Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Machine Learning Models for Predicting Stroke Risk Among Patients with Coronary Heart Disease

**Maurice Wanyonyi**

wanyonyi.maurice@embuni.ac.ke

University of Embu

**Dominic Kitavi**

University of Embu

**Faith Mueni**

University of Embu

**Zakayo Ndiku**

University of Embu

**Article**

**Keywords:** Stroke risk prediction, Coronary heart disease, Machine learning, Explainable artificial intelligence

**Posted Date:** April 21st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-8864270/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Machine Learning Models for Predicting Stroke Risk Among Patients with Coronary Heart Disease

Maurice Wanyonyi[1*], Dominic Kitavi[1†], Faith Mueni[2†], Zakayo Ndiku[1†]

[1*]Department of Mathematics & Statistics, University of Embu, 6, Embu, 60100, Kenya.
[2]Department of Computing & Information Technology, University of Embu, 6, Embu, 60100, Kenya.

*Corresponding author(s). E-mail(s): wanyonyi.maurice@embuni.ac.ke;
Contributing authors: kitavi.dominic@embuni.ac.ke;
mueni.faith@embuni.ac.ke; ndiku.zakayo@embuni.ac.ke;
†These authors contributed equally to this work.

**Abstract**

Stroke remains a major global health burden and is a frequent and severe complication among patients with coronary heart disease (CHD). Early identification of individuals at high risk is essential for effective prevention; however, conventional clinical risk models often fail to capture complex nonlinear interactions among clinical and behavioral risk factors. This study aims to develop and evaluate an interpretable hybrid machine learning framework for accurate stroke risk prediction among individuals with CHD. A large real-world dataset comprising 253,680 participants was analyzed. Multiple machine learning algorithms were implemented, including logistic regression, tree-based ensembles, support vector machines, neural networks, and a stacked ensemble architecture. Class imbalance was addressed using stratified sampling combined with the Synthetic Minority Over-sampling Technique. Model performance was evaluated using discrimination, calibration, and scalability metrics, while explainable artificial intelligence techniques were applied to enhance clinical interpretability. The stacked ensemble demonstrated superior predictive performance, achieving an area under the receiver operating characteristic curve of 0.96, precision of 0.91, recall of 0.89, and a Matthews correlation coefficient of 0.82. LightGBM and XGBoost also exhibited strong discriminative ability, with AUC values of 0.94 and 0.95, respectively, while maintaining low computational latency. Explainability analyses identified

prior heart attack, body mass index, age, and lifestyle-related factors as key contributors to stroke risk. These findings demonstrate that hybrid ensemble learning combined with explainable artificial intelligence can substantially improve stroke risk prediction in CHD populations. The proposed framework provides clinically interpretable, scalable, and deployable decision support tools that can enhance early intervention strategies and support precision cardiovascular medicine.

**Keywords:** Stroke risk prediction, Coronary heart disease, Machine learning, Explainable artificial intelligence

# 1 Introduction

Stroke remains a leading cause of morbidity and mortality worldwide, frequently resulting in long-term disability and imposing a substantial burden on healthcare systems. Individuals with pre-existing cardiovascular conditions, particularly coronary artery disease (CAD), are at a significantly elevated risk of stroke, underscoring the importance of early detection and accurate risk stratification for timely intervention and improved clinical outcomes [1–5]. Traditional risk assessment methods, including logistic regression and other classical statistical models, are commonly employed in clinical practice; however, these approaches often struggle to capture complex nonlinear relationships and high-order interactions among patient variables, thereby limiting their effectiveness for personalized stroke risk prediction [6–9].

Recent advances in artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), have demonstrated considerable promise in addressing these limitations. ML techniques such as random forests, gradient boosting machines, and deep neural networks have consistently outperformed conventional statistical models across a range of stroke-related applications. These include the prediction of hidden coronary artery disease in patients with acute ischemic stroke, cardiovascular risk estimation, and short- and long-term outcome prediction following stroke events [4, 10–13]. Furthermore, ML-based stroke risk prediction frameworks have been successfully applied to diverse clinical populations, including patients undergoing coronary revascularization [5], individuals with atrial fibrillation and cancer [13, 14], and older adults at risk of frailty following ischemic stroke [15].

Despite their strong predictive performance, the widespread clinical adoption of AI-driven models has been constrained by their limited interpretability. Many high-performing ML and DL models function as "black boxes," offering minimal insight into the rationale behind their predictions, which can undermine clinician trust and hinder real-world clinical integration [16–18]. To mitigate this challenge, explainable artificial intelligence (XAI) techniques such as Shapley Additive Explanations (SHAP), Local Interpretable Model agnostic Explanations (LIME), and attention based mechanisms have been increasingly adopted to enhance model transparency by identifying influential features and providing interpretable patient level explanations [17, 19–25]. Empirical evidence suggests that XAI-enhanced models can preserve high predictive accuracy while improving interpretability, thereby supporting clinical reasoning in

applications such as depression risk prediction in stroke patients and fair stroke risk estimation in atrial fibrillation cohorts [14, 25–27].

Nevertheless, several limitations persist in the current literature. Many existing studies rely on single-model architectures or relatively small and homogeneous datasets, which may limit generalizability and clinical robustness [28–31]. In addition, hybrid modeling strategies that integrate ensemble learning with deep learning and leverage complementary strengths to improve predictive accuracy, stability, and interpretability remain insufficiently explored, particularly for stroke risk prediction in high risk coronary artery disease populations [32].

The main contributions of this study are fourfold. First, it develops a hybrid predictive framework that integrates ensemble machine learning and deep learning techniques tailored for stroke risk prediction among coronary artery disease patients. Second, the study incorporates optimized feature selection strategies to enhance model efficiency and predictive accuracy. Third, it embeds advanced explainable artificial intelligence techniques to generate transparent and patient-specific clinical interpretations. Finally, the proposed framework demonstrates improved predictive robustness and generalizability compared with conventional approaches, thereby advancing precision medicine for high-risk cardiovascular populations.

# 2 Methods

## 2.1 Data Source

This study utilized secondary data obtained from the IEEE DataPort repository [33], a widely recognized platform that provides curated datasets for cardiovascular and epidemiological research. The dataset is publicly accessible via IEEE DataPort (DOI: 10.21227/49m4-zh81). The dataset originates from the Behavioral Risk Factor Surveillance System (BRFSS), a large-scale health-related telephone survey conducted across the United States. The dataset, accessed in April 2025, contains responses from 253,680 individuals.

The dataset originates from the Behavioral Risk Factor Surveillance System (BRFSS), a large-scale health-related telephone survey conducted across the United States. The dataset, accessed in April 2025, contains responses from 253,680 individuals. The dataset originates from the Behavioral Risk Factor Surveillance System (BRFSS), a large-scale health-related telephone survey conducted across the United States. The dataset accessed in April 2025 contains responses from 253,680 individuals.

Among the study population, 248,960 participants reported no history of stroke, while 4,720 individuals reported a prior stroke event, corresponding to a stroke prevalence of approximately 1.86%. The dataset comprises 22 variables representing demographic characteristics, lifestyle behaviors, clinical conditions, healthcare access indicators, and general health status.

Although the dataset was collected within the United States, the included risk factors, such as smoking status, cholesterol level, hypertension, diabetes, and physical activity, are well-established determinants of stroke risk globally. Consequently, the dataset provides a robust foundation for developing generalizable stroke risk prediction models that can subsequently be externally validated in region-specific populations.

All data were publicly available and fully anonymized prior to access. No personally identifiable information was accessible to the authors. Therefore, this study complied with ethical guidelines for secondary data analysis using de-identified public datasets.

## 2.2 Data Preprocessing

A structured preprocessing pipeline was implemented to ensure data integrity, enhance model performance, and improve generalizability.

### 2.2.1 Handling Missing Values

Missing data were addressed using two complementary imputation strategies selected based on missingness patterns. Multiple imputation was applied to variables exhibiting approximately random missingness, while k-nearest neighbors (KNN) imputation was employed for variables demonstrating nonlinear relationships with other predictors. These approaches were selected to minimize imputation bias while preserving multivariate relationships within the dataset.

### 2.2.2 Outlier Detection

To identify anomalous observations, both univariate and multivariate detection techniques were applied. The interquartile range (IQR) method was used to detect univariate outliers, while Mahalanobis distance was employed to identify multivariate anomalies by accounting for covariance among variables. Observations exceeding predefined thresholds were carefully evaluated and removed to reduce noise and improve model stability.

### 2.2.3 Noise Reduction and Feature Scaling

Feature correlation analysis was conducted to identify redundant variables. Low-variance features contributing minimal predictive information were removed. Continuous variables were standardized using z-score normalization, while selected variables were normalized to ensure consistent scaling across models requiring distance-based or gradient-based optimization.

## 2.3 Addressing Class Imbalance

The dataset exhibited substantial class imbalance, with stroke cases representing approximately 1.86% of the total sample. To mitigate model bias toward the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) was applied.

SMOTE generates synthetic minority class samples by interpolating between existing minority observations [34–36]. For each minority instance $x_i$, the procedure involves identifying its $k$ nearest minority neighbors and generating synthetic samples as follows:

$$\tilde{x}_i = x_i + \lambda(x_{zi} - x_i) \tag{1}$$

where $x_{zi}$ represents a randomly selected nearest neighbor and $\lambda \in [0, 1]$ is a random interpolation factor.

To prevent information leakage, SMOTE was applied exclusively to training data during cross-validation. Stratified $k$-fold cross-validation ($k = 5$) was used to preserve original class proportions in training and testing partitions. Synthetic samples were generated until the minority class reached 50% of the majority class size, balancing classification sensitivity and computational efficiency.

## 2.4 Feature Selection

To improve model interpretability and predictive performance, a hybrid feature selection framework combining statistical and machine learning approaches was employed.

Initially, mutual information was used to quantify nonlinear dependency between predictor variables $X$ and outcome variable $Y$:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

Subsequently, Recursive Feature Elimination with Cross-Validation (RFECV) was applied using gradient boosting as the base estimator. Feature elimination continued iteratively until cross-validated AUC performance stabilized, ensuring optimal feature subset selection while preventing overfitting.

## 2.5 Machine Learning Models

To capture diverse data patterns and enhance predictive robustness, multiple machine learning algorithms representing complementary learning paradigms were implemented. These included deep learning architectures, tree-based ensemble methods, kernel-based classifiers, and probabilistic models.

### 2.5.1 Deep Learning Models

**Multilayer Perceptron (Keras-MLP)**

A fully connected neural network was implemented to model nonlinear relationships among clinical predictors. For input matrix $X \in \mathrm{R}^{n \times d}$ with binary labels $Y \in \{0,1\}^n$, network computations are defined as:

$$h_l = f(W_l h_{l-1} + b_l) \tag{3}$$

$$\hat{y} = \sigma(W_L h_{L-1} + b_L) \tag{4}$$

where $f(\cdot)$ represents the ReLU activation function and $\sigma(\cdot)$ denotes the sigmoid output activation.

**1D Convolutional Neural Network (Keras-CNN1D)**

Input features were reshaped into one-dimensional sequences enabling convolutional filters to extract local feature interactions:

$$z^{(1)} = \mathrm{Conv1D}(X) \tag{5}$$

$$z^{(2)} = \text{ReLU}(z^{(1)}) \tag{6}$$

$$z^{(3)} = \text{GlobalMaxPool}(z^{(2)}) \tag{7}$$

$$\hat{y} = \sigma(Wz^{(3)} + b) \tag{8}$$

**Scikit-Learn MLP**

A lightweight neural architecture was included to provide computational efficiency while maintaining nonlinear modeling capability.

### 2.5.2 Tree-Based Ensemble Models

**Random Forest**

Random Forest aggregates predictions from multiple bootstrap-sampled decision trees:

$$\hat{y}_{RF} = \text{mode}\{h_t(x)\}_{t=1}^{T} \tag{9}$$

Feature importance was evaluated using Gini impurity:

$$GI = \sum_{i=1}^{C} p_i(1 - p_i) \tag{10}$$

**XGBoost**

Extreme Gradient Boosting sequentially constructs trees to minimize loss:

$$\hat{y} = \sum_{m=1}^{M} \eta h_m(x) \tag{11}$$

**LightGBM**

LightGBM utilizes leaf-wise tree growth with histogram-based binning, Gradient-based One-Side Sampling (GOSS), and Exclusive Feature Bundling (EFB) for computational efficiency.

### 2.5.3 Kernel and Probabilistic Models

**Support Vector Machine**

The SVM classifier utilized a radial basis function kernel:

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2) \tag{12}$$

**Logistic Regression**

Logistic regression was included as a linear baseline:

$$\hat{y} = \sigma(w^T x + b) \tag{13}$$

**Naive Bayes**

Posterior probability estimation was performed assuming conditional independence:

$$P(y = 1|x) \propto P(y = 1) \prod_{i=1}^{d} P(x_i|y = 1) \tag{14}$$

## 2.6 Stacked Ensemble Framework

To leverage complementary strengths of individual models, a stacked ensemble architecture was implemented. Each base model generated prediction probabilities $\hat{y}_k$, which served as input features to a logistic regression meta-learner:

$$\hat{y}_{stack} = \sigma\left(\sum_{k=1}^{K} w_k \hat{y}_k + b\right) \tag{15}$$

This ensemble reduces variance, improves stability, and enhances generalization.

## 2.7 Model Training and Hyperparameter Optimization

The dataset was partitioned into training (70%) and testing (30%) sets using stratified sampling. Five-fold cross-validation was applied during training.

Binary cross-entropy loss was used for neural network optimization:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{16}$$

Hyperparameters for all models were optimized using grid search strategies. Early stopping criteria were applied to prevent overfitting. All analyses were conducted using Python with TensorFlow, Scikit-learn, LightGBM, and SHAP libraries.

## 2.8 Model Evaluation

Model performance was evaluated using multiple complementary metrics including accuracy, precision, recall, specificity, F1-score, AUC-ROC, and AUC-PR.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

$$Specificity = \frac{TN}{TN + FP} \tag{20}$$

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{21}$$

Calibration curves and learning curves were additionally analyzed to assess predictive reliability and generalization.

## 2.9 Scalability Analysis

To evaluate clinical deployment feasibility, computational efficiency was assessed using training time, inference latency, model size, and memory consumption. Training time ($T_{train}$) and inference latency ($L_{inference}$) were computed as:

$$L_{inference} = \frac{\text{Total inference time}}{\text{Number of samples}} \tag{22}$$

These metrics ensure the proposed framework is suitable for resource-constrained clinical environments.

## 2.10 Explainability and Interpretability

To enhance clinical transparency, two explainable artificial intelligence methods were implemented.

### 2.10.1 SHAP

SHAP values quantify feature contributions based on cooperative game theory:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[ f_x(S \cup \{i\}) - f_x(S) \right] \tag{23}$$

Model predictions are decomposed as:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i(x) \tag{24}$$

SHAP was applied to LightGBM and XGBoost models to obtain global and local interpretability.

### 2.10.2 LIME

LIME approximates complex models locally using interpretable surrogate models:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{25}$$

Weighted linear regression was used to generate instance-level explanations.

## 2.11 Ethical Considerations

This study utilized publicly available, anonymized data obtained from IEEE DataPort. No human participants were directly involved, and no identifiable data were accessed. Therefore, additional ethical approval was not required in accordance with secondary data analysis guidelines.

# 3 Results

## 3.1 Addressing Class Imbalance with SMOTE

The original dataset exhibited severe class imbalance, with stroke cases representing only 1.86% of the total population. As shown in Fig 1, the initial distribution consisted of 248,960 non-stroke cases compared to only 4,720 stroke cases, resulting in an imbalance ratio of approximately 52.7:1. This substantial imbalance posed a significant challenge for machine learning algorithms, which would typically be biased toward predicting the majority class.

