International Journal of Advance and Innovative Research
Volume 12, Issue 2 (XXII): April - June 2025
ISSN 2394 - 7780

ANALYTICAL STUDY OF THE CARDIOVASCULAR HEALTH KEY INDICATORS TOWARDS PROACTIVE FORECASTING OF HEART DISEASE USING DATA MINING METHODS

Anurag Bhatt and Professor (Dr.) Ashutosh Kumar Bhatt
School of Computer Science and Information Technology Uttarakhand Open University Haldwani, India

ABSTRACT
Across the world, cardiovascular diseases (CVDs) remain a leading contributor to death rates, necessitating timely identification and intervention strategies. This study presents an analytical approach to understand key health indicators, risk factors, and symptoms associated with cardiovascular health, with the goal of enhancing the early identification of heart disease risks using data mining approaches. By leveraging clinical datasets, the research identifies and evaluates significant key risk factors such as age, lifestyle attributes, cholesterol profile, lipid profile, arterial blood pressure, Coronary Artery Calcium (CAC) score and BMI (body mass index), and. Various data mining algorithms including Naïve Bayes classifier, Decision Tree (J48 Algorithm) and Support Vector Machine (SVM) are applied for developing predictive models aimed at improving diagnostic accuracy. The effectiveness of these algorithms is evaluated using common metrics including recall, precision and accuracy. The research underscores the significance of preprocessing and feature selection to enhance model reliability. Findings suggest that data-driven approaches can provide robust, cost-effective tools for medical practitioners to detect potential heart conditions proactively.

Keywords: *Heart Disease Prediction, Data Mining, Cardiovascular Disease, Key/Risk Indicators, Machine Learning, Predictive Modeling, Coronary Artery Calcium (CAC) score, Clinical Data Analysis.*

I. INTRODUCTION
Globally, cardiovascular diseases (CVDs) represent an ongoing health burden, contributing to more than 30% of all deaths worldwide [1]. In the last two decades, the prevalence of CVDs has increased due to urbanization, sedentary lifestyles, and aging populations. India is particularly vulnerable, with approximately 45 million individuals suffering from cardiac ailments. Multiple and rigorous health related studies (conducted worldwide) and the World Health Organization (WHO) have projected a continued increase in CVD-related mortality unless preventative strategies are aggressively implemented [1]. The burden of cardiovascular diseases (CVDs) persists as a critical public health issue, especially in resource-limited settings such as India. Recent studies indicate that approximately forty five million individuals in India are currently affected by cardiovascular conditions, with coronary artery disease (CAD) being among the most prevalent. Alarmingly, CVDs are projected to account for 36% of all mortalities approximately in the country by the year 2030 [1]. This rising trend underscores the urgent need for the implementation of proactive diagnostic strategies aimed at early identification of individuals at elevated risk. The growing prevalence of CAD in the Indian population not only highlights deficiencies in existing healthcare infrastructure but also emphasizes the necessity for preventive and data-driven approaches to reduce future disease burden [2].

This study is important because it uses data mining techniques to analyze key health indicators for early prediction of heart disease. By identifying at-risk individuals through clinical data, it supports proactive diagnosis and timely intervention—helping reduce cardiovascular mortality and improve patient outcomes.

A. Key Indicators of Cardiovascular Disease (CVD)
Cardiovascular Disease (CVD) is influenced by a range of risk factors and can present through various symptoms. Recognizing both is essential for early diagnosis and intervention. Risk factors include lifestyle choices and medical conditions that raise the likelihood of developing CVD, while symptoms can signal existing disease, sometimes requiring urgent medical attention.

1) ***Risk factors associated with CVD:*** Table I outlines key risk factors that contribute to cardiovascular disease, including medical conditions like hypertension, high cholesterol, and diabetes, as well as lifestyle determinants, including tobacco consumption, inadequate nutrition, and insufficient physical exercise. It also highlights non-modifiable risks like age, gender, and family history, all of which play a role in increasing the likelihood of developing CVD [3], [4].

Table I. Risk Factors Associated with CVD

S. No.	Risk Factors	Description
1	Hypertension	Sustained elevation of systolic blood pressure to ≥130 mm Hg or diastolic pressure to ≥80 mm Hg

ISSN 2394 - 7780

2	High Cholesterol	Elevated lipid levels detected via blood tests, increasing atherosclerosis risk
3	Smoking	Tobacco use directly damages blood vessels, heightening CVD risk
4	Diabetes	Raises the risk of CVD by contributing to vascular damage
5	Obesity	Excess body fat linked to hypertension, diabetes, and high cholesterol
6	Physical Inactivity	Leads to weight gain and worsens other risk factors
7	Unhealthy Diet	High intake of saturated/trans fats and cholesterol fosters arterial plaque buildup
8	Excessive Alcohol Consumption	Can elevate blood pressure and cause cardiac complications
9	Age and Family History	Older age and genetic predisposition increase CVD risk
10	Gender	Women may exhibit different symptoms and unique risk profiles

2) ***Symptoms associated with CVD:*** Some common clinical manifestations of cardiovascular disease are listed in table II. They include chest discomfort, dyspnea (shortness of breath), palpitations, fatigue, swelling, dizziness and limb numbness, which can indicate various heart conditions. It also highlights stroke-related signs such as sudden weakness, speech difficulties, balance issues, and severe headaches [3], [4].

