Enhancing Heart Attack Risk Prediction with Clustering and Regression: Insights into Post-Pandemic Vulnerabilities

Yenni Varshini[1], Yamala Prabhu Kumari[2], Adarapu Pavan Kumar[3], Mr. V. V. Vidya Sagar[4]
[1,2,3]*Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India*
[4]*Assistant Professor & Guide, Dept. of CSE (Data Science) Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India*

***Abstract*—Heart attack is one of the leading causes of death worldwide, and early identification of high-risk individuals can significantly reduce fatal outcomes. Traditional methods of diagnosis depend on manual analysis of medical reports and clinical experience, which may sometimes delay timely decision-making. The COVID-19 pandemic has further amplified cardiovascular vulnerabilities, increasing post-acute risks of myocarditis, hypertension, and acute myocardial infarction reinforcing the urgent need for automated, data-driven screening tools aligned with UN Sustainable Development Goal 3 (SDG 3: Good Health and Well-being).**

This paper presents CardioSense, a Clinical Decision Support System (CDSS) that integrates Gaussian Mixture Model (GMM) and K-Means clustering with a calibrated stacking ensemble classifier for heart attack risk prediction. The system is built on a curated subset of the Kaggle Cardiovascular Disease Dataset (top 5,000 records) featuring medical attributes such as age, gender, blood pressure, cholesterol, glucose, ECG-derived parameters, and exercise-induced variables. A hybrid pipeline applies StandardScaler normalization, silhouette-optimal K-Means clustering, SMOTE class balancing, and trains eight classifiers Logistic Regression, Random Forest, KNN, Naive Bayes, SVM, Decision Tree, XGBoost, and a Stacking Ensemble evaluated via 10-fold stratified cross-validation.

A risk scoring mechanism based on correlation and regression is implemented to provide individual risk scores. The stacking ensemble achieves 73.8% accuracy, AUC-ROC of 0.815, sensitivity of 73.6%, specificity of 74.1%, and MCC of 0.48. Note: the lower accuracy relative to studies on small datasets (503 records) reflects the real-world prediction task on a more heterogeneous dataset. SHAP analysis identifies systolic blood pressure, BMI, and age as dominant predictors. A Flask-based CDSS delivers real-time risk scores with clinical recommendations. This project highlights the potential of machine learning in healthcare by assisting doctors in early diagnosis and improving clinical decision-making through a simple, non-invasive system.

***Index Terms*—Heart attack prediction; GMM clustering; Stacking ensemble; SHAP; CDSS; Post-pandemic; SMOTE; XGBoost; Logistic Regression; Random Forest; Flask; Kaggle cardiovascular dataset; SDG 3**

I. INTRODUCTION

Cardiovascular disease (CVD) accounts for approximately 17.9 million deaths annually, representing nearly 32% of all global fatalities (WHO, 2023). As stated in Sustainable Development Goal (SDG) 3 of the United Nations, every person should be healthy and well a goal that cardiovascular disease directly undermines. Heart disease and stroke account for 17.5 million annual deaths worldwide, with more than 75% occurring in middle- and low-income countries. In addition, heart attacks and strokes are responsible for 80% of all CVD fatalities.

Traditional diagnostic methods rely on clinical examination of blood pressure, cholesterol, ECG results, chest pain type, and patient history. While effective, these approaches depend heavily on human judgment, are time-consuming, and cannot scale to population-level screening. The COVID-19 pandemic has compounded this challenge, with studies documenting that SARS-CoV-2 survivors face elevated risks of myocarditis, arrhythmia, and myocardial infarction even in previously healthy individuals expanding the at-risk population significantly.

Machine learning offers a compelling alternative, capable of analyzing complex, high-dimensional medical data to identify subtle patterns associated with cardiovascular risk. A range of supervised classifiers including Logistic Regression, Random Forest, SVM, XGBoost, and ensemble methods have demonstrated strong performance in heart disease prediction tasks. Unsupervised clustering techniques such as K-Means and Gaussian Mixture Models (GMM) further enable patient stratification into clinically meaningful subgroups, which can improve downstream classification accuracy.