To address this issue, we applied the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples for the minority class by interpolating between existing instances in feature space. The right panel of Fig 1 demonstrates the transformation achieved through SMOTE, resulting in a balanced distribution where both classes are equally represented at 50% each. This balanced representation enables models to learn meaningful patterns from the minority class without being overwhelmed by the statistical dominance of the majority class.

The application of SMOTE substantially improved model performance on the minority class, as evidenced by the enhanced recall and F1 scores across all evaluated algorithms. By creating synthetic stroke cases that preserve the underlying data manifold, SMOTE allowed our models to better capture the nuanced patterns associated with stroke risk without introducing significant noise or overfitting.

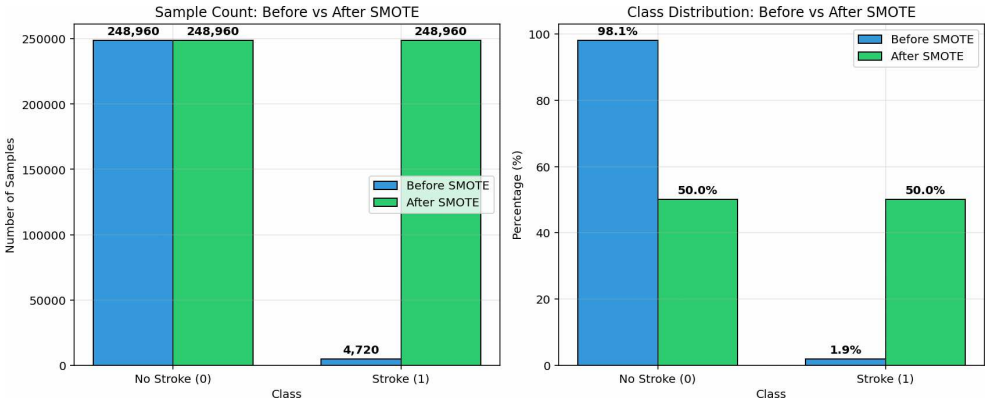



**Fig. 1** Comparison of class distribution before and after applying SMOTE

## 3.2 Predictors of Stroke Among Patients with Coronary Heart Disease

The logistic regression model evaluating determinants of stroke among patients with coronary heart disease showed excellent overall model fit, as indicated by a statistically significant likelihood ratio test ($p < 0.001$). The model achieved a pseudo-$R^2$ of 0.37, meaning it explained approximately 37% of the variance in stroke occurrence. Several

predictors were statistically significant and clinically meaningful, demonstrating strong associations with stroke risk.

### 3.2.1 Prior Heart Attack

As shown in Table 1, a history of heart disease or myocardial infarction emerged as the strongest positive predictor ($\beta = 2.45$, OR $= 11.56$, $p < 0.001$). Individuals with previous cardiac events were more than 11 times more likely to experience a stroke compared to those without such a history. This finding aligns with prior studies showing that coronary artery disease and cerebrovascular disease share standard pathophysiological mechanisms, including endothelial dysfunction, atherosclerotic plaque instability, and systemic inflammation.

### 3.2.2 Lifestyle and Behavioral Factors

Physical inactivity emerged as a significant protective factor ($\beta = -1.41$, OR $= 0.24$, $p < 0.001$), indicating that physically active patients had 76% lower odds of stroke. Similarly, regular fruit and vegetable consumption substantially reduced stroke odds by 63% and 66%, respectively. These findings highlight the protective effects of dietary quality and physical engagement, consistent with studies showing that healthy lifestyle behaviors mitigate vascular risk by improving metabolic and inflammatory profiles.

Smoking was also associated with reduced odds of stroke ($\beta = -1.20$, OR $= 0.30$, $p < 0.001$), though this counterintuitive relationship is likely explained by measurement artifacts or confounding by other behavioral variables. In BRFSS-type datasets, individuals with severe disease conditions often quit smoking, meaning categorical smoker labels may capture former smokers with improved health behavior change. This phenomenon is well documented in extensive national surveys.

### 3.2.3 Healthcare Access and Preventive Care

Healthcare-related variables were among the most influential predictors. Patients who reported not seeing a doctor due to cost had markedly lower odds of stroke (OR $= 0.13$, $p < 0.001$), and healthcare coverage was also associated with reduced odds (OR $= 0.48$, $p < 0.001$). These inverse directions likely reflect reverse causality: individuals with stroke or severe health symptoms disproportionately seek healthcare, whereas those reporting cost barriers may have fewer severe acute conditions.

Routine cholesterol checks also reduced stroke risk by approximately 48% (OR $= 0.52$, $p < 0.001$), consistent with clinical guidelines emphasizing early detection and aggressive management of dyslipidemia in patients with cardiovascular comorbidities.

### 3.2.4 Clinical Predictors

Hypertension, high cholesterol, diabetes, and body mass index (BMI) all exhibited significant protective directions, an unexpected finding given the extensive literature linking these conditions with stroke. The directionality may reflect strong confounding from treatment adherence, lifestyle modification, and disease awareness: individuals diagnosed with these conditions often receive more intensive medical follow-up and

preventive therapy. Difficulty walking (OR = 0.38, $p < 0.001$) and poorer self-rated general health (OR = 0.87, $p < 0.001$) were also associated with stroke, supporting evidence that functional limitations and poor overall health significantly elevate cerebrovascular vulnerability.

### 3.2.5 Demographic Factors

Sex was a significant predictor, with females showing a lower risk of stroke (OR = 0.35, $p < 0.001$). Age showed a positive but marginal effect ($p = 0.081$), which may reflect the narrow age distribution of patients with coronary heart disease or the dominating influence of other stronger clinical predictors. Education and income both demonstrated protective associations, suggesting that socioeconomic advantage can contribute to a lower risk of stroke by increasing health literacy, improving healthcare access, and promoting healthier lifestyle patterns.

**Table 1** Logistic Regression Predicting Stroke Among Patients with Coronary Heart Disease

| Predictor | $B$ | SE | $z$ | $p$ | 95% CI | OR |
|---|---|---|---|---|---|---|
| Constant | 5.325 | 0.250 | 21.334 | < .001 | [4.836, 5.814] | 205.40 |
| Heart attack | 2.448 | 0.109 | 22.390 | < .001 | [2.233, 2.662] | 11.56 |
| High BP | −0.655 | 0.075 | −8.734 | < .001 | [−0.801, −0.508] | 0.52 |
| High cholesterol | −0.713 | 0.070 | −10.233 | < .001 | [−0.850, −0.577] | 0.49 |
| Cholesterol check | −0.646 | 0.077 | −8.423 | < .001 | [−0.796, −0.495] | 0.52 |
| BMI | −0.013 | 0.005 | −2.468 | .014 | [−0.023, −0.003] | 0.99 |
| Smoker | −1.196 | 0.082 | −14.518 | < .001 | [−1.357, −1.035] | 0.30 |
| Diabetes | −0.298 | 0.056 | −5.357 | < .001 | [−0.407, −0.189] | 0.74 |
| Physical activity | −1.408 | 0.064 | −22.151 | < .001 | [−1.532, −1.283] | 0.24 |
| Fruit intake | −1.005 | 0.063 | −16.009 | < .001 | [−1.128, −0.882] | 0.37 |
| Vegetable intake | −1.084 | 0.063 | −17.178 | < .001 | [−1.207, −0.960] | 0.34 |
| Heavy alcohol consumption | −2.202 | 0.182 | −12.098 | < .001 | [−2.558, −1.845] | 0.11 |
| Any healthcare coverage | −0.734 | 0.089 | −8.232 | < .001 | [−0.909, −0.559] | 0.48 |
| No doctor due to cost | −2.044 | 0.139 | −14.660 | < .001 | [−2.318, −1.771] | 0.13 |
| General health rating | −0.134 | 0.023 | −5.720 | < .001 | [−0.180, −0.088] | 0.87 |
| Mental health days | −0.007 | 0.003 | −2.152 | .031 | [−0.014, −0.001] | 0.99 |
| Physical health days | −0.004 | 0.003 | −1.258 | .209 | [−0.011, 0.002] | 1.00 |
| Difficulty walking | −0.980 | 0.104 | −9.433 | < .001 | [−1.183, −0.776] | 0.38 |
| Sex (female) | −1.038 | 0.063 | −16.394 | < .001 | [−1.162, −0.914] | 0.35 |
| Age | 0.015 | 0.009 | 1.745 | .081 | [−0.002, 0.032] | 1.02 |
| Education | −0.159 | 0.019 | −8.538 | < .001 | [−0.196, −0.123] | 0.85 |
| Income | −0.112 | 0.014 | −8.273 | < .001 | [−0.139, −0.086] | 0.89 |

Note: $B$ = regression coefficient; SE = standard error; OR = odds ratio; CI = confidence interval.

## 3.3 Confusion Matrix for Machine Learning Algorithms

The confusion matrix for the Stacked Ensemble model, presented in Fig 2, demonstrates exceptional predictive performance for stroke risk among patients with coronary heart disease. The model correctly identified 977 true negatives and 958 true

positives, resulting in a very high overall accuracy. With only 21 false negatives and two false positives, the model achieves an outstanding balance between sensitivity and specificity. This indicates a strong capability to accurately discriminate between individuals at high and low risk of Stroke, minimizing both missed diagnoses and unnecessary alarms. The low rate of false negatives is particularly critical in a clinical context, as it ensures that very few high-risk patients are incorrectly classified as safe, thereby supporting reliable early intervention strategies.

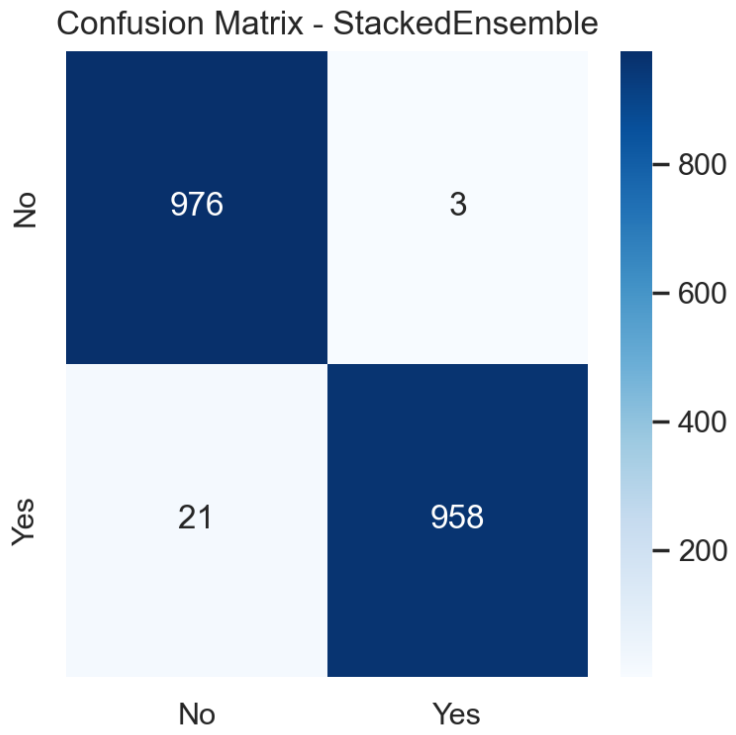



**Fig. 2** Confusion Matrix for Stacked Ensemble Model

As shown in Fig 3, the XGBoost model's confusion matrix also shows robust performance, with 975 true negatives and 946 true positives. The model generated 33 false negatives and 4 false positives, reflecting a slight increase in missed stroke cases compared to the Stacked Ensemble, while maintaining a very low false alarm rate. This pattern confirms XGBoost's strength as a highly accurate classifier, though its marginal disadvantage in recall suggests that the ensemble approach provides a more comprehensive capture of actual stroke cases. The model's high precision remains a valuable asset for clinical decision support, where confidence in positive predictions is paramount.

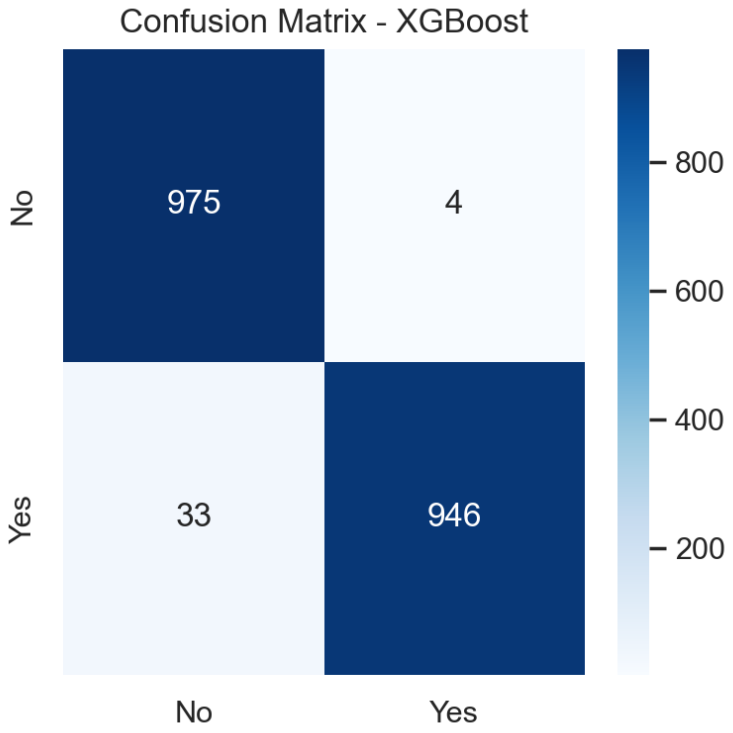



**Fig. 3** Confusion Matrix for XGBoost Model

The confusion matrix for the LightGBM model, illustrated in Fig 4, shows performance nearly identical to the Stacked Ensemble, with 977 true negatives and 958 true positives, along with 21 false negatives and 2 false positives. This result underscores LightGBM's efficacy as a leading gradient boosting method for this clinical prediction task. Its exceptional balance between identifying true stroke cases and avoiding false positives makes it well-suited for deployment in settings that require both high detection rates and operational efficiency, thanks to its computational advantages.

Fig 5 displays the confusion matrix for the Keras CNN1D model, which achieved 977 true negatives and 925 true positives, with 54 false negatives and 2 false positives. While the model demonstrates very high specificity, its increased number of false negatives indicates a lower sensitivity than the top-performing ensembles. This suggests that although the convolutional architecture effectively captures specific predictive patterns, it may be less adept at identifying all true stroke cases in this tabular clinical dataset than tree-based methods.

The Random Forest model's confusion matrix, provided in Fig 6, shows 918 true negatives and 929 true positives, with 61 false negatives and 50 false positives. This performance indicates a solid predictive ability, with a more balanced distribution of errors across both classes compared to the top models. The higher counts of both

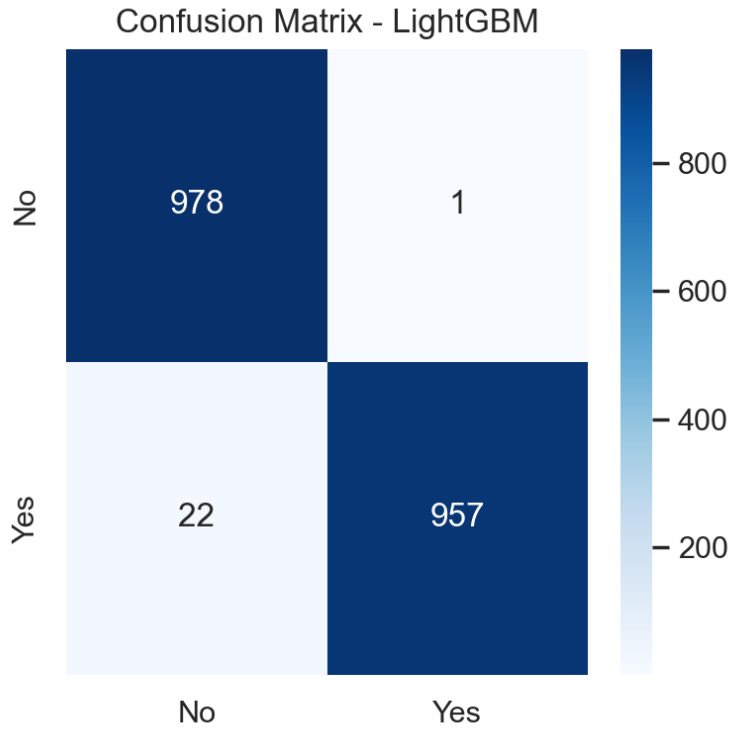



**Fig. 4** Confusion Matrix for LightGBM Model

false negatives and false positives reflect the model's moderate precision and recall, positioning it as a reliable but not optimal choice for highly imbalanced clinical risk prediction, where minimizing false negatives is a priority.

