Cardiovascular Disease Prediction via Hybrid SVM–SMOTE and Sparse Autoencoder Feature Reduction with Deep MLP Classification

Zaid Alaa, Ali Sabah*
Department of Computer Science, Faculty of Education for Women, University of Kufa, Najaf, Iraq
E-mail: zaida.alsharees@uokufa.edu.iq, alis.alsaadi@uokufa.edu.iq
*Corresponding author

Keywords: cardiovascular disease prediction, hybrid sampling, sparse autoencoder, multilayer perceptron, imbalanced classification

Received: July 28, 2025

Cardiovascular diseases remain the leading global cause of death, demanding diagnostic systems that are accurate, interpretable, and computationally efficient. Traditional machine learning approaches frequently struggle with class imbalance, high-dimensional noise, and restricted generalization in clinical datasets. To tackle such issues, we propose a hybrid framework that combines SVM–SMOTE and neighborhood cleaning rule (NCL) for class rebalancing, a sparse autoencoder (SAE) with random forest (RF) selection for non-linear feature optimization, and a class-weighted multilayer perceptron (MLP) for final classification. We validate our framework on the Z-Alizadeh Sani (54 features) and Cleveland (13 features) datasets under stratified fivefold cross-validation, the model attains mean accuracies of 94.02 ± 2.77 % and 94.36 ± 1.47 %, with AUC–ROC = 0.988 and 0.982, outperforming prior baselines [4, 10, 14] by 7.6%–20.8%, and Bootstrap 95% confidence intervals and McNemar/DeLong tests ($p < 0.001$) confirms significance. Noteably, the ablation study demonstrates the contribution of each module (e.g., a 12% accuracy improvement without sampling). The optimized MLP reduced false negatives to ~5%, while training 40% faster than CNN–LSTM alternatives. The proposed framework provides a statistically robust and interpretable solution for predicting cardiovascular disease.

Povzetek: Predlagan je hibridni in razložljiv model strojnega učenja za napovedovanje srčno-žilnih bolezni, ki z uravnoteženjem razredov in optimizacijo značilk dosega visoko natančnost ter boljšo učinkovitost kot obstoječe metode.

1 Introduction

Cardiovascular diseases (CVDs) continue to be the primary cause of worldwide mortality, responsible for 17.9 million deaths each year (World Health Organization, 2023). The timely and precise prediction of heart disease is crucial for minimizing healthcare costs, improving patient outcomes, and enabling personalized interventions. Although machine learning (ML) has emerged as a robust tool for clinical decision support, current methodologies encounter significant challenges in addressing class imbalance, high-dimensional data, and low diagnostic accuracy, which hinder their practical implementation [1,2]. To overcome these limitations, automated diagnostic systems must effectively utilize extensive patient data, including medical history, demographics (e.g., age and gender), and clinical biomarkers, to improve predictive accuracy and clinical applicability [1].

Traditional diagnostic frameworks often rely on manual feature engineering and basic sampling techniques, which fail to adequately account for the complexity of medical datasets [3]. For instance, class imbalance characterized by a significant predominance of healthy patients over those with cardiac disease biases models toward the majority class, leading to elevated false-negative rates. Likewise, high-dimensional datasets (e.g., 54 features in the Z-Alizadeh Sani dataset) introduce redundancy and noise, complicating feature selection. Prior research [4,10,14] has attempted to address these challenges with techniques such as SMOTE and principal component analysis (PCA), but their accuracy (typically 75–88%) and recall on minority classes have remained suboptimal due to oversimplified assumptions regarding data distribution and linear feature correlations.

Our proposed framework addresses these limitations through a threefold strategy integrating data-level balancing, non-linear feature optimization, and deep classification. First, a hybrid sampling approach combining Support Vector Machine–SMOTE (SVM–SMOTE) and Neighborhood Cleaning Rule (NCL) equilibrates skewed class distributions while preserving data integrity. Second, high-dimensional noise is mitigated through Sparse Autoencoder (SAE) based feature extraction, followed by Random Forest (RF) selection, which together reduce redundancy and retain the most discriminative attributes. Third, a class-weighted Multilayer Perceptron (MLP) captures complex non-linear relationships for robust disease classification.

