Eksakta : Berkala Ilmiah Bidang MIPA
VOLUME 27 NO 06 2026, pp 1130-1145
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol27-iss06/775

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

Article

Heart Attack History Classification Using XGBoost-Based Feature Selection and ADASYN

Article Info

Article history :
Received October 12, 2026
Revised November 11, 2026
Accepted November 16, 2026
Published December 30, 2026
In Press

Keywords :
Heart attack classification, XGBoost feature selection, ADASYN, SDG 3

Rahma Kurnia Dewi[1], A'yunin Sofro[2*]

[1]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, Indonesia
[2]Department of Actuarial Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, Indonesia

Abstract. Cardiovascular disease remains a major public health concern, highlighting the importance of data-driven approaches to understanding heart attack history. In the context of Sustainable Development Goal 3 (SDG 3), this study aims to classify individuals based on their reported history of physician-diagnosed heart attack using data from the fifth wave of the Indonesia Family Life Survey (IFLS5). The dataset comprises 32,173 individuals, including 315 positive cases (0.98%). The proposed approach combines XGBoost-based feature selection, Adaptive Synthetic Sampling (ADASYN), and seven classification algorithms. From ten initial predictors, four features were selected: Diabetes, HighCholesterol, BMI, and Weight_kg. Model evaluation used leakage-safe stratified 10-fold cross-validation and an independent holdout set, with a probability threshold of 0.50. Logistic Regression achieved the highest cross-validation ROC-AUC (0.603 ± 0.045), while SVM obtained the highest holdout ROC-AUC (0.6257). FDA approximation achieved the highest holdout recall (0.6190), but its precision was only 0.0161, resulting in 2,378 false-positive predictions. These findings indicate limited classification performance and a substantial false-positive burden. The study provides an exploratory analysis of reported heart attack history and should not be interpreted as a clinical diagnostic or future-risk prediction tool.

This is an open acces article under the CC-BY license.

This is an open access article distributed under the Creative Commons 4.0 Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2026 by author.

Corresponding Author :
Author A'yunin Sofro
Department of Actuarial Science, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Surabaya, Surabaya, Indonesia
Email : ayuninsofro@unesa.ac.id

1. Introduction

Cardiovascular disease (CVD) remains a major non-communicable disease burden worldwide and in Indonesia. Recent evidence indicates that cardiovascular mortality in Indonesia has changed substantially over the past three decades, while CVD remains associated with several modifiable risk factors, including hypertension, elevated body-mass index, and abnormal glucose levels [1]. Globally, CVD continues to account for a substantial proportion of deaths [2]. Hypertension is an established cardiovascular risk factor [3], while obesity and type 2 diabetes are associated with an increased risk of cardiovascular complications [4]. Hypertension has also been discussed in relation to other organ-related health conditions, including kidney fibrosis, highlighting its broader implications for health research [5]. These conditions highlight the importance of examining cardiovascular-related health outcomes using population-based data.

Machine learning (ML) has increasingly been investigated for cardiovascular disease classification and risk prediction. Machine learning has also been applied in other biomedical studies, such as breast cancer biomarker validation using metabolomics data [6], demonstrating its broader application in health-related research. Recent systematic reviews have examined the use of ML models for CVD risk prediction and heart disease classification, while also highlighting challenges related to data quality, model validation, and clinical applicability [7–8]. Several studies have applied feature selection to identify relevant predictors and reduce the dimensionality of cardiovascular datasets. For instance, Hartono et al. [9] compared K-Nearest Neighbor, Decision Tree, and Random Forest algorithms for heart disease classification, providing an example of the application of machine learning methods to cardiovascular-related data.

XGBoost-based feature importance has been used to identify relevant predictors for coronary artery disease and cardiovascular disease prediction [10–12]. Other gradient boosting algorithms have also been investigated for classification problems, including household food resilience classification during the COVID-19 pandemic [13]. Other studies have compared different feature selection algorithms and classification models for cardiovascular prediction [14–15]. These findings indicate that feature selection can support the development of more focused classification models. However, the selected predictors and model performance may vary depending on the characteristics of the dataset and the target outcome.

Class imbalance is another important challenge in cardiovascular classification, particularly when the positive outcome represents a relatively rare condition. Previous studies have investigated different resampling strategies to improve classification performance on imbalanced cardiovascular datasets [16–18]. One commonly used approach is Adaptive Synthetic Sampling (ADASYN), which generates synthetic minority-class observations by assigning greater sampling emphasis to minority observations that are more difficult to learn [19]. ADASYN and related adaptive oversampling approaches have also been investigated in medical prediction tasks [20]. Nevertheless, increasing the number of minority-class observations does not automatically guarantee better generalization. The effects of resampling should therefore be considered together with the validation procedure and the original class distribution of the evaluation data.

The choice of classifier also influences classification performance because different algorithms represent relationships between predictors and outcomes in different ways. Neural Network models have been applied to cardiovascular and cardiac disorder classification [21–22]. Flexible Discriminant Analysis (FDA) provides a flexible discriminant approach by allowing nonlinear relationships through

transformations of the predictor space [23] and has been applied in coronary artery disease classification [24]. Conventional methods, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, and Support Vector Machine, also provide different approaches to modeling classification boundaries. The application of classification algorithms has also been explored in other contexts, such as COVID-19 vaccine sentiment classification using K-Nearest Neighbor and FastText [25]. Comparing these models can provide insight into their relative performance under the same data preparation and evaluation procedures.

The Indonesia Family Life Survey (IFLS) provides population-based information that has been used in studies examining health conditions and their associated factors among Indonesian individuals [26–30]. IFLS5 includes information on demographic characteristics, health conditions, and physical measurements collected across several provinces in Indonesia. Previous studies have used IFLS5 to investigate multimorbidity and cardiovascular-related risk factors, demonstrating its applicability for population-based health research [28–30]. In this study, the response variable represents whether an individual has ever reported receiving a doctor-diagnosed heart attack. Therefore, the analysis concerns the classification of previously reported heart attack history rather than the prediction of future heart attack occurrence. Since the outcome is obtained from retrospective survey responses, the classification results should be interpreted in relation to the reported diagnosis and the information available in the survey.

