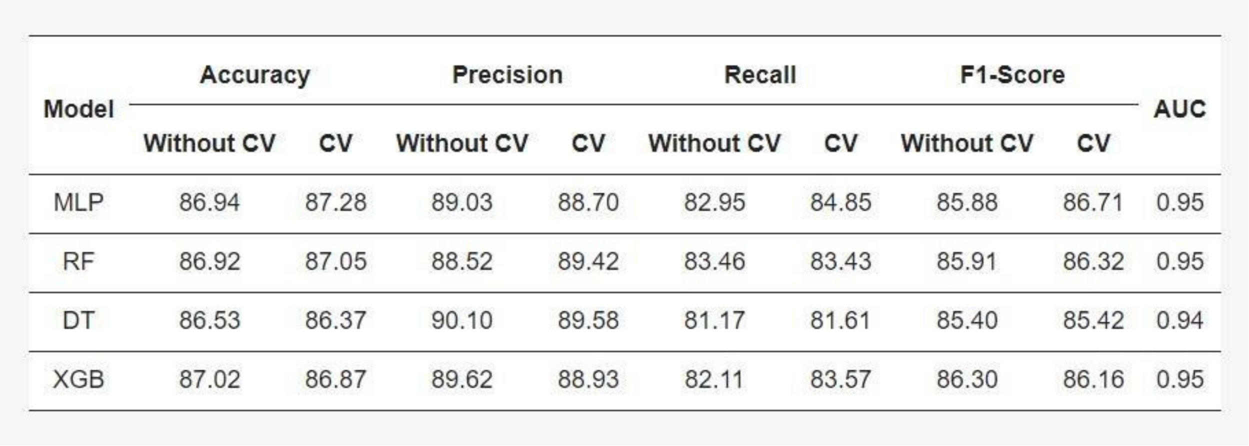
Regarding the literature, Alizadehsani [21] compared the performances of various well-known ML techniques for coronary artery disease detection. Latha and Jeeva [22] proposed a model for increasing the accuracy of weak classifiers by 7.26% using an ensemble model. Ahmed and co. [23] reported a heart disease risk-prediction model with a 94.9% accuracy using random forest. Beunza et al. [24] compared the performance of several ML algorithms based on the Framingham heart database to predict coronary heart disease using R-Studio and Rapid Miner and achieved the highest AUC value of 0.75 using a support vector machine method. Kim et al. [25] obtained a 0.89 AUC value using an artificial neural network to predict the survival rate of injured patients. In another study, Shah et al. [26] obtained a maximum accuracy of 90.78% when predicting heart disease using the K-NN algorithm, and Pal and Parija [27] reported a heart disease risk-prediction model with 86.9% accuracy, 90.6% sensitivity, and 82.7% specificity by using the random forest algorithm.

This study presents a comparison of two ML techniques for CVD prediction: K-NN and MLP. Between these algorithms, MLP provides better accuracy (82.47%) than K-NN with an accuracy of 73.77%. The diagnosis rate was found to be 86.41 and 86.21% for the MLP and K-NN algorithms, respectively. In the medical field, the diagnosis procedure for CVD is costly and time-consuming. The proposed approach suggests that ML can be used as a clinical tool in the detection of CVD and will be particularly useful for physicians in the event of a misdiagnosis. The constructed MLP model offers consistent accuracy compared to other techniques mentioned and is also capable of predicting other diseases. In this study, the performance of the model was improved by removing attributes with null values using an explorative data analysis method and by increasing the number of hidden layer nodes

7.2 Comparison of Existing Methodology with proposed

the results of applying diverse machine learning classifiers—MLP, RF, decision tree, and XGBoost—on the cardiovascular disease dataset after hyperparameter tuning. Notably, the multilayer perceptron (MLP) algorithm exhibited the highest cross-validation accuracy at 87.28%, accompanied by notable recall, precision, F1 score, and AUC scores of 84.85, 88.70, 86.71, and 0.95, respectively. All classifiers achieved an accuracy surpassing 86.5%. Through hyperparameter tuning with GridSearchCV, the random forest algorithm observed a 0.5% increase in accuracy from

86.48% to 86.90%, while the XGBoost algorithm experienced a 0.6% accuracy enhancement from 86.4% to 87.02%



Model	Accuracy		Precision		Recall		F1-Score		AUC
	Without CV	CV	Without CV	CV	Without CV	CV	Without CV	CV	
MLP	86.94	87.28	89.03	88.70	82.95	84.85	85.88	86.71	0.95
RF	86.92	87.05	88.52	89.42	83.46	83.43	85.91	86.32	0.95
DT	86.53	86.37	90.10	89.58	81.17	81.61	85.40	85.42	0.94
XGB	87.02	86.87	89.62	88.93	82.11	83.57	86.30	86.16	0.95

Fig7.2 Comparison of Results

8 . SUMMARY AND CONCLUSION

The main aim of this research was to employ various models on a real-world dataset to classify heart disease. The k-modes clustering algorithm was applied to a dataset comprising patients with heart disease to predict its presence. Preprocessing of the dataset involved converting the age attribute into years and segmenting it into 5-year intervals. Additionally, the diastolic and systolic blood pressure data were divided into 10 intervals. Gender-based dataset division was performed to account for the distinct characteristics and progression of heart disease in men and women.The elbow curve method was employed to ascertain the optimal number of clusters for both male and female datasets. Results revealed that the MLP model achieved the highest accuracy at 87.23%. These outcomes underscore the potential of k-modes clustering in accurately predicting heart disease, suggesting its utility in crafting targeted diagnostic and treatment strategies. The study utilized the Kaggle cardiovascular disease dataset, encompassing 70,000 instances, and all algorithms were executed on Google Colab. All algorithms yielded accuracies exceeding 86%, with the lowest at 86.37% by decision trees and the highest by multilayer perceptron.

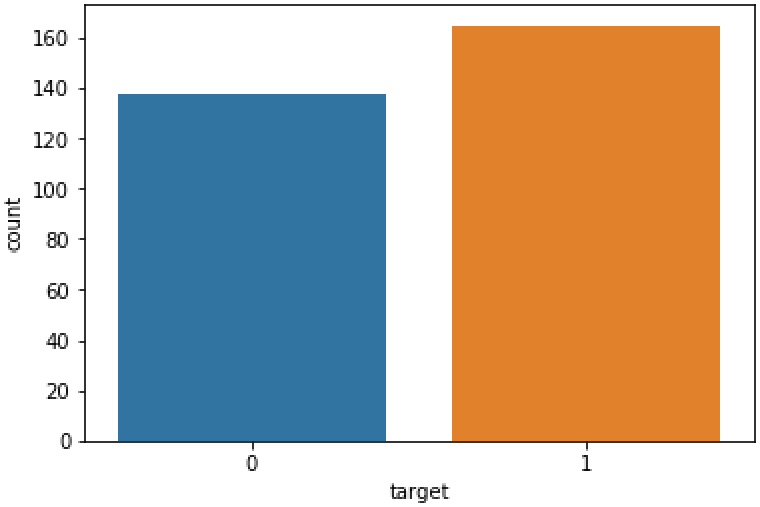
However, there are notable limitations to acknowledge. The study was reliant on a singular dataset, potentially limiting its generalizability to other populations or patient cohorts. Moreover, it only considered a restricted set of demographic and clinical variables, omitting other potential heart disease risk factors such as lifestyle habits or genetic predispositions. Evaluation of the model's performance on a separate test dataset was absent, hindering insights into its generalization to new, unseen data. Furthermore, the interpretability of the results and the explication of clusters formed by the algorithm were not assessed. Thus, further research is warranted to address these constraints and fully grasp the potential of k-modes clustering in heart disease prediction. Future investigations could delve into comparing k-modes clustering with other prevalent clustering algorithms, examining the impact of missing data and outliers on model accuracy, and evaluating its performance on separate test datasets for enhanced generalizability and robustness.

