Predicting Cardiovascular Disease with Machine Learning: An Explainable AI Approach

Fathima Zaineb Ismath[1], **Cristina Turcanu**[1], **Drishty Sobnath** [1]

[1] Heriot-Watt University Dubai
fathimazainab106@gmail.com, cristina.turcanu@hw.ac.uk, d.sobnath@hw.ac.uk

Abstract

Cardiovascular disease affects a huge number of individuals globally. Early detection and accurate risk prediction can reduce its impact. Traditional risk factors drive the urgency of developing predictive models that can effectively identify individuals at high risk. This study explores multiple machine learning techniques, including logistic regression, random forests, ensemble model and deep learning algorithms to develop an effective and explainable Cardiovascular Disease (CVD) risk prediction system. A key innovation in this work is the integration of risk stratification and Explainable AI (XAI) techniques to improve model transparency and interpretability in predictions, enabling healthcare professionals to understand the rationale behind model decisions. This is critical for gaining clinical trust and promoting the adoption of AI-driven diagnostic tools in healthcare settings.

Introduction

This research explores the prediction of cardiovascular disease and the assessment of patient risk using machine learning (ML) techniques. The heart is one of the most essential organs in the human body, responsible for ensuring continuous blood circulation. According to the World Health Organization, cardiovascular diseases are one of the leading causes of mortality worldwide, accounting for approximately 17.9 million deaths annually (Organization 2019). In the United States alone, an individual experiences a heart attack every 40 seconds (for Disease Control and Prevention 2024). The increasing prevalence of risk factors such as diabetes, obesity, and sedentary lifestyles has made the early detection of CVD more critical. This study explores the application of multiple machine-learning techniques to enhance early cardiovascular disease diagnosis. Predictive models can help healthcare professionals in identifying individuals at high risk and enable the development of personalised treatment strategies. By improving early detection, these models can also help to reduce the financial strain on healthcare systems while improving patient outcomes. Furthermore, this research integrates Explainable Artificial Intelligence to enhance the interpretability of model predictions, ensuring that physicians can confidently utilise machine learning tools in clinical decision-making.

Related Work

Recent progress in CVD prediction have utilised machine learning techniques, ranging from traditional supervised models to deep learning and XAI methods. Several studies have explored different approaches to enhance predictive accuracy and interpretability in healthcare applications. In 2019, Krishnani et al. investigated supervised learning models, comparing Random Forest, Decision Tree, and K-Nearest Neighbours for CVD prediction (Krishnani, Gupta, and Rao 2019). Their results demonstrated that Random Forest achieved the highest accuracy (96.8%) on the Framingham dataset. However, the study lacked a broader analysis of alternative feature selection techniques and model interpretability methods such as Shapley Additive Explanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME). Two popular explainable AI techniques are LIME and SHAP. The behaviour of the trained ML model can be explained by these techniques. LIME offers the user localised insights, while SHAP provides a more comprehensive overview—a crucial feature for complicated models. Similarly, Latha et al. evaluated ensemble techniques such as bagging, boosting, and stacking, showing that ensemble classifiers significantly improved weak classifiers' performance (Latha and Jeeva 2019). However, the study had a trade-off between accuracy and interpretability. Deep learning techniques have also been explored for CVD risk stratification. Schlesinger and the team highlighted the benefits of deep learning models for clinical risk assessment, employing methods like SHAP and Gradient-weighted Class Activation Mapping (Grad-CAM) for enhanced interpretability (Schlesinger and Stultz 2020). Kavitha et al. proposed a hybrid model that combined Decision Tree and Random Forest, demonstrating superior accuracy over individual models but lacking an in-depth evaluation of the benefits of hybridization (Kavitha, Prasad, and Sundaram 2021). Recent efforts in explainable AI have further refined model interpretability. Some authors utilized SHAP and LIME to improve trust in ML-driven predictions, advocating for a balance between accuracy and transparency in medical AI applications (Guleria, Thomas, and Kim 2022); (Bizimana, Chen, and Ravi 2024). These studies highlight the growing need for interpretable and scalable CVD prediction models, bridging the gap between performance and clinical usability.

Methodology

The research methodology follows a structured pipeline designed to enhance the robustness and clinical interpretation. A critical analysis of previous studies identified research gaps, emphasizing the need for integrating explainable AI and risk stratification into machine learning models. This study employs a two-way approach, exploring both traditional and deep learning techniques for CVD prediction. Figure 1 illustrates the methodology for developing the proposed CVD prediction model, starting with data acquisition and processing, followed by training and testing all six developed models. The best-performing model, selected after hyperparameter tuning, was integrated with XAI and deployed via a Streamlit-based application for risk stratification.