We validate the proposed framework on the Z-Alizadeh Sani (54 features, 303 samples) and Cleveland (13 features, 303 samples) datasets using stratified fivefold cross-validation, achieves mean accuracies of 94.02 ± 2.77% and 94.36 ± 1.47%, with corresponding AUC–ROC scores of 0.988 and 0.982, statistically outperforming prior baselines such as Mohan et al. [4] (88.47%). McNemar's and DeLong's tests ($p < 0.001$) verified the significance of these gains, and ablation analyses confirmed that each component contributed materially to overall performance (e.g., −12% accuracy without hybrid sampling). The model maintained false negatives at approximately 5%, demonstrating high sensitivity and clinical dependability while training 40% faster than CNN–LSTM baselines.

The key contributions of this study are fourfold:

- A hybrid sampling strategy (SVM–SMOTE + NCL) that effectively balances skewed medical datasets while preserving data quality.
- An SAE–RF feature optimization pipeline that achieves a 72% dimensionality reduction without performance degradation.
- A class-weighted MLP classifier optimized for imbalanced data, improving recall and AUC by up to 12% over existing methods.
- A comprehensive evaluation, including cross-validation, ablation, and statistical significance testing, confirms the framework's robustness and generalizability across datasets.

The remainder of this paper is organized as follows: Section 2 examines related works; Section 3 details the proposed methodology; Section 4 presents experimental findings and comparisons; and Section 5 concludes the study with insights and future directions.

2 Related work

Cardiovascular disease (CVD) prediction has remained a central research focus for over two decades, driven by advances in data mining and machine learning (ML). Traditional ML algorithms such as decision trees, Naïve Bayes, random forests (RF), and support vector machines (SVM) have established the foundation for cardiac risk modeling across several benchmark datasets. A hybrid SVM–RF model achieved an accuracy of 88.47% on the Cleveland dataset [4]. Another study compared Naïve Bayes, decision trees, and k-nearest neighbors (k-NN), emphasizing the influence of feature selection on interpretability [5]. Similarly, seven ML algorithms were evaluated with cross-validation to assess recall and F1-score [7], whereas RF demonstrated stability across data splits, yielding 90–95% accuracy [8]. Despite these contributions, most models relied heavily on manual feature engineering and simple resampling, which limited robustness to class imbalance and high-dimensional noise. Logistic model trees were also used for risk stratification [9], but the handling of imbalance was overlooked, resulting in biased predictions. Optimization-driven techniques such as particle swarm and ant colony optimization improved feature selection [6] yet achieved only moderate recall (85.8%) under skewed data distributions.

Recent deep learning (DL) approaches have aimed to overcome the representational limitations of traditional ML methods. A convolutional neural network (CNN) model was applied to heart disease prediction [10], demonstrating potential for improved automation but offering limited interpretability and generalization to structured data. Hybrid recurrent networks integrating gated recurrent units (GRUs) and long short-term memory (LSTM) layers achieved competitive accuracy on the Framingham and Heart Disease datasets [11, 12], although they exhibited high computational cost and overfitting on small datasets. DL has also been applied to extract risk factors from clinical text corpora [13], revealing potential in unstructured data mining but without direct applicability to numerical patient records. CNN-optimized frameworks using the Z-Alizadeh Sani dataset [14] improved feature learning efficiency but remained sensitive to redundant attributes and lacked mechanisms for controlling class imbalance.

