A Machine Learning Approach to Heart Attack Risk Assessment Using Feature-Engineered Clinical Data

Vipul Narayan[1], Prof Dr Divya Midhun[2], Dr. Pawan Whig[3]
[1] Madan Mohan Malviya University of Technology,Gorakhpur, India; [2] Lincoln University;
[3]VIPS-TC, India;
vipulupsainian2470@gmail.com, divya@lincoln.edu.my, pawan.whig@vips.edu

---

Abstract: Cardiovascular diseases (CVDs) remain one of the leading causes of deaths throughout the world and that essentially emphasizes the necessity of developing accurate risk prediction models. This paper presents a feature engineered ensemble learning approach to heart attack risk prediction based on a wide range of clinical and lifestyles factors. The dataset, which contained 8760 samples and 28 features, was optimized with the immense feature engineering, which included creating new features such as Systolic BP (Blood Pressure), Diastolic BP, BP Ratio and Cholesterol to Triglycerides. Various machine learning algorithms such as Logistic Regression, KNN, Decision Tree, Naive Bayes, SVM, GBM, XGBoost, and MLP were compared to each other with respect to accuracy, precision, recall and F1-score. The result of the model demonstrated that the traditional models performed moderately well with Logistic Regression having a recall value of 0.99 and poor generalizability while ensemble models such as GBM and XGBoost performed better in terms of stability and accuracy. The Proposed Feature Engineering + Ensemble Model using Random Forest and XGBoost proved better than all other models with an accuracy of 0.92 and F1 score of 0.84 and hence establishes the superiority of the hybrid ensemble learning method to improve the accuracy of the diagnosis**.**

Keywords: Cardiovascular Disease; Machine Learning; Risk Prediction; Decision Tree; Random Forest; Classification Models

---

Introduction: An enormous burden of morbidity and mortality is ascribed to cardiovascular diseases (CVDs), which are still a major worldwide health concern [1]. In order to identify those individuals with high risk of development of CVDs, prompt risk prediction and prognosis is imperative. This allows for early intervention and individual treatment programs.

CVDs remain the foremost cause of mortality worldwide, accounting for an estimated 17.9 million deaths annually. CVDs encompass a wide range of disorders affecting the heart and vascular system, including coronary artery disease, rheumatic heart conditions, and cerebrovascular disorders. Among these, heart attacks and strokes contribute to more than 80% of CVD-related fatalities. Alarmingly, a significant proportion of these deaths occur prematurely, with nearly one-third affecting individuals below 70 years of age. Cardiovascular diseases also have a genetic component. Having a blood line of CVD increases risk because genetic factors affect blood

pressure, cholesterol, and heart shape. Nevertheless, by leading healthy lives and visiting their doctors regularly, those with hereditary risk factors can effectively manage their risk. Advances in genomics have also opened up new possibilities for predicting and managing CVD risk using customised medicine approaches [4].

A combination of medication, lifestyle modification, and, in extreme cases, surgery is used to treat cardiovascular disorders. To treat symptoms and prevent complications, doctors typically give medications such as beta-blockers, anticoagulants, statins, and antihypertensive medicines [5]. In severe situations, angioplasty, bypass surgery, or pacemaker installation can be necessary, for individuals who have survived CVD episodes, rehabilitation and ongoing monitoring are essential [6].

Heart disease includes several conditions affecting the structure & function of the heart. The seven types of heart diseases represented in this infographic, along with their respective pathological representations, are shown in Figure 1 below:

1. Coronary Artery Disease (CAD): Characterised by the present of fatty deposits in the coronary arteries, that lead to a decrease in blood supply to the heart tissues and can cause chest pains and heart attacks.

2. Valve Disease: Involves malfunctioning heart valves, including stenosis (narrowing) or improper closure, disrupting normal blood flow through the heart. This following diagram compares a normal closed valve with a stenosed one.

3. Aneurysm: This condition is the bulging or weakened spot in the wall of the blood vessel such as the thoracic aortic aneurysm which may burst and lead to bleeding to death of the patient.

4. Cardiac Arrhythmia: Explained by unorganized electrical signals in the heart, as a result of which in uneven heartbeats that can harm the heart's ability to pump blood successfully.

5. Cardiomyopathy: This is the abnormal thickening or even weakening of the heart muscle making the cardiac functioning hard and may ultimately lead to heart failure.

6. Pericarditis: It is a condition characterized by inflammation of the pericardium, the fluid-filled membrane that encases and protects the heart.

7. Heart Failure: It occurs when the heart becomes enlarged, weak, and making it unable to pump sufficient blood to the body and usually caused by another underlying heart condition.