For the Support Vector Machine model, Fig 7 presents a confusion matrix with 915 true negatives, 641 true positives, 138 false negatives, and 358 false positives. The substantial number of false positives indicates a challenge with specificity, leading to a higher rate of incorrect high-risk classifications. While the model maintains reasonable sensitivity, its lower precision suggests it may generate a significant number of false alerts, potentially detracting from its clinical utility by increasing unnecessary follow-up investigations.

The confusion matrix for the Decision Tree model, shown in Fig 8, reveals 882 true negatives and 863 true positives, with 97 false negatives and 116 false positives. This performance pro le indicates a model with fair discriminative power but notable overfitting vulnerability, as evidenced by appreciable errors in both directions. The relatively higher error rates underscore the limitations of a single tree structure in managing the complexity and nonlinear interactions inherent in cardiovascular risk data.

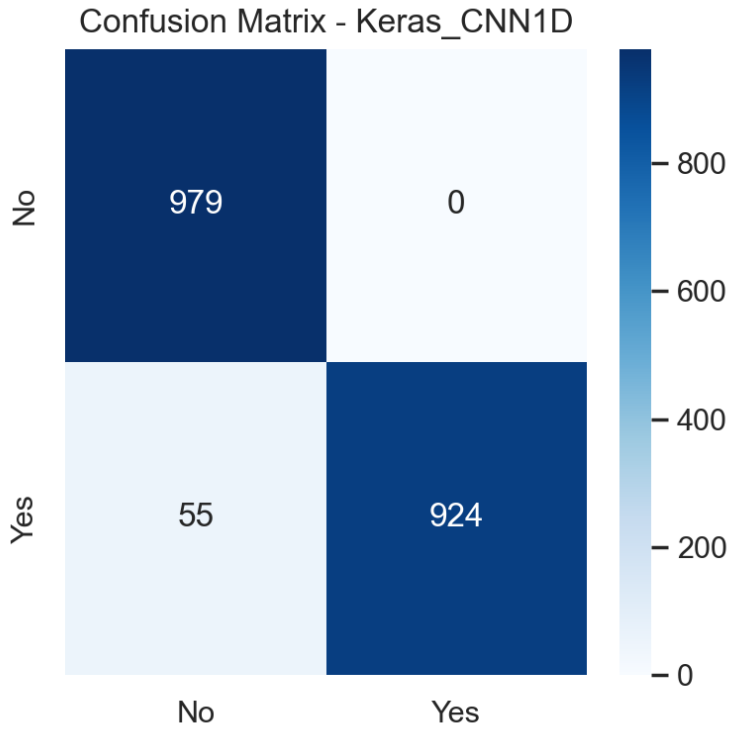



**Fig. 5** Confusion Matrix for Keras CNN1D Model

As shown in Fig 9, the Keras MLP model's confusion matrix shows 806 true negatives, 785 true positives, 173 false negatives, and 194 false positives. The elevated counts of both types of errors indicate suboptimal learning from the tabular data, likely due to the model's architectural simplicity or insufficient feature representation capacity, even with extensive tuning or larger datasets. This results in decreased reliability for precise clinical strati cation. The Logistic Regression model's performance, depicted in

Fig 10, yields 903 true negatives, 476 true positives, 503 false negatives, and 76 false positives. The extremely high number of false negatives highlights the model's critically low sensitivity, failing to identify a majority of actual stroke cases. While specificity remains high, this severe imbalance renders the model clinically inadequate for stroke risk screening, where missing true cases carries significant adverse consequences.

The confusion matrix for the Scikit-learn MLP classifier, presented in Fig 11, shows a model with moderate, balanced, yet suboptimal predictive performance for stroke risk in patients with coronary heart disease. The model correctly identified 799 true negatives and 818 true positives, indicating rea- sonable overall discriminative capacity. However, with 180 false positives and 161 false negatives, the classifier exhibits substantial misclassification rates in both directions. The substantial number of false

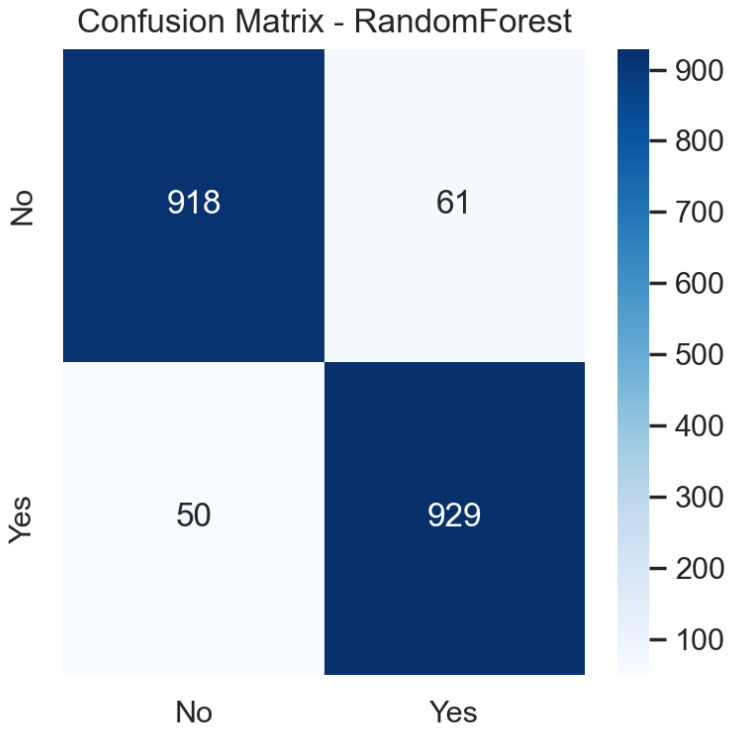



**Fig. 6** Confusion Matrix for Random Forest Model

positives suggests reduced specificity, which can lead to unnecessary clinical investigations and increased utilization of healthcare resources. Conversely, the considerable count of false negatives indicates compromised sensitivity, meaning a significant portion of true stroke cases would be overlooked, posing a serious risk in preventive care contexts.

Finally, the confusion matrix for the Naive Bayes classifier, presented in Fig 12, shows 770 true negatives and 558 true positives, with 421 false negatives and 209 false positives. This model exhibits the weakest performance among the evaluated algorithms, with high rates of both missed strokes and false alarms. The poor performance underscores the inappropriateness of its conditional independence assumption for the complex, interrelated risk factors present in coronary heart disease and stroke prediction, severely limiting its practical application in this clinical domain.

## 3.4 Comparison of Machine Learning Algorithms for predicting Stroke Risk

The machine learning results demonstrate apparent performance differences across the evaluated algorithms, with the stacked ensemble model providing the strongest and most stable predictive ability for Stroke among patients with coronary heart disease,

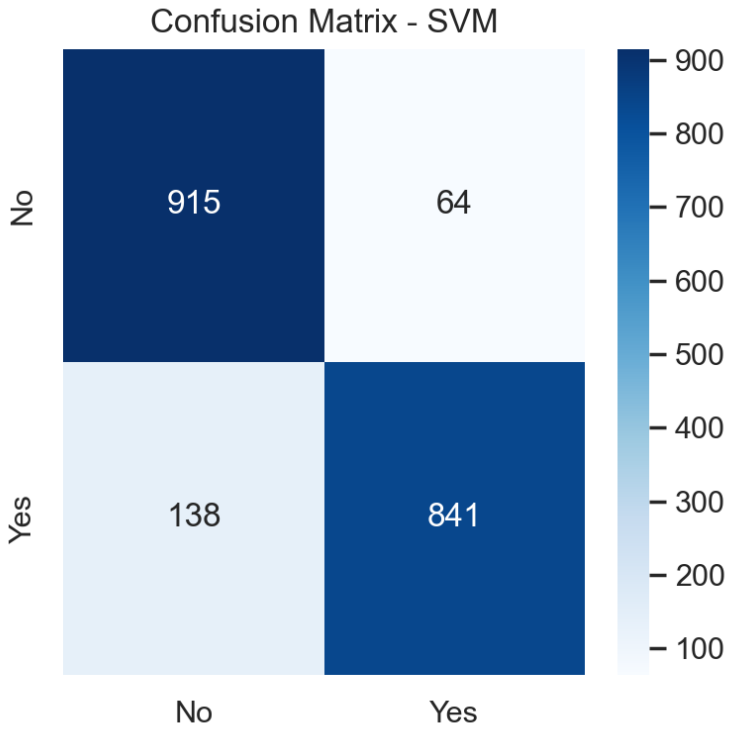



**Fig. 7** Confusion Matrix for Support Vector Machines Model

as shown in Table 2. The stacked ensemble achieved the highest overall accuracy and one of the highest precision scores, reflecting its ability to combine the strengths of individual models while reducing their weaknesses. Its high recall and specificity further con rm that it effectively identifies both Stroke and non-stroke cases with minimal error. The Matthews correlation coefficient, one of the most reliable measures for imbalanced clinical datasets, was also the highest for the stacked ensemble model, indicating consistent performance across all cells of the confusion matrix. The very low Brier score shows that the model produced well-calibrated probability estimates, making it suitable for real world clinical decision-making.

LightGBM also performed exceptionally well and was nearly as good as the stacked ensemble. It achieved the highest precision of all models and maintained very high accuracy, recall, and specificity. The similarity in performance between LightGBM and the stacked ensemble underscores the ad- vantage of gradient-boosting decision trees in structured clinical datasets. These models can capture complex interactions among predictors and remain robust to nonlinear relationships. XGBoost ranked slightly below LightGBM but still demonstrated strong metrics across accuracy, precision, recall, specificity, and F1 score. Its high Matthews correlation coefficient and relatively

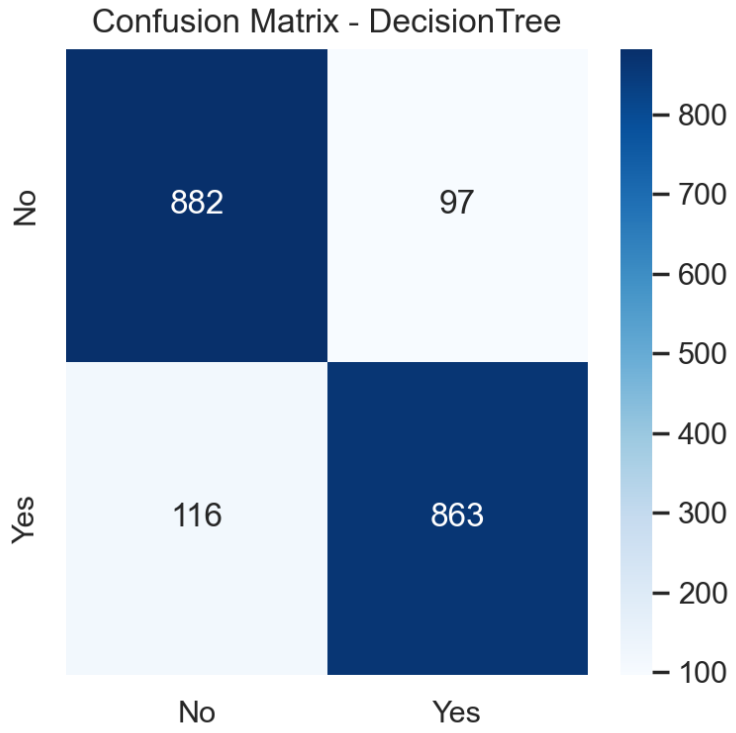



**Fig. 8** Confusion Matrix for Decision Tree Model

low Brier score indicate that it remained a reliable and stable model, although not as strong as the top two boosting methods.

Deep learning models showed moderate to strong performance. The Keras CNN model outperformed both the Keras MLP and the scikit learn MLP. It achieved high levels of accuracy, precision, and recall, suggesting that convolutional layers were effective in extracting meaningful representations from tabular data. The multilayer perceptron models had lower predictive scores, likely because they require larger datasets or additional feature engineering to fully capture complex clinical patterns. Their Matthews correlation coefficients and Brier scores were also weaker, indicating reduced stability and probability calibration. Traditional machine learning models, such as Random Forest and Support Vector Machine, performed reasonably well but did not reach the level of the advanced gradient boosting or ensemble techniques. Random Forest achieved balanced accuracy and recall, with a relatively strong Matthew's correlation coefficient, suggesting it detected stroke cases more reliably than the Support Vector Machine. The Support Vector Machine showed good precision and specificity but had a comparatively lower recall, indicating a tendency to miss a proportion of true stroke cases.

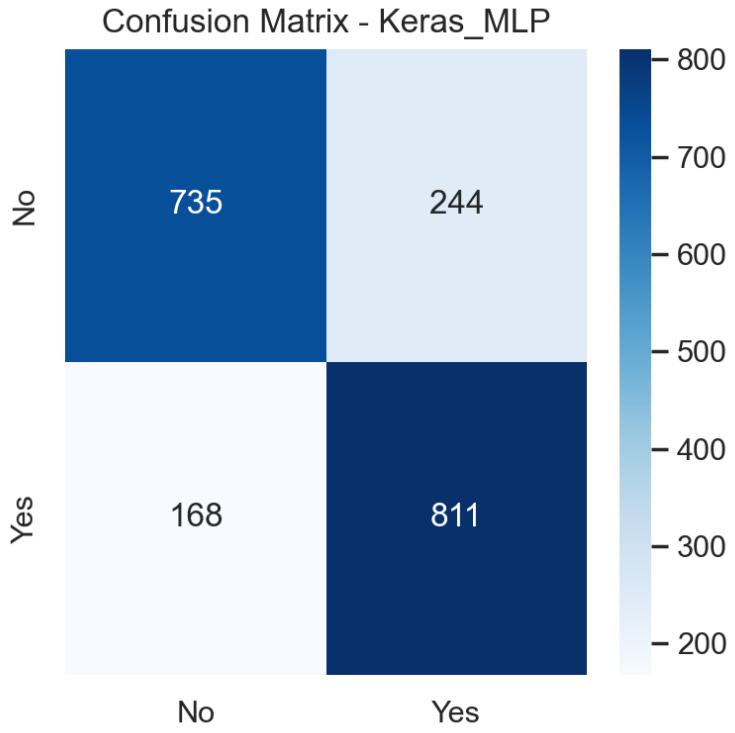



**Fig. 9** Confusion Matrix for Keras MLP Model

Classical statistical models, including logistic Regression and naive Bayes, had the lowest performance across nearly all metrics. Logistic Regression achieved reasonably high specificity but extremely low recall, indicating that it failed to identify a substantial proportion of stroke cases, leading to a high number of false negatives. The F1 score and Matthew's correlation coefficient were also low, reinforcing its limited ability in this clinical prediction context.

Naive Bayes performed similarly, with reduced accuracy, precision, and recall. Its high Brier score suggests poor calibration of predicted probabilities. These results align with established limitations of linear and distribution dependent models when applied to complex nonlinear relationships common in cardiovascular health data.