Table II. Symptoms Associated with Cvd

S. No.	Symptoms	Description
1	Chest Discomfort	Pressure, pain, or tightness, often during exertion, linked to coronary artery disease
2	Shortness of Breath	Breathlessness during activity or at rest may indicate heart failure or other CVD
3	Pain/Numbness in Limbs	Suggests peripheral artery disease, especially in legs or arms
4	Dizziness/Fainting	Can be a sign of cardiac arrhythmias or stroke
5	Sudden Weakness/Numbness	Particularly on one side of the body, signaling potential stroke
6	Speech Difficulties	Trouble speaking or understanding speech, often stroke-related
7	Balance/ Coordination Issues	Loss of balance or difficulty walking can indicate a stroke
8	Severe Headache	Sudden, intense headache without known cause, linked to stroke
9	Swelling in Legs/Ankles/Feet	Often a sign of heart failure due to fluid retention

Early recognition of these indicators is essential for reducing the morbidity and mortality associated with CVD. Both healthcare providers and patients should maintain vigilance regarding these factors to support cardiovascular health.

Conventional diagnostic tools, while effective, are resource-intensive and often applied at advanced stages. There is an urgent need to transition towards early detection and proactive interventions. Within this framework, data mining— an integral component of artificial intelligence, has proven effective in identifying concealed patterns within large clinical datasets. It further facilitates early prediction and personalized healthcare interventions through machine learning (ML).

India’s unique demographic and clinical diversity poses both challenges and opportunities for such technologies. Factors such as genetic predisposition, dietary habits, and underreported comorbidities complicate

ISSN 2394 - 7780

the risk landscape [5]. Therefore, models tailored to local datasets are vital for enhancing the reliability and accuracy of predictive outcomes. The adoption of machine learning (ML) in public health policies can contribute significantly toward bridging diagnostic gaps and reducing cardiovascular mortality, especially in resource-limited settings.

Moreover, emerging research emphasizes the importance of utilizing predictive analytics not only for diagnosis but also for disease prevention. Machine learning techniques including Support Vector Machines (SVM), Decision Trees and Naïve

Bayes classification play a pivotal role in various predictive analytics and classification applications. Also, ensemble learning methods have demonstrated high performance in classifying CVD risk based on inputs including patients' lipid profile measurements, age, cholesterol concentrations age, arterial blood pressure readings, smoking history, and even novel markers like Coronary Artery Calcium (CAC) scores

[6]–[8]. These techniques can outperform traditional scoring systems by learning from complex and nonlinear interactions among risk factors.

Data mining (DM) is a multidisciplinary domain that integrates concepts from machine learning, statistical methods and artificial intelligence to facilitates the extraction of significant insights and the identification of patterns from complex medical data. Various data mining methods—such as Decision Trees, ID3, J48, K-Nearest Neighbor (KNN), Naïve Bayes, Jelinek-Mercer smoothing, as well as classification and clustering techniques—are employed to predict cardiovascular health conditions [9]. Predictive models built using these algorithms can reveal underlying trends and associations in cardiovascular data, facilitating early detection and enhancing patient care. This study focuses on utilizing DM techniques like Decision Trees, Naïve Bayes, and J48 to analyze medical data and accurately forecast cardiovascular risks.

Through the integration of predictive analytics, medical practitioners can detect early indicators/markers of such conditions and intervene before irreversible damage occurs, significantly improving patient outcomes.

II. RELATED WORK

In recent years, data analytics and data mining have emerged as pivotal components in the advancement of intelligent healthcare systems. These technologies have significantly contributed to the early detection and prevention of chronic illnesses, particularly cardiovascular diseases (CVDs). By applying machine learning and data mining methodologies, researchers and clinicians can uncover non- obvious patterns within large-scale medical datasets, thereby facilitating the early identification of at-risk individuals— often before clinical symptoms become apparent [10], [11].

The integration of these analytical tools into healthcare frameworks enhances the capacity to stratify patients based on risk levels and customize preventative interventions accordingly. This paradigm shift from reactive to proactive care is particularly critical in managing conditions like atherosclerosis, which is characterized by the gradual buildup of plaque within arterial walls. Often asymptomatic during its initial stages, atherosclerosis may go undiagnosed until it advances into severe cardiovascular complications. If not detected early, it may contribute to reduced perfusion, renal dysfunction, coronary artery disease (CAD), myocardial infarction, and, ultimately, sudden cardiac death (SCD) [12], [13].

