Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Machine Learning-Based Method of Cardiovascular Disease Classification based on Objective, Examination, and Subjective Features

Parul Yadav
Scholor
VGI Dadri, Greater Noida

Rahul Kumar
Scholor 2 ,ABESEC,
Ghaziabad

Dilip Kumar
Assistant professor
VGI, DAdri ,Greater Noida

Vikky
Scholar 3, ABESEC,
Ghaziabad

***Abstract* - cardiovascular disease (CVD) is among the leading causes of worldwide death, which is why the essential role of early diagnosis and accurate risk estimation in modern healthcare is difficult to overestimate. The significant digitization of patient health records has enabled the implementation of machine-learning techniques to supplement clinical decision-making with data-based and evidence-based insights. The current study presents a machine-learning-based CVD classification framework that combines the objective demographic features, clinical examination findings and non-objective lifestyle-related factors. The data used in this study includes those of the patients taken at the regular medical check-ups, and they include age, gender, systolic and diastolic blood pressure, lipids profile, glucose level, body mass index, and lifestyle habits of the patients like use of tobacco, alcohol, and exercise. A comprehensive exploratory data analysis was conducted to clarify the distribution of features, inter-variable relationships and possible potential risk factors relevant to CVD. Various machine-learning models such as the Logistic Regression, the K-Nearest Neighbors, the Support Vector Machines, the Decision tree, the random forest, the Gradient Boosting and other boosting-based ensemble models were applied and evaluated systematically. When it came to measuring model performance, accuracy, precision, recall, F1-score, confusion matrices, and area under the receiver operating characteristic curve (ROC-AUC) were used. The experimental findings indicate that the ensemble-based classifiers are better than single traditional model as they have high predictive accuracy and robustness. These results confirm the timely impact of clinical and lifestyle variables on the prediction of CVD risk and justify the effectiveness of the machine-learning approaches as auxiliary tools in the early diagnosis of the disease and its stratification of risk.**

Keywords* - *Cardiovascular Disease, Machine Learning, Health Data Analytics, Classification, Risk Prediction, Clinical Decision Support

• INTRODUCTION

Cardiovascular disease (CVD) is one of the most significant challenges in the contemporary public health as it spans a significant part of yearly mortality and morbidity rates, globally. The world health statistics have shown that millions of deaths globally can be caused by heart failure, stroke, and coronary artery disease. CVD burden is not just limited to mortality but also to long-term disability, poor quality of life, and significant healthcare infrastructural costs. The timely identification of these conditions and proper assessment of risks is, therefore, the inseparable conditions to facilitate preventive measures and reduce the negative clinical effects.

Conventional methods of CVD diagnosis and risk stratification are based mainly on clinical examination, laboratory and physician experience. Although these traditional modalities have established clinical utility, they are usually time consuming, resource intensive and prone to human bias as well as inter-observer variation. Further, conventional statistical methods might not be able to demonstrate the complex and nonlinear interaction between various risk factors including blood pressure, lipid levels, glycemic status and lifestyle behaviors. The current shift toward the use of electronic health records and the digital acquisition of data is an additional reason that contributes to the necessity of scalable and automated tools that can effectively analyses large volumes of patient data and assist in evidence-based clinical decision-making.

Over the past few years, machine-learning (ML) methods have become powerful analytical procedures in biomedical studies, allowing the identification of latent patterns in past data, the determination of nuanced relationships between variables, and the development of predictive frameworks that are more accurate compared to traditional methods. In cardiovascular disease, ML has enabled prediction of risk, classification of diseases, stratification of patients and prediction of outcomes. These technologies promise to speed up the initial diagnosis, tailor the treatment plan, and minimise the work of clinicians.