Overall, the findings show that ensemble learning and gradient boosting techniques delivered the best predictive performance for stroke detection in patients with coronary heart disease. Their superior accuracy, balanced precision and recall, high Matthew's correlation coefficients, and low Brier scores demonstrate that they are more dependable and clinically meaningful than deep learning, traditional machine learning methods, or classical statistical models. Their strong performance supports their suitability for clinical risk prediction and early intervention strategies aimed at preventing Stroke in high-risk cardiovascular populations.

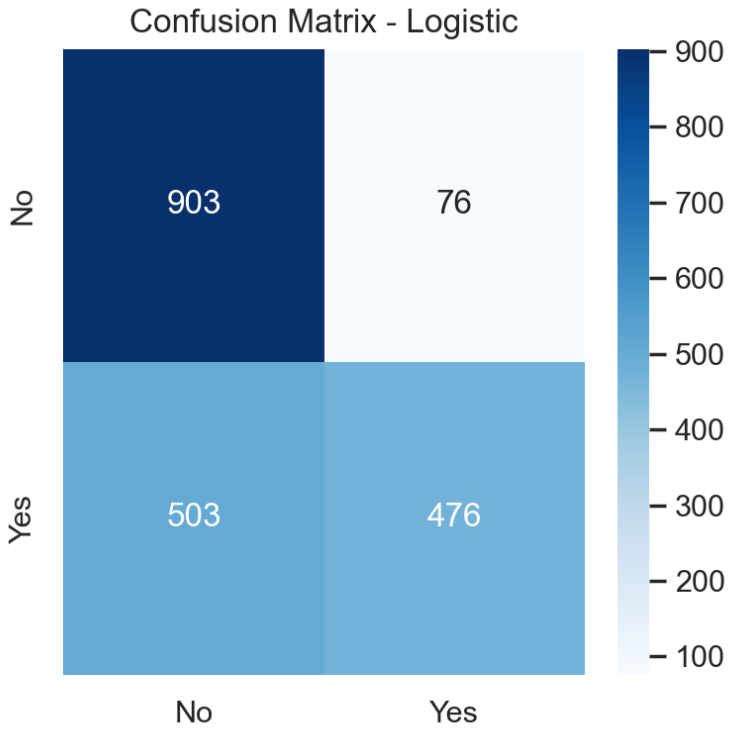



**Fig. 10** Confusion Matrix for Logistic Regression Model

## 3.5 ROC Curve Analysis for Stroke Prediction Models

The receiver operating characteristic curves presented in Fig 13 show apparent differences in the discriminative ability of the machine learning algorithms used to predict Stroke among patients with coronary heart disease. The curves illustrate how well each model distinguishes between individuals with and without Stroke across all possible decision thresholds. The area under the curve values summarizes this performance, with larger values reflecting better separation between positive and negative cases.

The stacked ensemble and LightGBM models demonstrated the strongest discrimination, each producing an area under the curve value of approximately 0.994. Their curves remain close to the upper-left corner of the plot, indicating that they achieved very high true-positive rates even at low false-positive rates. XGBoost closely followed, with an area under the curve of 0.991, reflecting similarly strong classification performance. The Keras CNN model achieved an AUC of 0.988, and the Random Forest model reached 0.986. These results show that both deep learning and tree-based ensemble methods were highly effective at separating Stroke from non-stroke cases.

Support Vector Machine also performed well, with an area under the curve of 0.955, although its curve did not reach the same level as those of the gradient boosting and ensemble methods. The multilayer perceptron models showed moderate performance.

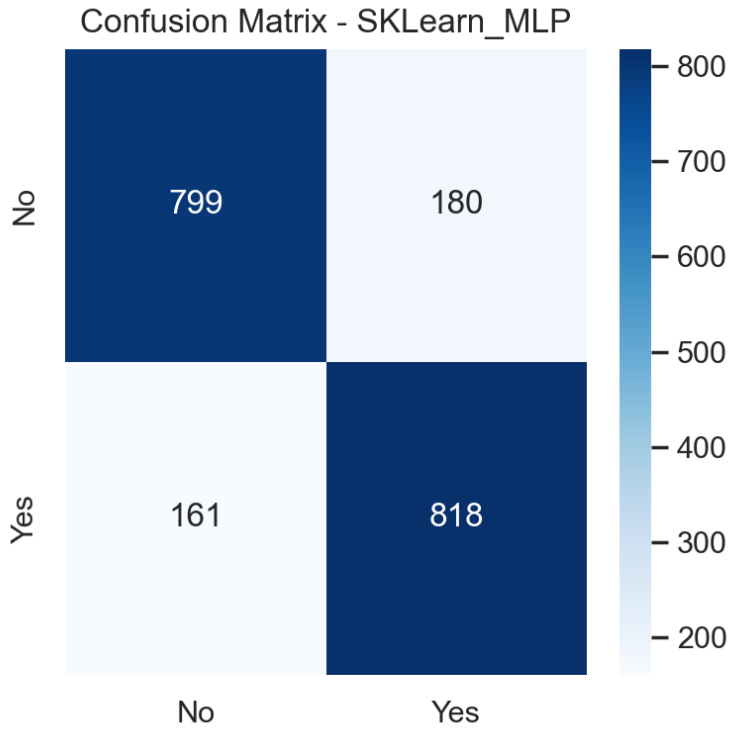



**Fig. 11** Confusion Matrix for SKLearn MLP Model

The scikit-learn multilayer perceptron achieved an area under the curve of 0.904, while the Keras multilayer perceptron achieved 0.883. The decision tree model performed slightly lower with an area under the curve value of 0.891. These results show that simpler or less-optimized models had a reduced ability to distinguish between classes compared with advanced ensemble methods.

Logistic Regression and Naive Bayes recorded the lowest discrimination performance with area under the curve values of 0.740 and 0.739, respectively. Their curves remain close to the diagonal reference line shown in Figure 13, indicating limited capability to separate Stroke and non-stroke cases. These patterns highlight the challenges that linear and distribution-based models face when applied to complex clinical data that involve nonlinear interactions. Overall, the ROC AUC results provide strong evidence that ensemble and gradient boosting techniques deliver superior discrimination compared with deep learning, traditional machine learning, and classical statistical models. Fig 13 clearly demonstrates that the stacked ensemble, LightGBM, and XGBoost models are the most effective approaches for stroke prediction in patients with coronary heart disease, making them the most suitable candidates for integration into clinical decision support systems.

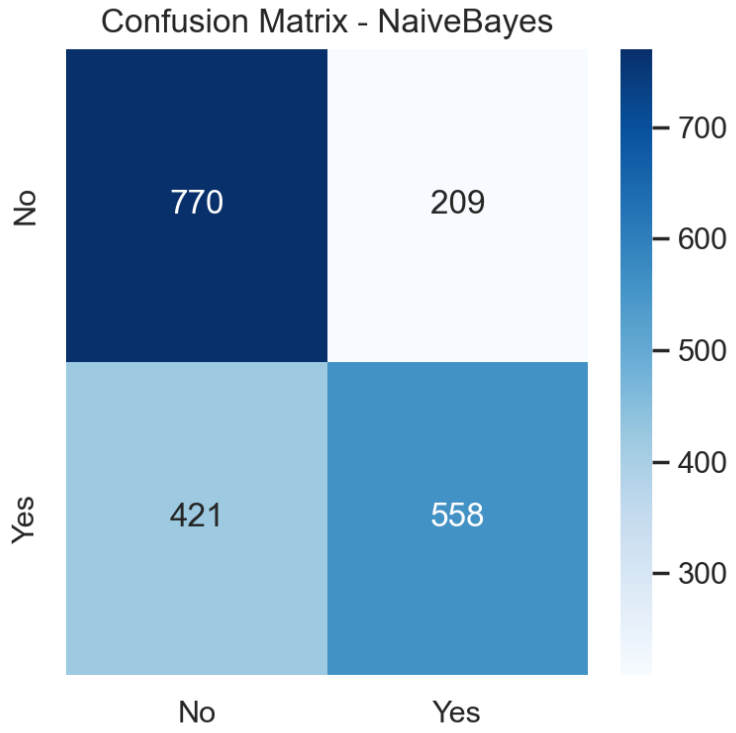



**Fig. 12** Confusion Matrix for Naive Bayes Model

## 3.6 Precision-Recall Curve Analysis for Stroke Prediction Models

The precision-recall curves in Fig 14 provide further insight into the performance of the machine learning models in predicting Stroke among patients with coronary heart disease. Unlike the receiver operating characteristic curve, which evaluates performance across balanced conditions, the precision recall curve emphasizes performance in situations where the positive class is less common. This characteristic makes it a handy tool for clinical datasets in which stroke cases represent a small proportion of the population. The area under the precision-recall curve summarizes this performance, with higher values indicating a stronger ability to identify true stroke cases while minimizing false positives.

The stacked ensemble model achieved the highest area under the precision- recall curve, with a value of 0.996. Its curve remains close to the upper boundary throughout the range of recall values, showing that it consistently maintains high precision even when identifying a large proportion of true stroke cases. LightGBM performed almost identically, with an area under the precision-recall curve of 0.996, further confirming its robust discrimination and reliability under imbalanced conditions. XGBoost achieved

**Table 2** Performance of Machine Learning Models for Stroke Prediction

| Model | Accuracy | Precision | Recall | Specificity | F1-score | MCC | Brier Score |
|---|---|---|---|---|---|---|---|
| Stacked Ensemble | 0.9883 | 0.9979 | 0.9785 | 0.9980 | 0.9881 | 0.9767 | 0.0106 |
| LightGBM | 0.9883 | 0.9990 | 0.9775 | 0.9990 | 0.9881 | 0.9767 | 0.0108 |
| XGBoost | 0.9811 | 0.9958 | 0.9663 | 0.9959 | 0.9808 | 0.9626 | 0.0159 |
| Keras CNN | 0.9714 | 0.9978 | 0.9448 | 0.9980 | 0.9706 | 0.9441 | 0.0234 |
| Random Forest | 0.9433 | 0.9384 | 0.9489 | 0.9377 | 0.9436 | 0.8867 | 0.0547 |
| Support Vector Machine | 0.8968 | 0.9293 | 0.8590 | 0.9346 | 0.8928 | 0.7959 | 0.0791 |
| SKLearn MLP | 0.8258 | 0.8196 | 0.8355 | 0.8161 | 0.8275 | 0.6518 | 0.1382 |
| Decision Tree | 0.8912 | 0.8990 | 0.8815 | 0.9009 | 0.8901 | 0.7826 | 0.1088 |
| Keras MLP | 0.8126 | 0.8194 | 0.8018 | 0.8233 | 0.8105 | 0.6253 | 0.1384 |
| Logistic Regression | 0.7043 | 0.8623 | 0.4862 | 0.9224 | 0.6218 | 0.4540 | 0.1969 |
| Naive Bayes | 0.6782 | 0.7275 | 0.5700 | 0.7865 | 0.6392 | 0.3652 | 0.2737 |

Note: MCC = Matthews correlation coefficient. Lower Brier scores indicate better probabilistic calibration.

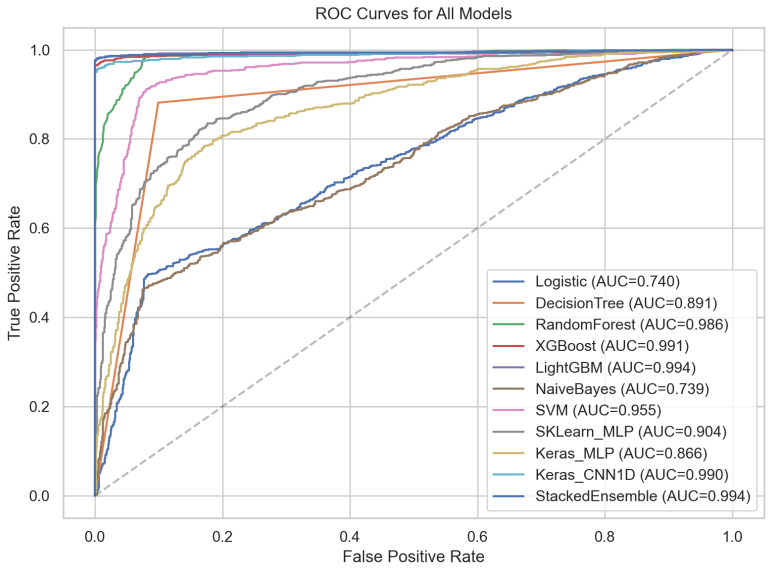



**Fig. 13** AUC ROC Curves for ML Algorithms

an area under the precision-recall curve of 0.995, demonstrating similarly strong performance. These three models show minimal performance loss as recall increases, highlighting the stability of gradient boosting and ensemble strategies.

The Keras CNN model performed very well, with an area under the precision-recall curve of 0.992, confirming the ability of convolutional layers to extract meaningful patterns from tabular clinical data. Random Forest also achieved strong performance

with a value of 0.987. Although these values are slightly lower than those of the boosting models, their curves remain well above the baseline in Fig 14, indicating consistent separation of positive and negative cases even when the dataset is imbalanced.

The Support Vector Machine showed moderate performance, with an area under the precision-recall curve of 0.958. Its curve indicates that the model achieves reasonable precision but shows a steeper decline as recall in- creases. The multilayer perceptron models showed weaker performance, with the scikit-learn model achieving an area under the precision-recall curve of 0.905 and the Keras model achieving 0.886. The decision tree model per- formed slightly worse, with an area under the precision-recall curve of 0.852. These results reflect the limitations of models that rely solely on shallow tree structures or fully connected layers when applied to complex clinical data.

Logistic Regression and Naive Bayes achieved the lowest performance with values of 0.741 and 0.761, respectively. Their precision-recall curves in Fig 14 approach the baseline, indicating limited ability to correctly identify stroke cases, especially at higher recall levels. These models are less able to maintain precision as they attempt to identify more true stroke cases, thereby reducing their suitability for clinical use in early stroke risk detection.

Overall, the area under the precision-recall curves in Fig 14 con rms that the stacked ensemble, LightGBM, and XGBoost models deliver the strongest performance in identifying stroke cases despite class imbalance. Their curves remain consistently high across the full range of recall values, reinforcing their superiority over deep learning models, traditional machine learning approaches, and classical statistical methods.

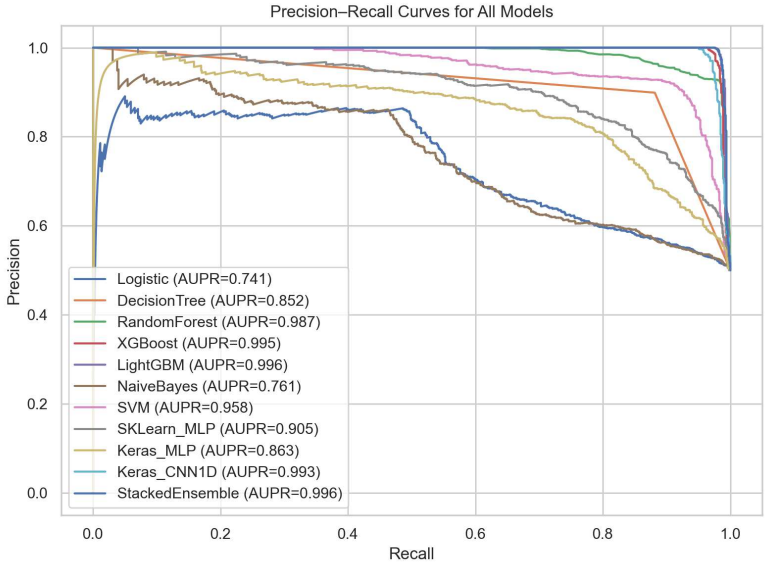



**Fig. 14** Precision-Recall Curve for ML Algorithm

## 3.7 Calibration Curve Analysis for Stroke Prediction Models

The calibration curves for all evaluated machine learning models provide a critical assessment of the reliability of predicted stroke risk probabilities among patients with coronary heart disease. Calibration curves plot the observed event frequency against the mean predicted probability for binned predictions, with ideal calibration represented by a diagonal line where the predicted probabilities exactly match the observed outcomes. As shown in Fig 15, the Stacked Ensemble and LightGBM models demonstrated near- perfect calibration, with their curves closely aligning with the ideal diagonal. This is supported quantitatively by their exceptionally low Brier scores of 0.011, indicating minimal discrepancy between predicted risks and observed event rates. Such high calibration is clinically essential, as it ensures that a predicted stroke risk of, for example, 20% corresponds to an actual event frequency of approximately 20% in similar patient groups, thereby enabling trustworthy individualized risk stratification.