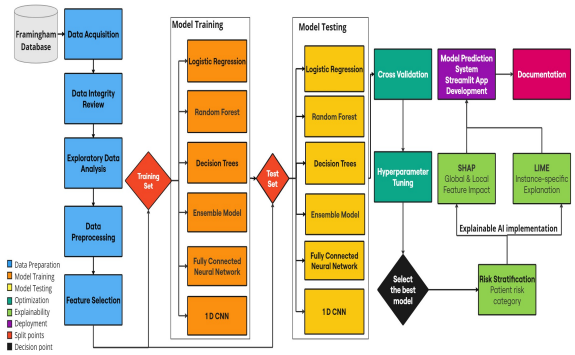

Figure 1: Research methodology

Implementation

Dataset

Data Source: The dataset used in this study is the Framingham Heart Study (FHS) dataset (Bhardwaj 2022) which provides an extensive record of cardiovascular risk factors. Compared to the Cleveland UCI dataset (UCI Machine Learning Repository 2023), the FHS dataset includes lifestyle attributes such as smoking and stroke history, offering a more holistic representation of cardiovascular risk. The dataset consists of 15 predictor variables and one target variable (TenYearCHD) represented as a binary outcome. For clarity, Table 1 encapsulates the dataset attributes with their respective descriptions and data type.

Data Preprocessing

Categorical variables such as male, currentSmoker, prevalentStroke, and prevalentHyp were already encoded in binary format, eliminating the need for further transformation. The education column was excluded from the dataset due to its negative correlation with the target variable. The dataset was split into training (80%) and testing (20%) subsets using stratified sampling to maintain the same proportion of CVD-positive cases across both sets. The dimensions of the full dataset, along with the respective partitions, were validated before model training. Stratified K-Fold Cross-Validation (K = 5) was implemented. StandardScaler was used to normalize numerical features with varying scales, ensuring uniform contributions across different variables. Standardization was particularly crucial for distance-based models such as Logistic Regression and Feedforward Neural Networks (FNN), as it optimises gradient-based learning and accelerates convergence. The scaler was fitted exclusively on the training data and then applied to both training and testing sets to prevent data leakage.

Attribute	Data Type	Description
age	int64	Age of the patient (in years)
male	int64	Gender of the patient (0 = Female, 1 = Male)
education	float64	Level of education
BPMeds	float64	Whether the patient is on blood pressure medication
prevalentStroke	int64	History of stroke (0 = No, 1 = Yes)
prevalentHyp	int64	History of hypertension
diabetes	int64	Whether the patient is diagnosed with diabetes
currentSmoker	int64	Current smoker status
cigsPerDay	float64	Number of cigarettes smoked per day
totChol	float64	Total serum cholesterol level (mg/dL)
sysBP	float64	Systolic blood pressure (mmHg)
diaBP	float64	Diastolic blood pressure (mmHg)
BMI	float64	Body Mass Index (BMI) of the individual
heartRate	float64	Resting heart rate (beats per minute)
glucose	float64	Blood glucose level (mg/dL)
TenYearCHD	int64	CHD occurrence within 10 years (0 = No, 1 = Yes)

Table 1: Dataset Attributes Description

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the structure and characteristics of the dataset, ensuring readiness for the modelling phase. Key statistical properties were examined, including age distribution, gender composition, smoking habits, and systolic blood pressure levels. The dataset consists of individuals aged 32 to 70, with a median age of 49 years. Nearly 50% of the population were current smokers. The systolic blood pressure values ranged between 110 and 180 mmHg, highlighting potential hypertension risks. For further analysis, risk level distribution was assessed and revealed a significant class imbalance, with only 15% of individuals developing CHD within ten years. Additionally, trends in cholesterol, blood pressure, and glucose levels were explored, showing a strong correlation between age and increasing cardiovascular risk factors. A scatter plot analysis of smoking behaviour and cholesterol levels indicated that while heavy smokers exhibited higher cholesterol levels in some cases, no direct correlation was present. A correlation heatmap was generated to visualize feature relationships, identifying age and systolic blood pressure as the strongest predictors of CVD. Additionally, strong associations were observed between prevalent hypertension and blood pressure levels, as well as diabetes and glucose levels, confirming well-documented medical risk factors.

Model Architecture and Training

This study implemented and compared a diverse range of ML models, including traditional, ensemble, and deep learning models to evaluate CVD prediction from multiple methodological perspectives. Six different models were developed in this study. Each model, along with its implementation and justification is discussed briefly below. Balanced class weighting strategies were employed in all models to ensure adequate representation of CVD-positive cases while training.

Traditional Models

Logistic Regression (LR): It is a widely used linear classification algorithm. It was implemented with stratified K-Fold cross-validation ($K = 5$) to ensure robustness across training subsets. To address class imbalance, a weighted loss function was applied, assigning a higher penalty to misclassified positive cases. The model was trained with an adaptive decision threshold fine-tuned to maximise recall.