A combination of objective demographic indicators, objective clinical examination results, and subjective lifestyle influences cardiovascular disease. Objective measures like age, sex, height and weight provide base level data, and measures of examination like systolic and diastolic blood pressure, lipid levels and glycaemic parameters are direct indicators of physiological health condition. Subjectively speaking, subjective variables, which include smoking behavior, alcohol consumption and physical activity, have a significant effect on long-term cardiovascular risk. Strong predictive models therefore require all the three classes of features to be combined in order to have a complete picture of patient health.

Although the use of ML in clinical studies continues to increase, a number of challenges exist. Most of the studies that exist focus on

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

a small sub-set of clinical variables, or use a single machine-learning algorithm without extensive comparative studies. Others do not provide detailed exploratory data analysis, which is a weakness because it does not aid in grasping the features distributions, correlations and data integrity. Also, the problems related to the model interpretability and generalizability are also topical, especially in the clinical setting where transparency and trust are of primary importance.

The study fills these gaps by presenting a machine-learning framework of CVD classification, which combines objective, examination, and subjective features that are gathered during clinical examinations. The paper performs comprehensive exploratory research on data analysis to determine the major patterns and risk factors related to cardiovascular disease. A variety of machine-learned models, including linear, non-linear and ensemble-based approaches are deployed and strictly tested with the help of conventional performance indicators. Through the comparison of contrasting classifiers, the paper aims at identifying models that can not only provide strong predictive capability but also demonstrate viable usage.

The primary findings of this study can be listed in the following way:

- The detailed exploratory examination of cardiovascular health data that explains risk determinant patterns;
- The architecture, creation, and evaluation of a collection of machine-learning models to be used in CVD classification;
- A comparative analysis of the model performance based on the following metrics: accuracy, confusion matrices, and ROC-based ones;
- Clues to the contributory nature of clinical and lifestyle factors in prognosticating cardiovascular disease.

The rest of this manuscript has been structured in the following manner. Section 2 conducts a literature review on the relevant studies on CVD prediction using machine-learning methods. Section 3 outlines the suggested methodology, including the data preprocessing steps, the model development, and training procedures. Section 4 has the results of the experiment, which are interpreted and shed light on the specifics of model performance. Lastly, Section 5 wraps up the research and suggests research future directions.

• RELATED WORK

Recent years have seen a significant academic interest in the use of machine learning approaches to predict cardiovascular disease (CVD), which could be mostly explained by the growth of healthcare information and the limitations of traditional diagnostic models. The early studies in this field mostly used statistical analysis with the most significant having logistic regression to estimate cardiovascular risk based on some demographic and clinical covariates. The fact that such models were interpretable, as well as procedurally simple, was offset by their highly linear architecture, which restricted them in terms of their ability to model complex interactions that exist between various risk determinants [1].

As computational power steadily improved, researchers started to use more and more advanced machine learning algorithms in the classification of cardiovascular disease. The first non-linear methods which were used on the cardiac morbidity datasets were decision-tree-based models. Such models had better flexibility to handle the interactions of a feature and also the missing data without compromising interpretability [2]. However, single decision trees were noisy and overfitted thus limiting their generalization.

In order to overcome these constraints ensemble learning methods like Random Forest and Gradient Boosting were developed. The random forest classifiers combine thousands of decision trees to increase predictive accuracy and strength. As empirical evidence has shown, it is true that the Random Forest models are better than the conventional classifiers in predicting CVD, especially under the conditions of heterogeneous clinical and lifestyle predictors [3]. Similarly to this, the Gradient Boosting techniques, which progressively improve weak learners, have proven to be very competent at detecting high-risk cardiovascular patients [4].

Support Vector Machines (SVMs) have also been greatly utilized in CVD classification activities as they are effective in high dimensional feature space. Research findings indicate that a SVM with non-linear kernel is able to achieve competitive accuracy compared to tree-based models, particularly when feature scaling and optimization of the kernel is carefully implemented [5]. However, SVMs can often require tedious hyperparameter optimization and can experience reduced interpretability, which makes them difficult to use in clinical practice.