Although previous studies have investigated ML-based cardiovascular prediction, feature selection, and class imbalance handling, the classification of doctor-diagnosed heart attack history using highly imbalanced population-based survey data remains a relevant methodological problem. In particular, the use of retrospective survey data requires careful interpretation of the target outcome, while the severe class imbalance requires an evaluation procedure that preserves the original class distribution in validation and testing observations. Data preprocessing, feature selection, and resampling must also be performed without allowing information from validation observations to influence model training. Data leakage can lead to overly optimistic estimates of model performance when information from evaluation data is inadvertently used during model development [31–34].

This study also aligns with the Sustainable Development Goals (SDGs), particularly SDG 3, which aims to ensure healthy lives and promote well-being for all at all ages. By examining the classification of physician-diagnosed heart attack history using machine learning, this study contributes to the broader effort to support cardiovascular health research and evidence-based health assessment. Although the study does not directly evaluate clinical interventions or predict future heart attacks, its findings may provide insights into the use of health survey data and machine learning methods for identifying patterns associated with a history of heart attack.

This study develops a machine learning framework that combines XGBoost-based feature selection, ADASYN, and seven classification models to classify doctor-diagnosed heart attack history using IFLS5 data. The initial predictor set consists of ten demographic, metabolic, anthropometric, and blood-pressure variables selected based on their relevance to cardiovascular health and their availability in the analytical dataset. XGBoost is then used to rank the predictors based on their contribution to the fitted model, and the selected variables are used as inputs for the classification stage. The models are evaluated using stratified 10-fold cross-validation, in which preprocessing, feature selection, and ADASYN are fitted independently within each training fold. The final evaluation is conducted using an untouched holdout test set that retains the original class distribution. The objective of this study is to compare the classification performance of seven models under severe class imbalance and to examine the consistency between internal validation and holdout performance. The findings are intended to provide a methodological assessment of population-based heart attack history classification rather than a clinical diagnostic tool.

2. Methods

The overall research procedure, from data preparation to model evaluation, is illustrated in Figure 1. The workflow consists of data partitioning, leakage-safe cross-validation, model development, and independent holdout testing.

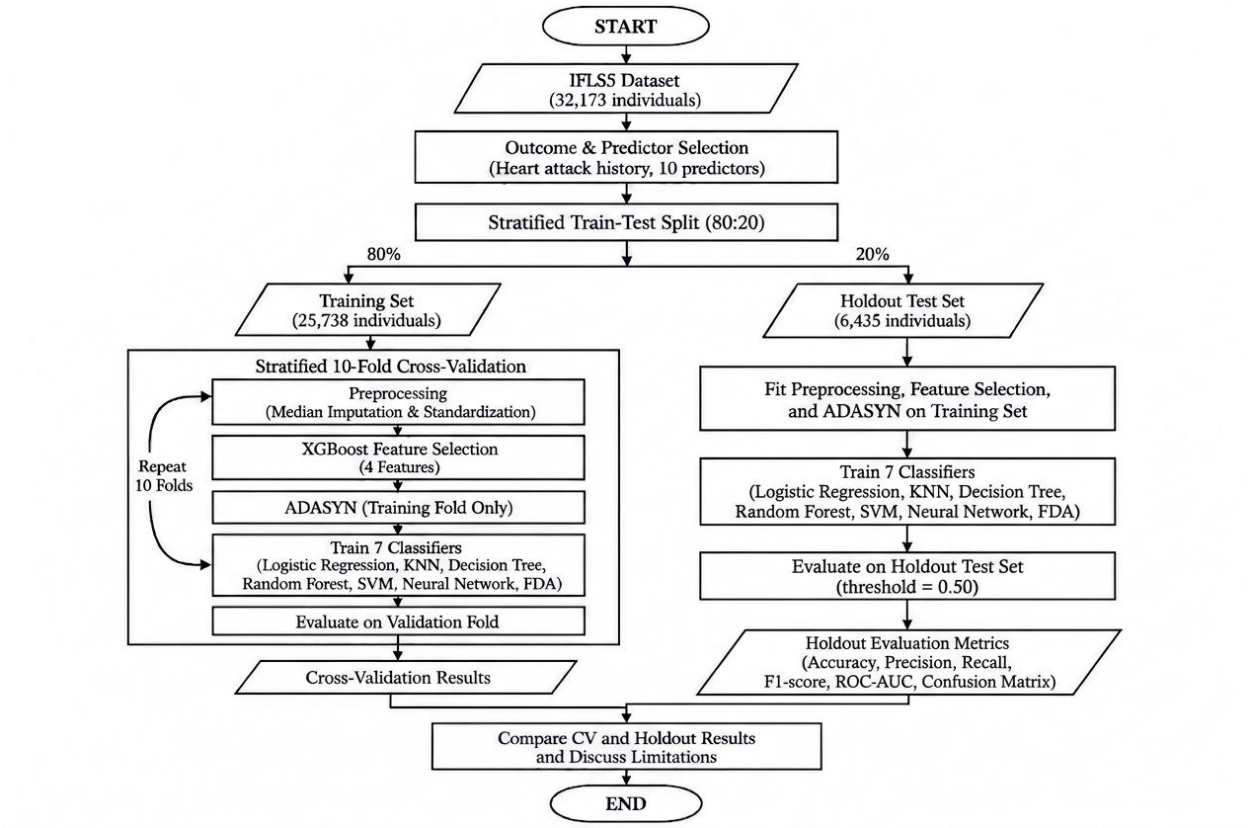


Figure 1. Research Flowchart of the Proposed Classification Framework

2.1. Materials

The data used in this study were obtained from the fifth wave of the Indonesia Family Life Survey (IFLS5), a large-scale population-based survey covering 13 provinces in Indonesia [26–28]. IFLS has been used in previous population-based studies to examine health conditions and their associated factors among Indonesian individuals [28–30]. The use of IFLS5 in this study allows the analysis to examine doctor-diagnosed heart attack history using individual-level demographic and health-related information collected through a population-based survey.

The analytical dataset consisted of 32,173 individual observations with ten initial predictor variables and one binary response variable. The response variable, HeartAttack, indicates whether an individual had ever received a doctor-diagnosed heart attack. Thus, the outcome represents a previously reported diagnosis and does not indicate the occurrence of a future heart attack. The dataset contained 315 positive cases and 31,858 negative cases, corresponding to a positive-class prevalence of 0.98%.