Random Forest (RF): This model consisted of 200 decision trees. It was configured with Gini impurity as the splitting criterion. To prevent overfitting, each tree was limited to a maximum depth of 10, and the minimum number of samples required for a split was set to 10.

Decision Trees (DT): This classifier was designed to partition the dataset hierarchically using Gini impurity. The depth of the tree was restricted to 10 to mitigate overfitting and pruning techniques were applied to remove nodes with minimal information gain.

Ensemble Model: An ensemble learning approach was adopted, combining the strengths of multiple classifiers to enhance predictive performance. The models used and justification for using are given below:

Gradient Boosting (200 estimators, learning rate = 0.05) for sequential model refinement. Random Forest for handling non-linearity in tabular data. Logistic Regression for interpretability and linear relationships.

A soft-voting strategy was used to aggregate predictions, producing a final probability score that balances the contributions of all base models.

Deep Learning Models

Feedforward Neural Network (FNN)

FNN is a deep learning model consisting of multiple layers of neurons that learn hierarchical representations of input features. Deep learning models are highly prone to over-fitting, especially when trained on imbalanced medical datasets. To mitigate this as part of model development, various regularization and optimization strategies were incorporated to improve generalisation and training stability. LeakyReLU was applied in the hidden layers instead of standard ReLU as the activation function to prevent vanishing gradients, ensuring better learning capacity for diverse medical datasets. The final output layer of the trained model used a Sigmoid activation function to provide probability-based predictions for binary classification. Figure 2 visualizes the architecture of the developed FNN model. It consists of an input layer (256 neurons), followed by three hidden layers (128, 64, and 32 neurons respectively), and an output layer with one neuron for binary classification. Each layer is fully connected, allowing the network to learn complex relationships between features. Table 2 summarises the neural network architecture, stating the type of layers, the output shape and the number of parameters in each layer.

Layer Type	Output Shape	Parameters
Dense	(None, 256)	3,840
Layer Normalization	(None, 256)	512
LeakyReLU	(None, 256)	0
Dropout	(None, 256)	0
Dense	(None, 128)	32,896
Layer Normalization	(None, 128)	256
LeakyReLU	(None, 128)	0
Dropout	(None, 128)	0
Dense	(None, 64)	8,256
Layer Normalization	(None, 64)	128
LeakyReLU	(None, 64)	0
Dropout	(None, 64)	0
Dense	(None, 32)	2,080
Dense	(None, 1)	33
Total Parameters	-	**144,005**

Table 2: Summary of the Neural Network Architecture

Regularization Techniques: Dropout Layers (40% and 30%) were added to prevent over-fitting and enhance generalization. Layer Normalization was applied after each dense layer to stabilise activations, improving training convergence. Higher penalties were assigned to misclassified CVD-positive cases, addressing class imbalance.

Optimization: The Adam optimizer was chosen for its adaptive learning capabilities, with a learning rate of 0.0005, empirically determined for stable convergence. Training was halted if validation loss did not improve for 10 consecutive epochs. This ensures efficient training without over-fitting.

Training Configuration: The model was trained using a batch size of 64 for a maximum of 100 epochs. The decision threshold was optimized to prioritise recall, reducing false negatives, which is critical for identifying high-risk CVD patients.

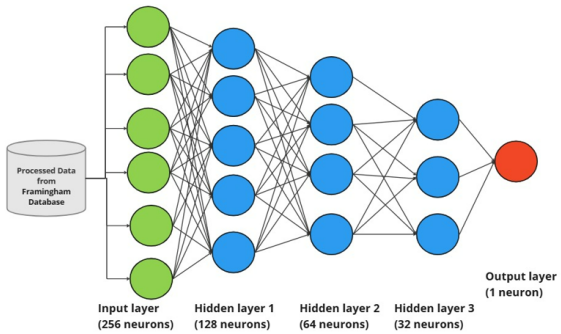

Figure 2: Feedforward Neural Network Model Architecture

One-Dimensional Convolutional Neural Network (1D-CNN) Convolutional Neural Networks (CNNs) are widely recognised for their ability to perform pattern recognition and feature extraction. While traditionally used for image and sequential data, 1D-CNNs have demonstrated their effectiveness in analysing structured tabular datasets by capturing local feature dependencies (Khan Mamun and Elfouly 2023). This study incorporates a 1D-CNN model to improve cardiovascular disease prediction by learning feature representations from patient health records.