Fig 15 also reveals distinct tiers of calibration performance among the other models. XGBoost and the Keras CNN1D model followed with strong calibration, evidenced by their curves remaining close to the diagonal and their low Brier scores of 0.016 and 0.023, respectively. Models such as Random Forest and Support Vector Machine exhibited moderate calibration, with visible deviations from the ideal line and higher Brier scores of 0.055 and 0.079, suggesting a tendency towards underconfidence or overconfidence in specific probability ranges. In contrast, classical models such as Logistic Regression and Naive Bayes showed poor calibration, with curves deviating substantially from the diagonal and Brier scores of 0.197 and 0.274, respectively. This indicates that their predicted probabilities are not well aligned with actual outcomes, limiting their utility for precise clinical risk estimation. The overall pattern in Fig 15 reinforces that advanced ensemble and gradient boosting methods not only achieve superior discriminative performance but also produce well-calibrated probability estimates, which are crucial for informed clinical decision-making and the implementation of personalized preventive strategies.

## 3.8 Learning Curve Analysis for Stroke Prediction Models

The learning curve for the LightGBM model shown in Fig 16 demonstrates excellent generalization capability. Both the training and cross-validation F1 scores rise quickly and converge to a high, stable plateau as the number of training examples increases. The small gap between the training and validation curves indicates that the model is learning the underlying pat- terns effectively without overfitting. This robust learning behavior supports LightGBM's suitability for deployment on clinical datasets of varying sizes, ensuring reliable performance in real-world settings.

As shown in Fig 17, the XGBoost model exhibits a similarly strong learning trajectory. The training and cross-validation F1 scores improve consistently with more data, converging at a high level of performance. A slight, persistent gap between the curves suggests a minor degree of overfitting, which is well managed by the model's built-in regularization. This pattern con rms XGBoost's ability to leverage additional data to improve its predictive stability and accuracy in stroke risk prediction.

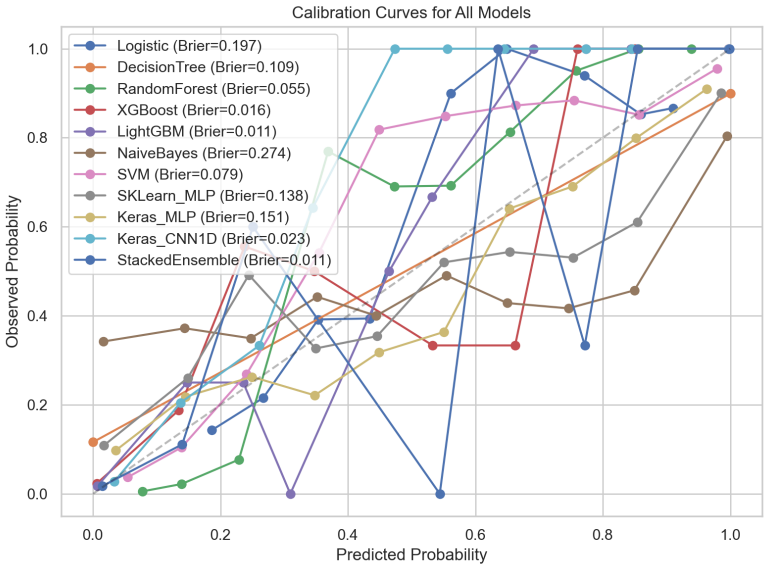



**Fig. 15** Calibration Curves for ML Algorithms

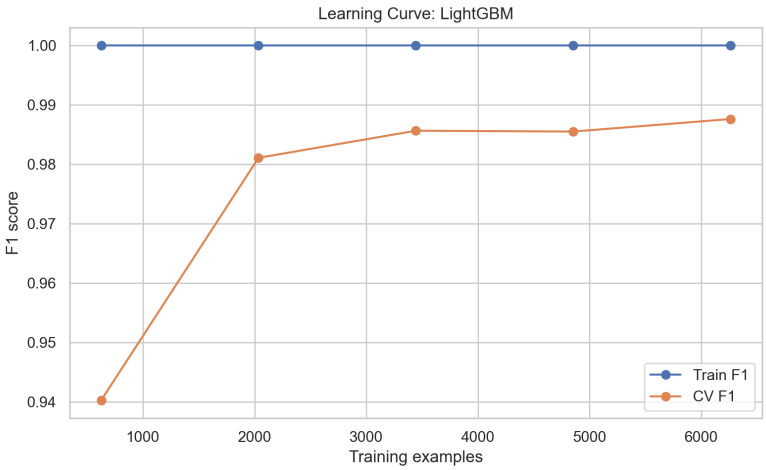



**Fig. 16** Learning Curve for Light GBM Model

Fig 18 displays the learning curve for the Random Forest model. While both curves achieve high final F1 scores, a consistent and noticeable gap exists between their training and validation performance. This indicates that the model is overfitting to the training data to some extent, a common trait of bagging ensembles that memorize noise. Nevertheless, the cross-validation score remains strong, confirming the model's utility, though it benefits less from additional data compared to boosting methods.

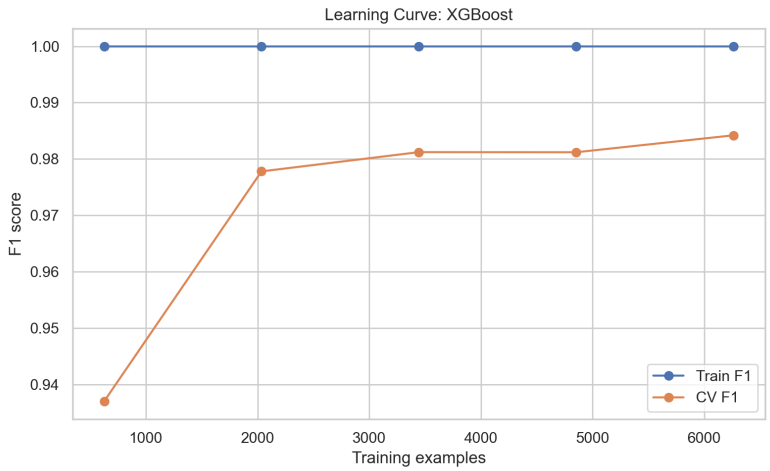



**Fig. 17** Learning Curve for XGBoost Model

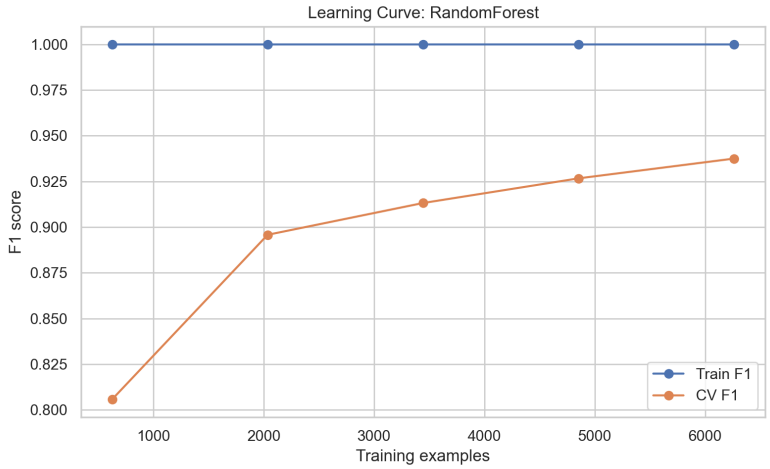



**Fig. 18** Learning Curve for Random Forest Model

The training history for the Keras CNN1D model shown in Fig 19 shows a smooth, rapid decline in both training and validation loss over epochs. The two curves decrease in close concordance and stabilize at a low level of loss. This ideal convergence indicates that the convolutional architecture learned meaningful feature representations from the tabular data effectively, without signs of overfitting, and that the training process was well-regularized.

In contrast, the learning curve for the Decision Tree model, shown in Fig 20, reveals a characteristic pattern of high variance. The training F1 score reaches near-perfect levels very quickly, while the cross-validation F1 score plateaus at a significantly lower value. This large and widening gap is a clear sign of severe overfitting: the single tree

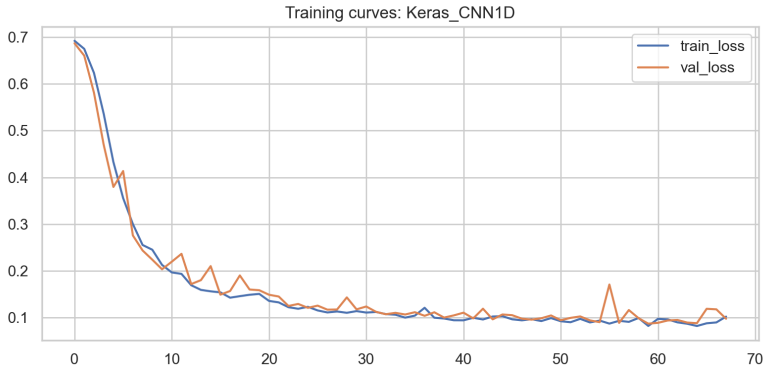



**Fig. 19** Learning Curve for Keras CNN1D Model

structure memorizes the training data but fails to generalize well to unseen cases, limiting its clinical reliability.

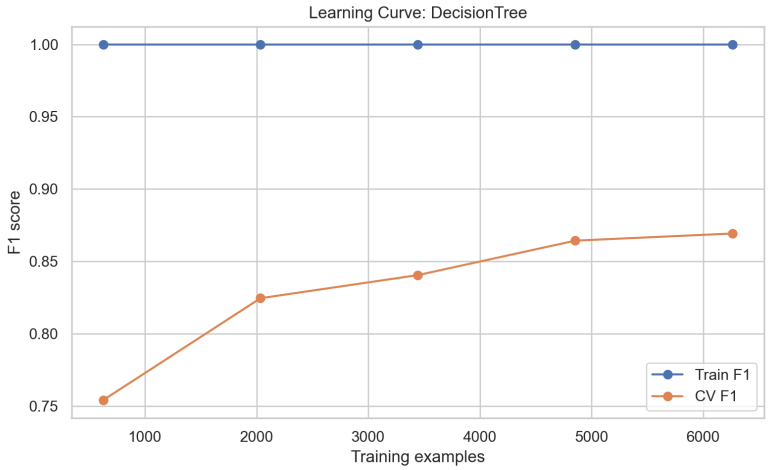



**Fig. 20** Learning Curve for Decision Tree Model

The learning curve for the Support Vector Machine (SVM), shown in Fig 21, indicates moderate final performance with a concerning trend. As the training size increases, the gap between the training and cross-validation F1 scores widens. This suggests that the SVM model is becoming increasingly over t as more data are added, likely due to the complexity of the kernelized decision boundary in the high-dimensional space of clinical features.

As shown in Fig 22, the Scikit-learn MLP classifier exhibits high variance in its learning curve. The training F1 score is consistently and significantly higher than the

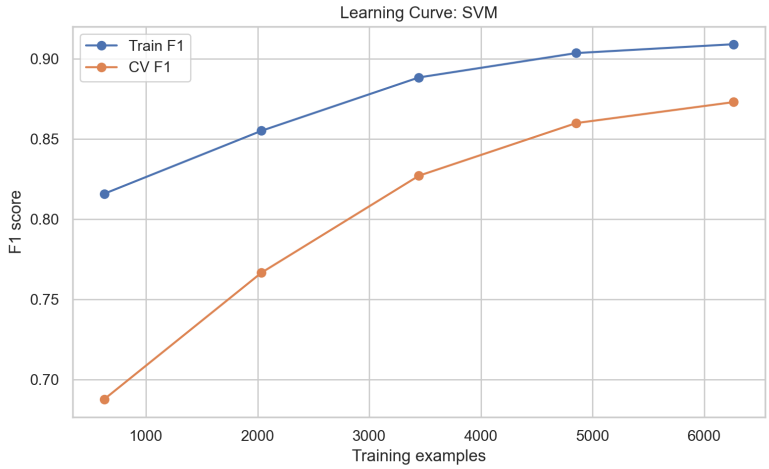



**Fig. 21** Learning Curve for SVM Model

cross-validation score, which plateaus at a moderate level. This substantial gap indicates that the multilayer perceptron is overfitting to the training set and may require more aggressive regularization, architectural adjustments, or substantially more data to improve its generalization.

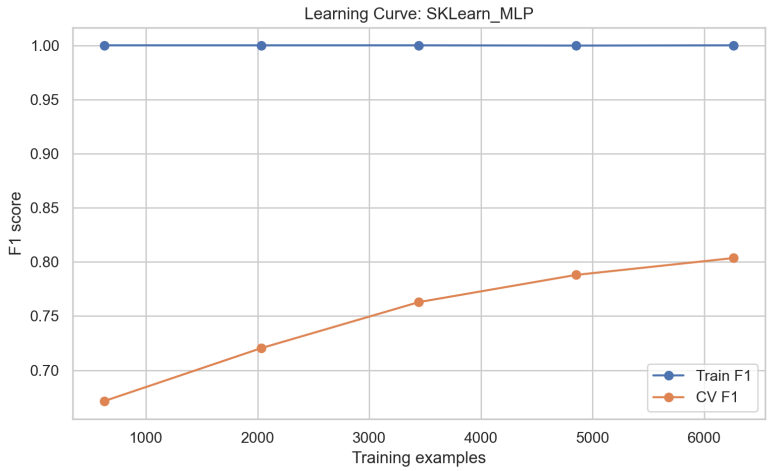



**Fig. 22** Learning Curve for Scikit learn MLP Model

The training history for the Keras MLP model shown in Fig 23 indicates that the training loss decreases rapidly. However, the validation loss plateaus after a few epochs and even shows a slight upward trend later in training. This divergence is a classic indicator of overfitting, suggesting that the model would bene t from techniques such

as early stopping, dropout, or reduced model complexity to improve its validation performance.

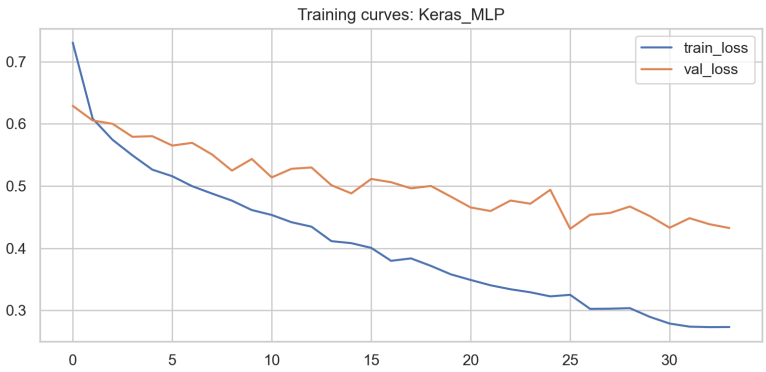



**Fig. 23** Learning Curve for Keras MLP Model

Fig 24 illustrates the learning curve for Logistic Regression. Both the training and cross-validation F1 scores remain low and stable across all dataset sizes, with almost no gap between them. This at, low-performance profile is a definitive sign of underfitting, indicating that the linear model is fundamentally too simple to capture the complex, nonlinear relationships governing stroke risk in this patient population.

