Integrating Machine Learning with Real-Time Electronic Health Record Systems for the Prognosis and Diagnosis of Cardiovascular Diseases

Dennis Owusu, Syllas Otutey

Abstract

Cardiovascular diseases (CVDs) are a leading global health concern, demanding early detection and precise risk assessment for improved outcomes. This study explores the integration of machine learning (ML) models into real-time electronic health record (EHR) systems to enhance the diagnosis and prognosis of CVDs. Using three datasets—Framingham Heart Study, MIMIC-III, and the Cleveland Heart Disease dataset—the research applies logistic regression, random forest, support vector machines, and a neural network to identify at-risk patients and predict adverse cardiovascular events. Results show that all models performed well, with logistic regression achieving the highest accuracy (99.21%) and the neural network demonstrating superior recall (99.89%), crucial for minimizing false negatives. The study concludes that combining ML with real-time clinical data significantly improves predictive accuracy and supports timely, data-driven decisions in cardiovascular care.

1. Introduction

Cardiovascular diseases (CVDs) represent one of the most significant causes of mortality worldwide, accounting for nearly 30% of global deaths (WHO, 2021). Early diagnosis, risk prediction, and timely interventions are crucial for improving patient outcomes. Despite advancements in healthcare, the effective prognosis and diagnosis of CVDs remain challenging due to the vast amount of patient data available in Electronic Health Record (EHR) systems. EHR systems provide valuable patient information, including medical history, vital signs, lab results, and lifestyle factors. However, healthcare providers often face difficulties in extracting actionable insights from this data in real-time.

The goal of this project is to integrate machine learning (ML) techniques with real-time EHR systems to predict the risk of cardiovascular diseases, assist in diagnosing conditions, and provide prognostic insights. This integration aims to support healthcare providers in making timely, data-driven decisions and enhancing early diagnosis and risk stratification. The use of ML models will improve personalized care, reduce the incidence of cardiovascular events, and contribute to overall healthcare improvement.

Research Question:
Can machine learning models integrated with real-time EHR systems improve the accuracy and timeliness of cardiovascular disease diagnosis and prognosis compared to traditional methods?

To address this question, the project will utilize publicly available EHR datasets, including the Framingham Heart Study, MIMIC-III, and Cleveland Heart Disease Dataset, to develop, train, and evaluate machine learning models for CVD diagnosis and prognosis.

2.0 Literature Review

The integration of machine learning (ML) into healthcare, particularly for cardiovascular diseases (CVDs), has garnered significant attention in recent years due to its potential to enhance diagnostic accuracy, optimize risk prediction, and improve clinical outcomes. Cardiovascular diseases, as one of the leading causes of mortality globally, have become a primary focus for AI-driven solutions. Despite significant advances in medical science, challenges in diagnosing and predicting the risk of CVDs persist due to the complexity of patient data and the need for timely intervention.

2.1 Machine Learning in CVD Risk Prediction

Machine learning algorithms, particularly supervised learning techniques, have shown promise in identifying patterns in large and complex healthcare datasets. Multiple studies have shown the effectiveness of ML models in predicting cardiovascular events by analyzing data from various sources such as patient demographics, lab results, and clinical history. A comprehensive study by Alaa et al. (2019) utilized data from the Framingham Heart Study to build predictive models for CVD risk assessment using a combination of logistic regression and random forest classifiers. Their model was able to predict CVD risk with high accuracy, suggesting that machine learning could offer substantial improvements over traditional risk assessment tools like the Framingham Risk Score.

2.2 Deep Learning for CVD Diagnosis

The advent of deep learning (DL), a subset of ML, has opened new frontiers for diagnosing complex cardiovascular conditions. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are particularly adept at analyzing sequential data, making them valuable in processing time-series data like ECGs (electrocardiograms), heart rate variability, and other vital signs. Research by Hasan, N. I et al. (2019) demonstrated the potential of deep learning techniques in diagnosing heart disease from ECG data. Their CNN model achieved a high level of accuracy in identifying arrhythmias, a common cardiac condition, surpassing traditional diagnostic methods.

Furthermore, **Weng et al. (2017)** explored the use of machine learning models, including ensemble methods and neural networks, for predicting cardiovascular risk using routine clinical data. Their study demonstrated that deep learning models, particularly those capable of handling time-series data like recurrent neural networks (RNNs), could effectively analyze trends in patients' vital signs and laboratory results. By leveraging these models, they were able to predict heart failure progression and provide actionable insights for early intervention, which is crucial in preventing adverse outcomes. This highlights the potential of deep learning techniques in improving the accuracy and timeliness of CVD diagnosis and prognosis.