Figure 3 shows the 1D CNN model architecture, highlighting its convolutional layers and fully connected layers. It has 3 convolutional layers which are responsible for feature extraction, a flattening layer, a dense dropout layer to prevent overfitting and an output layer with a single neuron, producing the final output. The initial stage includes 3 Conv1D layers with 256, 128 and 64 filters, each using kernel sizes of 5 and 3 and the ReLU activation function to introduce non-linearity. Batch normalization is applied after each convolutional layer to stabilize training and improve convergence. To mitigate over-fitting, dropout regularization is employed at rates of 0.3 and 0.4 across different layers.
The network transitions into fully connected layers comprising 256 and 128 neurons, further refining feature representations. The output layer consists of a single neuron with a sigmoid activation function, producing probability scores for binary classification.

The Adam optimizer, with an empirically determined learning rate of 0.0005, is used to minimise the binary cross-entropy loss function. The model is trained for 35 epochs using a batch size of 32, ensuring efficient learning while maintaining stability.
This approach balances interpretability, accuracy, and efficiency, making it suitable for real-world applications.

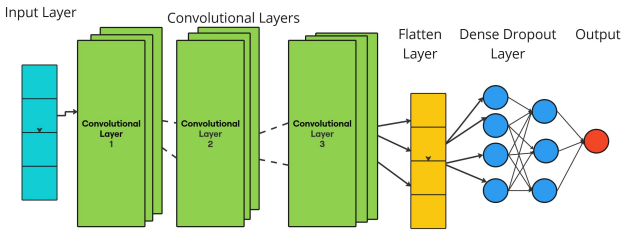

Figure 3: 1D CNN Model Architecture

Explainability and Risk Stratification

SHapley Additive exPlanations (SHAP) values were computed to provide a global explanation of feature importance. Age and systolic blood pressure were identified as the most influential predictors. Local explanations using LIME graph were generated for individual predictions, visualizing key factors contributing to model decisions (Lundberg and Lee 2017).
Risk Stratification: The Framingham Risk Score (FRS) framework was employed to categorise patients into 3 main categories:
Low Risk: Less than 10% probability of developing CVD in 10 years.
Intermediate Risk: 10% - 20% probability.
High Risk: Greater than 20% probability.
This classification enables targeted interventions for high-risk individuals.

Model Deployment

The final model was deployed via Streamlit, providing an interactive user interface for real-time CVD risk assessment. The app integrates SHAP visualizations, allowing users to understand the model's decision-making process, rather than just a binary outcome.

Hyper-parameter Optimization and Validation

For optimal performance, various optimization strategies were implemented. Hyperparameter tuning was conducted using GridSearchCV for traditional models, optimizing tree depth, learning rates, and class weighting. Deep learning models incorporated dropout regularization and early stopping mechanisms to prevent over-fitting. Stratified K-Fold Cross-Validation was applied to ensure model robustness across different training subsets.

Experimental Setup

The models were trained on a system equipped with an Intel Core i7-1165G7 processor and 16GB RAM. Although deep learning typically benefits from GPU acceleration, optimizations such as batch processing and early stopping ensure efficient training. The implementation utilised Python, leveraging key libraries:
Data Handling and Visualization: Pandas, NumPy, Matplotlib, Seaborn, Plotly were used for data preprocessing and visualising trends and distributions.
Machine Learning: TensorFlow, Keras were used for training and evaluating the deep learning models. Scikit-Learn provided tools for machine learning.
Explainability Tools: Primary tools used were SHAP and LIME.
Deployment: Streamlit was used for deployment.

Results and Evaluation

Testing Methodology

To assess the effectiveness of the proposed model, a structured evaluation approach was employed. Given the critical nature of healthcare applications, recall was prioritised as the primary metric to maximise the detection of high-risk patients. This approach ensures that potential CVD cases are not overlooked, aligning with the healthcare objective of minimising false negatives. The false positive cases can later be addressed through further medical testing.

Evaluation Metrics

Accuracy: Measures the proportion of correctly classified cases relative to the total number of cases.

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

Precision: Represents the proportion of correctly identified positive cases among all predicted positives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (Sensitivity): Indicates the model's ability to detect actual positive cases correctly, essential for minimising missed diagnoses.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score: The harmonic mean of precision and recall, balancing both metrics, useful in imbalanced datasets.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): Measures the model's ability to distinguish between classes, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR).