The ten initial predictors were selected to represent several groups of characteristics relevant to cardiovascular health. These variables included demographic characteristics (Sex and Age), metabolic and health conditions (Diabetes, HighCholesterol, and Hypertension), anthropometric measurements (BMI, Weight_kg, and Height_cm), and blood-pressure measurements (SystolicBP and DiastolicBP). The selection was intended to provide an initial set of demographic, metabolic, physical, and clinical-

related measurements available in the analytical dataset. The variables were selected as candidate predictors rather than as an exhaustive representation of all potential cardiovascular risk factors available in IFLS. The study compared seven classification models: Neural Network, Flexible Discriminant Analysis (FDA) approximation, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM).

2.2. Data Partitioning and Preprocessing

The dataset was divided into training and holdout test sets using stratified sampling with a ratio of 80:20. The training set consisted of 25,738 observations, while the holdout test set contained 6,435 observations. Stratification was applied to preserve the proportion of the target classes in both subsets. Missing numerical values were replaced using the median calculated from the training data [35]. Continuous predictors were then standardized using z-score normalization:

$$z_{ij} = \frac{(x *_{ij} - \mu_j)}{\sigma_j}$$

where $x *_{ij}$ represents the imputed value, while μ_j and σ_j denote the mean and standard deviation of predictor j in the training data. The same imputation and standardization parameters were applied to the holdout data. This procedure prevented information from the holdout set from influencing preprocessing [32–34].

2.3. XGBoost Feature Selection

XGBoost was used to rank and select relevant predictors before the classification stage [11–12],[15],[36]. Its objective function combines prediction loss with a regularization term:

$$L = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

where l represents the prediction loss and Ω penalizes model complexity. The model used 300 boosting trees, binary log-loss, and a fixed random seed of 42. Feature importance was evaluated using the gain criterion [36]. To account for the imbalanced training data, the positive-class weight was set according to the ratio between negative and positive observations:

$$\omega = \frac{N_0}{N_1} = \frac{25{,}486}{252} \approx 101.13$$

Based on the feature importance ranking, four predictors were retained: Diabetes, HighCholesterol, BMI, and Weight_kg. The number of predictors was therefore reduced from 10 to 4, and the selected predictors were used as inputs for all seven classification models.

The exclusion of Age, Sex, Hypertension, SystolicBP, and DiastolicBP should be interpreted as a result of the feature-ranking procedure within the fitted XGBoost model, rather than evidence that these variables are medically irrelevant to cardiovascular disease. Age and blood pressure are recognized cardiovascular risk factors, but their contribution to the gain-based feature ranking in this particular dataset was lower than that of the four retained predictors. Feature importance reflects the contribution of variables to the fitted model and does not establish causal relationships or determine the overall medical importance of a variable.

BMI and Weight_kg were both retained because they were among the four highest-ranked predictors according to the gain criterion. However, BMI is calculated using body weight and height, meaning that BMI and Weight_kg may contain overlapping information. Their simultaneous inclusion may therefore introduce redundancy in the selected predictor set. The potential overlap

between these variables was examined using Pearson correlation and the variance inflation factor (VIF). However, no alternative feature-combination experiment was conducted in this study. The retention of both variables in the current analysis should consequently be interpreted as a model-based selection result rather than evidence that the two variables provide entirely independent information.

2.4. Class Imbalance Handling Using ADASYN

Adaptive Synthetic Sampling (ADASYN) was applied to the feature-selected training data to address the severe class imbalance [19]. Unlike random oversampling, ADASYN generates synthetic minority-class observations according to the local difficulty of learning each observation. A synthetic observation is generated by interpolating between a minority-class observation and one of its minority-class neighbors:

$$x_{new} = x_i + \lambda(x_{i,nn} - x_i), \text{ where } 0 \leq \lambda \leq 1$$

where x_i is a minority-class observation, $x_{i,nn}$ is one of its selected nearest neighbors, and λ is a random interpolation coefficient. The model used five nearest neighbors and a fixed random seed of 42 [19]. Before resampling, the training set contained 252 positive and 25,486 negative observations. After ADASYN, the number of positive observations increased to 25,427, while the negative class remained unchanged. The resampled training set therefore contained 50,913 observations, with the positive class representing approximately 49.94% of the resampled data.

The substantial increase in minority-class observations was intended to reduce the imbalance during model training. However, the synthetic observations generated by ADASYN are interpolated samples and do not represent additional individuals from the original survey population. Consequently, the resampled data were used only for model training and were not interpreted as an increase in the actual number of individuals with a doctor-diagnosed heart attack. ADASYN was applied only to the training data, while the holdout test set retained its original class distribution.

The large increase in the minority class also requires careful assessment because synthetic observations may not fully represent the variability of the original minority-class population. Therefore, the effect of ADASYN should be examined by comparing the distributions of the original minority-class observations and the generated synthetic observations. Such an assessment may include comparisons of individual predictor distributions and pairwise relationships among the selected predictors. These checks are necessary to evaluate whether the resampling procedure produces plausible observations within the selected feature space.

2.5. Classification Models

Seven classification models were evaluated using the four predictors selected by XGBoost and the ADASYN-resampled training data. Each model represented a different classification approach. The model configurations used in this study were specified as follows. The Neural Network used a 4–64–32–1 architecture with ReLU activation in the hidden layers and a sigmoid output layer. The FDA approximation used a second-degree polynomial-basis transformation followed by Linear Discriminant Analysis (LDA). Logistic Regression used L-BFGS optimization, L2 regularization, balanced class weighting, and a maximum of 1,000 iterations. KNN used five nearest neighbors with uniform distance weighting. Decision Tree used the Gini impurity criterion and balanced class weighting. Random Forest used 300 trees and balanced class weighting. SVM used a radial basis function (RBF) kernel with a regularization parameter C of 1.0 and balanced class weighting.

The Neural Network was trained using the Adam optimizer [37], with L2 regularization of 0.0001, a learning rate of 0.001, a maximum of 1,000 iterations, and random seed 42. The hidden layers used the ReLU activation function [38]. The FDA approximation included linear terms, squared terms, and pairwise interaction terms from the four selected predictors [23].