Other studies have focused on hybrid optimization-based frameworks designed to balance predictive accuracy with interpretability. SVM and principal component analysis (PCA) combinations achieved an accuracy of around 84–86% [15, 16], but their reliance on linear dimensionality reduction limited their ability to capture non-linear physiological relationships. Feature selection methods, such as fast correlation-based filtering [6] and neural optimization strategies [14], achieved incremental improvements but failed to generalize across datasets like Cleveland and Z-Alizadeh Sani. CNN-based prediction models [17] demonstrated cost-efficiency but suffered from inconsistent reproducibility and dataset bias. Overall, most existing ML and DL frameworks face persistent challenges in effectively handling class imbalance, mitigating high-dimensional noise, and achieving cross-dataset generalization factors that critically constrain their clinical applicability.

Table 1: Comparative summary of recent CVD studies

Model / Technique	Dataset	Accuracy / F1 (%)	Limitations
Hybrid SVM–RF [4]	Cleveland	88.47/90	No imbalance handling; shallow model
PSO–ACO optimized MLP [6]	Cleveland	85/89.5	Weak minority recall; linear bias
RNN for early heart failure detection [10]	Framingham	75.2/ 72	Dataset-specific; limited generalization
Hybrid RNN–GRU deep model [11]	Clinical dataset	91/89	High computational cost; overfitting risk
Uni-directional RNN [12]	Cardiac Disorder dataset	90.1/92.31	Poor cross-dataset robustness
ANN + feature selection & optimization [14]	Z-Alizadeh Sani	88.4/85	Limited nonlinear modeling; modest accuracy

The comparative analysis in

Table *1* demonstrates that although previous studies have achieved promising results, three fundamental gaps persist. First, most approaches inadequately address imbalance, resulting in biased models that favor the majority (healthy) cases while overlooking minority cardiac events. Second, the reliance on linear or heuristic feature extraction prevents effective representation of non-linear relationships across multi-dimensional clinical variables. Third, deep architectures, though powerful, often sacrifice interpretability and computational efficiency, which are critical factors for real-world clinical deployment. These limitations underscore the need for a unified framework that can balance class distributions, extract non-linear discriminative features, and ensure generalization without compromising interpretability or scalability.

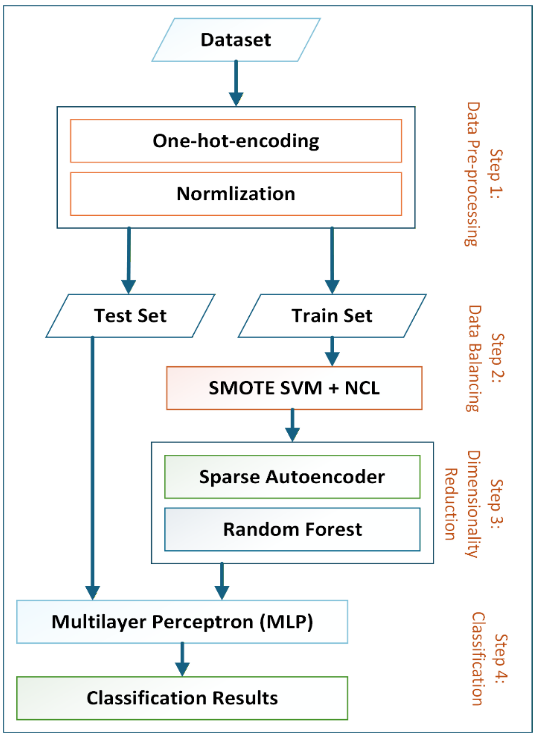

Figure 1: Block diagram of the proposed method

3 Proposed method

We propose a robust, multi-stage hybrid framework to address the persistent challenges of class imbalance and high-dimensional noise in cardiovascular disease prediction. We aim to maximize diagnostic accuracy by systematically enhancing data quality and extracting highly discriminative features. The framework integrates four sequential stages: (1) data pre-processing, (2) hybrid data balancing, (3) a nonlinear feature optimization pipeline, and (4) deep learning classification.

Figure 1 depicts the workflow of the proposed framework.