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \tag{5}$$

Confusion Matrix: Table 3 provides a concise overview of the Confusion Matrix, illustrating the classification outcomes.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 3: Confusion Matrix Representation

Model Type	**Recall**	**Accuracy**	**F1-Score**	**Precision**	**AUC-ROC**
Logistic Regression	64.9%	71.7%	41.0%	30.0%	72.5%
Random Forest	32.4%	79.2%	32.1%	31.9%	72.5%
Decision Tree	42.3%	68.3%	28.8%	21.9%	58.1%
Ensemble Model	25.2%	79.6%	27.3%	29.8%	71.4%
FNN	67.6%	64.1%	36.3%	24.9%	70.84%
1D CNN	72.9%	62.1%	36.9%	24.7%	69.7%

Table 4: Performance Metrics Comparison of Different Models

Results and Model Insights

Table 4 compares the performance metrics of the developed models used to evaluate each model's effectiveness in predicting cardiovascular disease. 1D-CNN model achieved the highest recall (72.9%) among the models evaluated, making it the most suitable for detecting high-risk patients. The FNN model followed closely in performance. In contrast, ensemble models and Random Forest exhibited higher accuracy but suffered from low recall, making them less effective for healthcare applications where detecting high-risk cases is a priority.

While traditional models like Logistic Regression and Random Forest demonstrated reasonable class separation, they exhibited lower recall, making them less suitable for identifying high-risk patients. The CNN was ultimately selected as the optimal model due to its ability to balance predictive accuracy with recall, ensuring that more high-risk cases were correctly identified.

Figure 4 compares the performance metrics of various models. It highlights that deep learning models (FNN, CNN) prioritise recall, improving high-risk CVD detection whereas traditional models emphasise accuracy but have lower recall scores. The ensemble model balances multiple metrics, showing moderate performance across all evaluation criteria.

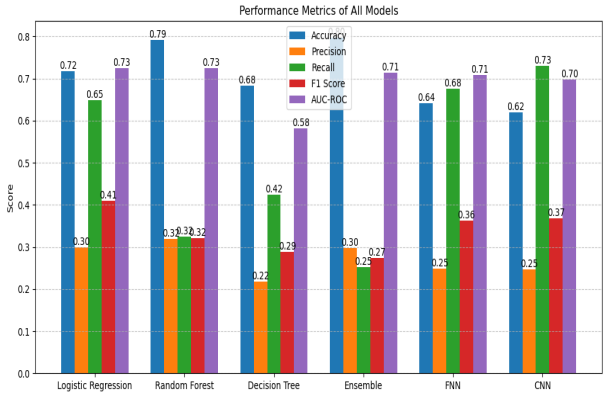

Figure 4: Comparison of Performance Metrics Across Models

ROC-AUC Curve Analysis

To further analyse model effectiveness, the AUC-ROC curves were plotted for all models. The curves in Figure 5 compare the AUC-ROC scores for various models, demonstrating their classification performance. It indicates that Logistic Regression and Random Forest classifiers achieved the highest AUC-ROC scores, followed by ensemble methods. These curves provide a comparative view of the model's discriminative ability.

Computational Efficiency

While model performance is a crucial aspect, computational efficiency also plays a vital role in real-world applications. Table 5 shows the training and prediction times for different models in seconds. It highlights the computational efficiency of each model, with Logistic Regression being the fastest and 1D CNN taking the longest time.

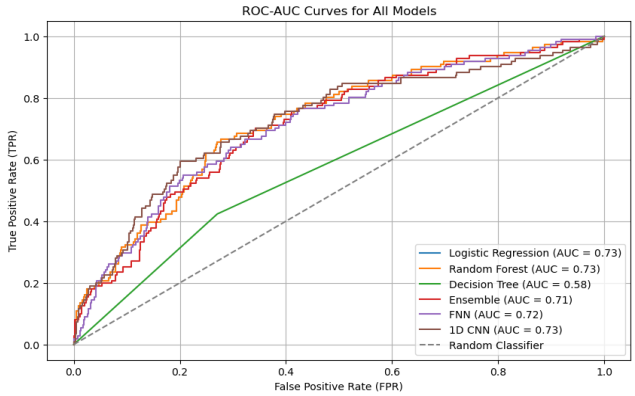

Figure 5: ROC-AUC Curve Comparison for Model Evaluation

Model Type	Training Time	Prediction Time
Logistic Regression	0.01	0.0045
Random Forest	1.36	0.1724
Decision Trees	0.03	0.0062
Ensemble Model	3.51	0.1204
fnn	11.27	0.7245
1D cnn	87.83	1.4495

Table 5: Training and Prediction Time for Different Models

Hyper-parameter Tuning

Hyperparameter tuning was applied to improve predictive performance by adjusting key parameters, including learning rate, depth of trees, and regularization techniques. To assess the impact of hyperparameter optimization, performance metrics were evaluated before and after tuning across different models. Table 6 summarises the model performance after tuning. FNN and Logistic Regression showed the most notable improvements in accuracy, increasing to 71.0% and 85.93%, respectively. However, some models exhibited a trade-off between recall and precision. For instance, Logistic Regression improved significantly in accuracy and precision (72.22%) but suffered a decline in recall (11.71%), reducing its sensitivity to detecting CVD cases. Similarly, Random Forest's recall dropped to 2%, despite an increase in precision. Conversely, models such as Ensemble Learning and Decision Trees maintained a balance between recall and precision post-tuning. Neural networks retained strong recall scores (59.4% and 46.85%), indicating their ability to detect positive cases while achieving a moderate boost in precision. Overall, hyperparameter tuning significantly enhanced accuracy and precision in most models, albeit at the cost of recall in some cases. The trade-off highlights the importance of balancing sensitivity and specificity when optimizing predictive models for CVD classification.