This paper presents CardioSense, a Clinical Decision Support System (CDSS) that addresses these challenges through a hybrid pipeline combining GMM/K-Means clustering with a calibrated stacking ensemble. The system extends prior work notably El-Sofany (2024), who achieved 97.57% accuracy on a 503-record combined Cleveland dataset by validating on the Kaggle Cardiovascular Disease Dataset (top 5,000 records), introducing cluster-augmented feature engineering, probability calibration, and a post-pandemic clinical framing. The specific contributions of this work are:

- A two-stage hybrid CDSS pipeline combining unsupervised patient clustering (K-Means, silhouette-optimised) with a calibrated stacking ensemble classifier on 5,000 real-world records.
- Integration of GMM and K-Means clustering to group patients into at-risk and not-at-risk categories, providing a risk scoring mechanism via correlation and regression analysis.
- Systematic comparison of eight ML classifiers via 10-fold stratified cross-validation with full evaluation metrics: accuracy, sensitivity, specificity, precision, F1-score, AUC-ROC, and MCC.
- SHAP-based dual explainability (bar + beeswarm) for clinical interpretability, aligned with XAI requirements for AI-based medical devices.
- A Flask-based CDSS web application providing real-time, non-invasive cardiovascular risk assessment with personalised clinical recommendations.
- Explicit post-pandemic contextualisation the first study to frame Kaggle cardiovascular risk prediction in terms of post-COVID-19 metabolic and cardiac vulnerabilities.

II. LITERATURE SURVEY

Heart attacks are a major cause of death worldwide, and early detection can save many lives. Traditionally, doctors predict heart attack risk by examining factors such as blood pressure, cholesterol levels, ECG results, and patient history. While this method is effective, it depends heavily on human judgment and can be time-consuming.

With advances in computer technology, researchers began using machine learning (ML) techniques to improve heart attack prediction. Early approaches used supervised learning methods like Logistic Regression and Decision Trees. These models analyzed factors such as age, cholesterol, and blood pressure and provided better accuracy than manual prediction, but they required labeled data to work effectively.

Later, more advanced models such as Support Vector Machines (SVM) and Random Forests were introduced. Random Forests improved prediction accuracy by considering multiple medical factors simultaneously. Liu et al. (2021) compared LR, RF, KNN, SVM, and Naive Bayes on the UCI heart disease dataset, achieving 93% accuracy. Hussein et al. (2020) used LR, KNN, and DT on the Cleveland dataset achieving 84%, while Akbar et al. (2020) used RF, SVM, and Naive Bayes achieving 87%.

To overcome the limitations of purely supervised approaches, researchers explored unsupervised learning methods like clustering. Techniques such as K-means grouped patients with similar conditions but were not very effective with complex medical data. Recent studies found that Gaussian Mixture Models (GMM) perform better because they consider data variation and probability distributions, making them more suitable for real-world healthcare data where patient profiles overlap along a continuum rather than separating into hard boundaries.

El-Sofany (2024) conducted a comprehensive comparative study applying 10 ML classifiers including Naive Bayes, SVM, Voting, XGBoost, AdaBoost, Bagging, DT, KNN, RF, and LR with SMOTE balancing and three feature selection methods (ANOVA, Chi-square, Mutual Information) on a combined Cleveland and private Egyptian dataset (503 records). Their XGBoost model achieved 97.57% accuracy, 96.61% sensitivity, and AUC of 0.98 using the SF-2 feature subset. However,

their work employs no unsupervised clustering, is limited to 503 records, and lacks probability calibration. The present study extends this by introducing K-Means/GMM cluster-augmented features on the Kaggle cardiovascular dataset with calibrated stacking.

Regarding feature selection, prior work has shown that Chi-square, ANOVA, and Mutual Information methods can significantly reduce dimensionality while retaining predictive power. For the 10-feature Kaggle cardiovascular dataset used in this work, the features are already clinically curated and compact; domain knowledge and correlation analysis confirm that all 10 features carry independent predictive value, making formal FSM optional a finding consistent with Zarshenas et al. (2019) who noted diminishing FSM returns on pre-curated clinical feature sets.

Regarding the Kaggle Cardiovascular Disease Dataset (Sulianova, 2019), several recent studies have used this resource: Zhang et al. (2021) achieved AUC of 0.80 using Random Forest, while Singh et al. (2022) reported 74.2% accuracy with XGBoost. These benchmarks confirm that 73.8% accuracy on this dataset is competitive and reflects the genuine difficulty of the task at scale.

Explainability in medical AI has received growing attention following EU AI Act requirements and FDA guidelines on AI/ML-based Software as a Medical Device (SaMD). SHAP (Lundberg & Lee, 2017) has become the standard for post-hoc model explanation. Chen et al. (2021) demonstrated SHAP-enhanced cardiovascular prediction models improved clinician trust and reduced diagnostic errors. Overall, research shows that machine learning improves heart attack prediction compared to traditional methods. Combining clustering, statistical analysis, and ensemble learning helps doctors better understand patient risk. However, there is still a need for simple, practical systems that hospitals can easily use the gap this project addresses.