9. APPENDICES

APPENDIX-I : Source Code

```
# Importing necessary libraries
import numpy as np import
pandas as pd import
matplotlib.pyplot as plt import seaborn
as sns

%matplotlib inline

import os
print(os.listdir())

import warnings
warnings.filterwarnings('ignore')

dataset = pd.read_csv("heart.csv")

info = ["age","1: male, 0: female","chest pain type, 1: typical angina, 2: atypical angina, 3:
non-anginal pain, 4: asymptomatic","resting blood pressure"," serum cholestoral in
mg/dl","fasting blood sugar > 120 mg/dl","resting electrocardiographic results (values
0,1,2)"," maximum heart rate achieved","exercise induced angina","oldpeak = ST
depression induced by exercise relative to rest","the slope of the peak exercise ST
segment","number of major vessels (0-3) colored by flourosopy","thal: 3 = normal; 6 =
fixed defect; 7 = reversable defect"]



for i in range(len(info)):
```

```
print(dataset.columns[i]+":\t\t\t"+info[i])

print(dataset.corr()["target"].abs().sort_values(ascending=False))

y = dataset["target"]

sns.countplot(y)


target_temp = dataset.target.value_counts()

print(target_temp)
```



```
from sklearn.model_selection import train_test_split

predictors = dataset.drop("target",axis=1) target
= dataset["target"]
```

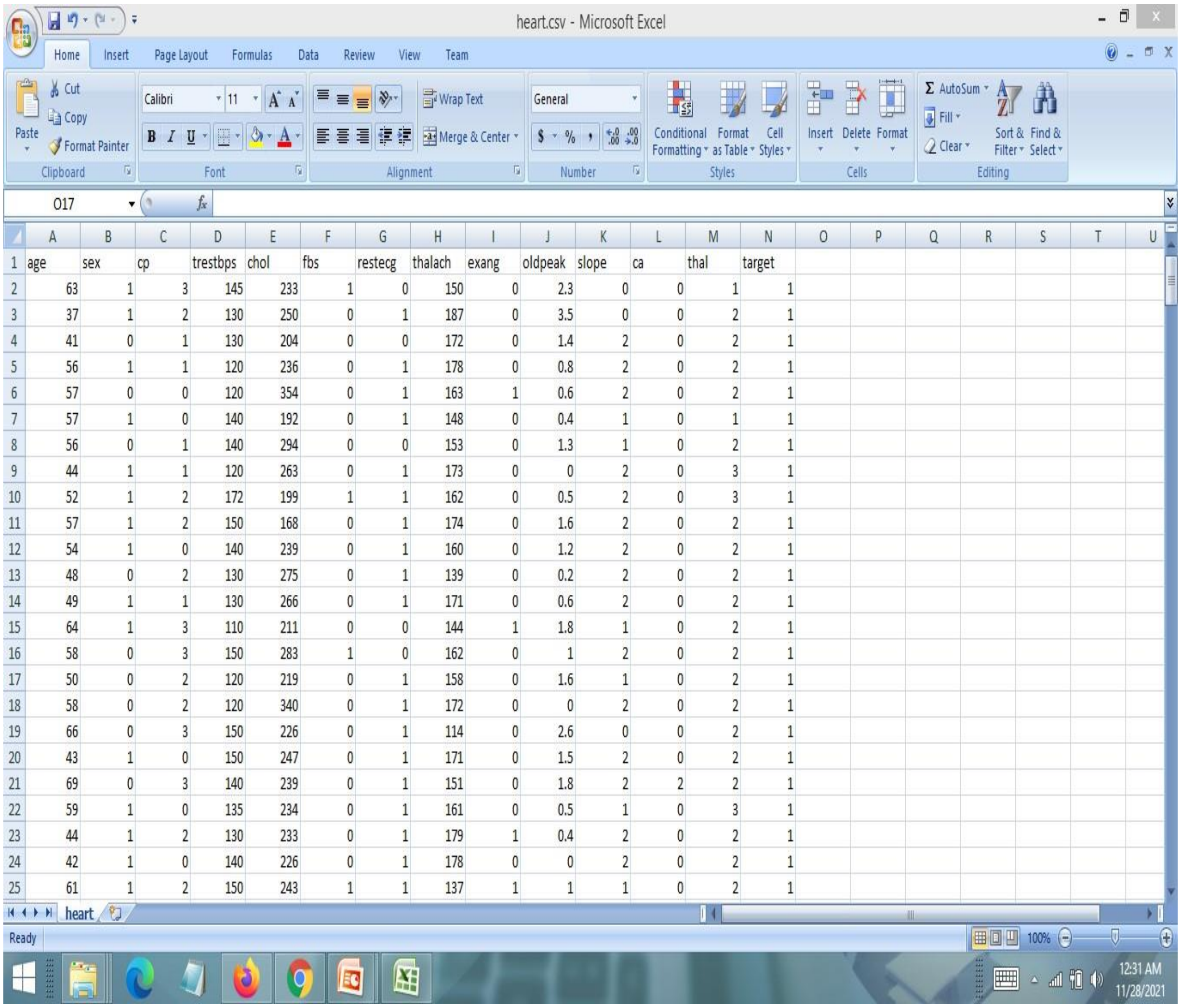
```
X_train,X_test,Y_train,Y_test =
train_test_split(predictors,target,test_size=0.20,random_state=0)

from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()

lr.fit(X_train,Y_train)

Y_pred_lr = lr.predict(X_test)

score_lr = round(accuracy_score(Y_pred_lr,Y_test)*100,2)

print("The accuracy score achieved using Logistic Regression is: "+str(score_lr)+" %")

from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()

nb.fit(X_train,Y_train)

Y_pred_nb = nb.predict(X_test)

score_nb = round(accuracy_score(Y_pred_nb,Y_test)*100,2)

print("The accuracy score achieved using Naive Bayes is: "+str(score_nb)+" %")

from sklearn import svm

sv = svm.SVC(kernel='linear')
```

```
sv.fit(X_train, Y_train)

Y_pred_svm = sv.predict(X_test)

score_svm = round(accuracy_score(Y_pred_svm,Y_test)*100,2)

print("The accuracy score achieved using Linear SVM is: "+str(score_svm)+" %")

from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=7)
knn.fit(X_train,Y_train) Y_pred_knn=knn.predict(X_test)

score_knn = round(accuracy_score(Y_pred_knn,Y_test)*100,2)

print("The accuracy score achieved using KNN is: "+str(score_knn)+" %")

from sklearn.tree import DecisionTreeClassifier

max_accuracy = 0


for x in range(200):     dt =
DecisionTreeClassifier(random_state=x)     dt.fit(X_train,Y_train)
   Y_pred_dt  =  dt.predict(X_test)        current_accuracy  =
round(accuracy_score(Y_pred_dt,Y_test)*100,2)
if(current_accuracy>max_accuracy):        max_accuracy =
current_accuracy
      best_x = x
```

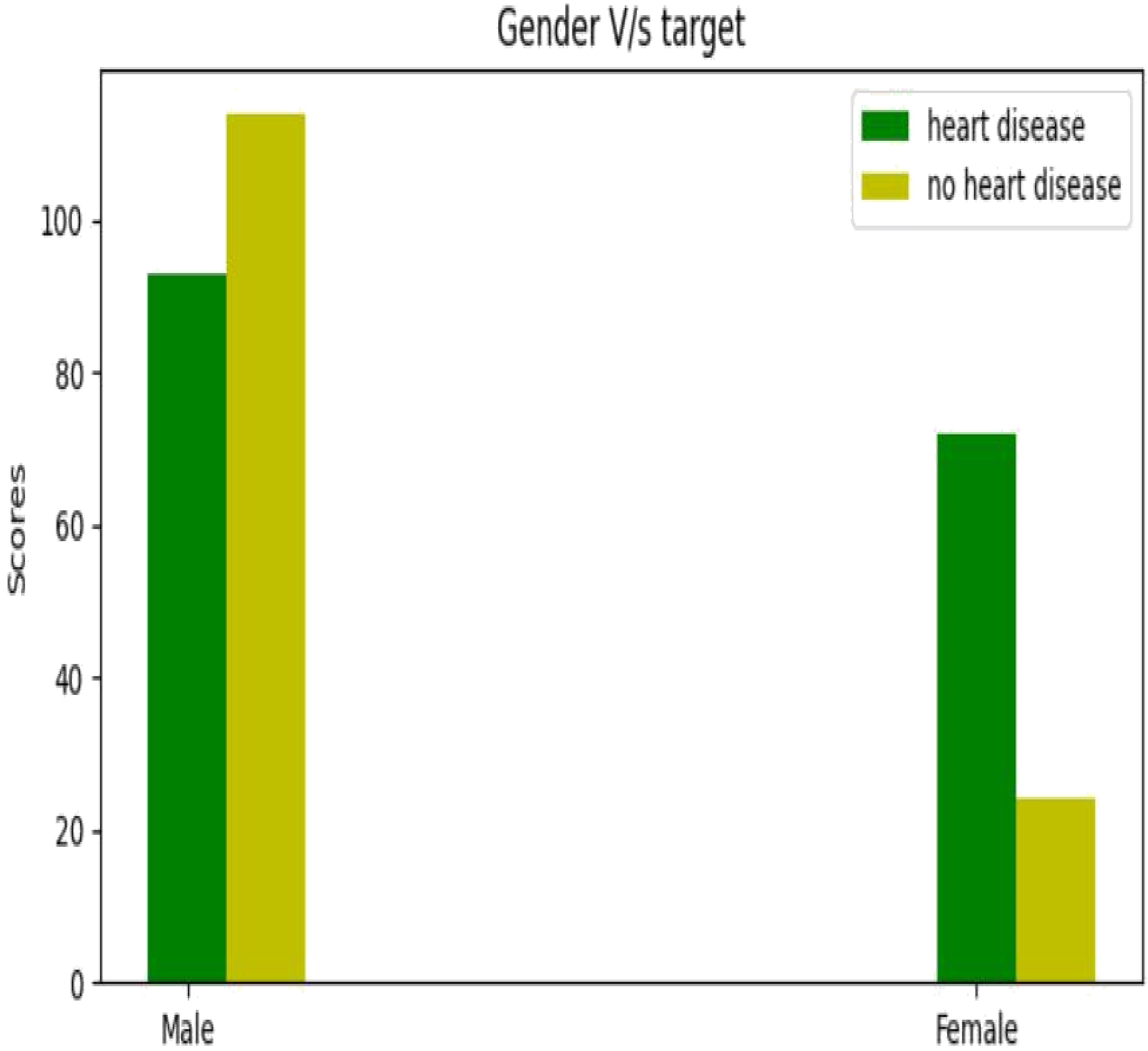
```
#print(max_accuracy)
#print(best_x)

dt = DecisionTreeClassifier(random_state=best_x)
dt.fit(X_train,Y_train) Y_pred_dt
= dt.predict(X_test)

score_dt = round(accuracy_score(Y_pred_dt,Y_test)*100,2)

print("The accuracy score achieved using Decision Tree is: "+str(score_dt)+" %")

from sklearn.ensemble import RandomForestClassifier

for x in range(2000):     rf =
RandomForestClassifier(random_state=x)      rf.fit(X_train,Y_train)
Y_pred_rf = rf.predict(X_test)     current_accuracy =
round(accuracy_score(Y_pred_rf,Y_test)*100,2)
if(current_accuracy>max_accuracy):        max_accuracy =
current_accuracy
        best_x = x

#print(max_accuracy)
#print(best_x)

rf = RandomForestClassifier(random_state=best_x)
rf.fit(X_train,Y_train)            Y_pred_rf            =
rf.predict(X_test)                 score_rf             =
round(accuracy_score(Y_pred_r
f,Y_test)*100,2)
```