External Validation

To evaluate model robustness on unseen data, external validation was conducted using the UCI Heart Disease dataset. As shown in Table 7, the CNN model achieved an accuracy of 73.93% and an AUC-ROC of 78.07%, demonstrating strong discriminative ability. The precision-recall balance suggests the model generalises well to external data, maintaining predictive power across datasets. However, the UCI dataset has some demographic limitations. It is derived primarily from the Cleveland Medical Centre in the U.S., with minimal representation of other ethnic or socioeconomic groups. This restricts generalizability to global populations. Originally collected in the 1980s to 1990s, the dataset consists of more male than female patients, which may affect model performance across genders. Key lifestyle factors such as smoking habits and family history of disease are absent.

Model Type	Recall	Accuracy	F1-Score	Precision	AUC-ROC
Logistic Regression	11.71%	85.93%	41.03%	72.22%	55.45%
Random Forest	02%	85%	04%	67%	72.53%
Decision Trees	43%	68%	29%	22%	58%
Ensemble Model	34.23%	78.83%	32,90%	31.67%	71.52%
FNN	59.46%	71.04%	38.37%	28.33%	72.11%
1D CNN	46.85%	74.45%	35.74%	28.89%	67.62%

Table 6: Performance Metrics Comparison of Different Models (After Hyper-parameter Tuning)

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
CNN	73.93%	70.27%	74.82%	72.47%	78.07%

Table 7: External Validation Results on UCI Dataset

SHAP Analysis

To interpret the predictions made by the model, SHAP values were used to assess the impact of each feature on the model output. As illustrated in Figure 6, it resolves the black-box issue of many ML models. SHAP provides a global explanation by quantifying how individual features contribute to the prediction, enhancing the transparency and interpretability of the model.

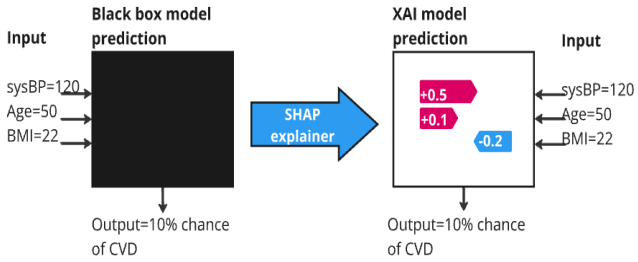

Figure 6: SHAP Explanaibilty over black box models

Figure 7 presents the global SHAP summary plot, which highlights the importance of different features. The x-axis represents the SHAP value, indicating the degree to which

each feature affects the model's prediction, while the y-axis lists the features in descending order of impact. Higher absolute SHAP values suggest greater influence on the model's decision-making process.

Positive SHAP values push the prediction toward a higher risk of cardiovascular disease, whereas negative SHAP values push it toward a lower risk. From the SHAP plot, *age* and *systolic blood pressure (sysBP)* were identified as the most significant predictors of cardiovascular disease, followed by *cigarettes per day* and *male*. Features such as *BMI and BP medication (BPMeds)* had comparatively lower contributions. The colour gradient indicates feature values, where red represents high feature values and blue denotes low values. For instance, older age and higher systolic blood pressure positively correlate with a higher risk of cardiovascular disease.

By leveraging SHAP, this study enhances model interpretability, ensuring that the decision-making process remains transparent for healthcare practitioners, thereby fostering trust in AI-driven medical diagnostics.