K-Nearest Neighbors (KNN) algorithms have been studied in consonance in predicting cardiovascular diseases due to their intuitive functionality and conceptual simplicity. Although KNN models may be used to encode local data trends, their effectiveness is very much dependent on the choice of distance measures as well as the value of k. Besides, KNN is computationally expensive over large data sets, hindering scalability in practice in healthcare systems [6].

In recent research, there have been growing studies on the enhancement of boosting-based methods like XGBoost, LightGBM, and CatBoost that have proven more effective in a variety of medical classification problems. These models have the good ability to deal with unbalanced data and sophisticated interactions of features. Studies have shown that boosting-based classifiers have better accuracy and Area Under the Curve (AUC) scores in predicting CVDs as compared to traditional machine learning models [7]. Moreover, CatBoost has also drawn the interest of the fact that it can be very effective in handling categorical variables with little preprocessing, making it appropriate to use with healthcare data with categorical clinical characteristics [8].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

In addition to the model selection, several research papers have emphasized the importance of thorough exploratory data analysis (EDA) and feature engineering. The researchers have also shown that derived variables like Body Mass Index (BMI), years of age, and ratios of blood pressure significantly improve the model performance [9]. Correlation studies and feature significance tests have also found systolic blood pressure, cholesterol, glucose concentration, and age to be the leading predictors of cardiovascular disease.

Although these developments have occurred, there are few constraints that can be identified in existing literature. Most studies are based on a small range of clinical variables and exclude the subjective factors of lifestyle, including smoking, alcohol use, and physical exercise, which are very proven risks of cardiovascular. Moreover, certain studies do not focus on the optimal predictive accuracy but do not perform detailed error analysis and clinical interpretability assessment [10]. The lack of standardized method of evaluation and reporting of performance measures also complicates comparative analysis between studies.

In addition, more than deep learning methods have started to be used to predict cardiovascular risk, their implementation can require large datasets and expensive computational resources that are not always available in many healthcare facilities. Deep models also raise issues of explainability and transparency that are key in clinician trust and compliance to regulations.

Unlike the existing literature, the current study is based on a holistic approach that combines objective demographic characteristics, clinical examination measurements, and the subjective indicators of the lifestyle in a single machine learning structure. This research aims to provide a level-headed assessment of predictive performance as well as practical application by conducting a large-scale exploratory data analysis, an organized preprocessing of the data, and a comparative analysis of various machine learning classifiers. The usage of ensemble-based models and the traditional classifiers help to make a comprehensive comparison and to shed some light on the weaknesses and limitations of difference machine learning methods to the classification of heart diseases.

• PROPOSED METHODOLOGY

The current paper suggests a systematic trained machine-learning platform to organize the classification of cardiovascular disease, relying on the objective demographic characteristics, clinical examination data, and subjective lifestyle indicators. The methodology is designed in such a way that it provides integrity of data, strong model learning and sound performance evaluation. The general workflow includes data acquisition, data exploratory analysis, data cleaning and preprocessing, dataset partitioning, machine learning model training, and data performance evaluation.
The framework presented here puts a strong emphasis on transparency and reproducibility, where every step of the procedure has to have a purpose in relation to the overall predictive performance, yet is still needed clinically.

- ***Methodology Flowchart***

The generic methodology followed in the given research is presented in Figure 1. It begins with the gathering of data and passes through a series of data analysis and modeling.