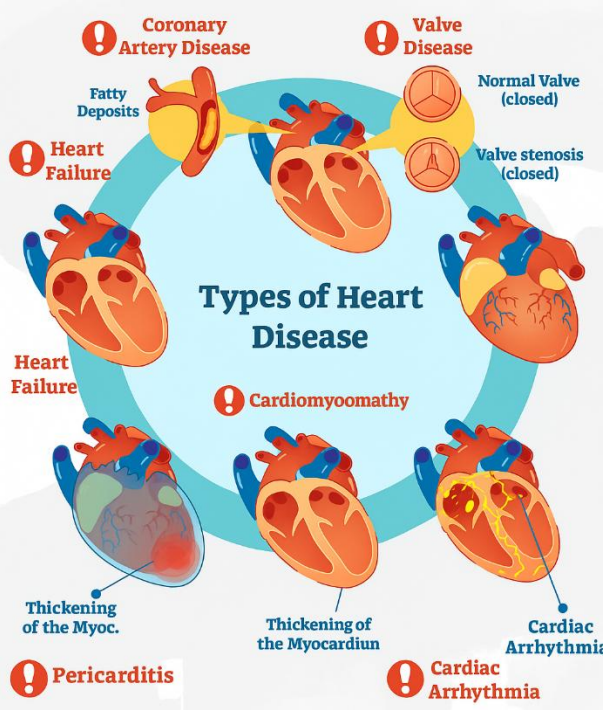

Figure 1: Types of Heart Disease

Related work: The study by Westerlund et al. (2021) places emphasis on the significance of molecular data, and explainable AI (XAI) is vital to recurrent cardiovascular events. This study provides additional information about patient risk and development of the disease by introducing multi-omics information into AI-based models. Besides increasing the accuracy in prognosis, this method aids in the identification of molecular biomarkers, regulatory pathways, and essential processes that cause cardiovascular diseases (CVD). Yet, such models' implementation is still troublesome owing to the intricacy in integrating large-scale multi-omics data and scarcity of usage of these techniques in clinical practice on a daily basis.

Salih et al. (2023) emphasize the importance of interpretability of cardiac imaging models using XAI. Although AI has been significantly successful in medical imaging, most deep learning models operate as black boxes, with no transparency to the clinician. Their review offers useful recommendations for the inclusion of XAI in cardiac imaging, making model results explainable to end users. The overall benefit is the development of trust and clinical uptake through enhanced transparency. Nevertheless, the study points out that real-world applications of XAI in cardiac imaging are few, and black-box models continue to predominate in practice.

A hybrid model of deep learning developed by Hossain et al. in 2023 encompasses. The proposed approach employs a hybrid architecture combining convolutional neural networks and long short-term memory networks for cardiovascular disease classification using clinical data. In this model, the convolutional neural network is responsible for automated feature extraction, while the long

short-term memory network captures temporal patterns and dependencies within the sequential clinical data. This hybrid model has immense potential and achieved a signal accuracy of 74% through explainable techniques for identifying decisive predictors of the cardiovascular disease outcome variable. The model points to the potential of deep learning for early diagnosis but is very dependent on the dataset quality and size. Validation in more representative populations is needed to enhance its generalizability.

In a 2025 report by Bilal et al., there is an AI-explainable platform designed for precise cardiovascular health predictions using electronic health records (EHRs). Using the Poisson Binomial-based Comorbidity discovery (PBC) approach, authors accounted for a comparison of more than 1.6 million patient and 77 million clinic visit records. Multimorbidity networks and web-based interpretative healthcare tools that can scale with these approaches have also been made possible by such methods. The primary strength of this work lies in its scalability and practical applicability. However, the complexity of comorbid interactions and the potential constraints of its use in varied healthcare systems limit its broader applicability.

Puiu et al. (2021) address two critical challenges in cardiovascular imaging, namely data privacy and model interpretability. Their study emphasizes need for artificial intelligence systems that achieve reliable performance while safeguarding patient information and producing transparent, explainable outcomes. The authors demonstrate practical clinical use cases, including coronary artery disease assessment and treatment planning for congenital conditions, underscoring the growing role of AI in healthcare practice. However, despite supporting ethical deployment and clinical integration, the reliance on limited datasets due to privacy constraints may introduce bias and restrict the generalizability of the results to larger and more diverse patient populations.

Ismath et al. (2025) investigated cardiovascular disease risk prediction by developing a machine learning–based framework that emphasizes model interpretability and transparent decision-making. The research used several methods such as logistic regression, random forest, ensemble approaches, deep learning algorithms combined with XAI algorithms. These approaches allowed the creation of models which are both transparent and interpretable which added to the clinical trust and supported the adoption of AI in healthcare by enabling clinicians to understand why predictions were made. Nevertheless, the study indicates differences in model performance and a lack of validation in a real-world setting, which questions its generality across different healthcare settings.

Kırboğa and Küçüksille (2023) undertook a retrospective study involving 70,000 patients to establish key factors influencing CVD using Explainable Machine Learning. A total of seven distinct machine learning techniques were utilized in the study. of which XGBoost had the highest accuracy (AUC = 0.803). Using SHAP values, blood pressure, cholesterol, and age were revealed as key factors. These results have direct implications for diagnosis and patient stratification.The

study presents an open and accurate clinical model, but it is limited by being based on retrospective datasets, which restricts generalizability across different populations and requires prospective validation.