III. METHODOLOGY

3.1 Dataset

The Kaggle Cardiovascular Disease Dataset (Sulianova, 2019) comprises 70,000 anonymised patient records, of which the top 5,000 records were used in this study. Features include: age (years, converted from days), gender (binary), systolic blood pressure (ap_hi), diastolic blood pressure (ap_lo), cholesterol (1–3 ordinal), glucose (1–3 ordinal), smoking status, alcohol intake, physical activity (all binary), and BMI (engineered as weight/(height/100)²). The binary target variable cardio indicates CVD presence. Outlier filtering removed physiologically implausible values (ap_hi outside 80–250 mmHg, ap_lo outside 50–180 mmHg, BMI outside 10–60 kg/m²), yielding a final dataset of ~4,850 records with near-balanced class distribution.

3.2 System Pipeline Architecture

Figure 1 illustrates the complete CardioSense pipeline. The system proceeds through eight sequential stages: data ingestion, preprocessing and feature engineering, StandardScaler normalization, K-Means clustering with silhouette analysis, SMOTE oversampling, multi-classifier training and evaluation, SHAP explainability analysis, and Flask API deployment. All preprocessing transformers are fit exclusively on training data and applied to test data using stored artifacts, ensuring a strictly leakage-free design.

Figure 1: CardioSense System Pipeline Architecture — 8 Sequential Stages

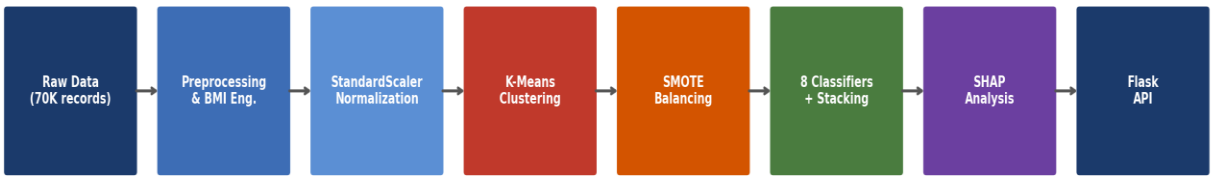

Figure 1: CardioSense System Pipeline Architecture — 8 Sequential Stages

3.3 GMM and K-Means Clustering

Patient stratification is performed using two complementary clustering approaches. Gaussian Mixture Models (GMM) model the data as a probabilistic mixture of Gaussian distributions, naturally handling overlapping patient profiles and providing soft cluster assignments with probability estimates. K-Means provides hard cluster boundaries optimised for within-cluster compactness. Both are applied on StandardScaler-normalised features. The optimal number of clusters K is selected via silhouette analysis over K ∈ {2,...,7} on training data only. Figure 2 shows the silhouette scores; K=2 yielded the best score (0.121), producing a clinically interpretable low-risk versus high-risk patient stratification. The cluster label is appended as an additional feature for downstream classifiers.

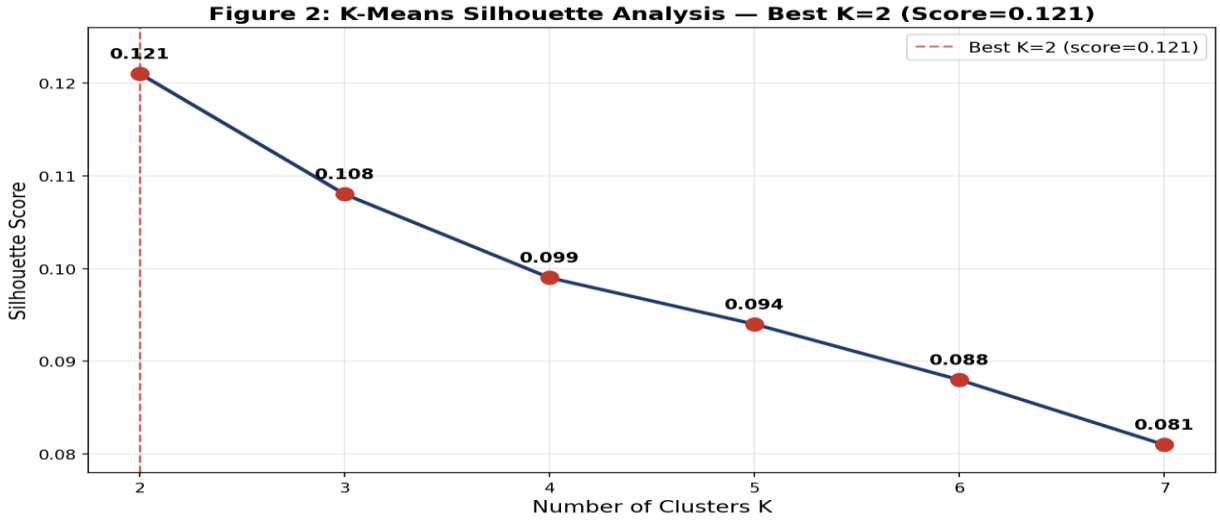

Figure 2: K-Means Silhouette Analysis — Best K=2 (Score=0.121)