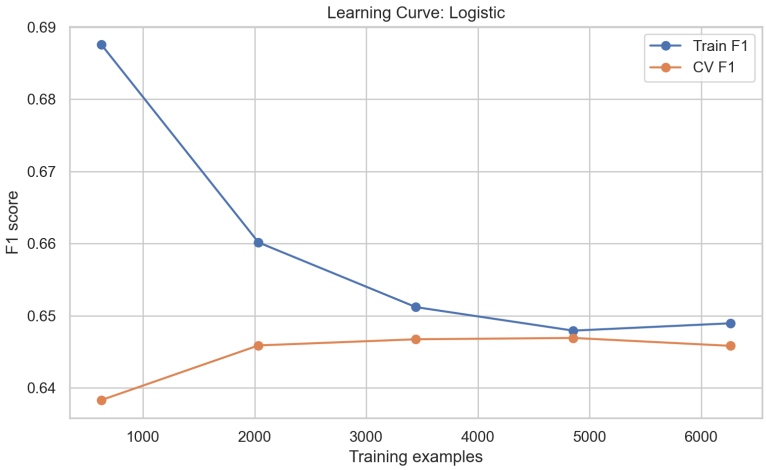



**Fig. 24** Learning Curve for Logistic Regression Model

## 3.9 Scalability Analysis for Machine Learning Model Deployment

Scalability was evaluated to determine the feasibility of deploying the trained models in real-world clinical environments, particularly those with constrained computational resources such as primary health facilities, mobile health systems, and edge devices. The assessment focused on four dimensions of computational efficiency: inference latency, training time, memory requirements during model training and inference, and final model size, as shown in Table 3. These metrics collectively indicate whether each model can operate effectively in settings with limited hardware capacity, response time, and storage.

Inference latency refers to the average time required for a model to generate a prediction for a single patient. Low latency is essential for use cases such as point-of-care screening, triage support, and mobile health ap- plications that require real-time responses. The results show that most models have microsecond-level inference times, with the Decision Tree and SKLearn_MLP emerging as the fastest learners. In contrast, the Keras_CNN1D and Support Vector Machine models demonstrated higher inference times, reflecting the computational intensity of convolutional and kernel-based operations.

Training time measures how long each model requires to learn from the dataset. Models with shorter training times are advantageous for scenarios requiring frequent retraining, such as incorporating new patient data or adapting to population-specific trends. Logistic Regression and Naive Bayes trained the fastest, completing in under one second, while deep learning models such as Keras_CNN1D required substantially longer training durations. These differences reflect the greater parameter complexity and optimization demands of neural architectures.

Training memory and inference memory quantify peak RAM usage during model learning and prediction generation, respectively. Naive Bayes and Decision Tree models consumed the least memory, making them suitable for budget-constrained environments, whereas neural networks, particularly Keras_MLP and Keras_CNN1D, required significantly more memory during training. Despite this, their inference memory remained modest, indicating that deployment is still feasible on mid-range systems.

Model size reflects storage requirements for saving and loading the trained model. This metric is critical for implementation in devices with limited disk capacity, such as mobile phones or embedded systems. Naive Bayes and Logistic Regression yielded the smallest models, occupying only a few kilobytes. At the same time, Random Forest produced the largest model, exceeding 30 megabytes due to the large number of trees and the storage of split parameters.

Overall, the scalability results demonstrate that lightweight models such as Logistic Regression, Decision Tree, and Naive Bayes are the most efficient for low-resource deployment. However, more complex models such as XGBoost and LightGBM maintain a strong balance between computational efficiency and predictive performance, making them suitable for both resource-limited and high-performance clinical environments. Deep learning models deliver strong predictive performance but require substantial computational resources during training, though their inference demands remain manageable.

**Table 3** Scalability Metrics for Machine Learning Models

| Model | Latency (s) | Train Time (s) | Train Mem (MB) | Infer Mem (MB) | Size (KB) |
|---|---|---|---|---|---|
| LightGBM | $2.15 \times 10^{-5}$ | 1.03 | 10.01 | 0.15 | 685.80 |
| XGBoost | $5.57 \times 10^{-6}$ | 1.60 | 9.06 | 0.07 | 400.21 |
| Keras CNN1D | $1.77 \times 10^{-4}$ | 49.15 | 44.47 | 0.63 | 135.77 |
| Random Forest | $4.81 \times 10^{-5}$ | 1.48 | 32.06 | 0.02 | 32392.67 |
| Support Vector Machine | $9.59 \times 10^{-4}$ | 27.20 | 20.40 | 0.02 | 2357.71 |
| SKLearn MLP | $1.22 \times 10^{-6}$ | 19.07 | 1.10 | 0.23 | 219.46 |
| Decision Tree | $5.81 \times 10^{-7}$ | 0.15 | 0.69 | 0.01 | 98.82 |
| Keras MLP | $1.73 \times 10^{-4}$ | 16.64 | 226.91 | 0.52 | 245.88 |
| Logistic Regression | $1.26 \times 10^{-6}$ | 0.04 | 3.16 | 0.04 | 1.36 |
| Naive Bayes | $1.77 \times 10^{-6}$ | 0.01 | 0.11 | 1.94 | 2.76 |

Note: Latency is measured per inference. Train Mem and Infer Mem indicate peak memory usage during training and inference, respectively.

## 3.10 Explainability and Interpretability

The integration of explainable artificial intelligence techniques is central to ensuring that the machine learning models developed in this study are transparent, trustworthy, and clinically meaningful. In particular, SHapley Additive exPlanations and Local Interpretable Model Agnostic Explanations were employed to provide both global and local insights into model behavior. Together, these methods enable a detailed examination of how individual features influence predictions, thereby facilitating clinical interpretation and supporting informed decision making.

### 3.10.1 Global Feature Importance

Figures 27 and 28 present the mean absolute SHAP values for the XGBoost and LightGBM models, respectively. These bar plots summarize the average contribution of each feature to the magnitude of the model output across the entire dataset. Notably, HeartDisease_1 and BMI consistently emerge as dominant predictors in both models. This finding aligns with established clinical evidence indicating that a history of heart disease and elevated body mass index are major risk factors for stroke among individuals with coronary heart disease. The strong agreement in feature rankings between the two highest performing models further supports the robustness and clinical plausibility of these predictors.

Figure 29 extends this analysis by comparing global feature importance across all evaluated models. The results demonstrate that tree-based models, including LightGBM, XGBoost, and Random Forest, consistently assign greater importance to clinically relevant variables such as heart disease and BMI than other modeling approaches. This comparison is particularly valuable for model selection, as it highlights not only predictive performance but also alignment with established clinical reasoning.

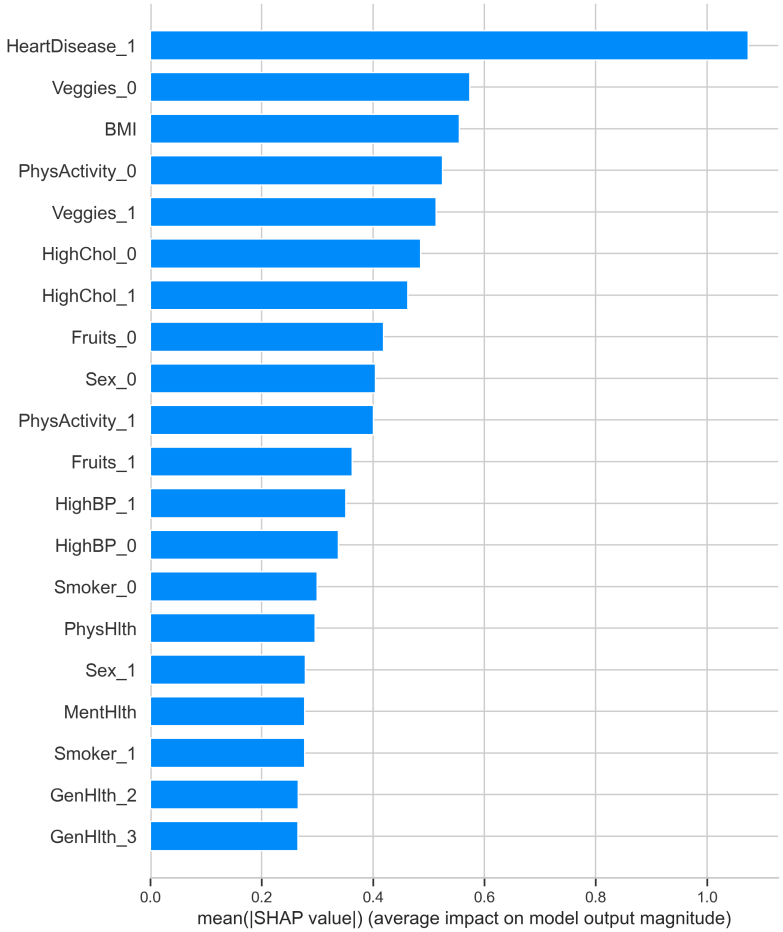



**Fig. 25** Global feature importance (mean absolute SHAP value) for the XGBoost model.

### 3.10.2 Global Model Output Distributions

Figure 30 illustrates a SHAP beeswarm plot that aggregates feature level contributions across all samples and models. The distribution and density of points reveal how each feature influences predictions throughout the population. Variables such as body mass index (BMI) and physical health days (PhysHlth) exhibit a wide range of SHAP values, reflecting complex and nonlinear relationships with stroke risk. Importantly, the strongest performing models, namely LightGBM and XGBoost, display more concentrated and structured attribution patterns, whereas lower performing models such as support vector machines and decision trees produce more diffuse and less interpretable effects.

A more detailed view of feature level behavior is provided in Figure 31, which focuses exclusively on the LightGBM model. This plot clearly demonstrates that higher values of `HeartDisease_0`, indicating no history of heart disease, are associated with

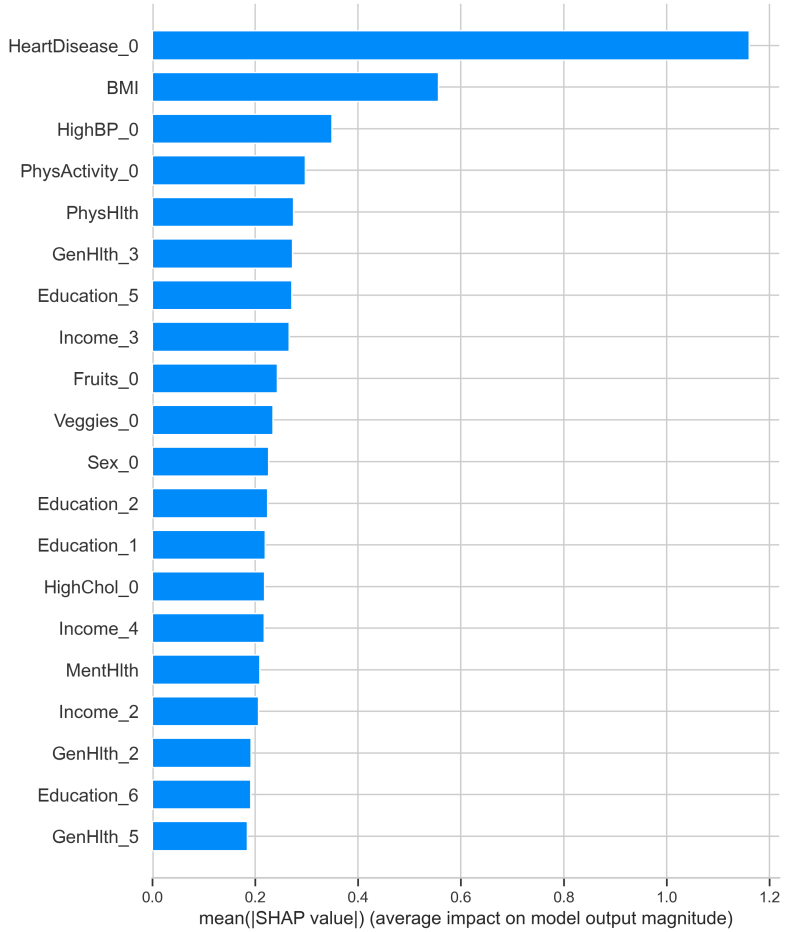



**Fig. 26** Global feature importance (mean absolute SHAP value) for the Light GBM model.

negative SHAP values and therefore lower predicted stroke risk. In contrast, increasing values of BMI are associated with positive SHAP values, indicating a greater contribution to predicted risk.

Together, these results provide direct insight into both the direction and magnitude of individual feature effects, reinforcing the interpretability and clinical relevance of the proposed modeling framework.

### 3.10.3 Local Explanations for Individual Predictions

Local explanation visualizations are essential for understanding individual predictions and represent a prerequisite for the clinical deployment of predictive models. Figures 32 and 33 present SHAP force plots for low risk predictions generated by the XGBoost and LightGBM models. These plots decompose each prediction by illustrating how individual features shift the model output from the baseline expectation $E[f(X)]$ to

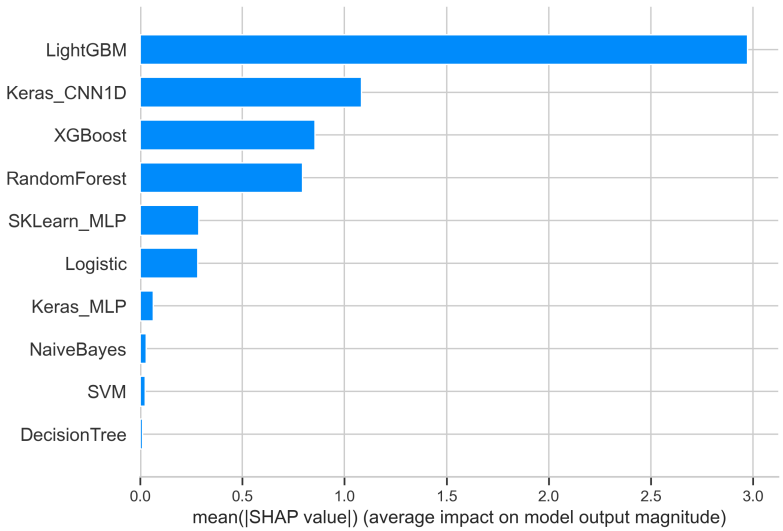



**Fig. 27** Comparison of global feature importance across all evaluated models

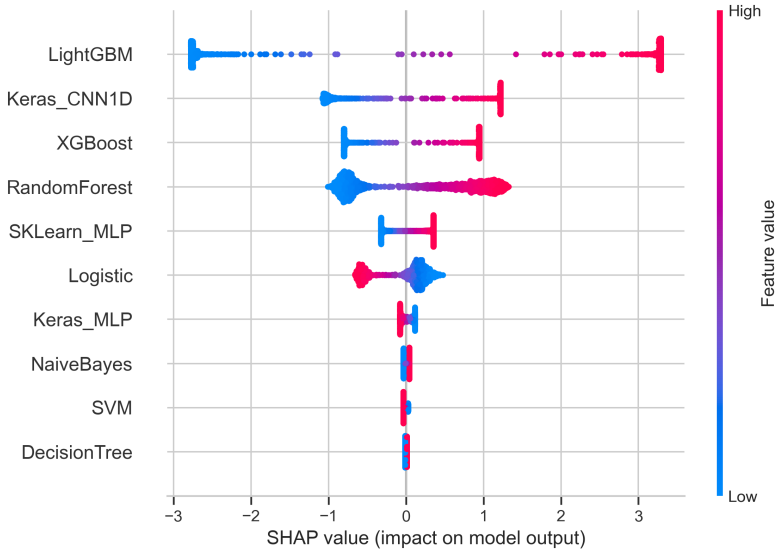



**Fig. 28** SHAP beeswarm plot showing feature effects across all models

the final predicted value $f(x)$. In the XGBoost example, features such as `HighChol_0` and `Veggies_0` contribute to a reduction in predicted risk, while a higher body mass index (BMI) partially offsets this effect. Nevertheless, the cumulative contribution remains indicative of a low risk classification.