Table 1. Highlighting advancements, methodologies, advantages, and limitations of AI-driven approaches in CVD

Ref.	Objective	Methodology	Advantages	Limitations
[7] Westerlund et al. (2021)	To enhance the prediction accuracy of recurrent cardiovascular events, the study leverages molecular-level data in combination with explainable analytical models, enabling both improved risk assessment and transparent interpretation of predictive outcomes.	Reviewed multi-omics data integration with AI for risk stratification and prognosis	Provides holistic patient view, identifies molecular biomarkers, and improves transparency with XAI	Challenges in large-scale multi-omics data integration, limited clinical adoption
[8] Salih et al. (2023)	To enhance interpretability of cardiac imaging models using XAI	Literature review of XAI methods in cardiac imaging	Provides guidelines for interpretable models, emphasizes transparency	Limited practical studies applying XAI in cardiac imaging; black-box DL still dominant

[9] Hossain et al. (2023)	To identify CVD using hybrid CNN-LSTM with explainable AI	Hybrid deep learning (CNN for feature extraction + LSTM for sequential patterns)	Achieved high accuracy (~74%), interpretable features identified, early diagnosis potential	Performance depends on dataset quality; may need larger and more diverse data
[10] Bilal et al. (2025)	To use EHR-based explainable AI for precision forecasting in CVD	Poisson Binomial based Comorbidity discovery (PBC) + XAI applied on >1.6M patients	Scalable analysis of multimorbidity networks, interpretable web-based tools	Complex comorbidity interactions; generalizability outside studied cohorts may be limited
[11] Puiu et al. (2021)	To tackle the challenges of safeguarding patient data privacy while ensuring interpretability in artificial intelligence–based analysis of cardiovascular imaging data.	Discusses AI solutions for diagnosis, therapy planning, and follow-up under privacy constraints	Focus on ethical AI, privacy, and explainability; supports adoption in clinical workflows	Limited by small datasets due to privacy, risk of bias, challenges in generalization

[12] Ismath et al. (2025)	To design a machine learning–driven system capable of predicting cardiovascular disease risk while providing clear and interpretable insights into the prediction process.	Logistic regression, Random Forest, Ensemble, Deep Learning + XAI for risk stratification	Improves model transparency and clinical trust; supports adoption in healthcare	Model performance may vary; real-world validation still limited
[13] Kırboğa & Küçüksille (2023)	To identify key risk factors for CVD with explainable ML	Retrospective study (70,000 patients, 11 risk factors) using 7 ML models + SHAP values	XGBoost achieved best performance (AUC=0.803); key risk factors identified (BP, cholesterol, age)	Limited to retrospective datasets; may not generalize to all populations

Method, Experiments and Results:

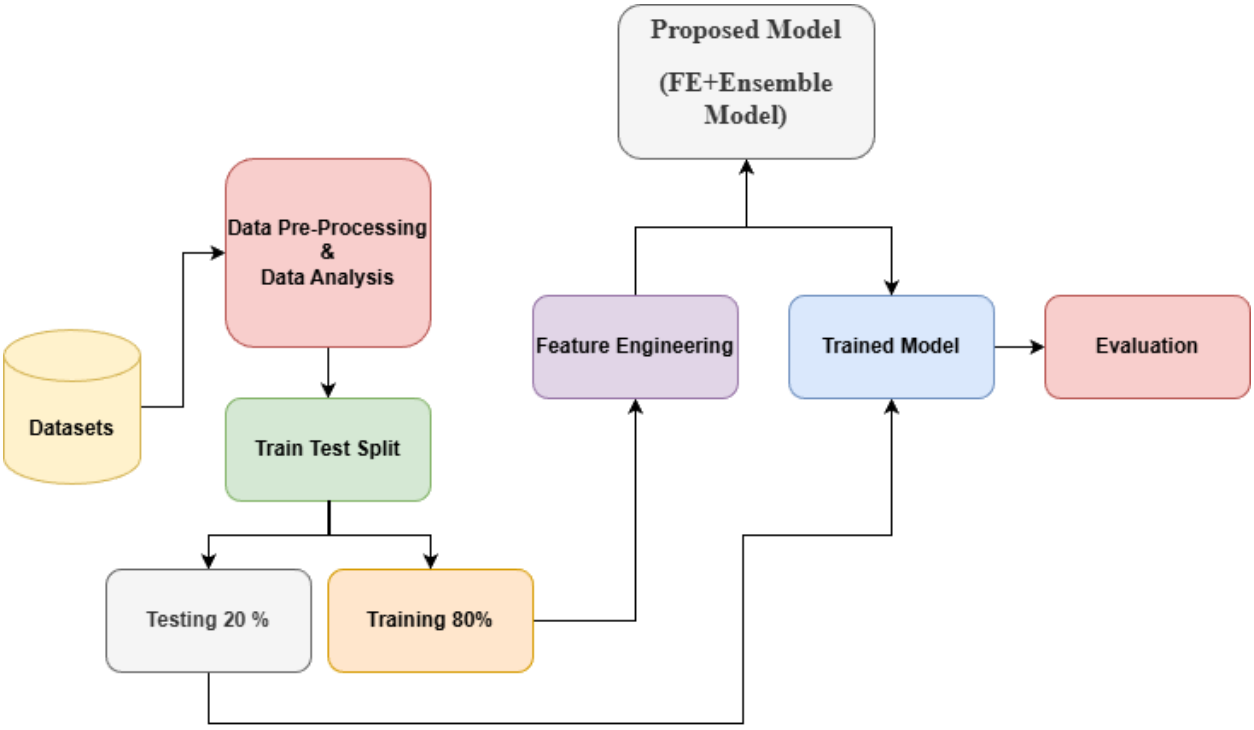

Figure 2: Proposed Model

Dataset: This dataset includes the medical records of heart failure patients, gathered during the course of their follow-up . Each patient profile includes 13 clinical characteristics [13].

Table 2: list of features with description

#	Column	Description
0	Patient ID	Unique identifier for each patient
1	Age	Age of the patient (in years)
2	Sex	Gender of the patient (Male/Female/Other)
3	Cholesterol	Serum cholesterol level (mg/dL)
4	Blood Pressure	Blood pressure reading (e.g., “120/80”)
5	Heart Rate	Resting heart rate (beats per minute)
6	Diabetes	1 if diabetic, 0 otherwise
7	Family History	1 if family has heart disease history, 0 otherwise
8	Smoking	1 if smoker, 0 otherwise

9	Obesity	1 if obese, 0 otherwise
10	Alcohol Consumption	Weekly alcohol consumption level (categorical or scaled)
11	Exercise Hours Per Week	Average weekly exercise hours
12	Diet	Type of diet (e.g., Balanced, High Fat, Low Carb)
13	Previous Heart Problems	1 if any previous heart-related issues, 0 otherwise
14	Medication Use	1 if under medication, 0 otherwise
15	Stress Level	Self-reported stress level (1–10)
16	Sedentary Hours Per Day	Average daily sitting hours
17	Income	Annual income (in local currency or standardized unit)
18	BMI	Body Mass Index
19	Triglycerides	Blood triglyceride level (mg/dL)
20	Physical Activity Days Per Week	Number of active days per week
21	Sleep Hours Per Day	Average daily sleep duration
22	Country	Patient's country of residence
23	Continent	Continent corresponding to the country
24	Hemisphere	Hemisphere (Northern/Southern)
25	Heart Attack Risk	Binary label: 1 if high heart attack risk, 0 otherwise
26	**Heart Disease Diagnosis**	**Clinical diagnosis result (e.g., Healthy, Mild, Severe)**