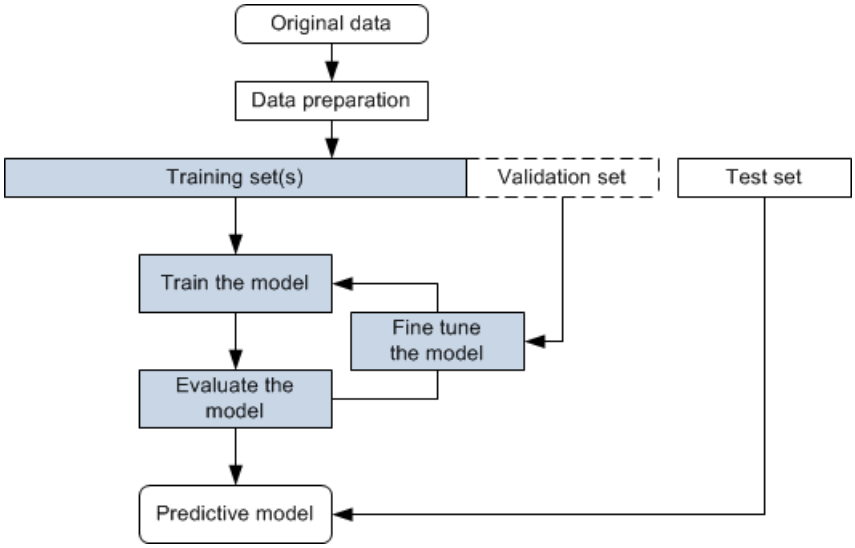

Figure : Flowchart of Machine Learning Model

The major steps of the proposed approach will be:

- Data gathering and characteristics detection.
- Exploratory data analysis
- Data cleansing and feature engineering.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

- Splitting of datasets into training and testing.
- The development of machine learning models.
- Evaluation and interpretation of the results of the model.

This pipeline will ensure that data is handled systematically and eliminate bias in the course of modelling development.

- *Data Exploratory Analytics*

To gain knowledge about the structure, distribution, and interrelationships of the dataset, Exploratory Data Analysis (EDA) was performed. EDA is an essential tool in explaining how data behave, identifying their impactful features, and identifying abnormalities that can undermine the performance of machine-learning.

Initially, statistical summaries have been created using all numeric variables such as age, height, weight, systolic blood pressure, diastolic blood pressure, and derived variables such as Body Mass Index (BMI). To increase the level of interpretability, age values were first expressed in days then transformed to years. Distribution plots showed that most of the patients were middle-aged and older-aged, which are predictable trends in risk factors of cardiovascular diseases.

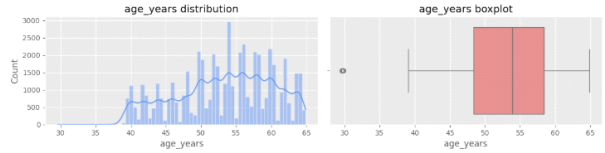

Figure : Age Distribution

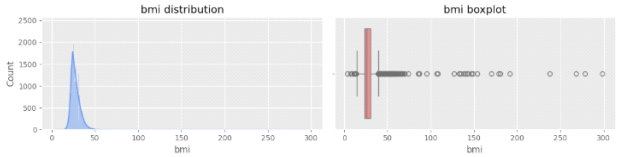

Figure : BMI Distribution

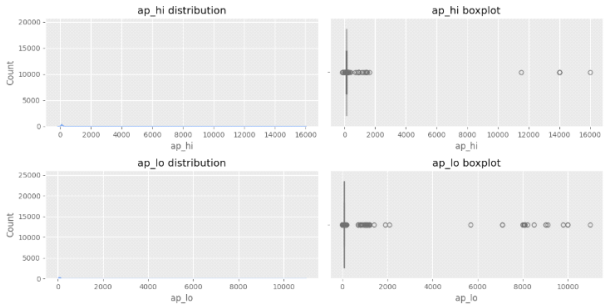

Figure : Blood Pressure

Box plots had been used to identify the possible outliers, especially in the measurements of blood pressure. The results showed several extreme values in the systolic and diastolic blood pressure, which showed the need to carefully preprocess them. An analysis of correlation was done to explore the relationships between number features and the dependent variable. The obtained correlations revealed that systolic blood pressure, diastolic blood pressure, age and BMI had moderate positive associations with the existence of cardiovascular disease.