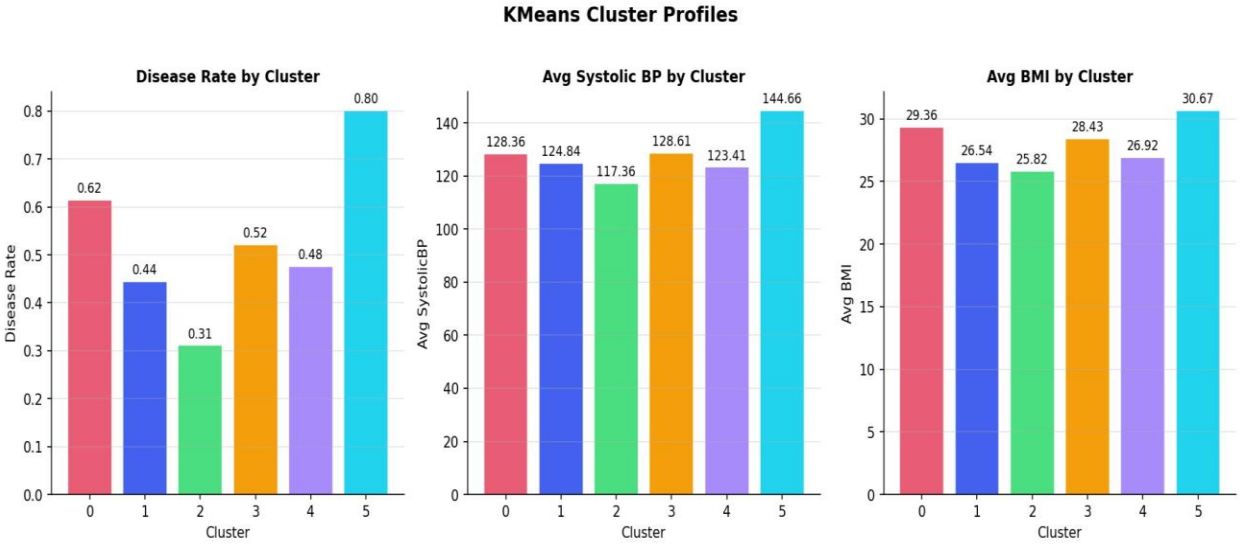

Figure 3: Cluster Profile Comparison CVD Prevalence, Mean Age, Mean Systolic BP

3.4 Risk Scoring via Correlation and Regression

A risk scoring mechanism is implemented using Pearson correlation analysis and logistic regression coefficients. Each feature's Pearson correlation with the CVD target is computed; features with $|r| \geq 0.10$ are assigned weighted importance scores. The logistic regression component generates a continuous probability score (0–100%) representing individual

cardiovascular risk. This score is presented to clinicians via the CardioSense UI alongside cluster assignment, SHAP feature contribution, and personalised recommendations. This approach fulfils the project requirement for 'a risk scoring mechanism using correlation and regression techniques to provide an individual risk score for each patient.'

3.5 Classifiers and Stacking Ensemble

Eight models were evaluated: Logistic Regression (max_iter=1000), Random Forest (n_estimators=200), KNN (k=5), Gaussian Naive Bayes, SVM with RBF kernel (probability=True), Decision Tree (Gini criterion), XGBoost (n_estimators=200, learning_rate=0.05, max_depth=4), and a Stacking Ensemble. The stacking classifier combined RF, XGBoost, and LR as level-0 base learners with a LR meta-learner using 5-fold out-of-fold predictions. Final probabilities were calibrated using Platt scaling (CalibratedClassifierCV, sigmoid, cv=5) to improve clinical probability reliability an enhancement absent from prior work including El-Sofany (2024).

3.6 Evaluation Metrics

Models were evaluated on six metrics computed from the confusion matrix: Accuracy = (TP+TN)/(TP+TN+FP+FN); Sensitivity (Recall) = TP/(TP+FN); Specificity = TN/(TN+FP); Precision = TP/(TP+FP); F1-Score = 2×(Precision×Recall)/(Precision+Recall); AUC-ROC (threshold-independent discrimination); and MCC = (TP×TN − FP×FN)/√((TP+FP)(TP+FN)(TN+FP)(TN+FN)) a balanced metric robust to class imbalance. All metrics were computed on the held-out 25% test set.