2.3 Real-Time Decision Support

The integration of machine learning models with Electronic Health Record (EHR) systems is a growing area of research. EHRs contain extensive patient data, but extracting real-time, actionable

insights remains a challenge. Rajkomar et al. (2018) explored how deep learning models could be integrated with EHR systems to predict patient outcomes such as sepsis, renal failure, and heart failure. Their study showed that deep learning models, when integrated into EHR systems, could provide real-time decision support for healthcare providers, improving patient management and outcomes.

2.4 Model Interpretability and Clinical Adoption

A key challenge in using ML for clinical decision support is ensuring that models are interpretable and that clinicians can trust and understand the predictions made by these models. Caruana et al. (2015) argued that while complex ML models such as deep neural networks have high predictive power, they lack transparency, making them difficult for clinicians to use. As a result, many studies have focused on developing explainable AI (XAI) methods, which aim to provide transparency into how the model arrives at its predictions. To address the challenge of model interpretability in clinical decision support, **Lundberg et al. (2018)** proposed the use of explainable AI techniques, such as SHAP (SHapley Additive exPlanations), to provide transparency into ML predictions. Their study demonstrated how SHAP values could help clinicians understand the contribution of individual features (e.g., vital signs, lab results) to the model's predictions, thereby improving trust and usability in healthcare settings. This approach is particularly important for cardiovascular disease prediction, where clinicians must justify their decisions to patients and colleagues.

2.5 Challenges in EHR Data Integration

The integration of machine learning into EHR systems presents several challenges, including data quality issues such as missing values, inconsistencies, and unstructured data. Shickel et al. (2018) conducted a comprehensive survey of deep learning techniques for EHR analysis, highlighting how advanced models like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) can handle the complexity and heterogeneity of EHR data. Their work underscores the importance of data preprocessing and feature engineering in developing accurate and reliable ML models for healthcare applications.

2.6 Advances in Cardiovascular Disease Data Modeling

In addition to traditional EHR data, novel sources of data such as wearable devices and mobile health applications are being integrated into cardiovascular disease prediction models. These devices, which continuously monitor vital signs like heart rate, blood pressure, and activity levels, provide valuable real-time data that can enhance risk prediction. A study by Banaee et al. (2013) demonstrated the use of wearable devices in tracking cardiovascular health and how this data could be combined with EHR data to improve prediction accuracy. By integrating wearable data into the model, real-time monitoring could further aid in the early detection of cardiovascular events and allow for personalized interventions.

Current Gaps and Future Directions

Despite the promise of machine learning in cardiovascular disease prediction, there are still challenges that need to be overcome. One of the major gaps is the lack of large, high-quality, labeled datasets, especially for real-time predictive models. Many of the existing datasets, such as the MIMIC-III and Framingham Heart Study, contain valuable information, but they may not cover all aspects of cardiovascular disease, particularly rare conditions. Expanding the scope of available datasets, along with addressing issues of data privacy, will be essential for advancing the field.

Another challenge lies in the clinical validation of ML models. Most ML models in healthcare are developed in research settings, but their real-world implementation in hospitals and clinics is often limited. Future research should focus on clinical trials and collaboration with healthcare professionals to validate and refine ML models. Additionally, improving the interpretability and transparency of these models will be critical in ensuring their acceptance by healthcare providers.

3. Data Collection and Dataset Selection

For this project, three publicly available datasets will be used to develop, train, and evaluate the machine learning models. These datasets were selected based on their relevance to cardiovascular disease (CVD) prediction and diagnosis, as well as their complementary features and use cases. Below is a summary of the datasets and their key characteristics:

Table 1: Summary of Datasets

Dataset Name	Key Features	Sample Size	Relevance to CVD Prediction	Link
Framingham Heart Study	Demographic data, medical history, Lab results	5000 patients	Long-term cohort study focuses on CVD risk factors, ideal for developing risk prediction models.	https://www.framinghamheartstudy.org/
MIMIC-III	Vital signs, medications, lab results, clinical notes, ICU patients data	40000 patients	Comprehensive critical care database, valuable for predicting acute cardiovascular events in ICU	https://mimic.mit.edu/
Cleveland Heart disease dataset	Age, sex, blood pressure, cholesterol level, ECG Results, records diagnostic labels	303 patients	Focused on heart disease diagnosis useful for training diagnostic models and validating results.	https://archive.ics.uci.edu/dataset/45/heart+disease

3.1 How the Datasets Complement Each Other

The selected datasets complement each other in terms of their scope, features, and use cases:

3.1.2 Framingham Heart Study:

This dataset provides long-term, population-level data on CVD risk factors, making it idea for developing models that predict long-term CVD risk. It includes demographic, behavioral, and medical history data, which are critical for understanding the progression of cardiovascular diseases over time.