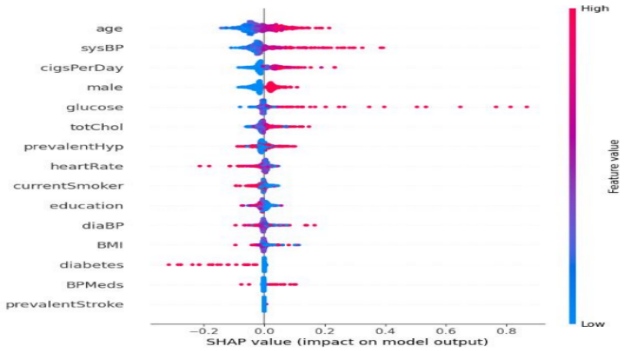

Figure 7: Global SHAP Analysis

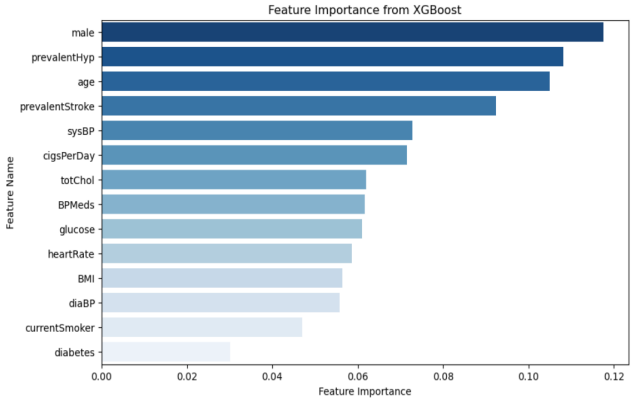

Figure 8: Feature importance graph

SMOTE Balanced Dataset

The Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset and improve the model performance. This section evaluates the impact of SMOTE on model performance compared to the original dataset. The original dataset exhibited an imbalance in CHD cases, with significantly more negative (non-CHD) cases than positive ones. After applying SMOTE, the dataset became more balanced, ensuring that models received equal representation of both classes during training. This improved the model's ability to generalize across different CHD risk levels. Table 8 shows the model performance after addressing the class imbalance using SMOTE Techniques. It is observed that applying SMOTE significantly improved model performance, particularly in terms of recall, precision and AUC-ROC, which are critical for early detection of CHD. The CNN model outperforms other models in recall, achieving the highest performance of 94%.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	65.9%	61.1%	87.1%	71.8%	73.1%
Random Forest	79.7%	78.8%	81.3%	80.0%	73.2%
Decision Tree	75.8%	74.9%	78.7%	75.6%	84.5%
Ensemble	**88.2%**	**88.2%**	88.0%	**88.1%**	**95.0%**
FNN	70.3%	63.7%	94.0%	76%	83.8%
CNN	73.5%	66.5%	**94.2%**	78.0%	86.9%

Table 8: Model Performance Metrics (SMOTE Applied)

Comparison with State-of-the-Art Models

To validate the model's performance, it was benchmarked against existing state-of-the-art models. Krishnan, Magalingam, and Ibrahim (2021) developed an enhanced RNN-GRU architecture with SMOTE balancing on the Framingham dataset, achieving a recall of 96%. Other similar studies leveraging recent state-of-the-art models on the UCI dataset were explored. Rahman et al. (2024) introduced a self-attention-based transformer achieving 96.5% accuracy, leveraging multi-head attention for improved feature learning. It focuses on accuracy, overlooking other important metrics such as recall and F1 score. Moreover, the proposed CNN model in this research integrates SHAP-based explainability alongside evaluation metrics such as recall. It achieves a competitive 94% recall on the Framingham dataset. Hence, it offers both predictive strength and clinical interpretability for real-world deployment.

Conclusion

Research Contributions and Findings

This study presents a structured and explainable approach to cardiovascular disease prediction, integrating machine learning, deep learning, and risk stratification techniques. The key findings of this research include:

The 1D-CNN model demonstrated the highest recall (73%), making it the most effective for identifying high-risk patients. Traditional classifiers such as Random Forest and Ensemble models exhibited high accuracy but struggled with recall, making them less suitable for healthcare applications. SHAP and LIME analysis provided insights into feature importance, confirming that age and systolic blood pressure were key predictors. The Framingham Risk Score was integrated to categorize patients into different risk levels.

Research Questions Addressed

The key research questions in this study were systematically addressed through feature analysis, model comparisons, and explainability techniques. The first question focused on identifying significant risk factors associated with cardiovascular diseases, which was explored using feature importance analysis from XGBoost and Chi-Squared tests. The findings confirmed that age, systolic blood pressure (sysBP) and total cholesterol levels (totChol) were among the most influential predictors, aligning with established clinical research.

The study also examined the most predictive features contributing to model accuracy and how they could be effectively identified and validated. Figure 8 illustrates the feature importance rankings derived from the XGBoost model, highlighting the most influential predictors in the dataset. Male, prevalent hypertension (prevalentHyp), and age were shown as the top three most important features influencing CVD risk. Prevalent stroke, systolic blood pressure (sysBP), and cigarette consumption (cigsPerDay) also contributed significantly to the predictions. This validation confirmed the reliability of machine learning in identifying key risk factors while maintaining a balance between accuracy and interpretability.