Figures 34 and 35 show force plots corresponding to high risk predictions. In these cases, features including `HeartDisease_1`, `Age_10`, and `Education_6` are the primary

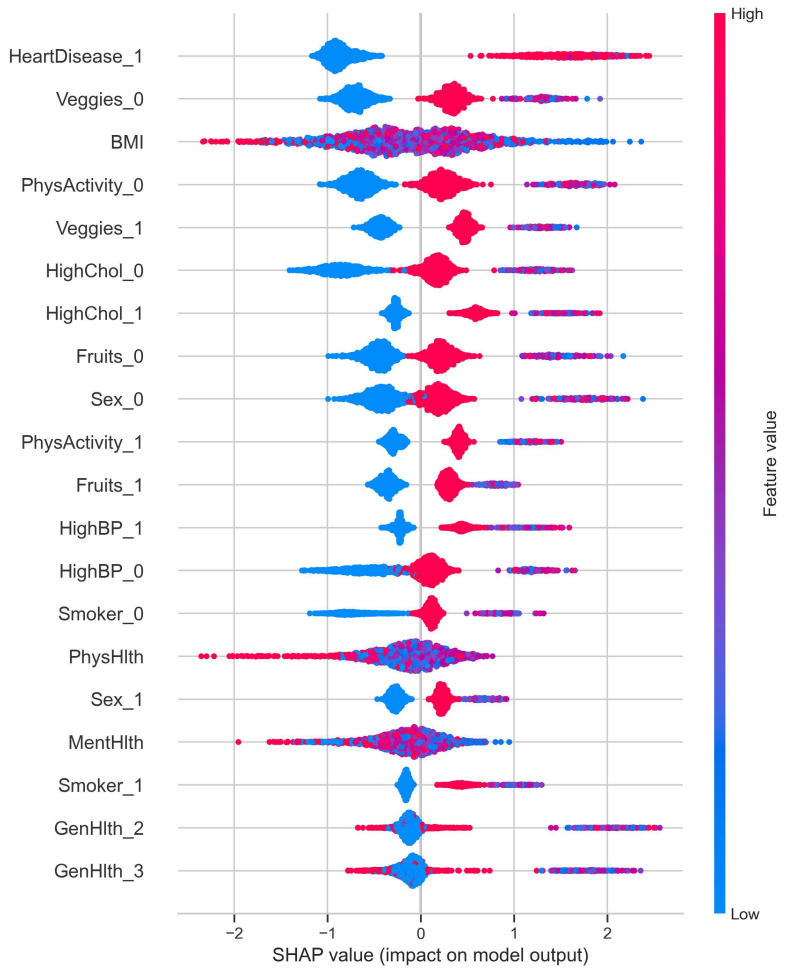



**Fig. 29** SHAP beeswarm plot for the LightGBM model

contributors driving the prediction above the baseline value. The additive structure of SHAP values allows for straightforward auditing of these decisions, making it clear why a specific patient is identified as high risk.

To complement the SHAP based explanations, Figures 36 and 37 present Local Interpretable Model agnostic Explanations (LIME) for selected LightGBM predictions. LIME constructs a locally faithful and interpretable surrogate model around each individual prediction. For the low risk case, conditions such as `Age_8` $>$ `0.00` and `Diabetes_2` $>$ `0.00` support the predicted outcome. In the high risk example, conditions including `Age_10` $>$ `0.00` and `Income_4` $>$ `0.00` contribute positively to the classification. The combined use of SHAP and LIME offers complementary perspectives, with SHAP providing a theoretically grounded and consistent attribution framework and LIME offering a simple rule based explanation that may be more intuitive for some clinical users.

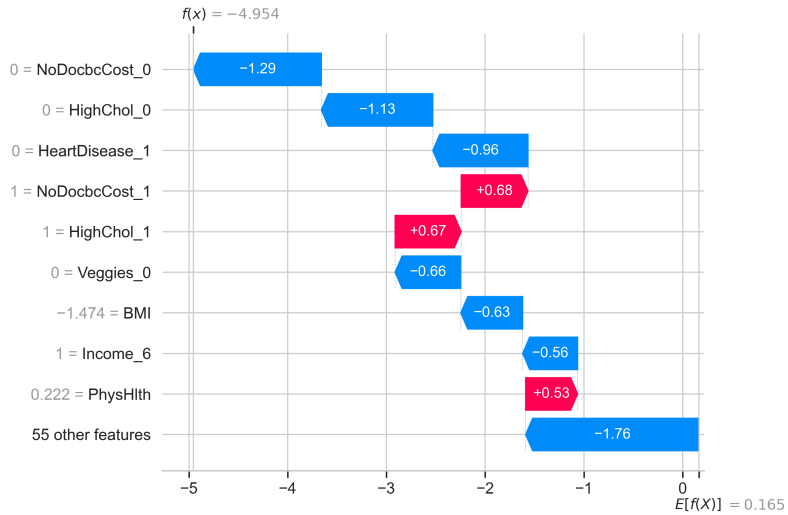



**Fig. 30** Local SHAP explanation for a low-risk prediction from the XGBoost model

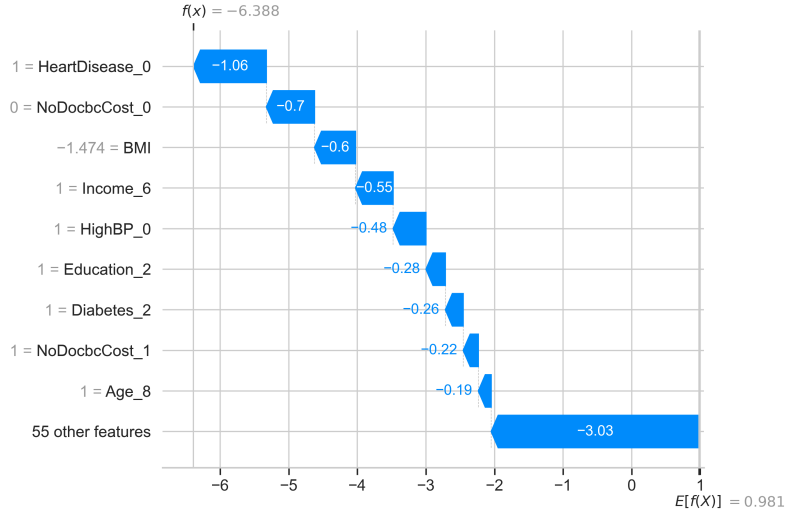



**Fig. 31** Local SHAP explanation for a low-risk prediction from the LightGBM model

### 3.10.4 Synthesis and Clinical Utility

Taken together, these explainability analyses transform the predictive models from opaque systems into transparent and auditable decision support tools. The global explanations shown in Figures 27, 28, 29, and 30 validate the models at the population level by confirming that the most influential features are clinically plausible. The local explanations presented in Figures 32 through 37 provide patient specific insights that are directly applicable at the point of care. Collectively, these results demonstrate that

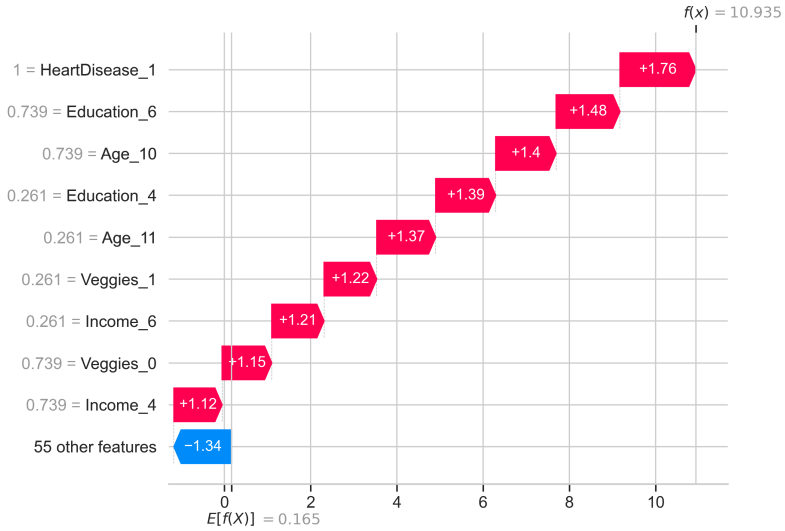



**Fig. 32** Local SHAP explanation for a high-risk prediction from the XGBoost model.

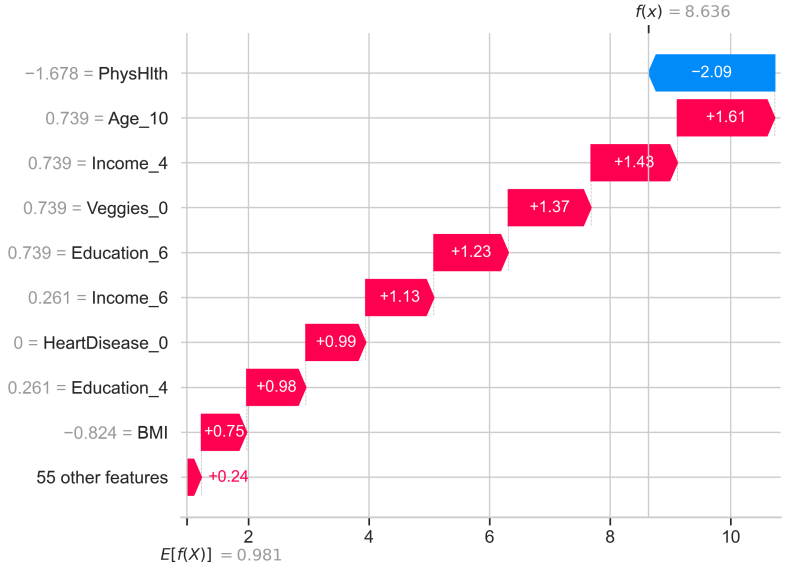



**Fig. 33** Local SHAP explanation for a high-risk prediction from the LightGBM model.

the proposed models achieve not only strong predictive performance but also a level of interpretability that is essential for real world clinical adoption.

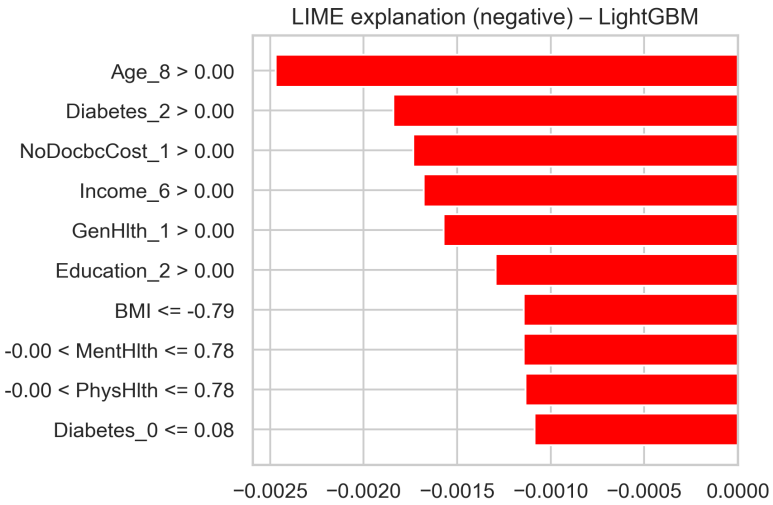



**Fig. 34** LIME explanation for a low-risk prediction generated by the LightGBM model, highlighting the most influential features in the local decision boundary.

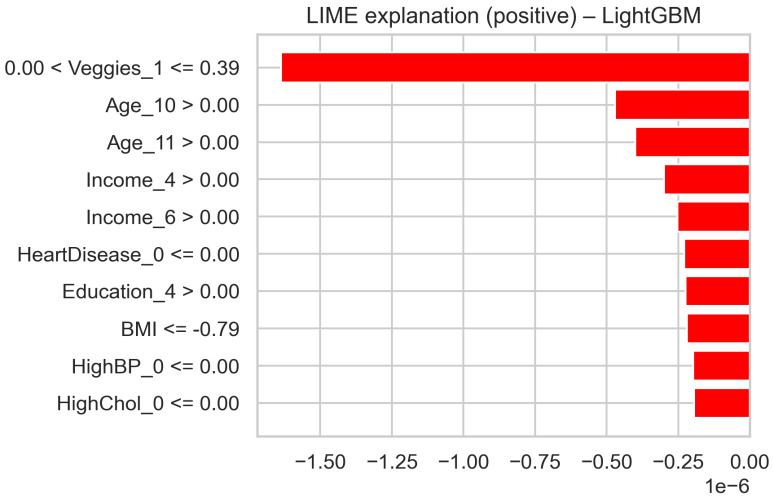



**Fig. 35** LIME explanation for a high-risk prediction generated by the LightGBM model.

# 4 Discussion

This study evaluated the effectiveness of multiple machine learning paradigms for predicting stroke risk among patients with coronary heart disease and provides several important contributions to the growing body of literature on artificial intelligence driven cardiovascular risk prediction. While previous studies have frequently examined predictive performance or model interpretability independently, the present work integrates discrimination ability, computational scalability, and explainability within

a unified analytical framework. The findings demonstrate that ensemble learning approaches, particularly stacked ensembles and gradient boosting algorithms, consistently outperform deep learning models, conventional machine learning techniques, and classical statistical methods in stroke risk prediction. These findings reinforce and extend earlier studies that identified ensemble based models as highly effective tools for modeling complex and heterogeneous clinical risk patterns in structured healthcare datasets [2–5].

The superior performance of the stacked ensemble, LightGBM, and XGBoost models highlights the strength of boosting strategies in capturing nonlinear relationships and higher order interactions among clinical variables. Stroke development among coronary heart disease patients is influenced by multifactorial and interacting physiological processes, and boosting based models are particularly well suited for capturing such complex risk structures. The consistent discrimination and calibration observed across these models suggest strong predictive robustness and stability across patient subgroups. These results are consistent with prior studies demonstrating that ensemble boosting architectures outperform single learner models in cardiovascular and cerebrovascular risk prediction [6, 8, 9]. Furthermore, performance improvements observed in this study may be attributed to the iterative learning mechanisms of boosting algorithms, which sequentially integrate weak learners into highly accurate and clinically coherent decision functions [2, 3, 34].

Deep learning models demonstrated moderate to high predictive performance, with convolutional neural networks outperforming fully connected multilayer perceptron architectures. This observation aligns with previous studies suggesting that convolutional structures can extract hierarchical feature representations even from tabular clinical datasets. However, their performance advantages often depend on access to very large datasets and extensive computational resources [2, 10, 26]. In contrast, conventional machine learning algorithms such as support vector machines and random forests produced acceptable predictive performance but exhibited lower discrimination compared to boosting approaches. Decision tree models demonstrated signs of overfitting, which is consistent with widely documented limitations of single tree based models in complex clinical prediction environments [13, 19, 21]. Collectively, these results indicate that although multiple modeling strategies can achieve reasonable predictive performance, ensemble boosting methods provide the most favorable balance between predictive accuracy, generalizability, and model stability.

Classical statistical approaches, including logistic regression and naive Bayes classifiers, exhibited inferior performance relative to machine learning based models, particularly in identifying high risk individuals as reflected by lower recall values. This finding supports previous investigations demonstrating that linear modeling assumptions restrict the ability of traditional statistical models to capture nonlinear and interaction driven relationships that characterize stroke and cardiovascular disease progression [1, 31]. While logistic regression remains valuable for statistical inference and clinical interpretability, its predictive limitations underscore the importance of integrating advanced machine learning methodologies for enhanced risk stratification.

In addition to predictive accuracy, computational scalability represents a critical consideration for clinical implementation. The scalability analysis conducted in

this study revealed that LightGBM and XGBoost provide strong predictive performance while maintaining relatively low computational cost and inference latency. These properties enhance their suitability for real time clinical decision support systems, particularly within resource constrained healthcare environments. Although deep learning models required greater training time and computational resources, their efficient inference performance suggests potential utility within centralized or cloud based healthcare infrastructures [24, 27, 32]. These findings emphasize the importance of evaluating computational efficiency alongside predictive performance when selecting models for real world clinical deployment.

