THESIS_REPORT__Copy_.pdf

International University of Business Agriculture and Technology

Document Details

Submission ID

trn:oid:::3618:125359567

Submission Date

Dec 30, 2025, 1:53 PM GMT+6

Download Date

Dec 30, 2025, 2:05 PM GMT+6

File Name

THESIS_REPORT__Copy_.pdf

File Size

476.9 KB

6 Pages

2,841 Words

18,033 Characters

85% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

29 AI-generated only 85%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)
16 – 18 April 2026, Chittagong, Bangladesh

Explainability-Aware Synthetic Data–Driven AutoML Framework for Heart Disease Detection

***Abstract*—This is a placeholder abstract. Briefly summarize the scope and contributions of your paper here. For example, ”This template provides a reusable structure for IEEE conference papers, including sections for abstract, introduction, methodology, results, tables, figures, and citations.”**

***Index Terms*—Placeholder, IEEE Template, Example Paper, Structure, Keywords**

I. Introduction

Heart disease remains a leading cause of mortality worldwide, frequently occurring without early warning and leading to severe clinical outcomes. Lifestyle-related factors such as physical inactivity, unhealthy diets, chronic stress, and metabolic disorders have substantially increased cardiovascular risk in recent years. Although modern diagnostic tools and treatment options are widely available, many heart disease cases are still detected at advanced stages, where interventions become more costly, invasive, and less effective. Conventional diagnostic practices rely heavily on hospital-based tests and clinical expertise, which may overlook subtle, patient-specific early symptoms and delay timely intervention.

The growing availability of electronic health records has enabled data-driven approaches for early heart disease prediction using machine learning techniques. By learning from historical clinical data, such models can support clinicians in identifying high-risk individuals at earlier stages and assist in preventive decision-making. However, the practical effectiveness of these approaches is often limited by common challenges in medical datasets, including small sample sizes, class imbalance between healthy and affected patients, incomplete records, and strict data privacy requirements. These constraints can reduce model reliability, increase bias, and limit their applicability in real-world clinical environments.

To address these challenges, this study presents an explainability-aware synthetic data–driven AutoML framework for early heart disease detection. The proposed approach integrates multi-source synthetic data generation with automated model selection and optimization to enhance data diversity while preserving patient confidentiality. By emphasizing synthetic data quality and interpretability, the framework improves predictive performance and offers a scalable and trustworthy solution for clinical decision-support systems.

II. Literate Review

Machine learning (ML) techniques have been widely and commonly applied to heart disease and early heart attack prediction by using structured clinical data. Early research demonstrated that supervised learning models can effectively support clinical decision-making .It has combined with appropriate preprocessing and feature selection. Many Studies based on widely used datasets such as the Cleveland heart disease dataset .It showed that traditional classifiers including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) provide reliable baseline performance for cardiovascular risk prediction. Selvakumar et al. (2023), for instance, evaluated multiple ML models on a dataset of 303 patient records and reported that Logistic Regression achieved the highest accuracy of 91.22% with a recall exceeding 93%, emphasizing its effectiveness in identifying high-risk patients. While these results confirm the potential of conventional ML models, the limited dataset size and absence of real-time validation restrict their venerability.

To improve predictive performance, several researchers have explored ensemble and hybrid machine learning techniques. Choudhary and Singh (2020) demonstrated that Decision Tree models often suffer from overfitting when applied independently, but performance can be improved through ensemble learning. By integrating AdaBoost with Decision Tree classifiers, their model achieved an accuracy of approximately 89%, indicating improved generalization. Building on this idea, Bhaduaria et al. (2025) proposed a stacked hybrid learning framework that combined Random Forest and XGBoost classifiers using clinical and lifestyle-related features. Their hybrid model achieved an accuracy of 89.13%, outperforming individual classifiers and highlighting the advantage of combining multiple learners to capture complex feature interactions. However, this framework was validated on a single secondary dataset, limiting its applicability to real-world clinical environments.