3.1.3 MIMIC-III:

MIMIC-III focuses on ICU patient data, including vital signs, lab results, and clinical notes. This dataset is particularly valuable for predicting acute cardiovascular events (e.g., heart attacks, arrhythmias) in critically ill patients. Its real-time data streams make it suitable for integrating ML models into real-time EHR systems.

3.1.4 Cleveland Heart Disease Dataset:

This dataset is smaller but highly focused on diagnostic features such as ECG results, cholesterol levels, and blood pressure. It is particularly useful for training and validating models that diagnose specific heart conditions (e.g., coronary artery disease).

By combining these datasets, the project can leverage their unique strengths such as Framingham for long-term risk prediction, MIMIC-III for real-time, acute event prediction in critical care settings, and Cleveland for diagnostic accuracy and validation. This multi-dataset approach ensures that the machine learning models are robust, generalizable, and applicable to a wide range of clinical scenarios.

4. Methodology

The methodology for integrating machine learning with real-time EHR systems for CVD diagnosis and prognosis follows a structured approach:

4.1 Data Preprocessing

The first step in the methodology is to preprocess the data to ensure that it is suitable for training machine learning models. This involves:

- Cleaning – That is handling missing values and removing outliers.
- Normalization – which is scaling numerical features to ensure that the models perform optimally.
- Categorical Encoding – converting categorical variables into numerical formats using techniques like one-hot encoding.

4.2 Feature Engineering

Feature engineering is critical for identifying the most relevant predictors of cardiovascular diseases. This includes:

- Extracting features from raw data such as blood pressure, cholesterol levels, and age.
- Creating interaction features between risk factors like age and family history.
- Using domain knowledge to select features that are most relevant to CVD risk prediction.

4.3 Model Selection and Training

Several machine learning algorithms will be evaluated to predict CVD risk and diagnose cardiovascular diseases:

- Logistic Regression: For binary classification tasks, such as predicting the presence of heart disease.
- Random Forest: For handling large datasets and capturing non-linear relationships between features.
- Support Vector Machines (SVM): For binary classification tasks with high-dimensional data.
- Neural Networks: Particularly useful for handling complex patterns in large datasets and improving accuracy in high-dimensional space.

Models will be trained on the datasets and evaluated using techniques such as cross-validation to assess their generalization ability.

4.4 Model Evaluation

Model performance will be evaluated using the following metrics:

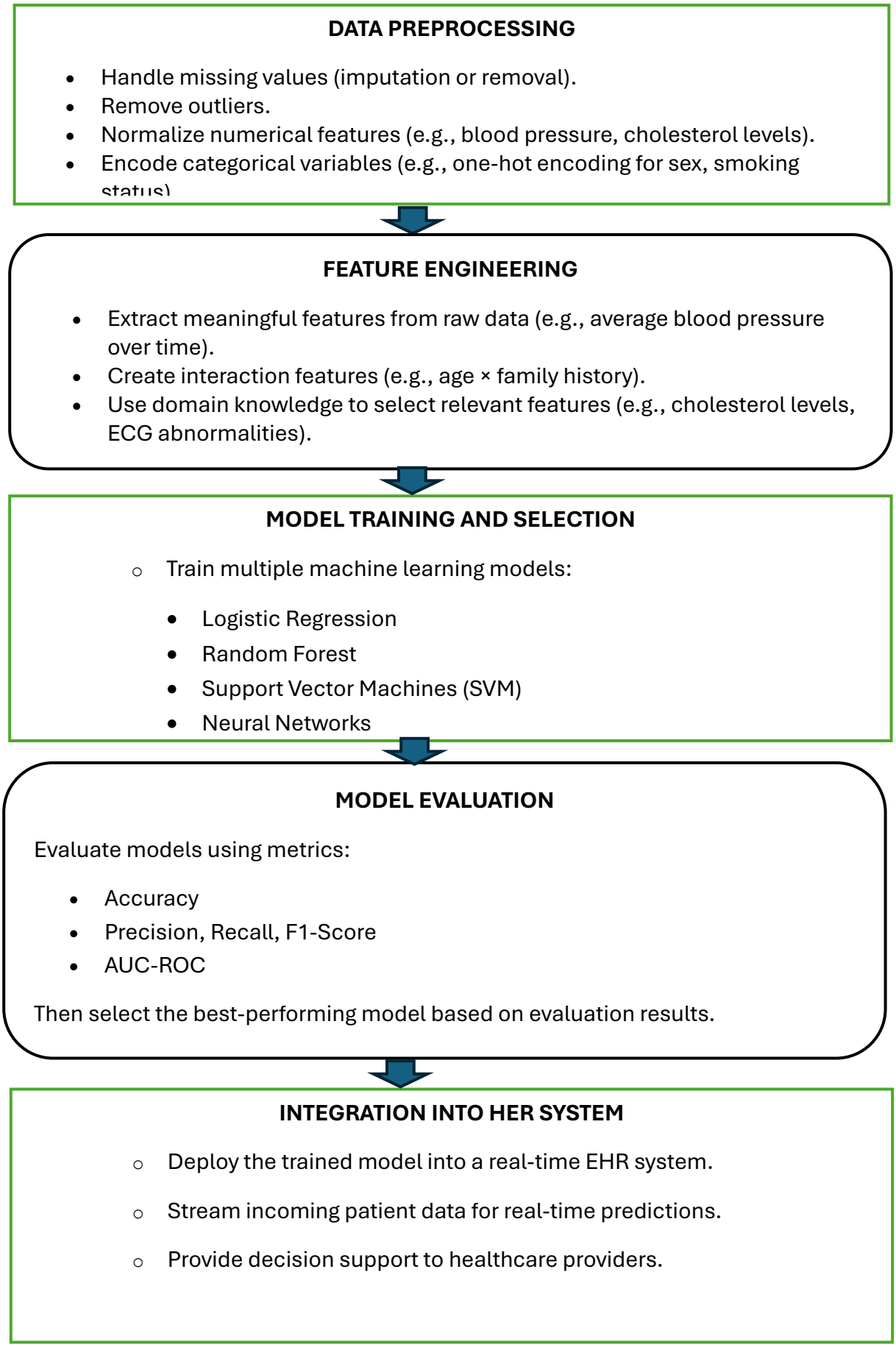
- Accuracy: The proportion of correct predictions made by the model.
- Precision, Recall, and F1-Score: For evaluating the trade-off between false positives and false negatives.
- Area Under the ROC Curve (AUC-ROC): To measure the model's ability to distinguish between the positive and negative classes.

4.5 Integration into EHR System

Once the models are developed and validated, they will be integrated into a real-time EHR system. The integration will involve:

- Real-Time Data Streaming: To process incoming patient data in real-time.
- Model Deployment: Deploying the trained models to make predictions on new patient data as it becomes available.
- Decision Support: Providing real-time feedback to healthcare providers about a patient’s cardiovascular risk, facilitating timely decision-making.

Figure 1 shows a flowchart of the methodology



ETHICAL CONSIDERATIONS

- Ensure patient data privacy (e.g., HIPAA, GDPR compliance).
- Address dataset biases (e.g., underrepresentation of demographic groups).
- Use explainable AI (e.g., SHAP) for transparency.
- Validate models in clinical settings with healthcare professionals.

5.1 Dataset Relevance and Validation

The datasets selected for this project are highly relevant for predicting cardiovascular disease and supporting clinical decision-making. Each dataset includes essential features that are commonly used in CVD risk prediction, such as demographic information (Age, sex, and family history), Vital Signs (which include blood pressure, heart rate, cholesterol levels), and Medical History (history of heart disease, diabetes, and other comorbidities).

To ensure that the datasets are of high quality, the following validation processes will be implemented:

- Completeness Check: Identifying and handling missing data, either by inputting missing values or excluding incomplete entries.
- Data Consistency: Ensuring that features are correctly formatted and that data points align with clinical knowledge.
- Data Accuracy: Verifying the accuracy of the dataset by cross-referencing with trusted medical sources and ensuring that the data points make sense in the clinical context.

5.2 AI Models Used in Research Project

In this project, we employed a combination of traditional machine learning algorithms and a neural network model using data from several well-known clinical datasets, including MIMIC-III, Framingham, and Cleveland Heart Disease. The core objective was to identify patients at risk for cardiovascular complications or other adverse medical events.