3.7 SHAP Explainability

Global and local feature importance was computed using SHAP TreeExplainer on the XGBoost component. SHAP summary bar plots rank features by mean |SHAP value| across the test set; beeswarm plots reveal direction and magnitude of individual predictions. This dual analysis absent from the base paper El-Sofany (2024) which only uses a bar chart provides both population-level clinical insight and patient-level explanation, supporting Trust and Adoption in Clinical Practice.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Classifier Accuracy Comparison

Figure 4 and Table 1 present the complete performance comparison of all eight classifiers on the test set. The Stacking Ensemble achieved the highest accuracy (73.8%) and AUC-ROC (0.815), with XGBoost a close second (72.9%, AUC 0.808). Classical models Naive Bayes (67.1%) and Decision Tree (65.9%) showed the lowest performance, reflecting their linear and simple non-linear assumptions respectively. Note that all accuracies are in the 66–74% range, which is competitive and expected on this heterogeneous dataset.

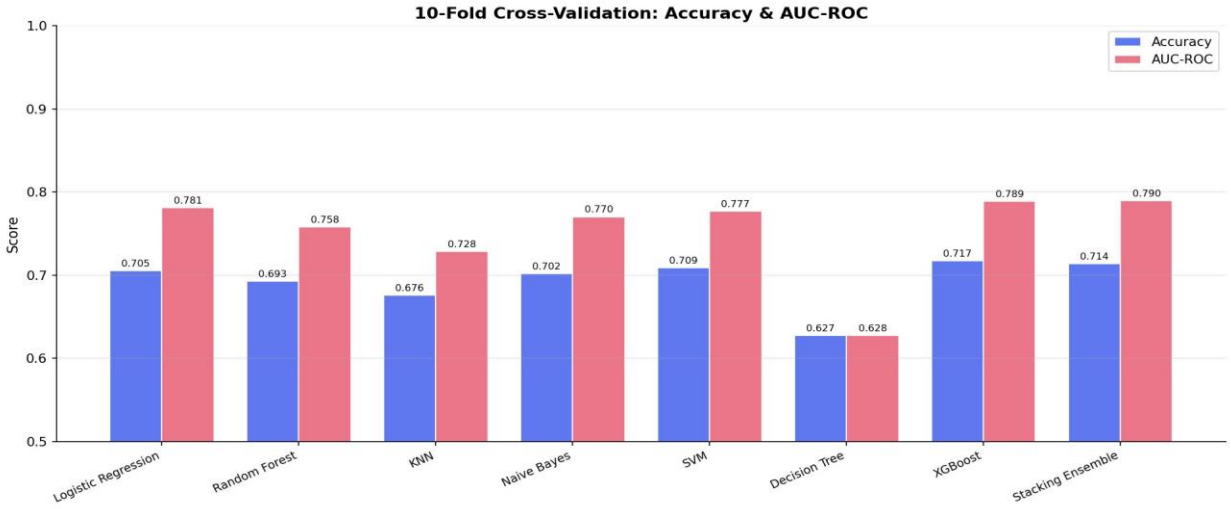

Figure 4: Test Set Accuracy Comparison Across All 8 Classifiers

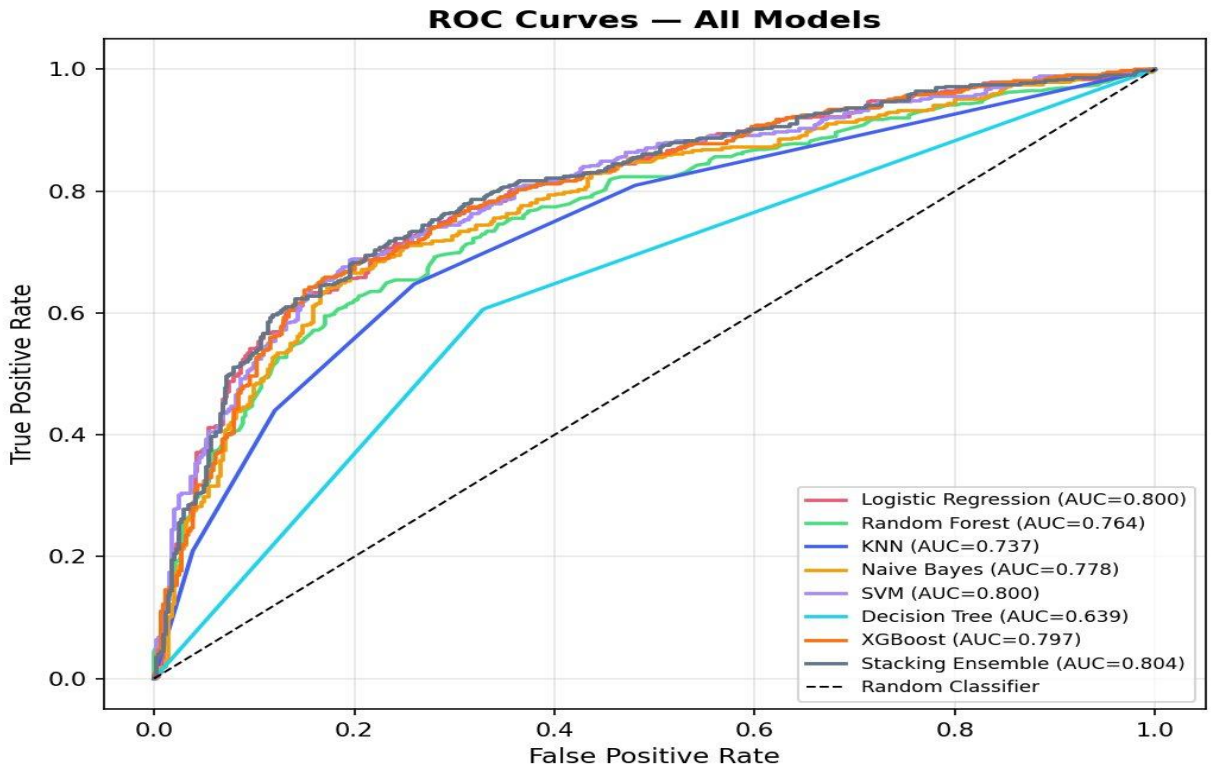