Further evidence supporting ensemble learning was provided by Alom et al. (2025), who conducted an extensive evaluation of advanced classifiers and ensemble techniques on a large Kaggle-based dataset. Their experiments showed that XGBoost achieved the highest accuracy of 97.88%, along with strong recall and ROC–AUC scores, demonstrating its robustness in handling complex and imbalanced data. Similarly, Teja and Rayalu (2025) performed a comparative analysis using combined datasets from Cleveland, Hungary, Switzerland, Long Beach, and Statlog, totaling over one thousand patient records. Their results indicated that Random Forest achieved stable performance with accuracy reaching up to 94% under 10-fold cross-validation, while ensemble methods such as Bagged Trees and XGBoost achieved accuracies above 93%. These studies collectively confirm that ensemble learning significantly enhances predictive accuracy and sta-

979-8-3315-4990-9/26/$31.00 ©2026 IEEE

bility across diverse datasets. Nevertheless, most ensemble-based approaches rely on fixed experimental settings and lack automated model selection strategies.

In addition to ensemble learning, several studies have focused on algorithm optimization and feature tuning to enhance classification performance. Kumar et al. (2024) evaluated multiple ML algorithms for heart attack prediction and found that SVM achieved the best overall performance, with an F1-score of 92.5%, precision of 93.9%, recall of 91.1%, and an AUC of 91.8%. These findings suggest that SVM is particularly well suited for binary heart disease classification tasks. Bouqentar et al. (2024) further emphasized the importance of feature engineering and hyperparameter tuning by evaluating six ML algorithms on the Cleveland and Statlog datasets. Their optimized models achieved accuracies of approximately 92% and 91.18%, respectively, using 10-fold cross-validation. Despite these improvements, such studies typically depend on structured and static datasets and do not adequately address class imbalance or scalability issues.

More recent research has introduced explainable and deep learning–based approaches to improve both prediction accuracy and clinical interpretability. Efat et al. (2025) proposed a Feature-Tuned Explainable SVM (FTX-SVM) framework that incorporated hyperparameter tuning, SMOTE-based class balancing, and cross-validation. Their model achieved an accuracy of 97.46%, outperforming several traditional and ensemble classifiers while also providing insights into key contributing clinical features. In contrast, Arooj et al. (2022) explored a deep learning approach using a deep convolutional neural network (DCNN) for heart disease detection. Evaluated on a dataset containing 1050 patient records, the DCNN achieved a validation accuracy of 91.7%, demonstrating the capability of deep learning models to capture complex nonlinear patterns in medical data. However, deep learning approaches often require high computational resources and lack transparency, which may limit their adoption in clinical practice.

Overall, the existing literature demonstrates that machine learning models—particularly ensemble-based and optimized classifiers—can achieve high accuracy in heart disease and heart attack prediction, with reported accuracies ranging from approximately 89% to 98%. Despite these advancements, most studies focus on individual models or fixed experimental pipelines, rely heavily on secondary datasets, and offer limited support for automation, scalability, and real-time deployment. Furthermore, challenges such as data imbalance, reproducibility, and systematic model selection remain insufficiently addressed. These limitations motivate the need for a synthetic data–driven AutoML framework that can automatically optimize model selection, improve generalization, and enhance robustness for early heart disease detection in practical clinical settings.

III. Methodology

We propose a unified methodology for heart disease classification that integrates data preprocessing, multi-source synthetic data augmentation, and ensemble-based automated machine learning to achieve robust and reproducible performance.

A. Dataset Description and Problem Formulation

This study uses the UCI Cleveland Heart Disease dataset, consisting of 303 patient records described by 13 clinically relevant attributes covering demographic, physiological, and diagnostic information. Each sample is represented as a feature–label pair (x_i, y_i), where $x_i \in \mathbb{R}^d$ denotes the clinical feature vector and $y_i \in 0, 1$ indicates the absence or presence of heart disease.

The learning objective is to estimate the probabilistic prediction function

$$f(x) = P(y = 1 \mid x), \tag{1}$$

which provides both classification and confidence estimation, a critical requirement for medical decision-support systems.

To prevent information leakage, the dataset is partitioned prior to any preprocessing or augmentation as

$$\mathcal{D} = \mathcal{D}train \cup \mathcal{D}test, \quad \mathcal{D}train \cap \mathcal{D}test = \emptyset, \tag{2}$$

using a stratified 70:30 split. All synthetic generation and model training are performed exclusively on $\mathcal{D}train$, while $\mathcal{D}test$ is reserved for final evaluation.

B. Data Preprocessing

Numerical features are standardized using Z-score normalization:

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{3}$$

where μ_j and σ_j are computed solely from the training data. This ensures numerical stability and prevents features with large ranges from dominating model learning.