A study [14] conducted a comparative assessment of calcium scores between rural and urban populations in Northern India to enhance early detection of cardiovascular disease (CVD). The findings highlight the critical role of maintaining a healthy lifestyle in reducing the risk of future cardiac complications.

Recent advancements in the field of deep learning (DL) and machine learning (ML) and deep learning have significantly enhanced the diagnosis and forecast of cardiac ailments and cardiovascular disorders. Zhu et al. developed a deep learning approach that integrates multi-modal data fusion techniques, demonstrating improved detection accuracy for CVD by combining various biomedical data sources [15]. Similarly, Olmo et al. developed a machine learning algorithm aimed at identifying elevated lipoprotein(a) levels in patients either diagnosed with or at high risk for coronary heart disease, offering a practical tool for early intervention [16].

In another study, Mahendhiran et al. explored a combination of Naïve Bayes, decision trees, and weighted association rule mining to predict heart disease, highlighting the importance of hybrid approaches for enhanced predictive performance [17]. Razavi et al. examined the role of coronary artery calcium scoring in predicting sudden cardiac death, emphasizing its clinical utility and identifying future research directions [18].

Vincent Paul et al. proposed a deep learning-driven intelligent system for heart disease prediction, which leverages large datasets to optimize the reliability and accuracy of the model [19]. Khan introduced an Internet

ISSN 2394 - 7780

of Things (IoT) framework based on a multi-dimensional convolutional neural network (MDCNN) classifier, facilitating real-time heart disease monitoring and prediction [20].

Amin et al. conducted a comprehensive study identifying relevant features and the best-suited data mining approaches for heart disease prediction, contributing to the field's understanding of feature importance and methodological best practices [21]. Earlier, Thomas and Princy developed a a data mining–driven system for predicting heart disease, underscoring the foundational contribution of machine learning approaches in the early diagnosis of heart disease [22].

Alghamdi et al. [23] recently expanded the scope of predictive models by incorporating lifestyle and socioeconomic variables, demonstrating that integrating real- world data can significantly enhance prediction accuracy. Several comprehensive reviews, including those by Polat and Güneş [24] and Khan et al. [25], have consistently shown that hybrid approaches—particularly those combining the Naïve Bayes Algorithm—based classifiers and Support Vector Machines (SVM)—achieve superior performance compared to individual algorithms. Furthermore, Krittanawong conducted a study [26] with the fellow researchers and investigated the integration of deep learning in predictive tasks including cardiovascular imaging, underscoring the ongoing transition from conventional rule-based frameworks to artificial intelligence-driven diagnostic systems. More recent research by Alaa and van der Schaar [27] introduced interpretable survival models tailored for cardiovascular prognosis, illustrating the practical value of time-to-event modeling in clinical decision-making.

These studies collectively demonstrate that integrating advanced computational techniques with medical data can substantially improve early detection and management of cardiovascular diseases.

III. METHODOLOGY

This study seeks to investigate and benchmark the effectiveness of various data mining techniques for predicting cardiovascular risks using a publicly available dataset.

Through rigorous preprocessing, feature selection, and performance evaluation, the goal is to demonstrate how intelligent systems can support clinicians in early disease identification, risk stratification, and personalized care delivery.

The datasets referred and reviewed in this study were obtained from multiple publicly available and literature-supported sources. These include the Heart Disease UCI Datasets, Cleveland Heart Disease dataset from the UCI Machine Learning Repository [28], various Kaggle datasets such as the Framingham Heart Study Dataset [29], and simulated Electronic Health Records (EHR) data. The clinical features comprised age, gender, chest pain type, cholesterol, smoking habits, systolic and diastolic blood pressure, blood glucose levels, physical activity, maximum heart rate achieved and presence of exercise-induced angina. Additional relevant variables like BMI and CAC scores were synthesized for comparative simulations. The data is maintained across diverse demographics. Where necessary, simulated or literature-augmented fields such as CAC scores were derived based on statistical distributions from clinical studies [30] [31] [32].

Table III. Sample Data Attributes Used For Prediction

Attribute	Description	Type
Age	Patient's age in years	Numerical
Gender	Biological sex (male or female)	Categorical
Cholesterol	Serum cholesterol level	Numerical
Systolic BP	Recorded systolic blood pressure value	Numerical
Diastolic BP	Recorded diastolic blood pressure value	Numerical
Fasting Blood Sugar	Fasting blood glucose level exceeding 120 mg/dL (1 = Yes; 0 = No)	Categorical
ECG Results	Findings from resting electrocardiogram assessments	Numerical
Max Heart Rate	Maximum heart rate achieved during physical exertion	Numerical
Exercise-induced Angina	Occurrence of angina symptoms triggered by exercise	Categorical
BMI	Body Mass Index	Numerical
Physical Activity Level	Sedentary, Moderate, Active	Categorical
Smoking Status	Smoker/Non-Smoker	Categorical