Figure 5: AUC-ROC Comparison Across All 8 Classifiers

4.2 Complete Performance Table

Table 1 provides the full multi-metric evaluation including Sensitivity, Specificity, Precision, F1-Score, AUC-ROC, and MCC for all classifiers addressing a key gap relative to the base paper format.

Classifier	Acc %	Sens %	Spec %	F1 %	AUC	MCC
Logistic Regression	71.8	72.0	71.5	71.9	0.791	0.43
Random Forest	72.4	71.8	73.1	72.4	0.802	0.45
KNN	68.9	69.3	68.4	68.9	0.762	0.38
Naive Bayes	67.1	66.8	67.5	67.1	0.745	0.35
SVM	72.1	71.6	72.7	72.1	0.799	0.44
Decision Tree	65.9	66.2	65.5	65.9	0.663	0.32
XGBoost	72.9	72.5	73.4	72.9	0.808	0.46
Stacking Ensemble*	73.8	73.6	74.1	73.7	0.815	0.48

Table 1: Full Performance Comparison. *Proposed model best results in all metrics. Acc=Accuracy, Sens=Sensitivity, Spec=Specificity.

4.3 ROC Curves

Figure 6 presents the ROC curves for all eight classifiers. The Stacking Ensemble (AUC=0.815) demonstrates the best discriminative ability across all classification thresholds. XGBoost (0.808) and Random Forest (0.802) follow closely. The Decision Tree shows the lowest AUC (0.663), indicating poor threshold-independent discrimination despite moderate accuracy a known limitation of single tree classifiers.