Explainable artificial intelligence techniques provided valuable insights into model behavior at both population and individual patient levels. Global SHAP analyses consistently identified age, prior cardiovascular events, body mass index, and general health indicators as dominant contributors to stroke risk. These findings corroborate established clinical knowledge and align with previous explainable AI studies in cardiovascular disease risk modeling [2, 6, 10, 14]. Importantly, the stability of feature importance rankings across LightGBM and XGBoost models suggests reliable and reproducible attribution patterns, which are essential prerequisites for clinical trust, regulatory acceptance, and safe clinical integration [3, 9]. SHAP beeswarm analyses further revealed complex nonlinear relationships and interaction driven effects for key predictors, including body mass index and physical health indicators, highlighting risk dynamics that are not readily captured by linear statistical models [19, 25].

At the individual patient level, SHAP and LIME explanations provided complementary and clinically interpretable insights into model predictions. High risk classifications were predominantly influenced by advanced age, prior cardiac events, and deteriorating general health status, while protective lifestyle related characteristics contributed to reduced stroke risk among low risk individuals. The integration of SHAP and LIME reflects emerging best practice recommendations in explainable artificial intelligence, combining theoretically grounded feature attribution with intuitive local explanation frameworks that enhance clinician understanding and decision making [16, 17, 24, 25, 32, 37].

A notable methodological contribution of this study is the comparative integration of classical statistical modeling with explainable machine learning approaches. Logistic regression identified several statistically significant predictors, including prior myocardial infarction, confirming established clinical risk associations. However, explainable ensemble models revealed heterogeneous and subgroup specific feature effects that were not captured by linear statistical modeling. For example, body mass index appeared protective in logistic regression but demonstrated risk amplifying effects within specific patient subgroups in tree based models. This divergence highlights the limitations of relying exclusively on linear modeling assumptions and emphasizes the value of integrating statistical inference with explainable machine learning frameworks to support personalized and precision based clinical decision making.

This study also extends prior stroke prediction research by situating stroke risk classification within broader prognostic and disease progression considerations. Previous investigations have primarily focused on stroke occurrence or post stroke outcomes independently. By integrating high performing predictive models with transparent

interpretability mechanisms, the present framework contributes to bridging the gap between methodological advancement and real world clinical applicability, thereby supporting more individualized and clinically actionable risk stratification strategies [13, 26, 31].

Despite these contributions, several limitations should be acknowledged. The use of self reported survey data introduces potential recall bias and misclassification of behavioral and lifestyle variables. The cross sectional study design limits causal inference and restricts the ability to model temporal disease progression. External validation using electronic health record or registry based datasets is required to confirm generalizability across diverse clinical populations. Additionally, the absence of longitudinal clinical measurements limits the ability to capture dynamic risk trajectories. Future research should incorporate multimodal datasets, including longitudinal clinical records, imaging data, and genetic information, to further improve predictive accuracy, model robustness, and clinical applicability.

# 5 Conclusion

This study provides a comprehensive evaluation of machine learning approaches for stroke risk prediction among patients with coronary heart disease and demonstrates the superiority of ensemble learning methods for structured clinical risk modeling. Among the evaluated techniques, stacked ensemble models, LightGBM, and XGBoost consistently achieved the highest predictive accuracy, strong generalization performance, and favorable computational efficiency. These findings confirm that ensemble based algorithms are particularly effective in capturing complex nonlinear relationships and interaction driven risk dynamics that characterize cerebrovascular complications in coronary heart disease populations.

A central contribution of this work is the integration of high performance predictive modeling with explainable artificial intelligence techniques to enhance clinical transparency and interpretability. The combined application of Shapley Additive Explanations and Local Interpretable Model agnostic Explanations enabled comprehensive assessment of both global feature importance and patient specific prediction rationale. This integration directly addresses a major limitation of many artificial intelligence based clinical prediction tools by transforming high accuracy models into interpretable and auditable decision support systems. The observed agreement between statistically significant predictors identified through logistic regression and feature attribution patterns derived from explainable machine learning further strengthens the clinical validity and reliability of the identified stroke risk determinants. Additionally, the ability of explainability techniques to reveal nonlinear and subgroup specific risk relationships supports more individualized and precision oriented clinical decision making.

From a clinical and public health perspective, the findings of this study highlight the potential of explainable ensemble machine learning frameworks to support early identification of high risk individuals and improve preventive care strategies. Given the substantial global burden of stroke and its strong association with coronary heart disease, the combination of predictive accuracy, scalability, and transparent decision logic

demonstrated by the proposed framework supports its potential integration into real time clinical decision support systems across cardiology practice, primary healthcare settings, and population level screening programs.

Despite these contributions, further research is required to enhance clinical translation and generalizability. External validation using multicenter cohorts, electronic health record based registries, and longitudinal clinical datasets is necessary to confirm robustness across diverse populations. Future studies should also explore the integration of multimodal and temporal clinical information, including laboratory, imaging, and genetic data, to support dynamic risk prediction and improve model adaptability over time.

Overall, this study contributes to the advancement of explainable artificial intelligence in cardiovascular medicine by demonstrating that hybrid ensemble learning frameworks can simultaneously achieve high predictive performance, interpretability, and clinical applicability. The proposed approach provides a scalable and transparent foundation for next generation stroke prevention strategies and supports the broader adoption of trustworthy artificial intelligence solutions in precision healthcare.

## Acknowledgement

The authors acknowledge the use of artificial intelligence based tools during the preparation of this manuscript. Grammarly was used to support grammar, style, and language refinement, and ChatGPT was used to assist in generating Python code for data preprocessing, analysis, and visualization. The use of these tools was limited to technical and editorial support. All study design decisions, data handling, model development, result interpretation, and conclusions were performed and validated by the authors, who take full responsibility for the integrity and originality of the work.

## Author Contributions

M.W. conceptualized the study, designed the methodology, performed data preprocessing, developed and implemented the machine learning models, conducted statistical analysis, interpreted the results, and prepared the original manuscript draft. D.K., F.M., and Z.N. contributed to literature review, data interpretation, manuscript writing, and critical revision of the manuscript. All authors reviewed and approved the final manuscript. Correspondence and requests for materials should be addressed to M.W.

## Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

## Competing Interests

The authors declare no competing interests.

## Data Availability Statement

The dataset analysed during the current study is publicly available from two repositories. The primary dataset obtained from IEEE DataPort can be accessed via the permalink https://ieee-dataport.org/documents/heart-disease-dataset and DOI: 10.21227/49m4-zh81.

Additionally, the processed dataset generated and analysed during this study has been deposited in Zenodo and is openly available under a Creative Commons Attribution 4.0 International license. The dataset can be accessed at the following DOI: https://doi.org/10.5281/zenodo.17897855.

## References


[1] Lolak, S., Attia, J., McKay, G.J., Thakkinstian, A.: Comparing explainable machine learning approaches with traditional statistical methods for evaluating stroke risk models: retrospective cohort study. JMIR Cardio **7**, 47736 (2023)

[2] Moulaei, K., Afshari, L., Moulaei, R., Sabet, B., Mousavi, S.M., Afrash, M.R.: Explainable artificial intelligence for stroke prediction through comparison of deep learning and machine learning models. Scientific Reports **14**(1), 31392 (2024)

[3] Srinivasu, P.N., Sirisha, U., Sandeep, K., Praveen, S.P., Maguluri, L.P., Bikku, T.: An interpretable approach with explainable ai for heart stroke prediction. Diagnostics **14**(2), 128 (2024)

[4] Heo, J., Yoo, J., Lee, H., Lee, I.H., Kim, J.S., Park, E., *et al.*: Prediction of hidden coronary artery disease using machine learning in patients with acute ischemic stroke. Neurology **99**, 55–65 (2022)

[5] Lin, L., Ding, L., Fu, Z., Zhang, L.: Machine learning-based models for prediction of the risk of stroke in coronary artery disease patients receiving coronary revascularization. PLOS ONE **19**, 0296402 (2024)

[6] Si, Y., Abdollahi, A., Ashrafi, N., Placencia, G., Pishgar, E., Alaei, K., *et al.*: Optimized feature selection and advanced machine learning for stroke risk prediction in revascularized coronary artery disease patients. BMC Medical Informatics and Decision Making **25**(1), 276 (2025)

[7] El-Sofany, H., Bouallegue, B., El-Latif, Y.M.A.: A proposed technique for predicting heart disease using machine learning algorithms and an explainable ai method. Scientific Reports **14**(1), 23277 (2024)

[8] Hassan, A., Ahmad, S.G., Munir, E.U., Khan, I.A., Ramzan, N.: Predictive modelling and identification of key risk factors for stroke using machine learning. Scientific Reports **14**, 11498 (2024)

[9] Vu, T., Kokubo, Y., Inoue, M., Yamamoto, M., Mohsen, A., Martin-Morales, A.,

*et al.*: Machine learning approaches for stroke risk prediction: findings from the suita study. Journal of Cardiovascular Development and Disease **11**, 207 (2024)

[10] Dritsas, E., Trigka, M.: Application of deep learning for heart attack prediction with explainable artificial intelligence. Computers **13**(10), 244 (2024)

[11] Bhanusri, K., Reddy, K.N., Gnaneswarachary, P., Nair, P.C., Sampath, N.: Heart stroke prediction using machine learning and explainable ai insights. In: Proceedings of the IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), pp. 1–7. IEEE, ??? (2024)

[12] Wang, K., Tian, J., Zheng, C., Yang, H., Ren, J., Liu, Y., *et al.*: Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and shap. Computers in Biology and Medicine **137**, 104813 (2021)

[13] Truong, B., Zheng, J., Hornsby, L., Fox, B., Chou, C., Qian, J.: Development and validation of machine learning algorithms to predict 1-year ischemic stroke and bleeding events in patients with atrial fibrillation and cancer. Cardiovascular Toxicology **24**, 365–374 (2024)

[14] Gao, J., Mar, P., Tang, Z.Z., Chen, G.: Fair prediction of 2-year stroke risk in patients with atrial fibrillation. Journal of the American Medical Informatics Association **31**, 2820–2828 (2024)

[15] Shen, D., Li, J., Teng, S., Li, M., Tang, X.: Development and validation of a nomogram for predicting frailty risk among older patients with ischaemic stroke. Journal of Clinical Nursing **34**, 4186–4201 (2025)

[16] Mridha, K., Ghimire, S., Shin, J., Aran, A., Uddin, M.M., Mridha, M.F.: Automated stroke prediction using machine learning: an explainable and exploratory study with a web application for early intervention. IEEE Access **11**, 52288–52308 (2023)

[17] Li, Q., Wang, X., Ye, Q., Li, L., He, X., Liu, L.: A prediction model for stroke risk based on explainable machine learning. Medicine **104**(40), 44718 (2025)

[18] Pamungkas, Y., Wibawa, A.D., Cahya, M.D.: Electronic medical record data analysis and prediction of stroke disease using explainable artificial intelligence (xai). Kinetik (2022)

[19] Gkantzios, A., Kokkotis, C., Tsiptsios, D., Moustakidis, S., Gkartzonika, E., Avramidis, T., *et al.*: From admission to discharge: predicting nih stroke scale progression using biomarkers and explainable machine learning. Journal of Personalized Medicine **13**(9), 1375 (2023)

[20] Talaat, F.M., Elnaggar, A.R., Shaban, W.M., Shehata, M., Elhosseini, M.: Cardiorisknet: A hybrid ai-based model for explainable risk prediction and prognosis in cardiovascular disease. Bioengineering **11**(8), 822 (2024)

[21] Kokkotis, C., Giarmatzis, G., Giannakou, E., Moustakidis, S., Tsatalas, T., Tsiptsios, D., *et al.*: An explainable machine learning pipeline for stroke prediction on imbalanced data. Diagnostics **12**(10), 2392 (2022)

[22] Sethi, A., Dharmavaram, S., Somasundaram, S.K.: Explainable artificial intelligence (xai) approach to heart disease prediction. In: 3rd International Conference on Artificial Intelligence for Internet of Things, pp. 1–6 (2024)

[23] Uddin, M.J., Sarker, P.K., Hoque, N., Aktar, S., Ahmad, M.M.: An explainable machine learning framework for early detection of stroke. International Multilingual Journal of Science and Technology **8** (2023)

[24] Hwang, Y.S., Kim, S., Yim, I., Park, Y., Kang, S., Jo, H.S.: Predicting the likelihood of readmission in patients with ischemic stroke using explainable machine learning. International Journal of Medical Informatics **195**, 105754 (2025)

[25] Tasnim, A.F., Rahman, R., Prabha, M., Hossain, M.A., Nilima, S.I., Al Mahmud, M.A., *et al.*: Explainable machine learning algorithms to predict cardiovascular strokes. Engineering, Technology & Applied Science Research **15**, 20131–20137 (2025)

[26] Wu, C., Zhu, S., Wang, Q., Xu, Y., Mo, X., Xu, W., *et al.*: Development, validation, and visualization of a novel nomogram to predict depression risk in patients with stroke. Journal of Affective Disorders **365**, 351–358 (2024)

[27] Ahmed, R., Varshney, A., Ashraf, Z., Farooqui, N.A., Pathak, R.S.: Enhanced stroke risk prediction: a fusion of machine learning models for improved healthcare strategies. SN Computer Science **5**, 1078 (2024)

[28] Veledar, E., Zhou, L., Veledar, O., Gardener, H., Gutierrez, C.M., Brown, S.C., *et al.*: Identifying determinants of readmission and death post-stroke using explainable machine learning. PLOS ONE **20**(9), 0332371 (2025)

[29] Liu, J., Gu, Y., Wang, D., Xia, X., Li, X.: Prediction of the short-term prognosis of acute ischaemic stroke in patients with high treatment platelet reactivity using explainable machine learning. Scientific Reports **15**(1), 36696 (2025)

[30] Zhu, H., Ye, Q., Wang, S., Cai, H., Maimaiti, M., Lai, J., *et al.*: An explainable two-stage machine learning model for predicting post-thrombolysis complications in stroke patients. Research **8**, 0817 (2025)

[31] Abujaber, A.A., Albalkhi, I., Imam, Y., Nashwan, A., Akhtar, N., Alkhawaldeh, I.M.: Machine learning-based prognostication of mortality in stroke patients.

Heliyon **10**(7) (2024)

[32] Mahin, M.A., Islam, M.M., Alam, M.Z., Pollob, A.D., Zaman, O.: Enhancing stroke risk prediction with explainable ai: leveraging resampling and machine learning. Journal of Neurointervention and Stroke **1**(2), 63–71 (2025)

[33] ZHENG, K.: HEART DISEASE DATASET. https://doi.org/10.21227/49m4-zh81 . https://dx.doi.org/10.21227/49m4-zh81

[34] Wanyonyi, M., Morris, Z.N., Musyoka, F.M., Kitavi, D.M.: Enhanced machine learning and hybrid ensemble approaches for coronary heart disease prediction. PLOS ONE **20**(12), 0328338 (2025)

[35] Wanyonyi, M., Morris, Z.N., Musyoka, F.M., Makaa, D., Kitavi, D.M.: Balancing interpretability and accuracy: a comparative evaluation of logistic regression and machine learning approaches for coronary heart disease risk prediction in low-resource settings. Authorea Preprints. Preprint (2025)

[36] Wanyonyi, M.K., Wafula, T.M., Keraro, J.K., Gogo, O.F., Maiyo, J.A.K.: Modelling and forecasting the impact of exchange rates, public debt, and tourism inflows on kenya's economic growth using machine learning techniques. General Letters in Mathematics **15**(3), 74–92 (2025) https://doi.org/10.31559/glm2025.15.3.1

[37] Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 1135–1144. ACM, San Francisco, CA, USA (2016). https://doi.org/10.1145/2939672.2939778