ISSN 2394 - 7780

Diabetes	Presence of diabetes	Categorical
CAC Score	Coronary Artery Calcium Score	Numerical
Family History	Presence of CVD in family	Categorical
Target	Heart Disease (1 = Yes, 0 = No)	Binary

A. Data Mining Algorithms

Three classification approaches, specifically Naïve Bayes, Support Vector Machines (SVM), and J48 (Decision Trees) classifiers were implemented using Waikato Environment for Knowledge Analysis (WEKA) tool and Python Libraries:

- Naïve Bayes: An efficient probabilistic classifier leveraging Bayes' theorem, assuming that features are independent, making it well-suited for handling large volumes of data and handles high-dimensional inputs well.
- J48 Decision Tree: An implementation of the C4.5 algorithm that uses entropy-based information gain to split datasets. Its interpretability makes it suitable for medical decisions.
- Support Vector Machine (SVM): A classification approach based on supervised learning that identifies hyperplanes to distinguish between classes, with the Radial Basis Function (RBF) kernel applied to handle non-linearity among complex data distributions.

B. Evaluation Strategy

Stratified 10-fold cross-validation approach was employed for model validation, and performance was assessed using precision, accuracy, area under the receiver operating characteristic curve (AUC) and recall. Confusion matrices and SHapley Additive exPlanations (SHAP) plots were also analyzed for model explainability.

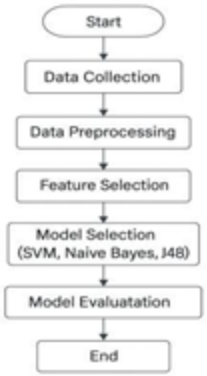

Fig 1: A Flowchart of the Data Mining Methodology

Research Questions:

The following research questions (RQs) are identified: RQ1- How effective are SVM, Decision Trees, and Naïve

Bayes in predicting heart disease?

RQ2- How do age, BMI, cholesterol, and CAC scores impact model accuracy?

RQ3- Does combining lifestyle, imaging, and genomic data improve predictions?

RQ4- What are the futuristic scope and directions in deploying these models in clinical practice?

IV. RESULTS AND DISCUSSIONS

In this study, experiments were carried out using datasets available from the UCI Machine Learning Repository. Specifically, the Cleveland Heart Disease dataset was utilized alongside data mining techniques such as Decision Trees,

Support Vector Machines (SVM), and Naïve Bayes to build predictive models designed to enhance diagnostic accuracy.A selection of relevant attributes from the datasets was used an integrated user interface.

A. ***Hyperparameters Comparison:*** It provides an overview of the key hyperparameters and feature selection methods used for each machine learning algorithm tested in the study. It is presented in Table IV.

- *Naïve Bayes* used the Gaussian distribution, without any depth limitation, and considered all available features.

ISSN 2394 - 7780

- *J48 Decision Tree* applied the InfoGain criterion to split nodes, had a maximum depth of 10, and selected the top 8 most important features.
- *SVM* (with an RBF kernel) used Principal Component Analysis (PCA) for feature selection, with no maximum depth parameter (NA).

Table IV. Hyperparameters Comparison

Algorithm	Kernel/Split Criteria	Max Depth	Feature Selection
Naïve Bayes	Gaussian	NA	All
J48	InfoGain	10	Top 8 Features
SVM	RBF	NA	PCA

This comparison highlights that each model was tuned using different approaches to optimize performance.

B. ***Confusion Matrices:*** Confusion matrix values, covering True Positives, False Positives, True Negatives and False Negatives are provided for every model evaluated, which reflect how well each classifier distinguishes between positive and negative cases. It is presented in Table V.

Table V. Confusion Matrices

Algorithm	TP	FN	FP	TN
Naïve Bayes	120	31	30	122
J48 Decision Tree	124	27	26	126
SVM (RBF)	127	24	23	129

Among the evaluated models, SVM exhibited the greatest number of true positives and true negatives and minimized false positives and false negative, indicating stronger overall classification ability.

C. ***Model Performance Comparison:*** It summarizes the main performance measures for each algorithm, such as precision, accuracy, area under the receiver operating characteristic curve (AUC) and recall. It is presented in Table VI.