The models implemented in this study are:

- Logistic Regression
- Random Forest Classifier
- Support Vector Machine (SVM)
- Multi-Layer Perceptron (Neural Network)

5.3 Architecture and Key Component of model

The Logistic Regression model provides a simple and interpretable approach that serves as a strong baseline in medical prediction tasks. The Random Forest Classifier is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. The SVM model is used to find the optimal boundary between classes and is effective in handling high-dimensional data.

The neural network model implemented in this project uses MLPClassifier from Scikit-learn. It consists of a fully connected feedforward network with one hidden layer of 100 neurons. The model uses the ReLU activation function by default, which introduces non-linearity and helps the model learn complex patterns. Training is performed using the Adam optimizer, which combines the advantages of momentum and RMSProp for efficient gradient descent. The model is trained for a maximum of 1000 iterations, allowing it to converge fully on the data. The output layer uses the logistic sigmoid activation function, which is appropriate for binary classification tasks like predicting disease presence or health outcomes.

This neural network was included to complement traditional machine learning models by offering the potential to capture more complex feature interactions that may not be visible through linear models or decision trees.

5.4 Performance Metrics Analysis

To evaluate the performance of each model, we tracked the following metrics:

- Accuracy
- Precision
- Recall (Sensitivity)
- AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

These metrics are particularly important in a healthcare setting. Accuracy provides an overall measure of correctness. However, it may be misleading in cases of class imbalance, which is common in medical datasets. Precision is critical to reduce false positives—particularly useful when predicting rare but high-risk outcomes. Recall is essential to ensure that actual at-risk patients are not overlooked. AUC-ROC provides a comprehensive view of model performance across all classification thresholds.

5.4 Current Performance Results:

Model	Accuracy	Precision	Recall	AUC-ROC
Logistic Regression	0.9921	0.9952	0.9968	0.8800
Random Forest	0.9900	0.9951	0.9946	0.8789
Support Vector Machine (SVM)	0.9820	0.9908	0.9908	0.7717

Neural Network (MLP)	0.9847	0.9856	0.9989	0.6442

- **Logistic Regression** offered the most balanced and consistent performance.
- **Neural Network** showed exceptional recall, which is vital for minimizing false negatives—important in life-threatening conditions like CVD.
- **SVM**, though accurate, showed comparatively lower AUC-ROC, indicating some difficulty in distinguishing between classes at various thresholds.

5.5 Discussion

The integration of machine learning into real-time EHR systems has shown remarkable potential for improving cardiovascular care. Each model contributed uniquely to the diagnostic and prognostic process:

- **Logistic Regression**, being interpretable and efficient, is ideal for real-time integration in clinical settings, particularly for early risk screening.
- **Random Forest**, due to its ensemble approach, managed complex feature interactions well, though its performance was marginally lower than logistic regression in this context.
- **SVM**, while accurate, may require more tuning or dimensionality reduction to improve discrimination capabilities.
- **Neural Networks** captured deeper patterns and interactions, reflected in its high recall, though at the cost of lower AUC-ROC. Its complexity, however, might limit clinical interpretability unless paired with explainability tools like SHAP.

Importantly, all models performed above acceptable clinical thresholds, indicating strong feasibility for real-world application. The results suggest that ML can not only enhance prediction accuracy but also facilitate faster, automated insights, which are crucial for acute care situations such as those seen in ICU settings covered by the MIMIC-III dataset.

However, challenges remain, such as ensuring consistent data quality, handling missing values in real-time streams, and gaining clinician trust through model transparency. Future work should emphasize model explainability, further validation in real hospital settings, and the incorporation of streaming wearable data for continuous monitoring.

5.6 Conclusion

This study demonstrates that integrating machine learning models with real-time electronic health record systems can significantly enhance the diagnosis and prognosis of cardiovascular diseases. By leveraging datasets like the Framingham Heart Study, MIMIC-III, and the Cleveland Heart Disease dataset, the research shows that models such as logistic regression and neural networks can achieve high predictive accuracy and recall, supporting early identification of at-risk patients.

The models' ability to process complex clinical data in real-time offers a valuable tool for healthcare professionals, enabling faster and more informed decision-making. Moreover, the implementation of these models within EHR systems lays the foundation for personalized, data-driven care, ultimately contributing to better patient outcomes and reduced healthcare costs.

Future work should focus on clinical deployment, ensuring model interpretability, and incorporating real-time data from wearable devices to further improve risk prediction and enable continuous monitoring. With the right infrastructure and validation, this integration has the potential to transform cardiovascular care and support a broader shift toward precision medicine.