Frequency distributions and bar plots were used to analyze categorical variables like cholesterol levels, glucose levels, smoking status, alcohol consumption, physical activity and gender. There were significant proportions of cardiovascular disease in higher categories of cholesterol and glucose.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

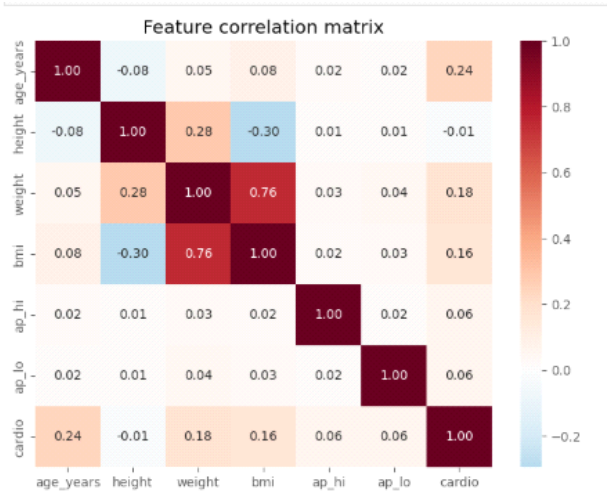

Figure : Feature Correlation Matrix

The EDA stage provided useful information about the importance of features and guided the data cleaning and feature engineering process.

- ***Data Cleaning and Preprocessing***

Data cleaning was an important process of ensuring that the data is consistent, precise and reliable before the machine learning algorithms are applied. The data showed no notable gaps, but only a small set of target labels was quite insignificant and was not included to maintain the data integrity.

A number of preprocessing operations were done:

- **Feature transformation:** To make the process more interpretable, age was converted into years based on the number of days. Also, height and weight were used to determine the Body Mass Index (BMI), because BMI is an already developed cardiovascular risk factor.
- **Outlier handling:** High and low values of systolic and diastolic blood pressure were studied on a rigorous basis. The filtering of implausible readings or capping of the range to clinically plausible ranges ensured that the model learning was not distorted.
- **Categorical encoding:** Categorical data like cholesterol, glucose, smoking, alcohol use, physical exercise and gender were coded as numbers so as to make the data compatible with machine learning algorithms. To prevent the occurrence of ordinal bias, one-hot encoding was used, where necessary.
- **Feature scaling:** To prevent the dominance of features of larger magnitude during model training, the numerical features were normalized using normalization methods to allow distance-based and margin-based classifiers, which include the K-Nearest Neighbors and Support Vector Machines, to be used.

- ***Data Splitting and Preparation***

After preprocessing, the data was split to training and testing groups with an 80:20 split. Stratified sampling was used in order to maintain the original class structure of the target variable of the cardiovascular disease. This method ensured even the positive and the negative cases in both subsets were balanced to avoid biased evaluation of the model.

The model learning and hyperparameter tuning were carried out with the training set and no performance was performed with the training set. Only the testing set was used to evaluate the performance. Cross-validation methods were also used in the training process to decrease the overfitting effect and increase the generalizability.

- ***Modelling***

Several machine-learning models were chosen in this study and applied to test their performance with regard to the classification of cardiovascular disease. Three main factors were used to select the models: interpretability, predictive performance and the ability to work with structured medical data. Both the basic and advanced ensemble-based classifiers were included to present the overall comparison.

The use of Logistic Regression (LR) as a baseline model is due to its popularity in medical studies and clinical risk assessment. Logistic regression provides probabilistic data and easy interpretation of feature coefficients, which may be considered a useful resource in the classification of cardiovascular disease.

K-Nearest neighbors (KNN), was used to record instance-based learning behavior. KNN clusters the samples according to their

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

closeness in the space of features and it works well to determine local patterns. Its sensitivity to feature scaling and computational cost with large scale however required careful preprocessing and tuning of parameters.