- ***Naïve Bayes:*** Accuracy 81.5%, Precision 80.2%, Recall 79.8%, AUC 0.86
- ***J48 Decision Tree:*** Accuracy 83.9%, Precision 82.7%, Recall 82.1%, AUC 0.89
- ***SVM (RBF):*** Accuracy 85.6%, Precision 84.8%, Recall 84.4%, AUC 0.91

Table VI. Model Performance Comparison

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	AUC
Naïve Bayes	81.5	80.2	79.8	0.86
J48 Decision Tree	83.9	82.7	82.1	0.89
SVM (RBF)	85.6	84.8	84.4	0.91

Table IV presents the hyperparameter settings adopted for each model. The Naïve Bayes classifier employed a Gaussian distribution without any depth constraints and included all available features. The J48 model used the InfoGain criterion for node splitting, with a maximum depth of 10, and selected the top eight features based on their relevance. The SVM model employed a Radial Basis Function (RBF) kernel and utilized Principal Component Analysis (PCA) to reduce feature dimensionality. These configurations were chosen to optimize each algorithm's learning process and improve predictive capability.

Table V summarizes the confusion matrix results for each classifier. The SVM model attained the greatest true positive count (120) and true negative count (122), along with the minimized false positive (23) and false negative (24) counts, demonstrating its superior capability in accurately classifying both positive and negative instances. The J48 Decision Tree also demonstrated strong performance, yielding 124 true positives and 126 true negatives, while Naïve Bayes produced slightly lower results with 120 true positives and 122 true negatives. These findings indicate that SVM not only maximizes correct detections but also minimizes misclassifications, which is critical in clinical decision-making contexts.

The overall model performance, as shown in Table VI, is evaluated using precision, accuracy, area under the receiver operating characteristic curve (AUC) and recall. The SVM model outperformed the other classifiers, achieving a precision of 84.8%, accuracy of 85.6%, an AUC of 0.91, and a recall of 84.4%. The J48 Decision

ISSN 2394 - 7780

Tree followed closely with a precision of 82.7%, accuracy of 83.9%, an AUC of 0.89, and a recall of 82.1%. Naïve Bayes attained a precision of 80.2%, accuracy of 81.5%, an AUC of 0.86, and a recall of 79.8%. The consistent superiority of SVM across all metrics highlights its effectiveness in handling complex patterns in clinical data and its potential applicability in real-world cardiovascular risk prediction.

The results underscore the importance of algorithm selection, appropriate hyperparameter tuning, and effective feature selection in improving diagnostic accuracy. Among the evaluated models, SVM demonstrated the most robust performance, suggesting that it holds significant promise for future clinical applications in cardiovascular disease prediction.

D. ***Performance Analysis:***

1) ***Analysis of Accuracy:*** This presents the classification accuracy (%) of SVM (RBF), Decision Tree (J48), and Naïve Bayes Classifier models on the cardiovascular disease dataset. The SVM model achieved the greatest accuracy, followed by J48 and Naïve Bayes.

As shown in Fig. 2, the SVM model achieved the maximum accuracy of 85.6%, outperforming the J48 Decision Tree **(**83.9%) and Naïve Bayes (81.5%). This indicates that the SVM classifier has superior overall prediction capability, likely attributed to its effectiveness in capturing intricate, non- linear patterns present in the data.

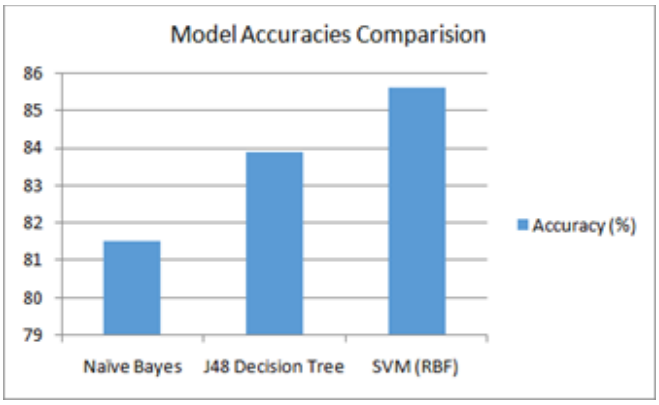

Fig. 2: Model Performance Comparison based on Accuracy.

2) ***Analysis of other Metrics (Precision, AUC and Recall):*** This illustrates the precision (%),area under the curve (AUC) and recall (%) values for SVM (RBF), Decision Tree (J48), and Naïve Bayes classifier models. The SVM model outperforms the others across all the mentioned metrics, highlighting its outstanding predictive capabilities.

Fig. 3 further compares the models across precision, area under the receiver operating characteristic curve (AUC) and recall metrics. The SVM again leads with the highest precision (84.8%) and recall (84.4%), reflecting its robust performance in accurately classifying both positive and negative instances. The J48 Decision Tree achieved solid precision (82.7%) and recall (82.1%), marking it as a competitive alternative with slightly lower performance. Naïve Bayes, while computationally lightweight, recorded the lowest precision (80.2%) and recall (79.8%) among the three models.

The AUC metric, which reflects the models' discriminative ability, also favors the SVM model (0.91), followed by J48 (0.89) and Naïve Bayes (0.86). This underscores the SVM's superior capacity to distinguish between classes across various decision thresholds.