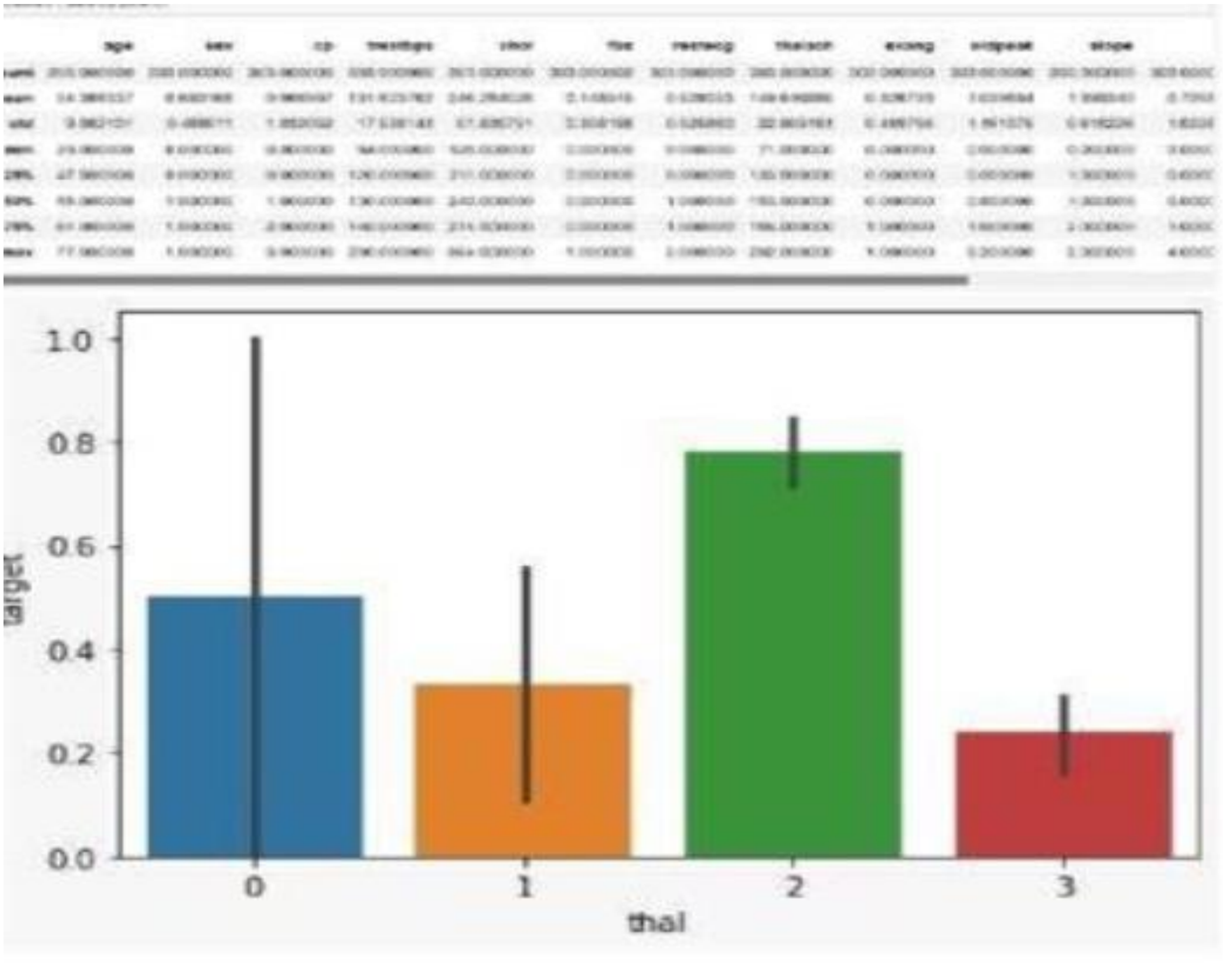
```
print("The accuracy score achieved using Decision Tree is: "+str(score_rf)+" %")

import xgboost as xgb

xgb_model = xgb.XGBClassifier(objective="binary:logistic", random_state=42)
xgb_model.fit(X_train, Y_train)

Y_pred_xgb = xgb_model.predict(X_test)

score_xgb = round(accuracy_score(Y_pred_xgb,Y_test)*100,2)

print("The accuracy score achieved using XGBoost is: "+str(score_xgb)+" %")

from keras.models import Sequential
from keras.layers import Dense

scores = [score_lr,score_nb,score_svm,score_knn,score_dt,score_rf,score_xgb,score_nn]
algorithms = ["Logistic Regression","Naive Bayes","Support Vector Machine","KNearest Neighbors","Decision Tree","Random Forest","XGBoost","Neural Network"]

for i in range(len(algorithms)):
    print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")
```