The model configurations were fixed for the comparative analysis. Based on the current experimental procedure, systematic hyperparameter optimization using Grid Search or Bayesian Optimization was not conducted. Therefore, the parameter settings should be interpreted as the configurations used in this study rather than as the optimal parameter values for each algorithm. The absence of systematic hyperparameter tuning is a limitation because the performance of each classifier may depend on its parameter configuration.

2.6. Leakage-Safe Cross-Validation

Internal model evaluation was conducted using stratified 10-fold cross-validation. To prevent information leakage, preprocessing, XGBoost feature selection, and ADASYN resampling were fitted independently within each training fold. The corresponding validation fold was not used in these operations and retained its original class distribution. This procedure ensured that the validation observations did not contribute to the estimation of preprocessing parameters, feature selection, or synthetic minority-class generation.The mean cross-validation score for each evaluation metric was calculated as:

$$M_{CV} = \frac{1}{10}\sum_{k=1}^{10} M_k$$

where M_k represents the metric obtained from the k-th fold. The mean and standard deviation across the ten folds were reported to describe model performance and its variation across folds. This validation procedure was designed to provide an evaluation that was independent of the data transformations performed within each training fold [31–34].

2.7. Holdout Test Evaluation

The final evaluation was conducted on the untouched 20% holdout test set, consisting of 6,435 observations, including 63 positive cases. No resampling was performed on the holdout data. Model performance was assessed using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). The evaluation metrics were calculated as follows:

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

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

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

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. ROC-AUC was used to assess the model's ability to distinguish between positive and negative cases across classification thresholds. Because the outcome was severely imbalanced, the metrics were interpreted jointly rather than relying on accuracy alone. Confusion matrices were also examined to assess the number of correctly and incorrectly classified observations, with particular attention to recall and precision.

For the calculation of class-based metrics and confusion matrices, the predicted probabilities were converted into class labels using a decision threshold of 0.5. This threshold was applied consistently across the classification models in the current analysis. However, a threshold of 0.5 is not necessarily optimal for a dataset with a positive-class prevalence of only 0.98%. A different threshold may produce a more suitable balance between precision and recall, depending on the intended application and the relative costs of false-positive and false-negative classifications.

The current evaluation did not apply cost-sensitive threshold optimization or select a decision threshold based on the precision-recall curve. Therefore, the reported precision, recall, F1-score, and confusion matrices represent model performance at the fixed threshold of 0.5. The threshold-dependent results should be interpreted with this limitation in mind. ROC-AUC, in contrast, evaluates discrimination across a range of thresholds and is not determined by the single threshold used to convert predicted probabilities into class labels.

3. Results and Discussion

The analysis compared seven classifiers for classifying doctor-diagnosed heart attack history using IFLS5. XGBoost selected four predictors, ADASYN was applied only to training data, and the models were evaluated using leakage-safe stratified 10-fold cross-validation and an untouched holdout test set. The following results describe feature selection, predictor association, resampling diagnostics, cross-validation, and holdout performance.

XGBoost reduced the initial ten predictors to Diabetes, HighCholesterol, BMI, and Weight_kg. Their gain-based importance scores were 0.1406, 0.1140, 0.1080, and 0.1032, respectively, representing a 60% reduction in the number of candidate predictors. The fold-wise selection-frequency analysis showed that Diabetes was selected in all ten folds, HighCholesterol in nine folds, BMI in seven folds, and Weight_kg in five folds. The lower selection frequency of Weight_kg suggests that its contribution was less stable across folds than that of Diabetes, although the final full-training-set ranking retained it.

Feature selection has been applied in previous studies to identify relevant predictors of cardiovascular disease [11–12],[14–15]. Hartono et al. [9] also examined several algorithms for heart disease classification. Similarly, this study used XGBoost gain importance to select four features. These were Diabetes, HighCholesterol, BMI, and Weight_kg. Their selection reflects the importance of these variables in the IFLS5 dataset.

Table 1. Feature importance ranking computed by XGBoost

Rank	Feature Variable	Importance Score
1	Diabetes	0.1406
2	HighCholesterol	0.1140
3	BMI	0.1080
4	Weight_kg	0.1032
5	DiastolicBP	0.0966
6	Age	0.0961
7	Height_cm	0.0912
8	SystolicBP	0.0895
9	Hypertension	0.0885
10	Sex	0.0723

Before resampling, the training set contained 25,486 negative and 252 positive observations. ADASYN increased the positive class to 25,427 observations while leaving the negative class unchanged, producing 50,913 training observations. The generated 25,175 synthetic samples were compared with the 252 original positive observations. Across the four standardized predictors, no synthetic values fell outside the original minority-class feature ranges. This marginal range check does not establish clinical plausibility, and the large synthetic-to-original ratio remains a limitation that warrants cautious interpretation.

Previous studies have investigated ADASYN and other resampling techniques to address class imbalance in cardiovascular datasets [16–18]. ADASYN generates synthetic minority samples based on their learning difficulty [19]. This approach increases positive-class representation during training. However, the results show that oversampling does not necessarily improve precision. The models still struggle to identify positive cases in the original imbalanced test data.

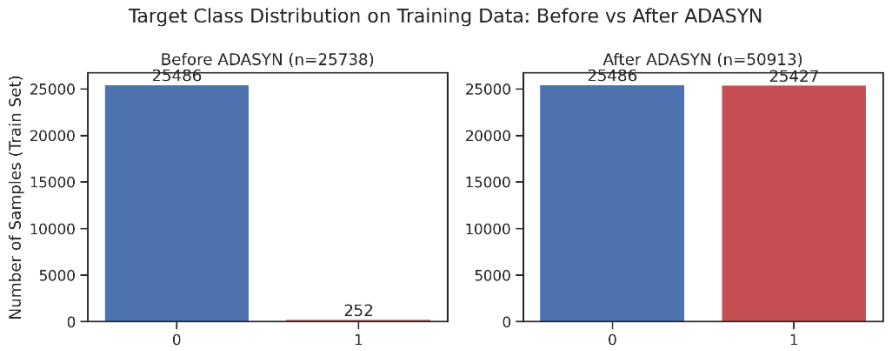


Figure 2. Target class distribution in the training data before and after ADASYN resampling.