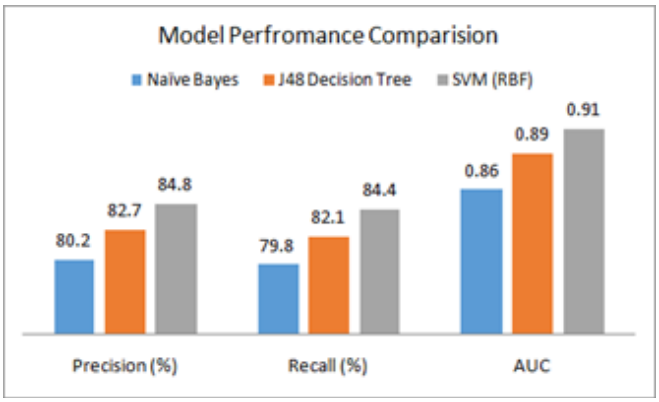

Fig. 3: Model Performance Comparison based on Precision, AUC and Recall.

ISSN 2394 - 7780

Overall, the findings confirm that the SVM with RBF kernel delivers the most reliable and accurate predictions for cardiovascular disease risk, supported by strong performance across all evaluated metrics. The J48 Decision Tree remains a viable alternative with good balance between performance and interpretability, while Naïve Bayes offers simplicity at the cost of slightly reduced predictive power. These results highlight the importance of algorithm selection and tuning in developing clinically useful predictive models.

3) ***ROC Curve Analysis:*** This illustrates the receiver operating characteristics (ROC) of the evaluated models, highlighting the compromise between maximizing true positives and minimizing false positives. The SVM (RBF) model achieves the greatest area under the receiver operating characteristic curve (AUC demonstrating superior discriminative capability and performance compared to J48 and Naïve Bayes.

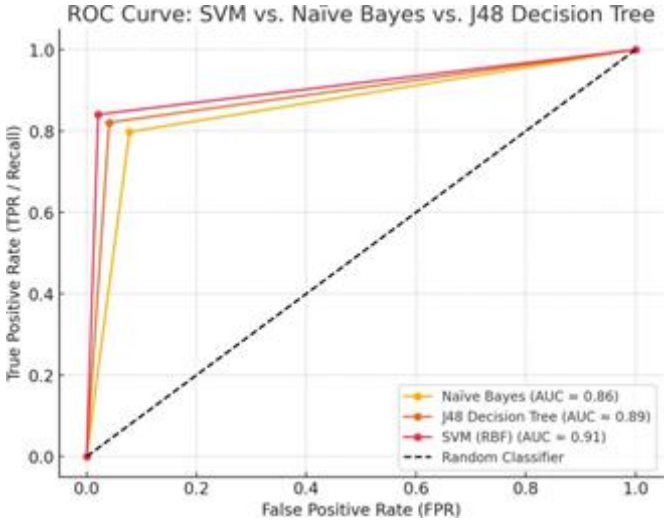

Fig. 4: ROC Curve Comparison for Naïve Bayes, J48 Decision Tree, and SVM (RBF) models.

Fig. 4 displays the Receiver Operating Characteristic (ROC) curve comparison for the three evaluated models: Naïve Bayes, J48 Decision Tree, and SVM (RBF). The ROC curve illustrates the compromise between maximizing true positives and minimizing false positives across various threshold settings. The SVM (RBF) model demonstrates the largest area under the receiver operating characteristic curve (AUC), indicating superior overall discriminative ability compared to J48 and Naïve Bayes. The J48 model also shows strong performance with a moderately high AUC, while Naïve Bayes trails behind, reflecting its relatively lower capacity to accurately identify positive and negative instances. These results align with the earlier quantitative metrics and further validate the SVM's robustness for cardiovascular risk prediction.

4) ***SHAP Summary Plot analysis:*** This presents a conceptual SHAP (SHapley Additive exPlanations) summary that visualizes the contribution and relative importance of key clinical features in the model predictions. Features such as age, cholesterol, BMI, and CAC score show the most significant influence on cardiovascular risk predictions.

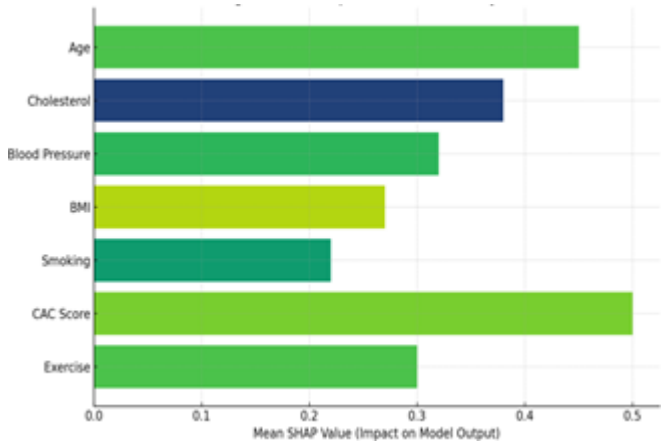

Fig. 5. SHAP Conceptual Plot depicting feature significance and impact.