Categorical attributes, including sex, chest pain type, electrocardiographic results, and the target variable, are treated as discrete variables. They are preserved as categorical during synthetic generation and encoded as integers for predictive modeling and SHAP-based explainability.

C. Multi-Source Conditional Synthetic Data Generation

To mitigate data scarcity and capture diverse feature dependencies, three synthetic data generators—CTGAN, TVAE, and Gaussian Copula—are trained independently on the real training data. Each learns a conditional distribution

$$p_k(x \mid y), \tag{4}$$

allowing class-aware modeling of feature relationships.

Conditional sampling enforces class balance by generating equal numbers of synthetic samples per class:

$$|\mathcal{D}syn^{(k)}(y = 0)| = |\mathcal{D}syn^{(k)}(y = 1)|. \tag{5}$$

This approach prevents imbalance amplification while leveraging complementary generative mechanisms.

D. Distributional and Statistical Validation

Synthetic datasets are compared with real training data using global statistical metrics. The Kolmogorov–Smirnov statistic is defined as

$$KS = \sup_x |F_{real}(x) - F_{syn}(x)|, \tag{6}$$

to assess marginal distribution similarity.

The Wasserstein distance,

$$W(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}(x,y) \sim \gamma[|x-y|], \tag{7}$$

captures distributional shape differences, while the Kullback–Leibler divergence,

$$DKL(P|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}, \tag{8}$$

quantifies information loss. These metrics are used for diagnostic analysis rather than direct sample filtering.

E. Synthetic Confidence Score and Instance-Level Quality Assessment

To evaluate individual synthetic samples, an instance-level Synthetic Confidence Score (SCS) is introduced. Structural realism is measured by proximity to real data clusters:

$$C(s_i) = \frac{1}{1 + d(s_i, \mathcal{C})}. \tag{9}$$

Statistical plausibility is assessed via feature deviation:

$$D(s_i) = \frac{1}{d} \sum_{j=1}^{d} \left| \frac{s_{ij} - \mu_j}{\sigma_j} \right|. \tag{10}$$

Explainability alignment is enforced using SHAP values computed from an XGBoost model trained on real data:

$$S(s_i) = 1 - \cos(\phi(s_i), \phi_{real}). \tag{11}$$

The final confidence score is defined as

$$\mathrm{SCS}(s_i) = \alpha C(s_i) + \beta D(s_i) + \gamma S(s_i), \tag{12}$$

with $\alpha + \beta + \gamma = 1$, emphasizing explainability consistency.

F. Dynamic Threshold-Based Filtering and Data Fusion

High-quality synthetic samples are retained using a generator-specific adaptive threshold:

$$\tau = \max(\mu_{SCS} - \sigma_{SCS}, \tau_{min}). \tag{13}$$

The final augmented dataset is constructed as

$$\mathcal{D}fused = \mathcal{D}train \cup \bigcup_k \mathcal{D}syn, filtered^{(k)} \cup \mathcal{D}SMOTE, \tag{14}$$

combining real data with filtered synthetic samples and SMOTE-based oversampling.

G. Model Training, Meta-Stacking, and Evaluation

Models are trained using AutoGluon with a tiered strategy. Individual datasets use efficient configurations, while the fused dataset employs stronger presets with stacking enabled. Predictions are combined via a logistic regression meta-learner:

$$p_{meta}(x) = \sigma\left(\sum_k w_k p_k(x)\right). \tag{15}$$

Interpretability consistency is assessed using the Explainability Stability Index:

$$XSI_{i,j} = \rho_{Spearman}(\mathrm{rank}(\phi_i), \mathrm{rank}(\phi_j)). \tag{16}$$

All performance metrics are computed on the held-out test set, ensuring privacy preservation:

$$\mathcal{D}test \cap \mathcal{D}syn = \emptyset. \tag{17}$$

IV. Results and Discussion

A. Overall Predictive Performance Analysis

The predictive performance of the proposed framework was evaluated on a held-out test set that remained completely unseen during training, synthetic data generation, and model selection. The proposed model achieved a ROC–AUC score of **0.906**, demonstrating strong discriminative capability for early heart disease prediction despite the limited size of the original dataset.