The models were evaluated using leakage-safe stratified 10-fold cross-validation. The mean ROC-AUC was highest for Logistic Regression (0.603 ± 0.045), followed by FDA approximation (0.585 ± 0.060). SVM achieved the highest mean recall (0.484 ± 0.136), while FDA approximation obtained the highest mean precision (0.016 ± 0.004) and F1-score (0.030 ± 0.007). However, Decision Tree recall was only 0.111 ± 0.079. The contrast shows that high accuracy in this severely imbalanced dataset does not necessarily indicate useful detection of positive cases.

Table 2. Leakage-safe 10-fold cross-validation results

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.695 ± 0.129	0.014 ± 0.003	0.432 ± 0.188	0.027 ± 0.006	**0.603 ± 0.045**
KNN	0.828 ± 0.084	0.013 ± 0.008	0.178 ± 0.068	0.024 ± 0.013	0.510 ± 0.043
Decision Tree	**0.892 ± 0.085**	0.012 ± 0.008	0.111 ± 0.079	0.021 ± 0.014	0.505 ± 0.033
Random Forest	0.888 ± 0.084	0.012 ± 0.008	0.111 ± 0.073	0.021 ± 0.013	0.523 ± 0.052
SVM	0.632 ± 0.102	0.013 ± 0.004	**0.484 ± 0.136**	0.026 ± 0.007	0.551 ± 0.056
Neural Network	0.670 ± 0.054	0.012 ± 0.003	0.384 ± 0.115	0.022 ± 0.006	0.530 ± 0.075
FDA approximation	0.707 ± 0.148	**0.016 ± 0.004**	0.435 ± 0.206	**0.030 ± 0.007**	0.585 ± 0.060

Note: Values are reported as mean ± standard deviation across ten folds. Bold indicates the highest value for each metric.

At the fixed probability threshold of 0.50, Decision Tree achieved the highest mean accuracy (0.892 ± 0.085), while FDA approximation obtained the highest mean precision (0.016 ± 0.004) and F1-score (0.030 ± 0.007). SVM achieved the highest mean recall (0.484 ± 0.136), whereas Logistic Regression yielded the highest mean ROC-AUC (0.603 ± 0.045). Precision remained low across all classifiers (0.012–0.016), reflecting the difficulty of identifying the rare positive class without producing many false-positive classifications.

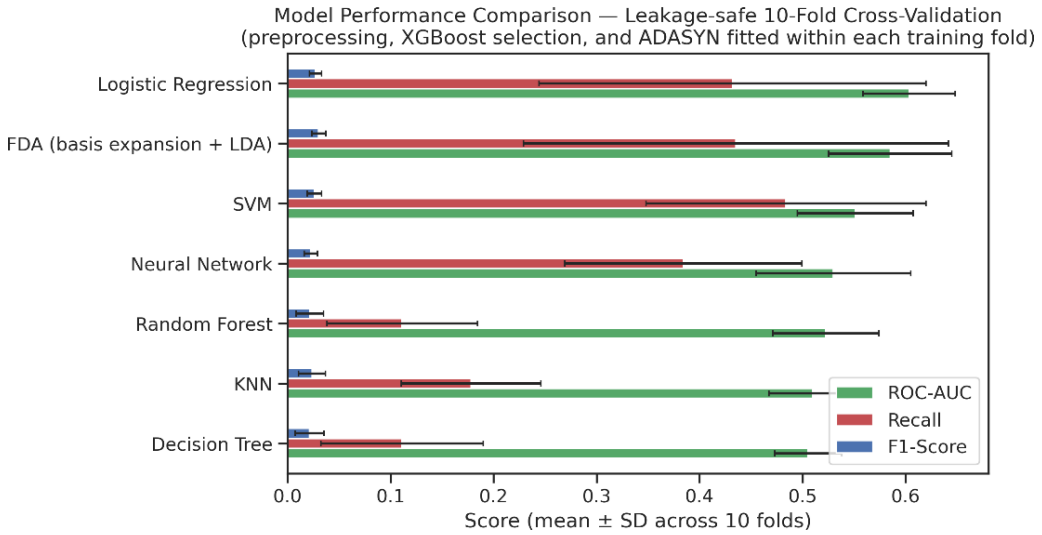


Figure 3. Comparison of recall, F1-score, and ROC-AUC across the seven models using leakage-safe 10-fold cross-validation.

On the untouched holdout test set, SVM achieved the highest ROC-AUC (0.6257), followed by Logistic Regression (0.6211) and FDA approximation (0.6190). FDA achieved the highest recall (0.6190), identifying 39 of 63 positive cases, but its precision was only 0.0161. Logistic Regression had a similar precision (0.0161) and recall of 0.5397. Random Forest achieved the highest accuracy (0.8709), but its recall was only 0.1270. These results indicate limited discrimination and show why accuracy should not be interpreted in isolation.

Table 3. Classification performance on the holdout test set

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Neural Network	0.7378	0.0120	0.3175	0.0232	0.5361
FDA approximation	0.6267	**0.0161**	**0.6190**	**0.0315**	0.6190
Logistic Regression	0.6733	**0.0161**	0.5397	0.0313	0.6211
KNN	0.7675	0.0136	0.3175	0.0260	0.5179
Decision Tree	0.8659	0.0086	0.1111	0.0160	0.4923
Random Forest	**0.8709**	0.0102	0.1270	0.0189	0.5298
SVM	0.6157	0.0153	0.6032	0.0298	**0.6257**

Note: Bold indicates the highest value for each metric.

The ROC curves show that SVM had the highest holdout ROC-AUC (0.6257), but the value still indicates only limited separation between positive and negative cases. The precision-recall curves provide a more direct view of positive-class retrieval under the 0.98% prevalence. Average precision ranged from approximately 0.010 to 0.014, which is close to the low baseline prevalence and indicates weak positive-class retrieval. The relatively low precision means that a large proportion of individuals classified as positive would be false alarms in a hypothetical screening setting.

Previous studies have demonstrated the potential of machine learning for cardiovascular disease prediction [7–9],[11–12]. However, their performance cannot be directly compared with this study. Differences in datasets, target definitions, predictors, and validation procedures may affect the results. This study focuses on physician-diagnosed heart attack history reported in IFLS5. The positive-class prevalence of only 0.98% may partly explain the low precision obtained.