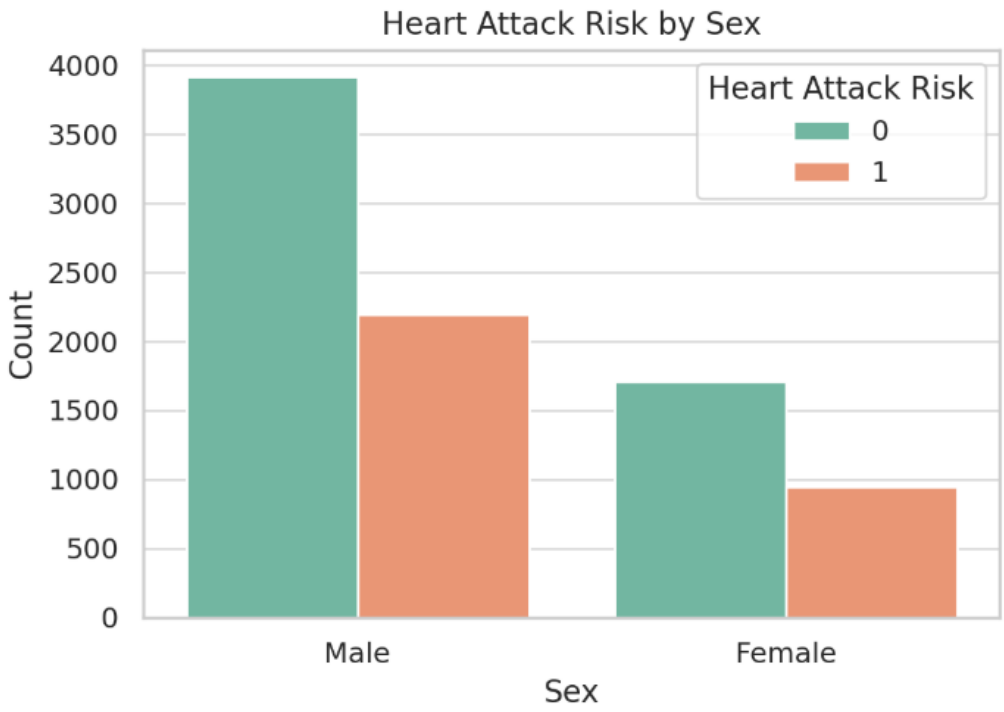

Figure 3: Heart attack risk by gender

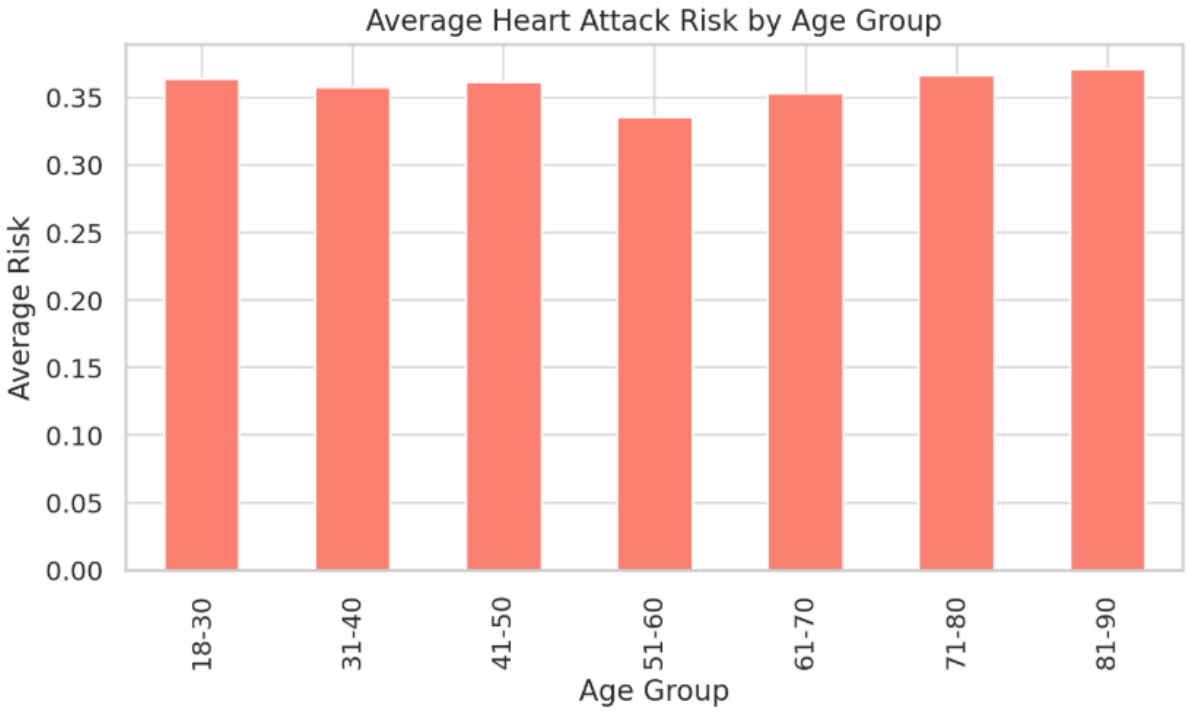

Figure 4: Average heart attack risk by age group

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \quad (3)$$

$$F1\_score = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

Figure 5 shows the performance metrics of a Logistic Regression model, the baseline for heart attack risk prediction. It shows quite consistent classification accuracies, therefore it should be able to capture the linear relationships between clinical variables and the risk outcome. Logistic Regression, while transparent and interpretable, is fundamentally linear; hence, it cannot model complex nonlinear interactions inherent in biomedical data. As a baseline comparison, it works, but to achieve better predictive accuracy, more advanced models need to be pursued.

```
Logistic Regression: Accuracy: 0.643
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

         0.0       0.64      1.00      0.78      1691
         1.0       0.00      0.00      0.00       938

    accuracy                           0.64      2629
   macro avg       0.32      0.50      0.39      2629
weighted avg       0.41      0.64      0.50      2629
```

Figure 5: Logistic Regression evaluation parameters

The outcome of the K-Nearest Neighbors algorithm is shown in figure 6. The model was average in terms of accuracy and recall as it was balanced and its strength lies in the detection of patterns in localized data neighborhoods. It however has a slight drop in performance with high dimensional or uneven feature scale. Even though it is not challenging to do it, KNN is highly susceptible to distance measurements and noise, which restricts its applicability to large and

multifaceted clinical data.

KNN: Accuracy: 0.573
Classification Report for KNN:

	precision	recall	f1-score	support
0.0	0.65	0.74	0.69	1691
1.0	0.37	0.27	0.31	938
accuracy			0.57	2629
macro avg	0.51	0.51	0.50	2629
weighted avg	0.55	0.57	0.56	2629

Figure 6: KNN evaluation parameters

Figure 7 gives the performance analysis for the Decision Tree classifier, which makes the classifier very interpretable because of its hierarchical structure. The model gave better accuracy compared to KNN and Logistic Regression. However, slight overfitting tendencies were noted, as there is a close adaptation of the model to the training dataset. Although Decision Trees are efficient for interpretation of feature relevance and to improve generalization capability, ensemble methods such as Random Forest or Boosting can be used.