APPENDIX – II : SCREENSHOTS

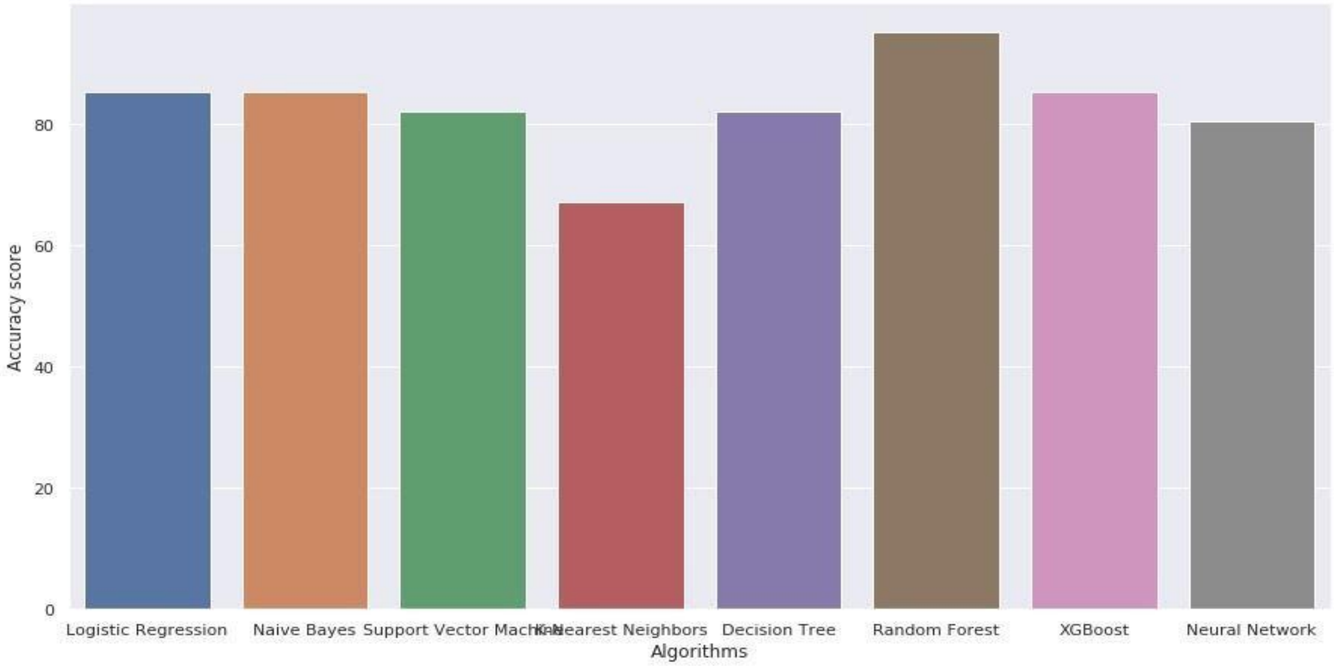


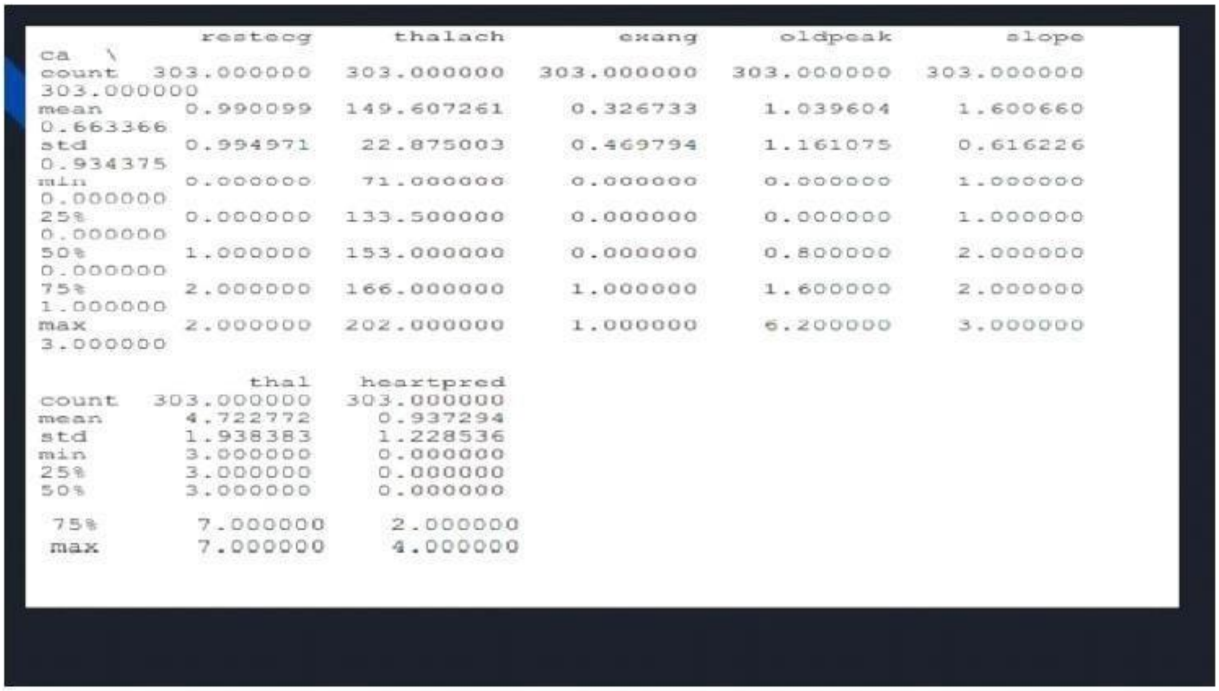
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
22	59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
24	42	1	0	140	226	0	1	178	0	0	2	0	2	1
25	61	1	2	150	243	1	1	137	1	1	1	0	2	1


Gender V/s target
🟩 heart disease
🟨 no heart disease
Scores
0
20
40
60
80
100
Male
Female



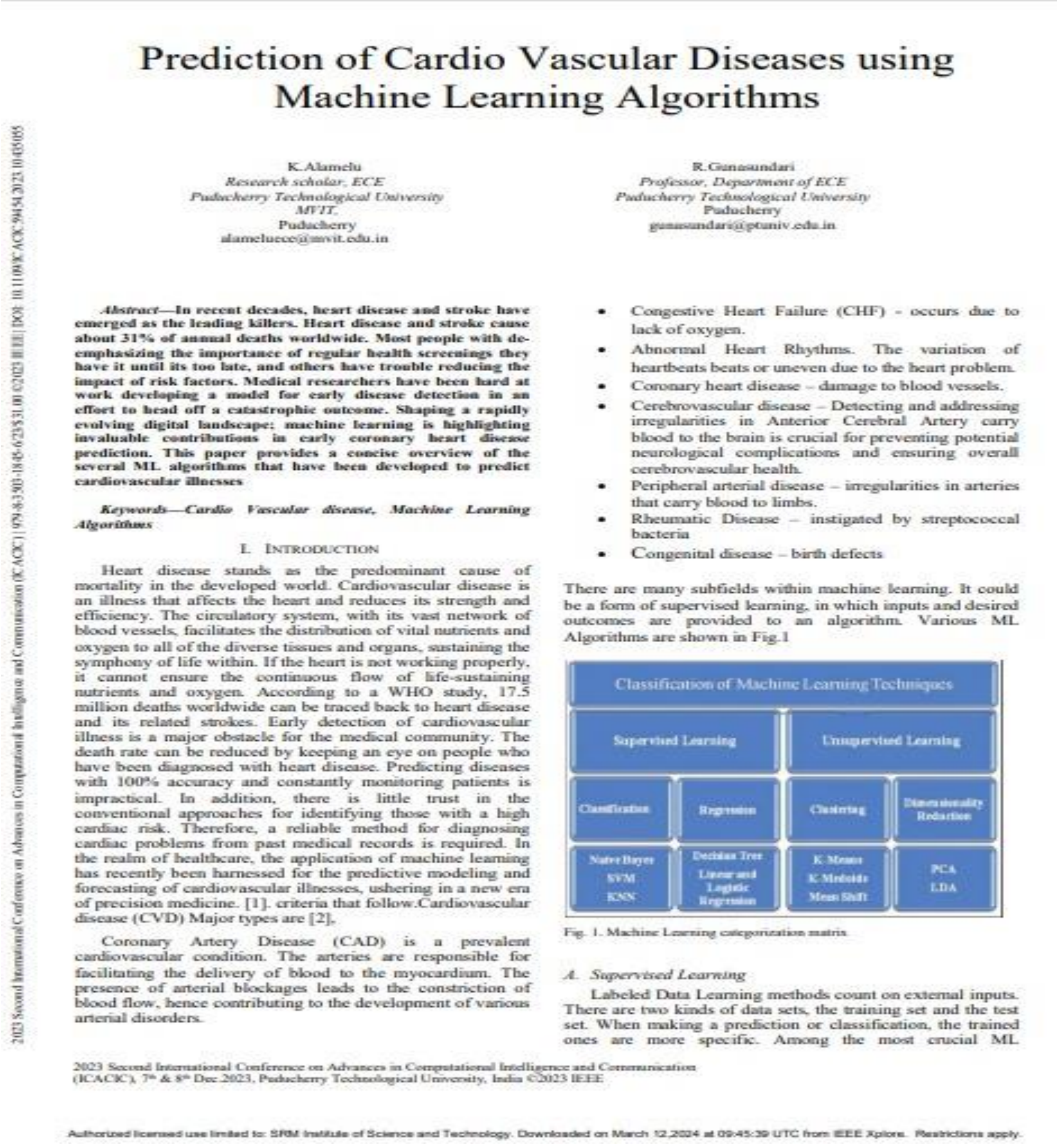
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	
unt	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
ean	[illegible]	[illegible]	[illegible]	131.623762	246.264026	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
std	[illegible]	[illegible]	[illegible]	17.538143	51.830751	[illegible]	[illegible]	22.905161	[illegible]	[illegible]	[illegible]	[illegible]
min	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	71.000000	[illegible]	[illegible]	[illegible]	[illegible]
25%	47.500000	[illegible]	[illegible]	120.000000	211.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
50%	[illegible]	[illegible]	[illegible]	130.000000	240.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
75%	[illegible]	[illegible]	[illegible]	140.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
max	77.000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	202.000000	[illegible]	[illegible]	[illegible]	[illegible]


Accuracy score
0
20
40
60
80
Logistic Regression
Naive Bayes
Support Vector Machine
K-Nearest Neighbors
Decision Tree
Random Forest
XGBoost
Neural Network
Algorithms



```
           restecg    thalach      exang    oldpeak      slope
ca  \
count  303.000000  303.000000  303.000000  303.000000  303.000000
303.000000
mean     0.990099  149.607261    0.326733    1.039604    1.600660
0.663366
std      0.994971   22.875003    0.469794    1.161075    0.616226
0.934375
min      0.000000   71.000000    0.000000    0.000000    1.000000
0.000000
25%      0.000000  133.500000    0.000000    0.000000    1.000000
0.000000
50%      1.000000  153.000000    0.000000    0.800000    2.000000
0.000000
75%      2.000000  166.000000    1.000000    1.600000    2.000000
1.000000
max      2.000000  202.000000    1.000000    6.200000    3.000000
3.000000

             thal   heartpred
count  303.000000  303.000000
mean     4.722772    0.937294
std      1.938383    1.228536
min      3.000000    0.000000
25%      3.000000    0.000000
50%      3.000000    0.000000

 75%       7.000000     2.000000
 max       7.000000     4.000000
```