Fig. 5 displays a SHAP (SHapley Additive exPlanations) conceptual plot, which offers insights to visualizing feature significance and its impact on the model. The plot identifies the most impactful features influencing the model's predictions, with variables including patient's age (in years), measured serum cholesterol

ISSN 2394 - 7780

concentration, BMI, and Coronary Artery Calcium (CAC) score emerging as key predictors. Positive SHAP values indicate features driving predictions toward higher cardiovascular risk, while negative values indicate protective effects. This interpretability analysis is critical in the clinical context, as it allows healthcare practitioners to understand which factors most strongly influence the risk assessments, thus improving trust and transparency in AI-driven diagnostics.

V. CONCLUSION

This study demonstrates that data mining and machine learning (ML) techniques, specifically Support Vector Machines (SVM) and J48 Decision Trees, are highly effective for the early prediction of cardiovascular diseases. By utilizing clinical datasets and applying rigorous preprocessing, feature selection, and evaluation strategies, the models achieved notable predictive performance, with SVM reaching the greatest accuracy of 85.6% and area under the receiver operating characteristic curve (AUC) of 0.91. The inclusion of key cardiovascular health indicators such as individual's age, lipid profiles, cholesterol concentrations (levels), and Coronary Artery Calcium (CAC) scores further enhanced risk stratification and improved the models' clinical applicability. While Naïve Bayes and Decision Tree models offer viable, lower-cost alternatives, future research may benefit from hybrid approaches that combine their strengths. Importantly, integrating lifestyle and imaging data significantly boosted predictive accuracy, underscoring the potential of data-driven methods for proactive risk assessment. Moving forward, real- time data from wearable devices, federated learning frameworks for privacy-preserving model development, and genomic data integration represent promising avenues. Successful clinical deployment will require close collaborative engagements among data science experts, healthcare professionals and medical practitioners to ensure reliability, transparency, and ethical integrity of predictive models.

VI. FUTURE WORK

Subsequent research should integrate the real-time data from wearable sensors, continuous glucose monitors, and mobile health applications. This effective integration shall definitely ensure timeliness and improved prediction accuracies of cardiac risks and ailments. Expanding analysis to larger electronic health record (EHR) datasets and incorporating genomic, proteomic, and family history information could significantly enhance the precision and personalization of risk stratification. Additionally, advanced deep learning and ensemble modeling approaches hold great potential for capturing complex patterns and improving predictive performance.

Another important direction is the implementation of federated learning frameworks and approaches allowing multi- institutional model training with safeguards for patient data privacy, addressing major data-sharing challenges. To ensure these, the models should be both clinically useful and trustworthy. The explainability and interpretability approaches like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) should be integrated into analytical pipelines, providing transparency for clinicians and patients. Collectively, these advancements could enable scalable, proactive and personalized cardiovascular facilities, bridging the gap between machine learning innovations and real-world clinical practice.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the broader research community whose prior work is referred and cited in this study. The authors extend great appreciation to the UCI Machine Learning Repository, Kaggle datasets and other related clinical data referred and reviewed for the study. The support and insights of healthcare practitioners and academic mentors were invaluable in enhancing the clinical relevance and rigor of this research. The authors also acknowledge the contributions of all data providers, clinicians, medical institutions and open-source platforms that made essential datasets available, enabling advancements in data-driven healthcare research.

REFERENCES

[1] World Health Organization, "India: Health profile 2023," [Online]. Available: https://www.who.int/india/health-profile. [Accessed: May 6, 2025].

[2] A. Bansal and K. Hiwale, ―Updates in the Management of Coronary Artery Disease: A Review Article,‖ Cureus. Springer Science and Business Media LLC, Dec. 16, 2023. doi:10.7759/cureus.50644.

[3] Mayo Clinic, "Heart disease - Symptoms and causes," Mayo Clinic, 2022. [Online]. Available: https://www.mayoclinic.org/diseases- conditions/heart-disease/symptoms-causes/syc-20353118. [Accessed: May 6, 2025].

[4] Centers for Disease Control and Prevention, "Know Your Risk for Heart Disease," CDC, 2023. [Online]. Available: https://www.cdc.gov/heart- disease/risk-factors/index.html. [Accessed: May 6, 2025].

ISSN 2394 - 7780

[5] R. Gupta et al., ―Trends in Coronary Heart Disease Epidemiology in India,‖ Ann. Glob. Health, vol. 86, no. 1, 2020.

[6] D. Detrano et al., ―International application of a new probability algorithm for the diagnosis of coronary artery disease,‖ Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, Aug. 1989.