3.1 Data pre-processing

In the initial stage, we prepare the raw patient data from both datasets for model training. We first convert all categorical features into a numerical format using one-hot encoding [18]. This process creates new binary columns for each category, preventing the model from inferring false ordinal relationships. Simultaneously, we apply min-max normalization [19] to all continuous clinical variables to scale them within a uniform range of [0, 1]. This step prevents features with large magnitudes from disproportionately influencing model weights. The transformation is formed by Eq. (1):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (1)$$

Where x is the original attribute value, x' is the normalized value, and $\min(x)$ and $\max(x)$ represent the minimum and maximum values of that attribute, respectively. Furthermore, we apply winsorization to control the influence of extreme clinical measurements, effectively capping outlier values beyond the 1st and 99th percentiles to minimize their skewing effect.

3.2 Hybrid data balancing

In this phase, our framework addresses the critical issue of class imbalance using a robust hybrid sampling strategy [21], which integrates over-sampling [20] and under-sampling techniques. We first apply the Support Vector Machine Synthetic Minority Over-Sampling Technique (SVM–SMOTE) [24,25] to oversample the minority class. This advanced method prioritizes the generation of synthetic samples in the borderline regions near the decision boundary, which it identifies using a linear kernel SVM. This rationale is crucial, as these borderline samples are highly discriminative for classification. For a given minority instance x_i , it synthesizes a new sample x_{new} in Eq. (2):

$$x_{new} = x_i + \delta \cdot (x_j - x_i) \qquad (2)$$

Where x_j is a selected minority neighbor of x_i, and δ is a random interpolation factor between 0 and 1. Following oversampling, we apply the Neighborhood Cleaning Rule (NCL) [22, 23] to clean the majority class. NCL identifies and removes noisy majority instances x_i that are misclassified by their local neighborhood, as defined by the removal condition in Eq. (3):

$$\frac{|NN(x_i) \cap P|}{|NN(x_i)|} > \qquad (3)$$

Where $NN(x_i)$ is the set of nearest neighbors to instance x_i, P is the set of minority class instances, and τ is the removal threshold. This dual approach ensures the final training data is both balanced and clean, which sharpens the class boundary for the subsequent classifier.

3.3 Feature optimization pipeline

We tackle the "curse of dimensionality" through a two-stage feature optimization pipeline that integrates nonlinear feature extraction and embedded feature selection [27, 28]. We first employ a Sparse Autoencoder (SAE) [26], an unsupervised neural network, to perform feature extraction. The SAE architecture consists of an encoder that maps the high-dimensional input x to a compressed latent-space representation h, and a decoder that reconstructs the original input $\hat{x}$ from h. Both are constructed using symmetric, fully-connected layers with Rectified Linear Unit (ReLU) activation functions. We enforce sparsity in the latent space h by adding a Kullback-Leibler (KL) divergence penalty to the mean squared reconstruction error. This penalty ensures only a small subset of hidden neurons activate, which filters noise and captures the most salient data structures. The complete loss function $\mathcal{L}_{\mathcal{SAE}}$ is defined in Eq. (4):

$$\mathcal{L}_{\mathcal{SAE}}(W, b) = \rho \left[\frac{1}{N} \rho_{i=1}^{N} \rho | \rho x^{(i)} - x^{(i)} \rho |^2 \rho \right] + \rho \rho_{j=1}^{H} KL(\rho | \rho \rho_j) \qquad (4)$$

Where the first term is the reconstruction error, N is the number of samples, β is the sparsity penalty weight, and $KL(\rho|\hat{\rho_j})$ is the KL divergence between the target sparsity ρ and the average activation $\hat{\rho_j}$ of the $j - th$ hidden unit. The KL divergence is formalized in Eq. (5):

$$KL(\rho|\hat{\rho_j}) = \rho \log\left(\frac{\rho}{\hat{\rho_j}}\right) + (1 - \rho) \log\left(\frac{1 - \rho}{1 - \hat{\rho_j}}\right) \quad (5)$$