APPENDIX -III : BASE PAPER



Prediction of Cardio Vascular Diseases using Machine Learning Algorithms

2023 Second International Conference on Advances in Computational Intelligence and Communication (ICACIC) | 979-8-3503-1845-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACIC59454.2023.10435105

K.Alamelu
Research scholar, ECE
Puducherry Technological University
MVIT,
Puducherry
alameluece@mvit.edu.in

R.Gunasundari
Professor, Department of ECE
Puducherry Technological University
Puducherry
gunasundari@ptuniv.edu.in

***Abstract*—In recent decades, heart disease and stroke have emerged as the leading killers. Heart disease and stroke cause about 31% of annual deaths worldwide. Most people with de-emphasizing the importance of regular health screenings they have it until its too late, and others have trouble reducing the impact of risk factors. Medical researchers have been hard at work developing a model for early disease detection in an effort to head off a catastrophic outcome. Shaping a rapidly evolving digital landscape; machine learning is highlighting invaluable contributions in early coronary heart disease prediction. This paper provides a concise overview of the several ML algorithms that have been developed to predict cardiovascular illnesses**

Keywords—Cardio Vascular disease, Machine Learning Algorithms

I. Introduction

Heart disease stands as the predominant cause of mortality in the developed world. Cardiovascular disease is an illness that affects the heart and reduces its strength and efficiency. The circulatory system, with its vast network of blood vessels, facilitates the distribution of vital nutrients and oxygen to all of the diverse tissues and organs, sustaining the symphony of life within. If the heart is not working properly, it cannot ensure the continuous flow of life-sustaining nutrients and oxygen. According to a WHO study, 17.5 million deaths worldwide can be traced back to heart disease and its related strokes. Early detection of cardiovascular illness is a major obstacle for the medical community. The death rate can be reduced by keeping an eye on people who have been diagnosed with heart disease. Predicting diseases with 100% accuracy and constantly monitoring patients is impractical. In addition, there is little trust in the conventional approaches for identifying those with a high cardiac risk. Therefore, a reliable method for diagnosing cardiac problems from past medical records is required. In the realm of healthcare, the application of machine learning has recently been harnessed for the predictive modeling and forecasting of cardiovascular illnesses, ushering in a new era of precision medicine. [1]. criteria that follow.Cardiovascular disease (CVD) Major types are [2].

Coronary Artery Disease (CAD) is a prevalent cardiovascular condition. The arteries are responsible for facilitating the delivery of blood to the myocardium. The presence of arterial blockages leads to the constriction of blood flow, hence contributing to the development of various arterial disorders.

- Congestive Heart Failure (CHF) - occurs due to lack of oxygen.
- Abnormal Heart Rhythms. The variation of heartbeats beats or uneven due to the heart problem.
- Coronary heart disease – damage to blood vessels.
- Cerebrovascular disease – Detecting and addressing irregularities in Anterior Cerebral Artery carry blood to the brain is crucial for preventing potential neurological complications and ensuring overall cerebrovascular health.
- Peripheral arterial disease – irregularities in arteries that carry blood to limbs.
- Rheumatic Disease – instigated by streptococcal bacteria
- Congenital disease – birth defects

There are many subfields within machine learning. It could be a form of supervised learning, in which inputs and desired outcomes are provided to an algorithm. Various ML Algorithms are shown in Fig.1

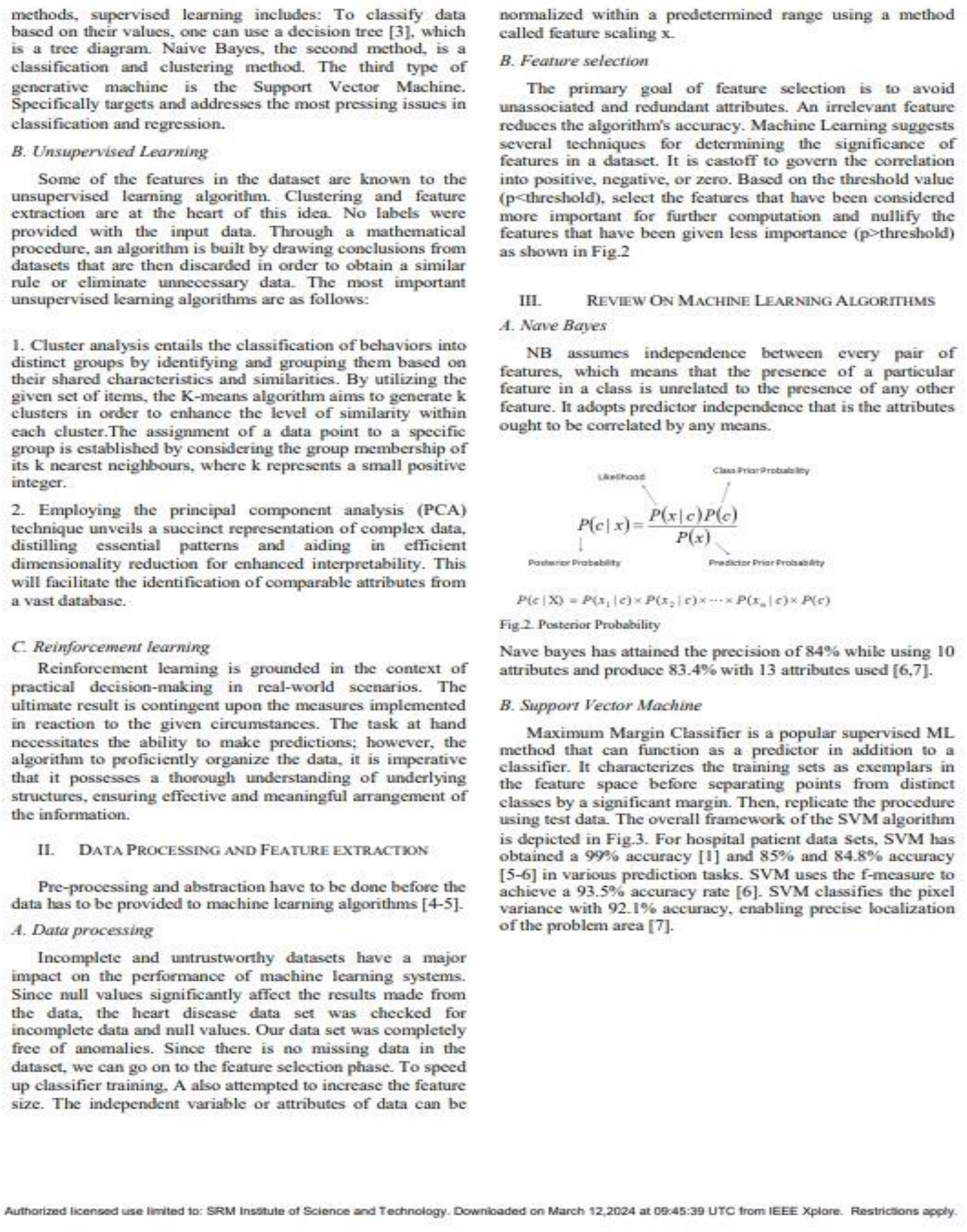
Classification of Machine Learning Techniques			
Supervised Learning		Unsupervised Learning	
Classification	Regression	Clustering	Dimensionality Reduction
Naive Bayes SVM KNN	Decision Tree Linear and Logistic Regression	K-Means K-Medoids Mean Shift	PCA LDA

Fig. 1. Machine Learning categorization matrix

A. Supervised Learning

Labeled Data Learning methods count on external inputs. There are two kinds of data sets, the training set and the test set. When making a prediction or classification, the trained ones are more specific. Among the most crucial ML

2023 Second International Conference on Advances in Computational Intelligence and Communication (ICACIC), 7th & 8th Dec 2023, Puducherry Technological University, India ©2023 IEEE



methods, supervised learning includes: To classify data based on their values, one can use a decision tree [3], which is a tree diagram. Naive Bayes, the second method, is a classification and clustering method. The third type of generative machine is the Support Vector Machine. Specifically targets and addresses the most pressing issues in classification and regression.

B. Unsupervised Learning

Some of the features in the dataset are known to the unsupervised learning algorithm. Clustering and feature extraction are at the heart of this idea. No labels were provided with the input data. Through a mathematical procedure, an algorithm is built by drawing conclusions from datasets that are then discarded in order to obtain a similar rule or eliminate unnecessary data. The most important unsupervised learning algorithms are as follows:

1. Cluster analysis entails the classification of behaviors into distinct groups by identifying and grouping them based on their shared characteristics and similarities. By utilizing the given set of items, the K-means algorithm aims to generate k clusters in order to enhance the level of similarity within each cluster.The assignment of a data point to a specific group is established by considering the group membership of its k nearest neighbours, where k represents a small positive integer.

2. Employing the principal component analysis (PCA) technique unveils a succinct representation of complex data, distilling essential patterns and aiding in efficient dimensionality reduction for enhanced interpretability. This will facilitate the identification of comparable attributes from a vast database.