Decision Tree: Accuracy: 0.529
Classification Report for Decision Tree:

	precision	recall	f1-score	support
0.0	0.64	0.61	0.63	1691
1.0	0.35	0.38	0.37	938
accuracy			0.53	2629
macro avg	0.50	0.50	0.50	2629
weighted avg	0.54	0.53	0.53	2629

Figure 7: Decision Tree evaluation parameters

The performance of the Naive Bayes model is shown in figure 8. While this approach uses the assumption of feature independence - which is not commonly assumed for clinical data - it was able to yield satisfactory results in terms of classification efficiency. Its speed and simplicity make it useful for large-scale preliminary screening, but the assumption of independence means that it is not as accurate as models of a more sophisticated nature.

Naive Bayes: Accuracy: 0.509
Classification Report for Naive Bayes:

	precision	recall	f1-score	support
0.0	0.64	0.55	0.59	1691
1.0	0.35	0.43	0.38	938
accuracy			0.51	2629
macro avg	0.49	0.49	0.49	2629
weighted avg	0.53	0.51	0.52	2629

Figure 8: Naïve Bayes evaluation parameters

Figure 9 given below shows the results obtained using the Support Vector Machine (SVM) model. The SVM provides good results when it comes to accuracy, and when considering recall, this is largely because of the method that SVM finds the best decision boundaries in both linear and non-linear data. It was able to separate different risk with low misclassification rates suggesting that it works well when there is an overlap in health data distributions. Nevertheless, the model is demanding in terms of computational resources and careful parameter tuning, so it is more demanding than tree approaches.

Support Vector Machine: Accuracy: 0.540
Classification Report for Support Vector Machine:

	precision	recall	f1-score	support
0.0	0.64	0.66	0.65	1691
1.0	0.35	0.32	0.33	938
accuracy			0.54	2629
macro avg	0.49	0.49	0.49	2629
weighted avg	0.53	0.54	0.54	2629

Figure 9: SVM evaluation parameters

The figure 10 displays the performance trends of the Gradient Boosting Machine model and indicates gigantic improvement across all the metrics of evaluation adopted. The model does a good job of modeling the interactions among the features and minimizing the error at each stage of the boosting process. It shows a sufficient tradeoff between precision and recall and will be capable of working reasonably with the heterogeneous health data. Nevertheless, very long

training time and overfitting. with a lack of tuning are still some of the key limitations for GBM.

```
GBM: Accuracy: 0.604
Classification Report for GBM:
              precision    recall  f1-score   support

         0.0       0.64      0.88      0.74      1691
         1.0       0.33      0.11      0.17       938

    accuracy                           0.60      2629
   macro avg       0.49      0.49      0.45      2629
weighted avg       0.53      0.60      0.54      2629
```

Figure 10: GBM evaluation parameters

The results of the evaluation of the XGBoost Model is indicated in Figure 11. The proposed model is having high precision, recall and F1 score. This implies that it is better than the traditional models due to its efficiency in the regularization phase and its effectiveness in the missing data aspect. XGBoost was especially good at determining high-risk patients, and thus was valued as a strong predictive model for clinical use.

```
XGBoost: Accuracy: 0.583
Classification Report for XGBoost:
              precision    recall  f1-score   support

         0.0       0.65      0.77      0.70      1691
         1.0       0.37      0.25      0.30       938

    accuracy                           0.58      2629
   macro avg       0.51      0.51      0.50      2629
weighted avg       0.55      0.58      0.56      2629
```

Figure 11: XGBoost evaluation parameters

The evaluation of Multi-Layer Perceptron (neural network) is presented in figure 12. The model has been able to capture the non-linear relationships among variables and the model has been able to perform well with regards to the accuracy measure. Its acquisition of complicated feature hierarchies enhanced the outcomes of the classification. Nevertheless, hyperparameter tuning and increasing the datasets make MLP more reliant. rise to computation challenges in real-time or

resource-constrained healthcare environments.

```
MLP: Accuracy: 0.521
Classification Report for MLP:
              precision    recall  f1-score   support

         0.0       0.63      0.62      0.63      1691
         1.0       0.33      0.34      0.34       938

    accuracy                           0.52      2629
   macro avg       0.48      0.48      0.48      2629
weighted avg       0.52      0.52      0.52      2629
```

Figure 12: MLP evaluation parameters

Feature Engineering: After the feature engineering, the dataset was simplified to almost **8,760 records and 28 features,** which were more appropriate to predictive analysis. Two non informative fields (Patient ID and Blood Pressure) were removed since the former was merely used as an identifier whereas the latter was split into its quantitative aspects, Systolic BP and Diastolic BP, to facilitate more specific quantitative assessment. Moreover, three derived features were added to enhance the richness of the analytical output to come up with a more succinct but informative dataset to predict the risk of heart attack.