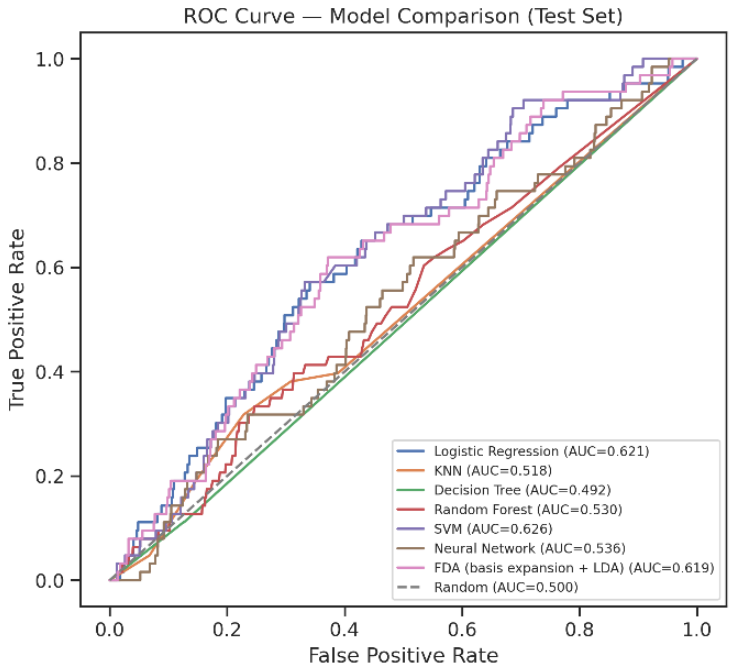


Figure 4. Receiver operating characteristic (ROC) curves of the seven classification models on the holdout test set.

The training-only OOF threshold analysis did not produce a consistent improvement on the holdout set. For example, the Logistic Regression threshold of 0.7722 yielded zero true positives on the holdout set, whereas the SVM threshold of 0.5065 yielded 36 true positives and 2,305 false positives. FDA's OOF-selected threshold of 0.6219 yielded six true positives and 369 false positives, compared with 39 true positives and 2,378 false positives at the fixed 0.50 threshold. These outcomes show that maximizing F1 on training OOF predictions did not guarantee better positive-case detection on the independent test set. The fixed 0.50 results are therefore retained as the primary comparison.

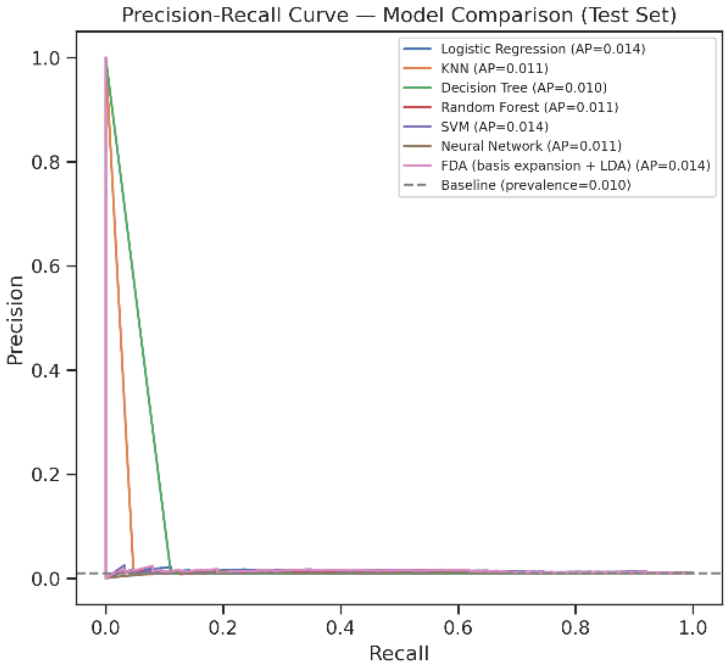


Figure 5. Precision-recall curves of the seven classification models on the holdout test set.

Table 4. Internal comparison of Random Forest ROC-AUC before and after correcting the validation procedure.

Evaluation Procedure	ROC-AUC	Interpretation
Feature selection and ADASYN before cross-validation	0.9601	Earlier leakage-prone estimate
Preprocessing, feature selection, and ADASYN within each training fold	0.523 ± 0.052	Leakage-safe 10-fold CV estimate

The confusion matrices provide further insight into the errors made by the models. FDA correctly identified 39 positive cases but produced 2,378 false positives. SVM identified 38 positive cases and produced 2,448 false positives. In comparison, Random Forest produced 776 false positives but missed 55 of the 63 positive cases. These results indicate that FDA achieved the highest recall, whereas Random Forest produced fewer false-positive classifications but failed to identify most positive cases. The confusion matrices for all seven models are presented together in Figure 6.