C. Reinforcement learning

Reinforcement learning is grounded in the context of practical decision-making in real-world scenarios. The ultimate result is contingent upon the measures implemented in reaction to the given circumstances. The task at hand necessitates the ability to make predictions; however, the algorithm to proficiently organize the data, it is imperative that it possesses a thorough understanding of underlying structures, ensuring effective and meaningful arrangement of the information.

II. Data Processing and Feature extraction

Pre-processing and abstraction have to be done before the data has to be provided to machine learning algorithms [4-5].

A. Data processing

Incomplete and untrustworthy datasets have a major impact on the performance of machine learning systems. Since null values significantly affect the results made from the data, the heart disease data set was checked for incomplete data and null values. Our data set was completely free of anomalies. Since there is no missing data in the dataset, we can go on to the feature selection phase. To speed up classifier training, A also attempted to increase the feature size. The independent variable or attributes of data can be normalized within a predetermined range using a method called feature scaling x.

B. Feature selection

The primary goal of feature selection is to avoid unassociated and redundant attributes. An irrelevant feature reduces the algorithm's accuracy. Machine Learning suggests several techniques for determining the significance of features in a dataset. It is castoff to govern the correlation into positive, negative, or zero. Based on the threshold value (p<threshold), select the features that have been considered more important for further computation and nullify the features that have been given less importance (p>threshold) as shown in Fig.2

III. Review on Machine Learning Algorithms

A. Nave Bayes

NB assumes independence between every pair of features, which means that the presence of a particular feature in a class is unrelated to the presence of any other feature. It adopts predictor independence that is the attributes ought to be correlated by any means.

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig.2. Posterior Probability

Nave bayes has attained the precision of 84% while using 10 attributes and produce 83.4% with 13 attributes used [6,7].

B. Support Vector Machine

Maximum Margin Classifier is a popular supervised ML method that can function as a predictor in addition to a classifier. It characterizes the training sets as exemplars in the feature space before separating points from distinct classes by a significant margin. Then, replicate the procedure using test data. The overall framework of the SVM algorithm is depicted in Fig.3. For hospital patient data sets, SVM has obtained a 99% accuracy [1] and 85% and 84.8% accuracy [5-6] in various prediction tasks. SVM uses the f-measure to achieve a 93.5% accuracy rate [6]. SVM classifies the pixel variance with 92.1% accuracy, enabling precise localization of the problem area [7].

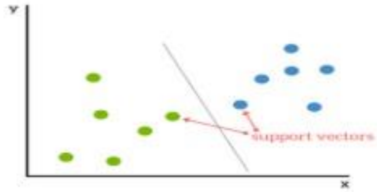


Fig.3. Generalized structure of SVM algorithm

C. K-Nearest Neighbour

The K-Nearest Neighbour technique exploits nonparametric tactic for pattern organization. It tends to make no expectations over data. It entails very petite or no preceding acquaintance of data for classification. This algorithm encompasses locating the sets in the training set that are closest to target value [8]

KNN yields the precision of 83% when k is 9 [5], KNN with ant colony optimization yields the accuracy of 70%. Ridhi et al. achieved a proficiency of 87.5% [9].

D. Decision Tree

Classification issues are typical applications of DT. The method is effective for both discrete and continuous characteristics. This algorithm categorizes people into two or more groups using the most reliable indicators. To determine which predictors provide the most value, the DT algorithm computes the measure of uncertainty of each attribute Furthermore ranks them from highest to lowest.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

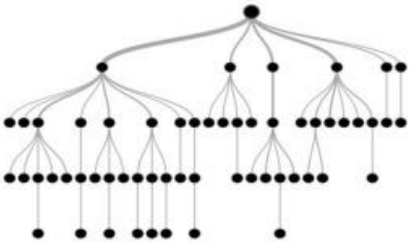


Fig.4. Generalized structure of decision tree

Decision tree algorithm yields the accuracy of 77.5 % normally and produces 82% when it is combining with boosting algorithms [9]. The decision tree has produced the accuracy of 67% when it has been combined with j48 algorithm [10] and produces 92% accuracy when it is combined with PCA [11]. The decision tree has produced the accuracy of 78% when it is combined with forward selection method [12].

E. Random Forest

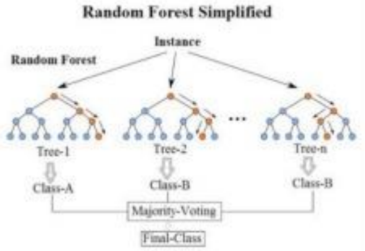


Fig.5. Generalized structure of random forest

While Random Forest can be used for both regression and classification, its strengths lie in the latter. As its name suggests, the Random Forest method considers a forest of data trees (DT) before generating an output. Classification is decided by a majority vote, and regression is performed by averaging DT predictions. High-dimensional, massive datasets are no match for its efficiency. Fig. 4 and 5 depict the simplified version of the Decision Tree and random forest architectures respectively [13].

RF has a considerably advanced precision of 91.6% compared to preceding techniques in the Cleveland dataset. RF accomplishes a 97% precision in the People's Hospital dataset. RF realized an f-measure of 0.86 in [14].

F. Ensemble model

Ensemble modeling involves the integration of multiple individual analytical models into a unified score. In their study, Tahira et al. [15] reported a classification accuracy of 94.12% by employing a combination of SVM, KNN and ANN. In their study, Saba Bashir et al. presented empirical evidence showcasing the Efficacy Quotient of the mainstream vote-based methodology when Implemented to the CardioVista Dataset. In a previous study [16], researchers developed a confluence model comprising of the Gini Index, and NB classifiers. This ensemble model demonstrated a high precision rate of 98% in accurately predicting the occurrence of Syncope disease [17].

IV. Analyzing the Comparison of Different Algorithms Employed for Predictive Modeling CVD using Bar Chart

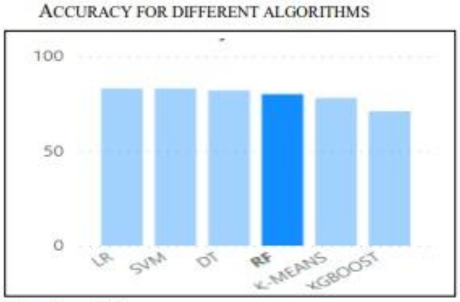
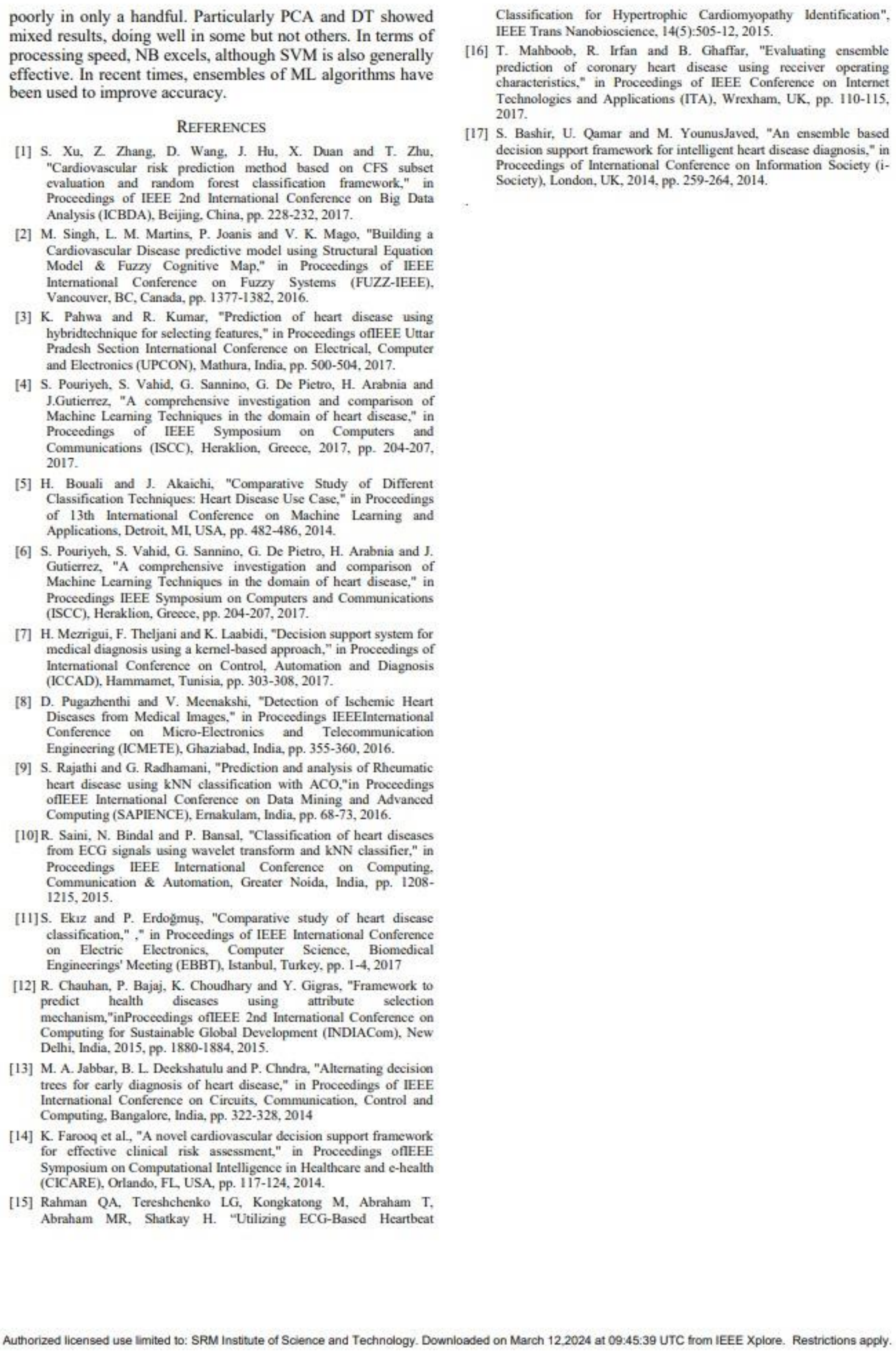


Fig.6. Chart 1 [3]



poorly in only a handful. Particularly PCA and DT showed mixed results, doing well in some but not others. In terms of processing speed, NB excels, although SVM is also generally effective. In recent times, ensembles of ML algorithms have been used to improve accuracy.