As shown in Figure 02, both the real-data-trained model and the synthetic-data-augmented model exhibit strong classification performance. However, the model trained using the proposed synthetic augmentation framework consistently outperforms the real-only model across a wide range of false positive rates. This indicates that the observed improvement is not limited to a specific operating threshold, but reflects superior ranking quality of predicted risk scores. Such behavior is particularly important in clinical screening applications, where reliable prioritization of high-risk patients is critical.

B. Comparison with Classical Machine Learning Models

To establish a strong and fair baseline, multiple classical machine learning classifiers were evaluated using identical preprocessing and evaluation protocols. These include linear models, distance-based methods, ensemble learners, and neural networks.

As shown in Table I, the Support Vector Machine (SVM) achieves the highest ROC–AUC among all classical classifiers, indicating that margin-based decision boundaries are effective for this dataset. However, the proposed framework surpasses even the best-performing classical model, demonstrating that the observed performance gains arise primarily from the proposed data-centric strategy rather than algorithm selection alone.

TABLE I: ROC–AUC comparison of classical machine learning models and the proposed framework on the test set

Model	ROC–AUC
Logistic Regression	0.8885
Support Vector Machine (SVM)	0.8896
K-Nearest Neighbors (KNN)	0.8701
Random Forest	0.8593
Gradient Boosting	0.8452
XGBoost	0.8355
LightGBM	0.8355
Extra Trees	0.8690
CatBoost	0.8561
Neural Network	0.8680
Proposed Framework	**0.906**

C. Explainability Stability Analysis

Beyond predictive accuracy, interpretability consistency is essential for clinical trust. To evaluate this aspect, the Explainability Stability Index (XSI) was computed using Spearman rank correlation between SHAP feature importance rankings across models trained on real and synthetic datasets.

Figure 04 reveals strong positive correlations between the real-data model and models trained on selected synthetic datasets, indicating preservation of core clinical reasoning patterns. In contrast, weaker correlations observed for certain synthetic configurations highlight potential explainability drift when synthetic data are not carefully controlled. Importantly, the proposed framework maintains explainability alignment while improving predictive performance.

D. Distributional Fidelity and Correlation Difference Analysis

To further examine the realism of synthetic data, correlation difference analysis was performed between real training data and synthetic datasets.

As illustrated in Figure 05, most feature-pair correlations exhibit minimal deviation, indicating that the synthetic data largely preserve inter-feature relationships observed in real patients. However, moderate deviations are observed for certain clinically relevant variables. These findings highlight the limitation of relying solely on global distributional similarity metrics and justify the need for quality-aware synthetic data generation.

E. Reinforced Explainability Stability Across Models

To further validate interpretability consistency, an additional explainability stability visualization is presented.

Figure 04 confirms that the explainability behavior of the proposed model remains stable across training on both real and synthetic data. The consistency observed across SHAP feature rankings suggests that the performance gains achieved through synthetic augmentation do not come at the cost of distorted or clinically implausible decision logic.

F. Summary of Key Findings

Overall, the experimental results demonstrate that while classical machine learning models provide strong baselines, their performance is constrained by limited data availability. The proposed synthetic augmentation framework successfully improves predictive performance, achieving a ROC–AUC of **0.906**, while preserving interpretability and maintaining realistic feature relationships. The combined evidence from ROC analysis, explainability stability evaluation, and correlation difference analysis confirms that the proposed approach constitutes a robust and trustworthy solution for early heart disease prediction under data-scarce conditions.

V. Discussion

The experimental results show that classical classifiers, including the best-performing SVM, are constrained by limited training data and fail to surpass a ROC–AUC of 0.89. The proposed framework overcomes this limitation through a data-centric design that integrates multi-source synthetic generation with explainability-aware instance filtering. This approach achieves a superior ROC–AUC of 0.906, demonstrating improved generalization without increasing model complexity. Importantly, explainability stability analysis confirms that predictive gains are obtained without distorting clinical reasoning. These findings highlight the contribution of the proposed framework as a trustworthy and effective solution for early heart disease prediction under data-scarce conditions.