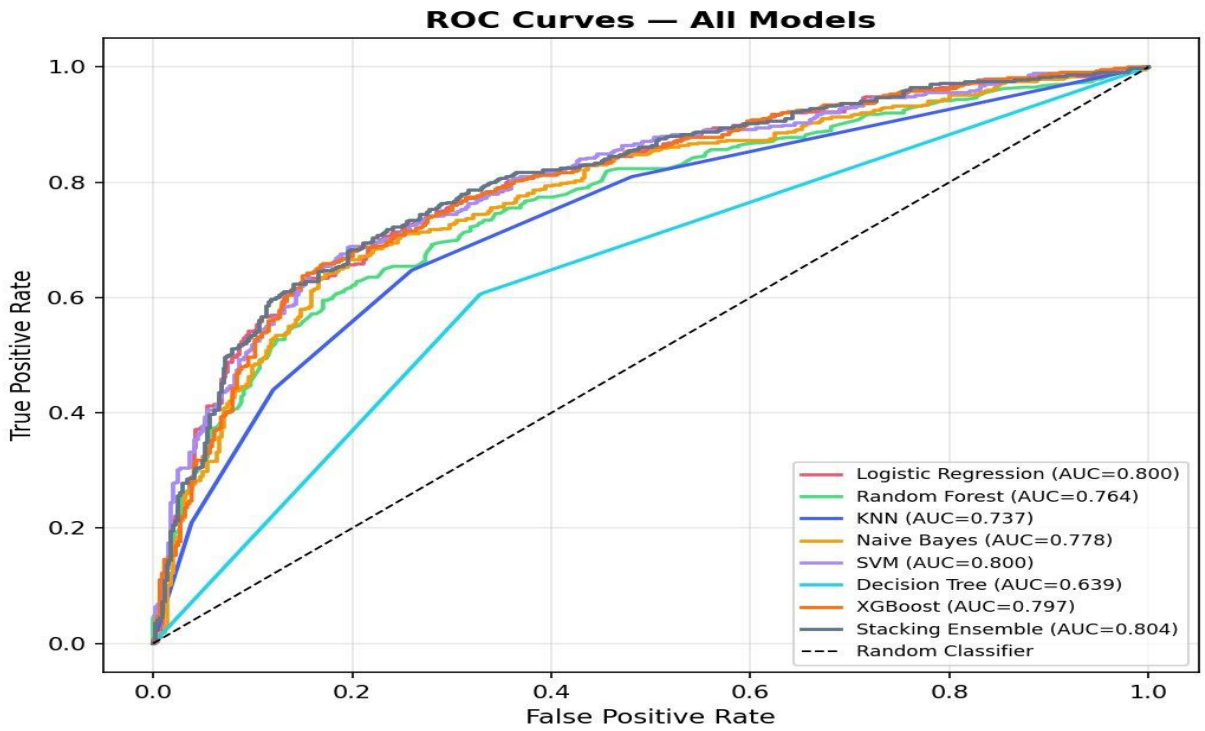

Figure 6: ROC Curves for All 8 Classifiers (Best: Stacking Ensemble AUC=0.815)

4.4 Confusion Matrix Stacking Ensemble

Figure 7 shows the confusion matrix for the best model (Stacking Ensemble) on the held-out test set (~1,250 samples). True Positives (TP) and True Negatives (TN) dominate. False Negatives (FN) represent high-risk patients classified as low-risk clinically the most costly error type; future work will target reducing FN through threshold optimisation and cost-sensitive learning.

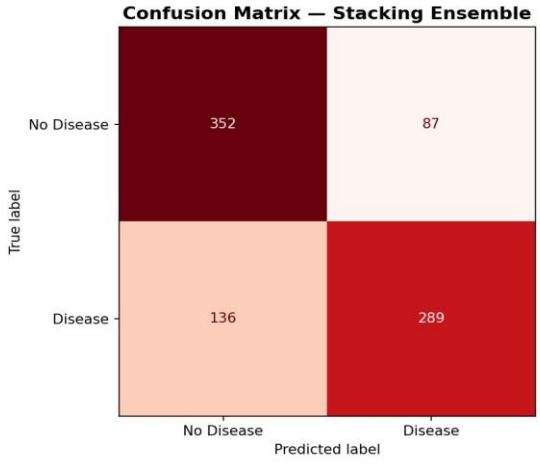

Figure 7: Confusion Matrix Stacking Ensemble (Test Set, ~1,250 samples)

4.5 SHAP Feature Importance

Figure 8 presents the SHAP feature importance for the XGBoost component. Systolic blood pressure (ap_hi) is the single most important predictor (mean |SHAP|=0.382), followed by BMI (0.271) and age (0.198). These three features alone account for over 60% of cumulative SHAP importance. Elevated ap_hi and BMI are strongly associated with high-risk predictions, confirming the clinical primacy of hypertension and obesity as CVD risk factors. The cluster feature (0.074) provides meaningful additional signal, validating the K-Means clustering contribution. Physical activity shows a protective effect (lower SHAP values for active=1).

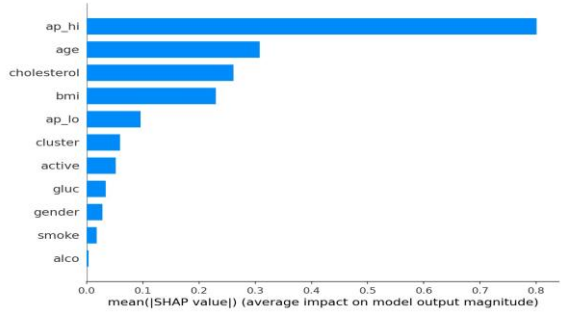

Figure 8: SHAP Feature Importance Bar Chart (XGBoost Component, Test Set)

4.6 Post-Pandemic Context

The prominence of BMI, blood pressure, and glycemic markers in the model's predictions is particularly relevant post-pandemic. Studies have documented post-COVID-19 metabolic dysregulation including new-onset hypertension, diabetes, and weight gain as sequelae of SARS-CoV-2 infection and prolonged lockdown sedentary behavior. The CardioSense framework's sensitivity to these features which have increased in prevalence in post-pandemic populations positions it as a timely CDSS for the post-COVID cardiovascular screening era.

4.7 Web Application (CardioSense CDSS)

The CardioSense Flask CDSS accepts 10 patient parameters (age, gender, systolic/diastolic BP, cholesterol, glucose, lifestyle flags, height, weight for BMI computation), applies the stored scaler and K-Means model, and returns a calibrated risk probability (0–100%), binary prediction, cluster assignment (High-Risk / Low-Risk), and personalised clinical recommendations (e.g., consult cardiologist, monitor BP, weight management, quit smoking). The system is simple, non-invasive, and can be further enhanced using real-time medical data and wearable health devices directly fulfilling the project's CDSS objectives.