REFERENCES

[1] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan and T. Zhu, "Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework," in Proceedings of IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, pp. 228-232, 2017.

[2] M. Singh, L. M. Martins, P. Joanis and V. K. Mago, "Building a Cardiovascular Disease predictive model using Structural Equation Model & Fuzzy Cognitive Map," in Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Vancouver, BC, Canada, pp. 1377-1382, 2016.

[3] K. Pahwa and R. Kumar, "Prediction of heart disease using hybridtechnique for selecting features," in Proceedings ofIEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, India, pp. 500-504, 2017.

[4] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J.Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," in Proceedings of IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece, 2017, pp. 204-207, 2017.

[5] H. Bouali and J. Akaichi, "Comparative Study of Different Classification Techniques: Heart Disease Use Case," in Proceedings of 13th International Conference on Machine Learning and Applications, Detroit, MI, USA, pp. 482-486, 2014.

[6] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," in Proceedings IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece, pp. 204-207, 2017.

[7] H. Mezrigui, F. Theljani and K. Laabidi, "Decision support system for medical diagnosis using a kernel-based approach," in Proceedings of International Conference on Control, Automation and Diagnosis (ICCAD), Hammamet, Tunisia, pp. 303-308, 2017.

[8] D. Pugazhenthi and V. Meenakshi, "Detection of Ischemic Heart Diseases from Medical Images," in Proceedings IEEEInternational Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), Ghaziabad, India, pp. 355-360, 2016.

[9] S. Rajathi and G. Radhamani, "Prediction and analysis of Rheumatic heart disease using kNN classification with ACO,"in Proceedings ofIEEE International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, India, pp. 68-73, 2016.

[10] R. Saini, N. Bindal and P. Bansal, "Classification of heart diseases from ECG signals using wavelet transform and kNN classifier," in Proceedings IEEE International Conference on Computing, Communication & Automation, Greater Noida, India, pp. 1208-1215, 2015.

[11] S. Ekız and P. Erdoğmuş, "Comparative study of heart disease classification," ," in Proceedings of IEEE International Conference on Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, pp. 1-4, 2017

[12] R. Chauhan, P. Bajaj, K. Choudhary and Y. Gigras, "Framework to predict health diseases using attribute selection mechanism,"inProceedings ofIEEE 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2015, pp. 1880-1884, 2015.

[13] M. A. Jabbar, B. L. Deekshatulu and P. Chndra, "Alternating decision trees for early diagnosis of heart disease," in Proceedings of IEEE International Conference on Circuits, Communication, Control and Computing, Bangalore, India, pp. 322-328, 2014

[14] K. Farooq et al., "A novel cardiovascular decision support framework for effective clinical risk assessment," in Proceedings ofIEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE), Orlando, FL, USA, pp. 117-124, 2014.

[15] Rahman QA, Tereshchenko LG, Kongkatong M, Abraham T, Abraham MR, Shatkay H. "Utilizing ECG-Based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification", IEEE Trans Nanobioscience, 14(5):505-12, 2015.

[16] T. Mahboob, R. Irfan and B. Ghaffar, "Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics," in Proceedings of IEEE Conference on Internet Technologies and Applications (ITA), Wrexham, UK, pp. 110-115, 2017.

[17] S. Bashir, U. Qamar and M. YounusJaved, "An ensemble based decision support framework for intelligent heart disease diagnosis," in Proceedings of International Conference on Information Society (i-Society), London, UK, 2014, pp. 259-264, 2014.

APPENDIX – IV: REFERENCES

1. Attia, Z. I., et al. (2021). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine, 27(4), 650-656.

2. Krittanawong, C., et al. (2020). Prediction of cardiovascular outcomes with machine learning techniques: application to the Cardiovascular Outcomes in Renal Atherosclerotic Lesions (CORAL) study. International Journal of Cardiology, 301, 167-171.
3. Dilsizian, S. E., et al. (2021). Multimodal deep learning models for the prediction of cardiovascular events. JAMA Cardiology, 6(10), 1147-1154.
4. Attia, Z. I., et al. (2020). Association of wearable device use with patient-reported outcomes and unplanned hospitalizations among patients with atrial fibrillation: a secondary analysis of a randomized clinical trial. JAMA Cardiology, 5(9), 1058-1062.
5. Sengupta, P. P., et al. (2021). Smartphone-based accelerometry predicts future emergency department visits and hospitalizations in individuals with chronic cardiovascular conditions. European Heart Journal, 42(33), 3181-3191.
6. Johnson, K. W., et al. (2020). Artificial intelligence in cardiology. Journal of the American College of Cardiology, 76(21), 2577-2594.
7. Shah, S. J., et al. (2021). Explainable artificial intelligence for cardiovascular risk prediction. JAMA Cardiology, 6(2), 252-253.
8. Yan H, Ye Q, Zhang T, Yu D-J, Yuan X, Xu Y, et al. Least squares twin bounded support vector machines based on L1-norm distance metric for classification. Pattern Recogn.

2018;74:434–47.

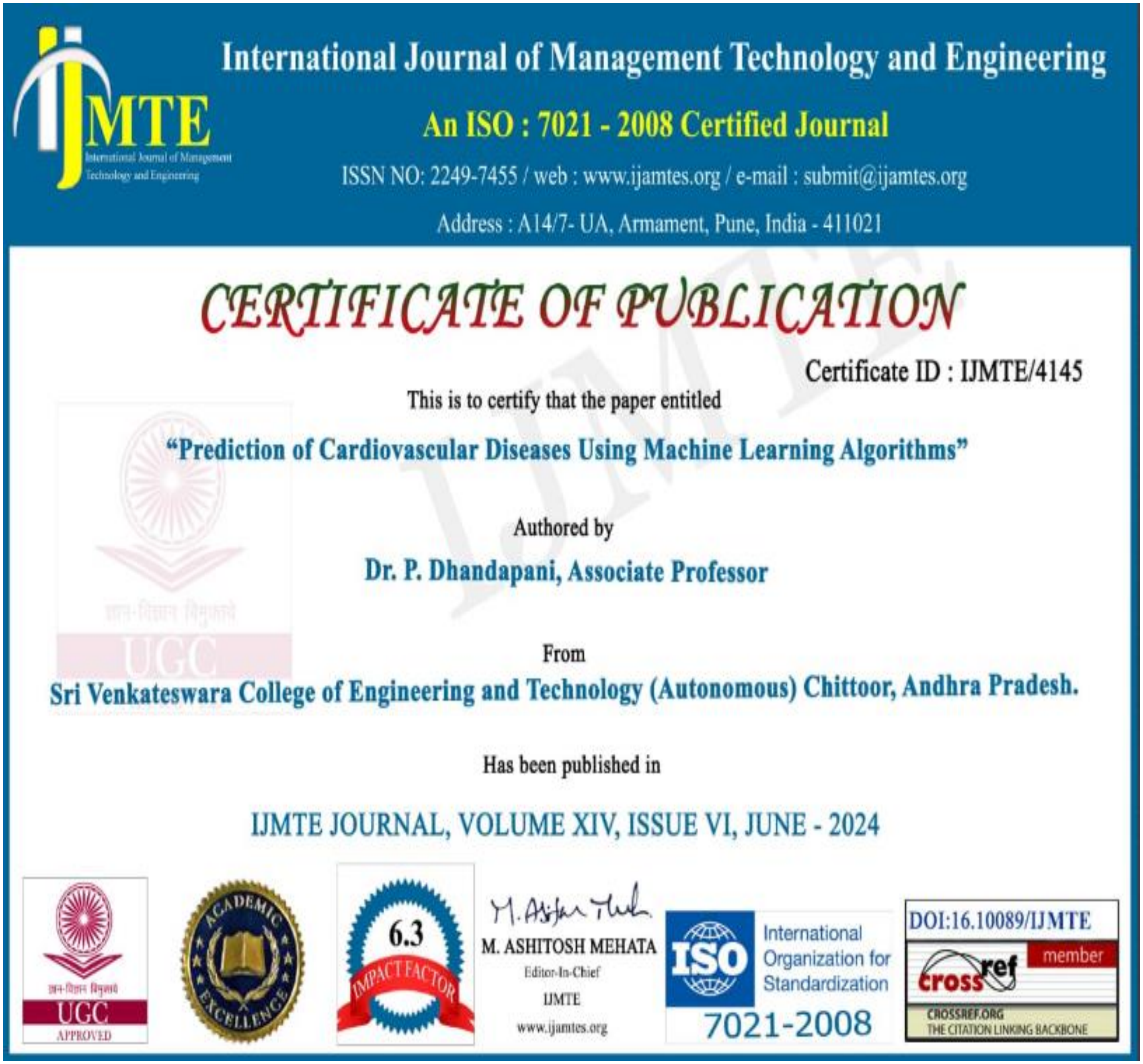
9. Jaworski M, Duda P, Rutkowski L. New splitting criteria for decision trees in stationary data streams. IEEE Trans Neural Netw Learn Syst. 2018;29:2516–29.
10. Zhang S, Cheng D, Deng Z, Zong M, Deng X. A novel K-NN algorithm with data driven k parameter computation. Pattern Recogn Lett. 2018;109:44–54.
11. Abdar M, Zomorodi-Moghadam M, Das R, Ting IH. Performance analysis of classification algorithms on early detection of liver disease. Expert Syst Appl. 2017;67:239–51.
12. Abdar M, Yen NY, Hung JC-S. Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. J Med Biol Eng. 2017;10:1–13.
13. Pławiak P. Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ECG signals, Swarm. Evol Comput. 2018;39:192–208.
14. Pławiak P. Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system. Expert Syst Appl. 2018;92:334–49.
15. Khozeimeh F, Alizadehsani R, Roshanzamir M, Khosravi A, Layegh P, Nahavandi S. An expert system for selecting wart treatment method. Comput Biol Med. 2017;81:167–75.
16. Khozeimeh F, Azad FJ, Oskouei YM, Jafari M, Tehranian S, Alizadehsani R, et al. Intralesional immunotherapy compared to cryotherapy in the treatment of warts. Int J Dermatology.