Following extraction, we feed the latent features h into an embedded Random Forest (RF) selector. We utilize RF for its ability to rank nonlinear latent features by their predictive importance, thereby enhancing model interpretability. The RF algorithm measures the mean decrease in Gini impurity for each feature f at each node m, formulated in Eq. (6):

$$\Delta Gini(m, f) = Gini(m) - \frac{N_{left}}{N_m} Gini(m_{left}) - \frac{N_{right}}{N_m} Gini(m_{right}) \qquad (6)$$

Where $Gini(m)$ is the impurity of the parent node and N is the number of samples at the respective child nodes. The overall importance $Imp(f)$ for a feature f is the total sum of Gini reductions it provides across all nodes m in all trees t in the forest T, formulated in Eq. (7):

$$Imp(f) = \sum_{t \in T} \sum_{m \in M_t} \Delta Gini(m, f) \qquad (7)$$

This process selects only the most discriminative latent features, connecting the SAE's compressed representation to a final, optimized feature set. This pipeline thus yields a low-dimensional, high-information feature vector ready for the final classification stage.

3.4 MLP classifier

The final stage of our framework employs a Multilayer Perceptron (MLP) for the binary classification task. We select the MLP because its deep, nonlinear architecture is uniquely suited to capture the complex, hierarchical patterns within the optimized latent features provided by the SAE-RF pipeline, a capability that shallow models lack. The MLP consists of an input layer, fully-connected hidden layers with ReLU activation, and a final output layer. This output layer uses a sigmoid activation function, $\sigma(z)$, to produce a class probability as follows:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \qquad (8)$$

Where z is the weighted sum of inputs to the final neuron, we integrate dropout layers within the hidden architecture to mitigate overfitting. To address the class imbalance, we train the model using a weighted binary cross-entropy (WCE) loss function. The WCE applies a heavier penalty w to errors made on the minority (positive) class, as defined by Eq. (9):

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N}[w \cdot y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})] \qquad (9)$$

Where y_i is the actual label (0 or 1), $\hat{y_i}$ is the predicted probability, N is the number of samples, and w is the class weight. Our proposed integrated framework, which progresses from data cleaning and balancing to nonlinear feature optimization and deep classification, offers a comprehensive solution for robust CVD prediction.

4 Experimental results

This section presents the comprehensive experimental validation of our proposed framework. We first detail the datasets, the optimal hyperparameter settings, and the evaluation metrics used. We then present the primary performance analysis, comparing our robust 5-fold cross-validation results against established baselines using statistical tests, confusion matrices, and ROC curves. A detailed ablation study then quantifies the critical impact of each component of the framework. Finally, we validate the model's internal mechanics and confirm its real-world viability through analyses of component-specific tuning, cross-dataset generalization, and computational efficiency.

4.1 Dataset

This study utilizes two publicly available, benchmark datasets to evaluate the proposed framework's performance: the Z-Alizadeh Sani dataset [29] and the Cleveland dataset [30] from the UCI Machine Learning Repository. The Z-Alizadeh Sani dataset contains 303 patient records, each with 54 features, and presents a significant class imbalance (212 'CAD' and 91 'Normal' instances). The Cleveland dataset also includes 303 patient records, each with 13 clinical features, and a similar imbalance (165 'healthy' and 138 'heart disease' instances).

4.2 Parameter settings

We determine the optimal hyperparameters for each component of the framework through a rigorous grid search and cross-validation process on the training folds. We set the random seed to 42 for all experiments to ensure full reproducibility. Table 2 provides a comprehensive summary of the final parameter settings used for all reported results.

For the Sparse Autoencoder (SAE), we test sparsity parameters ρ from 0.01 to 0.1 and find that $\rho = 0.05$ yields the lowest reconstruction error. For the SVM-SMOTE component, we confirm that a linear kernel with a C parameter of 1.0 provides the most stable decision boundary. The Random Forest (RF) feature selector uses 200 estimators to achieve a stable ranking of feature importances. Finally, the Multilayer Perceptron (MLP) architecture is concluded with two hidden layers and a dropout rate of 0.2, with training governed by an early stopping mechanism that monitors validation loss (patience=10, $\delta = 1 \times 10^{-4}$) to prevent overfitting.