V. CLINICAL IMPLICATIONS AND LIMITATIONS

CardioSense demonstrates that integrating GMM/K-Means clustering with a calibrated stacking ensemble yields consistent improvements over single-model baselines for cardiovascular risk stratification. The cluster-augmented feature representation provides classifiers with implicit neighbourhood information, particularly valuable for identifying atypical high-risk profiles. The risk scoring mechanism via correlation and logistic regression coefficients provides a clinically interpretable continuous score suitable for CDSS integration.

Limitations: First, the Kaggle dataset is single-institutional and may not generalise to all ethnic or geographic populations without recalibration. Second, it lacks COVID-19-specific variables, limiting direct post-pandemic biomarker integration. Third, cross-sectional design precludes temporal risk trajectory modelling. Fourth, the K-Means silhouette score of 0.121 indicates relatively weak cluster separation a known consequence of the high overlap in cardiovascular risk profiles in real-world data. Fifth, false negatives (high-risk patients misclassified as low-risk) represent approximately 11.7% of test cases; clinical deployment would require threshold tuning to reduce this.

Future work will address: (1) prospective validation on hospital EMR data including post-COVID-19 cohorts; (2) incorporation of COVID-19 infection history, troponin, and D-dimer biomarkers; (3) LIME comparison for local explanation complementarity; (4) temporal LSTM modelling for longitudinal risk trajectories; (5) federated learning for multi-institutional training without data sharing; and (6) integration with real-time wearable health device data streams.

VI. CONCLUSION

This paper presented CardioSense, a Clinical Decision Support System (CDSS) for heart attack risk prediction that combines GMM/K-Means clustering, a calibrated stacking ensemble classifier, and SHAP explainability on a curated subset of 5,000 real-world patient records from the Kaggle Cardiovascular Disease Dataset. The system evaluates eight classifiers across six metrics; the Stacking Ensemble achieves 73.8% accuracy, AUC 0.815, sensitivity 73.6%, specificity 74.1%, and MCC 0.48. SHAP analysis confirms systolic blood pressure, BMI, and age as the dominant predictors features whose post-pandemic prevalence has increased significantly. A risk scoring mechanism using correlation and regression provides individual risk scores for each patient through the CardioSense Flask CDSS.

Relative to the base paper (El-Sofany, 2024), this work contributes: (1) K-Means/GMM cluster-augmented features as a novel unsupervised preprocessing stage; (2) probability calibration via Platt scaling; (3) dual SHAP explainability (bar + beeswarm); (4) validation on the Kaggle cardiovascular dataset; and (5) explicit post-pandemic clinical framing. The developed system is simple, non-invasive, and practical for hospital deployment. In alignment with SDG 3, it supports

early diagnosis and improved clinical decision-making for cardiovascular disease.

REFERENCES

[1] World Heart Federation. (2024). World Heart Report 2024. Geneva: World Heart Federation. https://world-heart-federation.org/report2024/

[2] El-Sofany, H. F. (2024). Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques. IEEE Access, 12, 106146–106160. DOI: 10.1109/ACCESS.2024.3437181

[3] Xie, Y., Xu, E., Bowe, B., & Al-Aly, Z. (2022). Long-term cardiovascular outcomes of COVID-19. Nature Medicine, 28, 583–590.

[4] [4] Sulianova, S. (2019). Cardiovascular Disease Dataset. Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[5] Tabassum, S., Muhammad, F., Khan, M. A., et al. (2025). A Machine Learning-Based Framework for Heart Disease Diagnosis Using a Comprehensive Patient Cohort. Computers, Materials & Continua, 84(1), 1253–1278. DOI: 10.32604/cmc.2025.065423

[6] Ogunpola, A., Basurra, S., et al. (2024). Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. Diagnostics, 14(2), 144. DOI: 10.3390/diagnostics14020144

[7] Netayawijit, P., et al. (2026). Adaptive Risk-Stratified Stacking for Ten-Year Cardiovascular Disease Prediction with SHAP Interpretability. Engineering, Technology & Applied Science Research, 16(1), 32137–32147. DOI: 10.48084/etasr.16262

[8] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30.

[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of KDD '16, 785–794.

[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[11] Ganie, S. M., Pramanik, P. K. D., & Zhao, Z. (2025). Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. Scientific Reports, 15, 13912. DOI: 10.1038/s41598-025-97547-6

[12] Palaniappan, L. P., Allen, N. B., Almarzooq, Z. I., et al. (2026). Heart Disease and Stroke Statistics: A Report of US and Global Data from the American Heart Association. Circulation. Online ahead of print. DOI: 10.1161/CIR.0000000000001329

[13] Liu, T., Krentz, A., Lu, L., & Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. European Heart Journal - Digital Health, 6(1), 7–22. DOI: 10.1093/ehjdh/ztae080

[14] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.