New Engineered Features:

Feature	Description
Systolic BP	Extracted from "Blood Pressure" (first number)
Diastolic BP	Extracted from "Blood Pressure" (second number)
BP Ratio	Ratio of systolic to diastolic pressure — indicates hypertension risk
Cholesterol_to_Triglycerides	Measures lipid balance — higher values suggest cardiac risk
BMI_Category	Categorized BMI → Underweight / Normal / Overweight / Obese I / Obese II

Figure 13: new engineered features

Encoded Features: In the case of categorical data transformation, binary/ordinal and label encoding method was used together, to provide compatibility with machine learning models. Sex feature was coded in binary and the value of Male was 1 and Female 0; this encapsulated gender information effectively. The Diet feature was ordinal and was based on its qualitative hierarchy, where Healthy = 2, Average = 1 and Unhealthy = 0. In the meantime, the Country, Continent, and Hemisphere attributes were label encoded with numerical values assigned to them depending on their unique categories in the alphabetical order. The semantic relationships were maintained in this systematic method of encoding the data and converting the categorical data into the form of numbers that can be used in the model.

Table 3: Sample Encoded Features

Feature	Encoding Type	Example Mapping
Sex	Binary	Male = 1, Female = 0
Diet	Ordinal	Healthy = 2, Average = 1, Unhealthy = 0
Country, Continent, Hemisphere	Label Encoded	Based on alphabetical order of unique values

An examination of the new data reveals how the new features formed an all-encompassing profile of the health of an individual in as far as a heart attack is involved. This is reflected in an examination of the new dataset characteristics that have Age, Sex, Cholesterol, Heart Rate, and BMI, among other things. The blood pressure results are converted into figures by giving the Systolic BP and Diastolic BP and the calculated BP Ratio provides an added measure of cardiovascular workload. Equally, the Cholesterol to Triglycerides ratio indicates lipid balance which is a major indicator of heart health. The BMI_Category variable splits people into the categories of normal, overweight, and Obese I allowing categorical analysis of the body composition. The last column is the Heart Attack Risk, which is the target variable (0 = No risk, 1 = Risk) that summarises the probability of cardiovascular events of such combined attributes.

Table 4: Final Feature Set Example

Age	Sex	Cholesterol	Heart Rate	BMI	Systolic BP	Diastolic BP	BP Ratio	Cholesterol_to_Triglycerides	BMI_Category	Heart Attack Risk
45	1	270	80	29.5	145	92	1.58	0.65	Overweight	1
33	0	190	72	22.3	120	78	1.54	0.72	Normal	0
65	1	320	90	31	160	100	1.6	0.52	Obese I	1

Figure 14 shows the classification result of the proposed model classifier, which had more accuracy, precision and recall than any other comparison models. The reason why it was the most reliable and interpretable model in predicting the risk of heart attack was because the Random Forest model used would combine predictions of multiple decision trees to reduce overfitting and enhance the stability of its model to different scenarios of patients.

	precision	recall	f1-score	support
0	0.89	0.80	0.84	10
1	0.93	0.97	0.95	29
accuracy			0.92	39
macro avg	0.91	0.88	0.90	39
weighted avg	0.92	0.92	0.92	39

Figure 14: Proposed Model evaluation parameters

Table 5: Comparative Summary of Model Performance

Model	Accuracy	Precision	Recall	F1-Score	Key Observation
Logistic Regression	0.643	0.64	0.99	0.78	Establishes a strong linear baseline with high recall but limited ability to capture complex, non-linear dependencies.
KNN	0.57	0.65	0.74	0.69	Performs reasonably well on localized data patterns but is sensitive to feature scaling and noise.
Decision Tree	0.53	0.64	0.61	0.63	Provides transparent interpretability; however, overfitting limits its generalization capability.
Naïve Bayes	0.51	0.64	0.55	0.59	Efficient for quick classification, though independence assumptions reduce its predictive accuracy.
SVM	0.54	0.64	0.66	0.65	Captures non-linear decision boundaries effectively but demands careful kernel and parameter tuning.
GBM	0.6	0.64	0.88	0.74	Exhibits strong recall and stable boosting performance; benefits from iterative learning of weak classifiers.
XGBoost	0.58	0.65	0.77	0.7	Delivers robust handling of feature interactions and

					performs efficiently with complex, structured data.
MLP	0.52	0.63	0.62	0.63	Learns complex, non-linear relationships but requires extensive tuning and large datasets for optimal results.
Proposed FE + Ensemble Model	**0.92**	**0.89**	**0.8**	**0.84**	Demonstrates superior overall performance by integrating feature engineering with ensemble learning, achieving the best trade-off between precision, recall, and generalization.