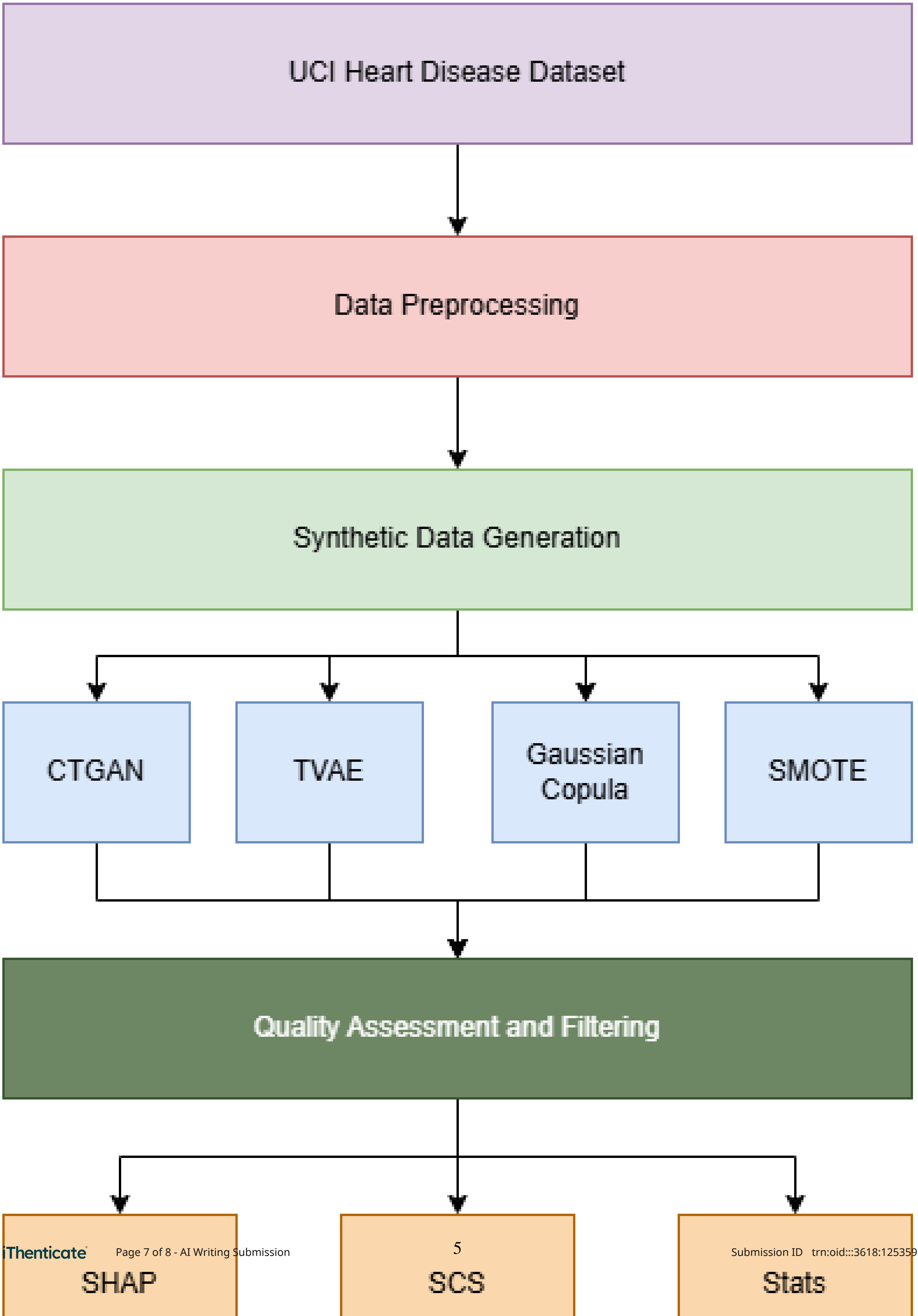
VI. Conclusion

This study presented an explainability-aware synthetic data–driven AutoML framework for early heart disease detection, aiming to overcome critical challenges associated with limited sample size, class imbalance, and strict privacy constraints in medical datasets. By integrating multi-source synthetic data generation methods, including CTGAN, TVAE, Gaussian Copula, and SMOTE, the framework enriches the training data while avoiding direct reliance on additional real patient records.

A key contribution of this work is the proposed Synthetic Consistency Score (SCS), which performs instance-level quality assessment of synthetic samples by jointly considering statistical similarity, structural consistency, and SHAP-based semantic alignment. This quality-aware filtering mechanism ensures that only clinically meaningful synthetic data are retained, mitigating explainability drift and reducing the risk of learning spurious patterns. As a result, models trained on the fused dataset demonstrate improved generalization, achieving a ROC–AUC of 0.906, which surpasses all classical machine learning baselines.