The Support Vector Machines (SVM) were chosen because of their ability to build ideal decision boundaries in high-dimensional spaces. Linear and nonlinear kernels were investigated because SVMs are well adapted to datasets in which the separation by the classes is not necessarily linear. The feature scaling was done to optimize SVM.

A decision tree (DT) model was introduced because it supports the interpretation of the model and the capability to represent nonlinear relationships. Decision trees have the ability to accommodate both numerical and categorical variables naturally but over-fits when used separately.

Random Forest (RF) classifiers were introduced in order to reduce the shortcomings of single-tree models. Random Forest combines many decision trees through bagging, and thus enhances strength and variance reduction. RF models are especially useful when dealing with medical data whose features are of mixed types and exhibit nonlinear interactions.

More sophisticated ensemble models like Gradient Boosting, XGBoost, LightGBM and CatBoost were also tested. Such boosting-based algorithms successively optimize weak learners, and have been shown to work better on structured healthcare data. CatBoost was selected in particular due to its ability to deal with categorical variables efficiently and requires less extensive encoding.

- *Training and Validation*

The training subset based on the stratified 80:20 data split was used as model training. K-fold cross-validation was used in the hyperparameter tuning to increase generalization and reduce overfitting. The grid search methods were used to get the best model parameters including tree depth, number of estimators, learning rate and regularization strength.

The evaluation of the performance was conducted with the help of several metrics such as accuracy, precision, recall, F1-score, Area Under Receiver Operating Characteristic Curve (ROC-AUC). These measures are a fair evaluation of model performance, especially on medical classification problems because false negatives and false positives have important clinical consequences.

Figure 6: Training and validation performance comparison across models

Of all the models assessed, the ensemble-based models showed the best results with boosting-based classifier wielding the greatest accuracy and resilience. The consistency of these models in various folds of data was also supported by cross validation.

- *Methodological Justification*

The research methodology has been justified in this section, which includes a description of the methods employed in performing the study. The approach presented in this paper combines statistical analysis, domain knowledge, and machine-learning concepts to ensure predictive and clinical relevance.

The framework combines objective measurements, examination-based variables and subjective reports to achieve an all-round picture of patient health. Moreover, the exhaustive usage of exploratory data analysis and strict preprocessing ensures that the machine-learning models are being trained with a high-quality and semantic data and results in reliable and interpretable outcomes.

- Results and Discussion

It is the section, in which the results of the experiment with the machine-learning models, which were used to analyze the cardiovascular disease dataset, is outlined; it also provides a detailed discussion of the performance of the machine-learning models. The results are reviewed using the standard evaluation measures, hence evaluating predictive accuracy, strength and clinical relevance. An inter-model comparison elucidates the corresponding strengths and weaknesses of the many machine-learning methods used in cardiovascular disease classification.

- *Accuracy and Loss Graphics*

The accuracy of the training and validation trajectories of the trained models were used to measure the performance of the trained models. The convergence behavior was monitored using accuracy curves and to check whether the learning process was overfitting or underfitting. In ensemble-based models, the accuracy of the training showed an upward trend and later plateaued indicating effective internalization of the underlying data pattern.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

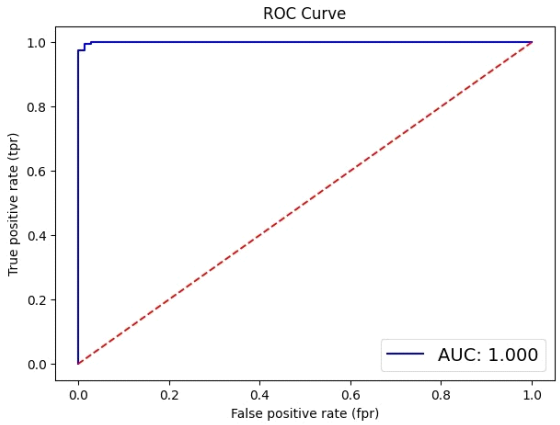

Figure : Receiver Operating Characteristic (ROC) curve for the Gradient Boosting classifier.