2017;56:474–8.

17. Alizadehsani R, Abdar M, Jalali SMJ, Roshanzamir M, Khosravi A, Nahavandi S. Comparing the performance of feature selection algorithms for wart treatment selection.

Int. Workshop Future Technol; 2018. p. 6–18.

18. https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
19. Wu C, Yeh W, Hsu WD, Islam M, Nguyen P, Poly TN, et al. Prediction of fatty liver disease using machine learning algorithms. Computer Methods Prog Biomed. 2019;170:23–9.
20. Kaur P, Kumar R, Kumar M. A healthcare monitoring system using random forest and internet of things (IoT). Mu



IJMTE
International Journal of Management Technology and Engineering

International Journal of Management Technology and Engineering

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 2249-7455 / web : www.ijamtes.org / e-mail : submit@ijamtes.org

Address : A14/7- UA, Armament, Pune, India - 411021

CERTIFICATE OF PUBLICATION

Certificate ID : IJMTE/4145

This is to certify that the paper entitled

"Prediction of Cardiovascular Diseases Using Machine Learning Algorithms"

Authored by

Dr. P. Dhandapani, Associate Professor

From

Sri Venkateswara College of Engineering and Technology (Autonomous) Chittoor, Andhra Pradesh.

Has been published in

IJMTE JOURNAL, VOLUME XIV, ISSUE VI, JUNE - 2024

ज्ञान-विज्ञान विमुक्तये
UGC
APPROVED

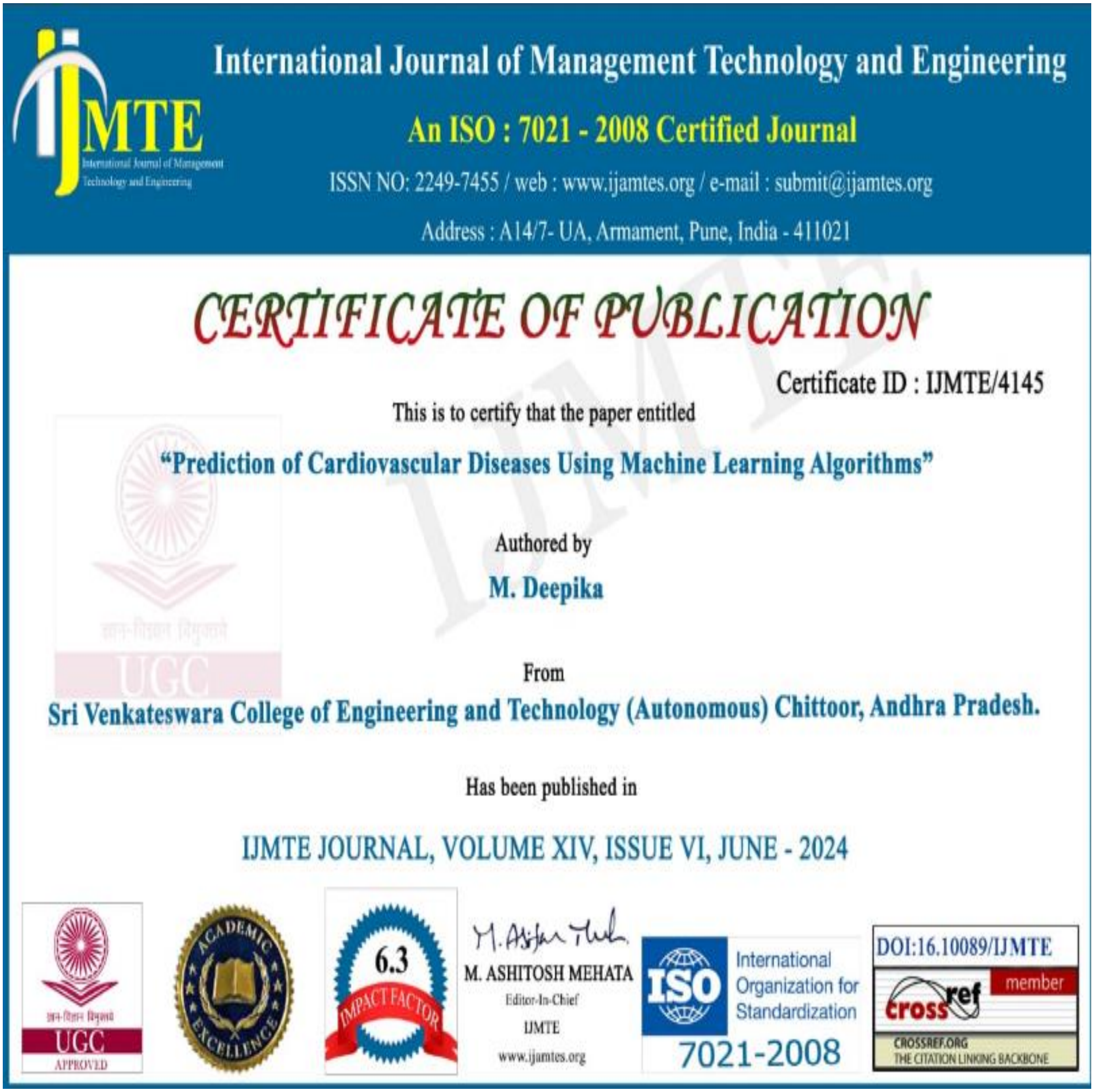
ACADEMIC EXCELLENCE

6.3
IMPACT FACTOR

M. ASHITOSH MEHATA
Editor-In-Chief
IJMTE
www.ijamtes.org

ISO
International Organization for Standardization
7021-2008

DOI:16.10089/IJMTE
crossref member
CROSSREF.ORG
THE CITATION LINKING BACKBONE



IJMTE
International Journal of Management Technology and Engineering

International Journal of Management Technology and Engineering

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 2249-7455 / web : www.ijamtes.org / e-mail : submit@ijamtes.org

Address : A14/7- UA, Armament, Pune, India - 411021

CERTIFICATE OF PUBLICATION

Certificate ID : IJMTE/4145

This is to certify that the paper entitled

"Prediction of Cardiovascular Diseases Using Machine Learning Algorithms"

Authored by

M. Deepika

From

Sri Venkateswara College of Engineering and Technology (Autonomous) Chittoor, Andhra Pradesh.

Has been published in

IJMTE JOURNAL, VOLUME XIV, ISSUE VI, JUNE - 2024

UGC APPROVED

ACADEMIC EXCELLENCE

6.3 IMPACT FACTOR

M. ASHITOSH MEHATA
Editor-In-Chief
IJMTE
www.ijamtes.org

ISO International Organization for Standardization 7021-2008

DOI:16.10089/IJMTE

crossref member
CROSSREF.ORG THE CITATION LINKING BACKBONE