Beyond predictive accuracy, the proposed framework emphasizes interpretability, fairness, and robustness—factors that are essential for real-world clinical adoption. The explainability analysis confirms that feature importance patterns remain stable after synthetic augmentation, while fairness and robustness evaluations indicate consistent model behavior across sensitive attributes and under input perturbations. Overall, the findings demonstrate that combining explainability-guided synthetic data generation with AutoML provides a robust, privacy-preserving, and clinically reliable solution for early heart disease prediction.

References


UCI Heart Disease Dataset
Data Preprocessing
Synthetic Data Generation
CTGAN
TVAE
Gaussian Copula
SMOTE
Quality Assessment and Filtering
SHAP
SCS
Stats

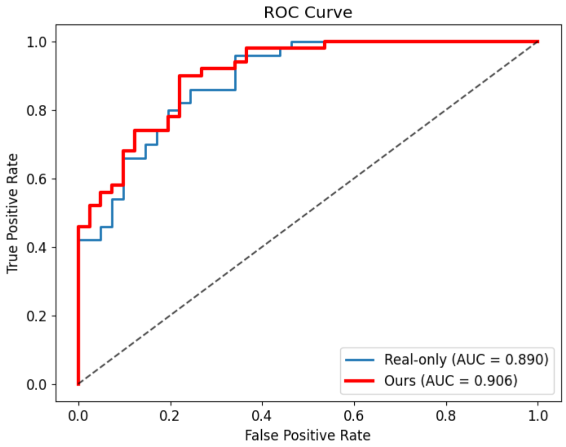


Fig. 2: ROC curve comparing real-data and synthetic-data based models

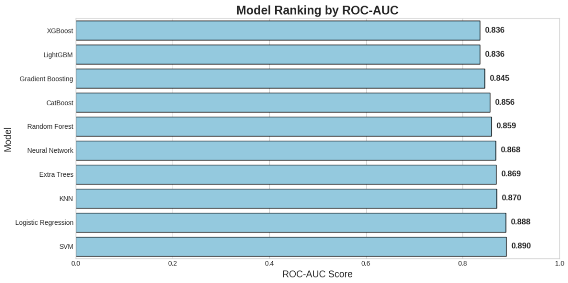


Fig. 3: Model Ranking by ROC

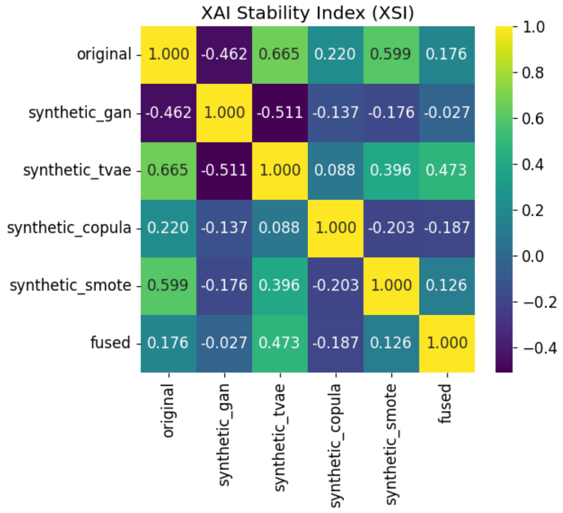


Fig. 4: Explainability Stability Index (XSI) heatmap

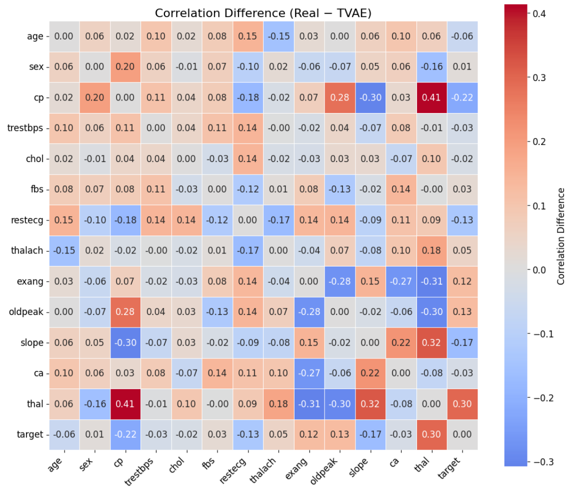


Fig. 5: Correlation difference heatmap between real and synthetic data