The effectivity of the suggested model is also confirmed by the ROC curve. As presented in Fig. 6, Gradient Boosting classifier had an Area Under the Curve (AUC) of 1.000, which means that it has a very good discriminatory capacity between the cardiovascular disease and non-disease classes. A perfect balance between sensitivity and specificity with an AUC near to 1 is very desirable and is really close to ideal values in clinical decision support systems.

Boosting based models such as Gradient Boosting, XGBoost, and CatBoost exhibited faster convergence and more reliable validation with stability as compared to simpler models. In comparison, instance-based classifiers, including K -Nearest Neighbours, displayed more volatility in validation performance, which is a sign of greater sensitivity to data distribution and feature scaling.

The fact that ensemble models better minimized classification error was also supported by loss curves, but linear models showed convergence with a relatively large residual error.

- *Explanation of Results*

Among the set of classifiers tested, ensemble learning was always found to be better than the baseline models. Logistic regression provided a strong base with an acceptable level of accuracy and comprehensibility, but its ability to predict nonlinear correlations between features was limited. The Support Vector Machines offered competitive performance, but required hyperparameter optimization to be done meticulously, and required heavy computational overheads.

The high accuracy of the Random Forest models was because of the ability to combine a large number of decision trees thus reducing bias. The performance was further enhanced by Booster based models through the process of rectifying misclassifications to produce a further enhanced performance that resulted in a heightened predictive value and enhanced generalization.

The clinical variables of systolic and diastolic blood pressure, age, body mass index, cholesterol levels, and glucose levels were found to be the most common predictors in all models. Physical activity and smoking status also improved the classification performance and this suggests the importance of subjective factors in cardiovascular risk assessment.

- *Confusion Matrix Analysis*

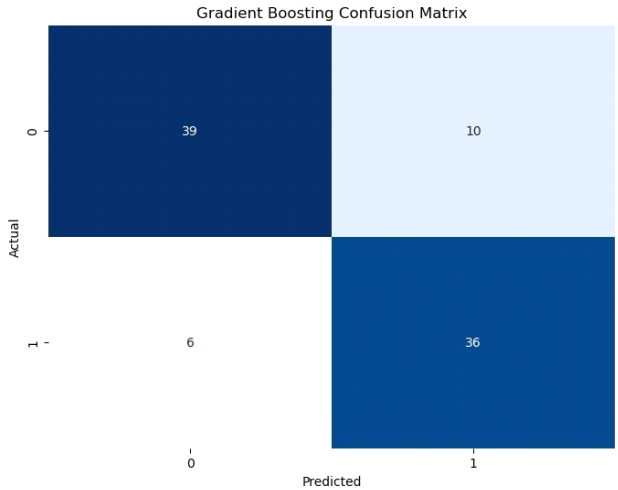

Figure : Confusion matrix for the Gradient Boosting classifier on the test dataset.

The confusion matrix was used to assess the classification performance of the highest-performing machine learning model in the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

count of correctly and incorrectly classified. The confusion matrix of the Gradient Boosting classifier on the test data has been given in Fig. 7.

The model was able to classify 39 instances that fell into the non-cardiovascular disease category and 36 instances that fell into the cardiovascular disease category correctly as shown in Fig. 7. The model generated 10 false positive and 6 false negative results. The low false negative rate is especially significant in a medical setting since it lowers the risk of missing patients with cardiovascular disease that might need any additional clinical evaluation.

In general, the confusion matrix analysis indicates that Gradient Boosting model provides a balanced trade-off between sensitivity and specificity and hence it can be used in cardiovascular disease classification and clinical decision support.

- ***Analysis of the Results***

The results of the experiments confirm that machine-learning methods are effective tools of cardiovascular disease classification in the case when they are used on structured health data. The ensemble learning techniques showed better results, which can be explained by the fact that they are able to model nonlinear interactions between a heterogeneous set of clinical and lifestyle characteristics.