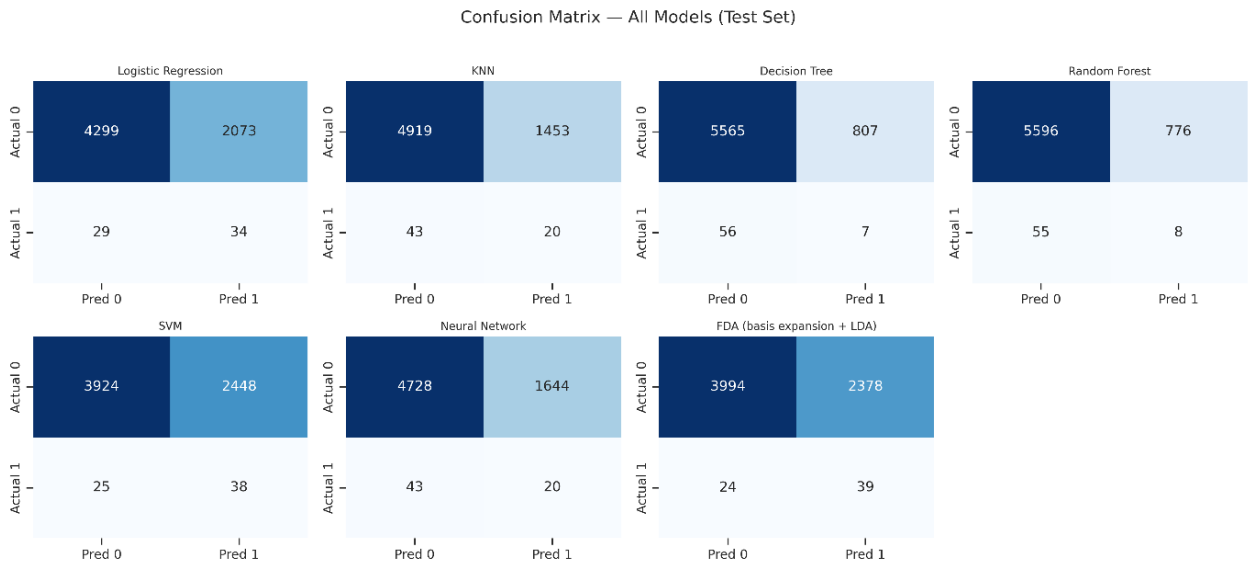


Figure 6. Confusion matrices of the seven classification models on the holdout test set at a probability threshold of 0.50.

The earlier Random Forest ROC-AUC of 0.9601 decreased to 0.523 ± 0.052 under the leakage-safe procedure, with a holdout ROC-AUC of 0.5298. This difference highlights the risk of data leakage during model evaluation, as discussed in previous studies [31–33]. Therefore, feature selection and ADASYN were performed within each training fold to maintain independent validation data, following fold-based validation principles [34].

Overall, the seven classifiers showed limited discrimination, with holdout ROC-AUC values ranging from 0.4923 to 0.6257 and low precision across models. These results may partly reflect the limited set of predictors available in the analytical dataset. Variables such as smoking history, family history, electrocardiographic findings, and detailed laboratory measurements were not included in the selected four-feature model; however, their absence cannot be established as the sole cause of the observed performance. Prior systematic reviews have similarly emphasized limitations in dataset coverage, external validation, and clinical applicability in ML-based heart disease research [7–8]. Future work should evaluate additional survey and clinical variables, alternative feature combinations, model calibration, external validation, and carefully justified imbalance-handling and threshold-selection strategies. The present findings remain exploratory and do not establish clinical diagnostic performance.

4. Conclusion

This study classified retrospectively reported doctor-diagnosed heart attack history using IFLS5 data containing 32,173 individuals, of whom 315 (0.98%) were positive. XGBoost selected four predictors from ten candidates: Diabetes, HighCholesterol, BMI, and Weight_kg. The selected BMI and Weight_kg variables were strongly correlated (r = 0.853), while their VIF values were 3.708 and 3.681, respectively. ADASYN increased the training positive class from 252 to 25,427 observations; distribution checks found no synthetic feature values outside the original positive-class ranges, although this does not establish clinical plausibility.

Under leakage-safe 10-fold cross-validation, Logistic Regression achieved the highest mean ROC-AUC (0.603 ± 0.045), while FDA achieved the highest mean recall (0.435 ± 0.206). On the untouched holdout set, SVM achieved the highest ROC-AUC (0.6257), and FDA achieved the highest recall (0.6190) with a precision of only 0.0161. The large number of false positives and limited discrimination

indicate that the models are not suitable for clinical diagnosis or screening in their current form. Further research should investigate broader predictors, alternative sampling and modeling strategies, calibration, and external validation.

The results also show that leakage-safe validation substantially reduced the earlier Random Forest ROC-AUC from 0.9601 to 0.523 ± 0.052. The difference demonstrates the importance of fitting data-dependent operations within training folds. The study contributes a methodological assessment of retrospective survey-based classification, but its findings should not be interpreted as prediction of future heart attack events or as evidence of clinical utility.

References

[1] Muharram, F. R., Multazam, C. E. C. Z., Mustofa, A., Socha, W., Andrianto, Martini, S., Aminde, L., and Yi-Li, C. (2024). The 30 Years of Shifting in The Indonesian Cardiovascular Burden: Analysis of The Global Burden of Disease Study. *Journal of Epidemiology and Global Health*, 14, 193–212.

[2] Chong, B., et al. (2025). Global burden of cardiovascular diseases: projections from 2025 to 2050. *European Journal of Preventive Cardiology*, 32(11), 1001–1015.

[3] Katsimardou, A., Imprialos, K., Stavropoulos, K., Sachinidis, A., Doumas, M., and Athyros, V. (2020). Hypertension in Metabolic Syndrome: Novel Insights. *Current Hypertension Reviews*, 16(1), 12–18.

[4] Lecube, A. (2024). Impact of obesity and diabetes on health and cardiovascular disease. *Atencion Primaria*, 56(12), 103045.

[5] Afifah, F., and Dewi Irawati Soeria Santoso (2024) Preventive Effect of Exercise on Kidney Fibrotic Pathway in Hypertension, *EKSAKTA: Berkala Ilmiah Bidang MIPA*, 25(3), 323–333.

[6] Ratnaningayu, N. D., Tedjo, A., and Panigoro, S. S. (2024). The Implementation of Machine Learning Algorithms for Breast Cancer Biomarker Validation in Metabolomics Studies, *EKSAKTA: Berkala Ilmiah Bidang MIPA*, 25(4), 468–483.

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

[8] Banerjee, T., and Paçal, İ. (2025). A systematic review of machine learning in heart disease prediction. Turkish Journal of Biology, 49(5), 600–634.

[9] Hartono, A., Dewi, L. A., Yuniarti, E., Putri, S. T. H., and Harahap, T. S. (2023). Machine Learning Classification for Detecting Heart Disease with K-NN Algorithm, Decision Tree and Random Forest, *EKSAKTA: Berkala Ilmiah Bidang MIPA*, 24(4), 513–522.

[10] Uddin, N., Hossain, M. S., Mahmud, M. Z., Ferdous, R., and Talukder, M. A. (2026). HPML-CVD: Hyperparameter Tuned Machine Learning Model to Predict Cardiovascular Disease. *Biomedical Materials & Devices*, 4(1), 861–875.