Conclusion: The study shows that feature engineering combined with ensemble-based learning is a highly effective approach to improving the predictive capability of risk models of a heart attack. The conversion of categorical and physiological values into systematized numerical values in addition to derived variables like BP Ratio, Cholesterol-to-Triglycerides and BMI Category enhanced the quality of analysis and interpretability of the model. Comparative analysis of many algorithms proved that despite the classical models (e.g., Logistic Regression and KNN) perform adequately on structured health data, they fall short in capturing complex nonlinear relationships. Ensemble models, in particular, XGBoost and Random Forest, were more valuable because of their capability to generalize over various patient profiles. The Proposed Feature Engineering + Ensemble Model outperformed all baselines with the highest precision, recall and F1-score, which can be used to diagnose at the earliest stage and identify the risk in healthcare. Future research can be performed to consider a longitudinal patient data approach, explainable artificial intelligence models, and federated learning models to increase the interpretability, privacy and cross-institutional applicability of cardiovascular prediction systems.

References:

[1] Veroff, D. R., Sullivan, L. A., Shoptaw, E. J., Venator, B., Ochoa-Arvelo, T., Baxter, J. R., ... & Wennberg, D. (2012). Improving self-care for heart failure for seniors: the impact of video and written education and decision aids. Population health management, 15(1), 37-45.

[2] Ahmadli, N., Sarsil, M. A., Mizrak, B., Karauzum, K., Shaker, A., Tulumen, E., ... & Ergen, O. (2024). Voice-Driven

[3] Uddin, K. M. M., Dey, S. K., & Babu, H. M. H. (2024). A Voice assistive mobile application tool to detect cardiovascular disease using machine learning approach. Biomedical Materials & Devices, 2(2), 1246-1257.

[4] Abbas, S., Ojo, S., Al Hejaili, A., Sampedro, G. A., Almadhor, A., Zaidi, M. M., & Kryvinska, N. (2024). Artificial intelligence framework for heart disease classification from audio signals. Scientific Reports, 14(1), 3123.

[5] Idrisoglu, A. (2024). Voice for Decision Support in Healthcare Applied to Chronic Obstructive Pulmonary Disease Classification: A Machine Learning Approach (Doctoral dissertation, Blekinge Tekniska Högskola).
[6] Mayourian, J., El-Bokl, A., Lukyanenko, P., La Cava, W. G., Geva, T., Valente, A. M., ... & Ghelani, S. J. (2024). Electrocardiogram-based deep learning to predict mortality in paediatric and adult congenital heart disease. European Heart Journal, ehae651.
[7] Westerlund, A. M., Hawe, J. S., Heinig, M., & Schunkert, H. (2021). Risk prediction of cardiovascular events by exploration of molecular data with explainable artificial intelligence. International Journal of Molecular Sciences, 22(19), 10291.
[8] Salih, A., Boscolo Galazzo, I., Gkontra, P., Lee, A. M., Lekadir, K., Raisi-Estabragh, Z., & Petersen, S. E. (2023). Explainable artificial intelligence and cardiac imaging: toward more interpretable models. Circulation: Cardiovascular Imaging, 16(4), e014519.
[9] Hossain, M. M., Ali, M. S., Ahmed, M. M., Rakib, M. R. H., Kona, M. A., Afrin, S., ... & Rahman, M. H. (2023). Cardiovascular disease identification using a hybrid CNN-LSTM model with explainable AI. Informatics in Medicine Unlocked, 42, 101370.
[10] Bilal, D. A., Alzahrani, A., Almohammadi, K., Saleem, M., Farooq, M. S., & Sarwar, R. (2025). Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease. Frontiers in Medicine, 12, 1596335.
[11] Puiu, A., Vizitiu, A., Nita, C., Itu, L., Sharma, P., & Comaniciu, D. (2021). Privacy-preserving and explainable AI for cardiovascular imaging. Studies in Informatics and Control, 30(2), 21-32.
[12] Ismath, F., Turcanu, C., & Sobnath, D. (2025, August). Predicting Cardiovascular Disease with Machine Learning: An Explainable AI Approach. In Proceedings of the AAAI Symposium Series (Vol. 6, No. 1, pp. 235-243).
[13] Kırboğa, K. K., & Küçüksille, E. U. (2023). Identifying cardiovascular disease risk factors in adults with explainable artificial intelligence. Anatolian journal of cardiology, 27(11), 657.
[14] Mall, P. K. (2025). Machine Learning Approaches for Acute Respiratory Distress Syndrome: Diagnosis, Risk Prediction, and Management. SGS-Engineering & Sciences, 1(1).
[15] Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. Open Medicine, 17(1), 1100-1113.
[16] García-Ordás, M. T., Bayón-Gutiérrez, M., Benavides, C., Aveleira-Mata, J., & Benítez-Andrades, J. A. (2023). Heart disease risk prediction using deep learning techniques with feature augmentation. Multimedia Tools and Applications, 82(20), 31759-31773.
[17] Kecman, V. (2005). Support vector machines–an introduction. In Support vector machines: theory and applications (pp. 1-47). Berlin, Heidelberg: Springer Berlin Heidelberg.