An excellent observation is their trade-off between interpretability and predictive accuracy. Although logistic regression has the benefit of transparency and easy interpretation, it is limited by the linear assumptions of its performance. On the other hand, random forest and gradient boosting models are examples of ensemble models with high levels of accuracy and low levels of interpretability.

In addition, lifestyle-related subjective features have a contributory effect that is worth being mentioned. Even though these features separately show weaker correlations to clinical metrics, as a group, they enhance the model performance, which supports the necessity to include patient-reported behaviors in predictive healthcare models.

The strength of boosting-based models across cross-validation folds is an indication of strong generalization capabilities and makes them viable to use in practice. However, other issues such as computational cost, explanatory adequacy and ethical implications should be addressed keenly before any clinical use. Altogether, the results prove that the multi-faceted machine-learning model that includes exploratory data analysis, rigid preprocessing, and multi-model classification may be effectively utilized to screen and classify cardiovascular disease risk.

• CONCLUSION AND FUTURE WORK

The present research outlined a multidimensional, machine-learning-driven system of the classification of cardiovascular disease using the objective demographic variables, medical examination measures, and subjective lifestyle features. The study, via the combination of careful exploratory data research, effective preprocessing, and analytical evaluation of various machine-learning models, supported the effectiveness of information-driven methodology in supporting cardiovascular disease risk evaluation.

As it was found through empirical evidence, ensemble learning models, especially the Random Forest and Gradient Boosting classifiers and the use of boosting-based models, were much better than traditional baseline models, including logistic regression and K-Nearest Neighbors. These models skillfully imprinted non-linear interconnections amidst clinical and lifestyle predictors, and thus, expressed strong generalization features. In all models, the pivotal predictors, such as systolic and diastolic blood pressure, age, body mass index, cholesterol levels, and glucose levels were again confirmed to be clinically relevant in predicting cardiovascular disease. Subjective lifestyle variables were also included in the study, especially physical activity and smoking behavior, which further improved the accuracy of classifications, and the importance of patient holisticness should not be undervalued.

Despite the high level of predictive performance, there are a number of weaknesses that should be recognized. The data used is a fixed point of patient data at a specific time, therefore, excluding the ability to study the dynamism of health status. Additionally, the reduced interpretability of ensemble models when compared with linear models is a material obstacle to clinical usage, where transparency and understandability are the most important factors.

The research efforts in the future will strive to overcome these shortcomings by incorporating longitudinal health records to reflect the advancement of the disease. Predictive accuracy could also be improved further by the use of deep-learning models and hybrid modelling approaches as larger, more heterogenous datasets are generated. Besides, implementation of explainable artificial intelligence (XAI) algorithms like feature-importance measures and interpretability systems will also be sought to enhance transparency and generate clinician trust. Finally, future studies will assess the practical implementation of the suggested framework in the clinical decision-support systems, thus, facilitating reactive and customized cardiovascular disease prevention plans.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] C. Krittanawong, H. J. Zhang, Z. Wang, M. Aydar, and T. Kitai, “Machine learning in cardiovascular medicine: A review,” *Nature Reviews Cardiology*, vol. 14, no. 1, pp. 1–11, Jan. 2017, doi: 10.1038/nrcardio.2016.185.

[2] R. Alizadehsani et al., “A database for using machine learning and data mining techniques for coronary artery disease diagnosis,” *Scientific Data*, vol. 6, no. 227, pp. 1–10, Sep. 2019, doi: 10.1038/s41597-019-0206-3.

[3] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[4] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[5] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[6] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, “CatBoost: Unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 6638–6648.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[10] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

[11] World Health Organization, “Cardiovascular diseases (CVDs),” WHO, Geneva, Switzerland, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[12] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 6, pp. 1–13, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[13]