By incorporating explainable AI techniques such as SHAP and LIME, the study ensured the model decisions were interpretable and trustworthy, aiding in more informed decision-making in clinical settings. Compared to existing research, the models developed in this study demonstrated good performance in some metrics. For example, the Logistic Regression model outperformed Suhatril et al. (2024) with higher recall (87%) compared to their 84%. Decision Trees had a similar AUC-ROC of 58%. Additionally, the study by Anderies et al. (2022) showed a recall of 63% for the decision trees model, whereas the model in this study achieved a notable recall of 78%, as recall was prioritized. Without SMOTE balancing, the logistic regression model achieved an AUC-ROC (72% vs. 70%), performing better Suhatril et al. (2024). These results were achieved without any data alteration. The findings may vary due to differences in data splits, feature selection, and preprocessing techniques. The results emphasize the importance of recall in CVD detection (Suhatril et al. 2024) and highlight the value of explainable AI using SHAP (Guleria, Thomas, and Kim 2022).

Limitations

The Framingham dataset used in this study presents a class imbalance, with only 15% of individuals developing Coronary Heart Disease (CHD) within ten years, which could lead to biased model performance. To mitigate this, SMOTE technique was applied as an experimental approach to balance the training set by generating synthetic samples. However, since synthetic data generation remains a topic of debate, the final model evaluation was conducted on the original imbalanced dataset to preserve clinical interpretability. Although external validation demonstrated the model's robustness, its reliance on the Framingham dataset limits its applicability across more diverse populations.

Future Work

Future research could enhance this study by incorporating richer data sources such as electrocardiogram medical imaging, genetic markers, and wearable sensor data to improve CVD risk assessment. Patient history, laboratory results, and lifestyle factors could be seamlessly integrated via a developed web or mobile application for real-time risk assessment, improving accessibility. Additionally, expanding the dataset to include a broader demographic representation across different regions would also improve model generalizability, ensuring unbiased predictions for real-world applications.

References

Anderies, A.; Tchin, J.; Putro, P.; Darmawan, Y.; and Gunawan, A. 2022. Prediction of Heart Disease UCI Dataset Using Machine Learning Algorithms. *Engineering, MAthematics and Computer Science (EMACS) Journal*, 4: 87–93.

Bhardwaj, A. 2022. Framingham Heart Study Dataset. Kaggle. Last accessed: 2025-03-08.

Bizimana, J.; Chen, L.; and Ravi, S. 2024. Transparency in AI-Based Medical Diagnosis: Balancing Accuracy and Explainability. *Health Informatics Journal*, 45(1): 50–70.

for Disease Control, C.; and Prevention. 2024. Heart Attack Facts Statistics.

Guleria, V.; Thomas, A.; and Kim, J. 2022. Explaining Black-Box AI Models in Healthcare: A SHAP and LIME Approach. *AI in Medicine*, 33(5): 210–225.

Kavitha, R.; Prasad, M.; and Sundaram, S. 2021. Hybrid Machine Learning Models for Cardiovascular Disease Prediction. *Biomedical AI Research*, 22(3): 155–168.

Khan Mamun, M. M. R.; and Elfouly, T. 2023. Detection of Cardiovascular Disease from Clinical Parameters Using a One-Dimensional Convolutional Neural Network. *Bioengineering (Basel, Switzerland)*, 10(7): 29.

Krishnan, S.; Magalingam, P.; and Ibrahim, R. 2021. Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction. *International Journal of Electrical and Computer Engineering*, 11: 5467–5476.

Krishnani, R.; Gupta, S.; and Rao, K. 2019. Supervised Learning Models for CVD Prediction. *International Journal of AI Research*, 34(5): 45–60.

Latha, C.; and Jeeva, S. 2019. Boosting and Stacking for Cardiovascular Disease Risk Prediction. In *Proceedings of the International Conference on AI in Medicine*, 1–7. AAAI Press.

Lundberg, S. M.; and Lee, S. I. 2017. A Unified Approach to Interpretable Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Organization, W. H. 2019. Cardiovascular Diseases Fact Sheet.

Rahman, A.; Alsenani, Y.; Zafar, A.; Ullah, K.; Rabie, K.; and Shongwe, T. 2024. Enhancing heart disease prediction using a self-attention-based transformer model. *Scientific Reports*, 14.

Schlesinger, D.; and Stultz, C. 2020. Deep Learning for Cardiovascular Risk Stratification Using SHAP and Grad-CAM. *Journal of Medical AI*, 28(7): 234–245.

Suhatril, R.; Syah, R.; Hermita, M.; Gunawan, B.; and Silfianti, W. 2024. Evaluation of Machine Learning Models for Predicting Cardiovascular Disease Based on Framingham Heart Study Data. *ILKOM Jurnal Ilmiah*, 16: 68–75.

UCI Machine Learning Repository. 2023. Statlog (Heart) Dataset. https://archive.ics.uci.edu/dataset/45/heart+disease. Accessed: 2023-10-31.