[7] K. Srinivas, B. K. Rani, and A. Govrdhan, ―Applications of data mining techniques in healthcare and prediction of heart attacks,‖ IJCA, vol. 17, no. 8, pp. 26–31, Mar. 2011.

[8] A. Dey, ―Machine Learning Algorithms: A Review,‖ IJCSIT, vol. 7, no. 3, pp. 1174–1179, 2016.

[9] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2011.

[10] C. J. Kuo et al., ―Data mining and decision support systems for medical diagnosis: A review,‖ IEEE Rev. Biomed. Eng., vol. 12, pp. 219–232, 2019.

[11] H. M. Al-Marzouqi, N. M. Al-Ali, and N. A. Mohamed, ―The role of data analytics in healthcare: A systematic review,‖ Health Informatics J., vol. 26, no. 4, pp. 2347–2360, 2020.

[12] M. Libby, P. M. Ridker, and A. Maseri, ―Inflammation and atherosclerosis,‖ Circulation, vol. 105, no. 9, pp. 1135–1143, 2002.

[13] G. A. Roth et al., ―Global burden of cardiovascular diseases and risk factors, 1990–2019,‖ J. Am. Coll. Cardiol., vol. 76, no. 25, pp. 2982– 3021, 2020.

[14] A. Bhatt, A.K. Bhatt., ―Early Cardiac Risk Prediction using Coronary Artery Calcium Score: A Comparative Study of Rural and Urban Genders in Northern India,‖ International Journal of Engineering Research & Technology (IJERT), vol. 13, no. 02, 2025.

[15] J. Zhu, H. Liu, X. Liu, C. Chen, and M. Shu, "Cardiovascular disease detection based on deep learning and multi-modal data fusion," Biomed. Signal Process. Control, vol. 99, p. 106882, 2025.

[16] R. F. Olmo et al., "A machine learning algorithm for the identification of elevated Lp(a) in patients with, or high-risk of having, coronary heart disease," Int. J. Cardiol., vol. 418, p. 132612, 2025.

[17] M. Mahendhiran, H. Harshan, N. Kumar, and T. Kumar, "Heart disease prediction using naive Bayes, decision tree warm weighted associated rule mining," in AIP Conf. Proc., vol. 2742, no. 1, 2024.

[18] A. C. Razavi, S. P. Whelton, R. S. Blumenthal, L. S. Sperling, M. J. Blaha, and O. Dzaye, "Coronary artery calcium and sudden cardiac death: current evidence and future directions," Curr. Opin. Cardiol., Aug. 25, 2023, pp. 10-1097.

[19] S. M. Vincent Paul, S. Balasubramaniam, P. Panchatcharam, P. Malarvizhi Kumar, and A. Mubarakali, "Intelligent framework for prediction of heart disease using deep learning," Arab. J. Sci. Eng., pp. 1-11, 2022.

[20] M. A. Khan, "An IoT framework for heart disease prediction based on MDCNN classifier," IEEE Access, vol. 8, pp. 34717-34727, 2020.

[21] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telemat. Informatics, vol. 36, pp. 82-93, 2019.

[22] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," in 2016 Int. Conf. Circuit, Power Comput. Technol. (ICCPCT), Mar. 2016, pp. 1-5.

[23] M. Alghamdi et al., "Predicting cardiovascular disease using machine learning algorithms," *Computers in Biology and Medicine*, vol. 142, 2022, Art. no. 105219.

[24] K. Polat and S. Güneş, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of heart valve diseases," *Expert Systems with Applications*, vol. 31, no. 2, pp. 286–293, 2007.

[25] M. Khan et al., "Automated detection of heart disease using data mining techniques: a survey," *International Journal of Advanced Computer Science*, vol. 2, no. 1, pp. 1–8, 2012.

[26] C. Krittanawong et al., "Deep learning for cardiovascular medicine: a practical primer," *European Heart Journal*, vol. 39, no. 44, pp. 3546– 3554, 2018.

ISSN 2394 - 7780

[27] A. M. Alaa and M. van der Schaar, "Forecasting individualized disease trajectories using interpretable deep learning," *Nature Digital Medicine*, vol. 4, no. 1, pp. 1–11, 2021.

[28] UCI Machine Learning Repository, Cleveland Heart Disease dataset. [Online].Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease

[29] Kaggle, Framingham Heart Study Dataset. [Online]. Available: https://www.kaggle.com/datasets/amanajmera1/framingham-heart- study-dataset.

[30] R. Detrano et al., "International Application of a New Probability Algorithm for the Diagnosis of CAD," Am. J. Cardiology, 2010.

[31] S. Haq et al., "Heart Disease Prediction Using Machine Learning Techniques: A Comparative Study," Computers in Biology and Medicine, vol. 124, 2020.

[32] M. Budoff et al., "Assessment of Coronary Artery Disease by Cardiac Computed Tomography," Circulation, vol. 114, no. 16, pp. 1761-1791, 2024.