[11] Hassan, M. M., Zaman, S., Rahman, M. M., Bairagi, A. K., El-Shafai, W., Rathore, R. S., and Gupta, D. (2024). Efficient prediction of coronary artery disease using machine learning algorithms with feature selection techniques. *Computers and Electrical Engineering*, 115, 109130.

[12] Cao, K., Liu, C., Yang, S., Zhang, Y., Li, L., Jung, H., and Zhang, S. (2025). Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model. *Scientific Reports*, 15(1), 12406.

[13] Syukri, N., Oktarina, S. D., and Rahardiantoro, S. (2026). Comparative Performance of Gradient Boosting Algorithms for Household Food Resilience Classification during The COVID-19, *EKSAKTA: Berkala Ilmiah Bidang MIPA*, 27(4), 479–494.

[14] Ullah, T., Ullah, S. I., Ullah, K., Ishaq, M., Khan, A., Ghadi, Y. Y., and Algarni, A. (2024). Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection. *IEEE Access*, 12, 16431–16446.

[15] Hasan, N., and Bao, Y. (2021). Comparing different feature selection algorithms for cardiovascular disease prediction. *Health and Technology*, 11, 49–62.

[16] Pahlevani, A., Kashefi, A., Arslan, E., Kussia, B. K., and Joloudari, J. H. (2025). Predicting coronary artery disease with ensemble-based feature selection and grid search-tuned SVM under class imbalance conditions. *Discover Data*, 3(1), 48.

[17] Rahardi, M., Asaddulloh, B. P., Aminuddin, A., Abdulloh, F. F., Saifudin, I., and Kusumawijaya, F. P. (2025). Optimizing Machine Learning Models for Class Imbalance in Heart Disease Prediction. Engineering, *Technology & Applied Science Research*, 15(3), 23599–23604.

[18] Enihe, R. O., Prasad, R., Ogwueleka, F. N., and Abdullahi, F. B. (2025). The effect of imbalance data mitigation techniques on cardiovascular disease prediction. *Journal of the Nigerian Society of Physical Sciences*, 7(2), 2385.

[19] He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence*), 1322–1328.

[20] Mohammedqasim, H., Jasim, A. A., Mohammedqasem, R., and Ata, O. (2024). Enhancing predictive performance in COVID-19 healthcare datasets: A case study based on Hyper ADASYN over-sampling and genetic feature selection. *Journal of Engineering Science and Technology*, 19(2), 598–617.

[21] Khan, A. H., Hussain, M., and Malik, M. K. (2021). Cardiac Disorder Classification by Electrocardiogram Sensing Using Deep Neural Network. *Complexity*, 2021, 5512243.

[22] Sonawane, J. S., and Patil, D. R. (2014). Prediction of Heart Disease Using Multilayer Perceptron Neural Network. *In 2014 International Conference on Information Communication and Embedded Systems (ICICES*), 1–6.

[23] Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible Discriminant Analysis by Optimal Scoring. *Journal of the American Statistical Association*, 89(428), 1255–1270.

[24] Fadaei, F., et al. (2023). Assessment of the Performances of Adaptive Neuro-Fuzzy Inference System and Two Statistical Methods for Diagnosing Coronary Artery Disease. *Journal of Medical Signals and Sensors*, 37, 46.

[25] Safrizal, A. N., and Agustian, S. (2024). Classification of Covid-19 Vaccine Sentiment Using K-Nearest Neighbor and Fasttext on Twitter, *EKSAKTA: Berkala Ilmiah Bidang MIPA*, 25(3), 362–371.

[26] Patria, B. (2022). The longitudinal effects of education on depression: Finding from the Indonesian national survey. *Frontiers in Public Health*, 10, 1017995.

[27] Siramaneerat, I., Astutik, E., Agushybana, F., Bhumkittipich, P., and Lamprom, W. (2024). Examining determinants of stunting in Urban and Rural Indonesian: a multilevel analysis using the population-based Indonesian Family Life Survey (IFLS). *BMC Public Health*, 24, 1371.

[28] Griselda, M., Alfian, S. D., Wicaksono, I. A., Wawruch, M., et al. (2023). Findings from the Indonesian family life survey on patterns and factors associated with multimorbidity. Scientific Reports, 13, 18607.

[29] Sujarwoto, Maharani, A., Praveen, D., Palagyi, A., Kumar, P. S. G., Abimbola, S., Tampubolon, G., and Patel, A. (2025). Healthcare access and socio-demographic determinants

of estimated 10-year risk of cardiovascular diseases in Indonesia: A population-based study. *PLOS ONE*, 20(8), e0318112.

[30] Lydia, A., Setiati, S., Soejono, C. H., Istanti, R., Marsigit, J., et al. (2021). Prevalence of prehypertension and its risk factors in midlife and late life: Indonesian family life survey 2014–2015. *BMC Public Health*, 21, 493.

[31] Lee, H.-T., Cheon, H.-R., Lee, S.-H., Shim, M., and Hwang, H.-J. (2023). Risk of data leakage in estimating the diagnostic performance of a deep-learning-based computer-aided system for psychiatric disorders. *Scientific Reports*, 13, 16633.

[32] Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 1–21.

[33] Tampu, I. E., Eklund, A., and Haj-Hosseini, N. (2022). Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data*, 9(1), 580.

[34] Bey, R., Goussault, R., Grolleau, F., Benchoufi, M., and Porcher, R. (2020). Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. Journal of the American Medical Informatics Association, 27(8), 1244–1251.

[35] Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F., and Dwivedi, G. (2022). A reinforcement learning-based approach for imputing missing data. *Neural Computing and Applications*, 34, 9701–9716.

[36] Lin, Z., Fan, Y., Tan, J., Li, Z., Yang, P., Wang, H., and Duan, W. (2025). Tool wear prediction based on XGBoost feature selection combined with PSO-BP network. *Scientific Reports*, 15, 3096.

[37] Kingma, D. P., and Ba, J. (2015). Adam: A Method for Stochastic Optimization. *In 3rd International Conference for Learning Representations (ICLR)*.

[38] Dahl, G. E., Sainath, T. N., and Hinton, G. E. (2013). Improving Deep Neural Networks for LVCSR Using Rectified Linear Units and Dropout. In 2013 IEEE International Conference on Acoustics, *Speech and Signal Processing*, 8609–8613.