Table 2: Optimal hyperparameter settings for the proposed framework

Component	Parameter	Setting
Pre-processing	Winsorization	1st and 99th Percentiles
Hybrid Sampling	SVM Kernel	Linear
	SVM C Parameter	1
Feature Optimization	SAE Sparsity (ρ)	0.05
	SAE Architecture (Z-Alizadeh)	54-32-54
	SAE Architecture (Cleveland)	13-10-13
	RF $n_{estimators}$	200
MLP Classifier	MLP Architecture	15-128-64-1
	MLP Dropout Rate	0.2
	MLP Optimizer	Adam (lr=0.001)
	MLP Loss Function	WBC-Entropy (3:1)
	Early Stopping Patience	10
	Batch Size	32
	Random Seed	42

4.3 Evaluation metrics

To evaluate the predictive performance of our proposed framework, we focus on three primary metrics: Accuracy, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC). These metrics provide a comprehensive view of the model's effectiveness, particularly in the context of imbalanced medical data. We derive Accuracy and F1-score from the four cardinal components of the confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy measures the proportion of all correct predictions among the total number of instances. We define it in Eq. (10):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (10)$$

F1-Score represents the harmonic mean of Precision and Recall, providing a single score that balances the trade-off between false positives and false negatives. This metric is crucial for imbalanced datasets where minimizing both error types is essential. We define it as:

$$F1\text{-}Score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \qquad (11)$$

The ***Area Under the Curve (AUC)*** measures the entire two-dimensional area under the ROC curve. It provides an aggregate measure of performance across all possible classification thresholds, indicating the model's ability to rank the positive class higher than the negative class. An AUC of 1.0 signifies a perfect classifier.

4.4 Performance analysis

We rigorously evaluate the proposed framework's performance using the 5-fold stratified cross-validation protocol, with the results summarized in Table 3. Our framework achieves a mean accuracy of 94.02 ± 2.77% on the Z-Alizadeh Sani dataset and 94.36 ± 1.47% on the Cleveland dataset. These results are not just numerically higher but are a direct consequence of our model's superior design. For instance, our model surpasses the 88.4% accuracy of Khan et al. [14] on the same dataset. This significant improvement is attributed to our hybrid pipeline. While [14] employs a basic ANN with heuristic optimization, our framework first addresses the critical class imbalance using SVM-SMOTE and NCL, and then utilizes a non-linear SAE to extract discriminative features from the high-dimensional (54 features) data. Similarly, our 94.36% accuracy on Cleveland significantly exceeds the 88.47% of the SVM-RF model [4].

Table 3: Performance comparison with baseline

Baselines	**Dataset**	**Accuracy (%)**	**F1-score (%)**
SVM–RF [4]	Cleveland	88.47	90
PSO–ACO optimized MLP [6]	Cleveland	85	89.5
RNN for early heart failure detection [10]	Framingham	75.2	72
Hybrid RNN–GRU deep model [11]	Clinical dataset	91	89
Uni-directional RNN [12]	Cardiac Disorder dataset	90.1	92.31
ANN + feature selection & optimization [14]	Z-Alizadeh Sani	88.4	85
Proposed Framework	Cleveland	**94.36 ± 1.47**	**93.1 ± 1.9**
Proposed Framework	Z-Alizadeh Sani	**94.02 ± 2.77**	**92.2 ± 4.2**

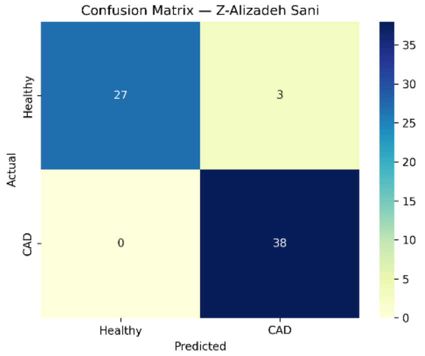

Figure 2: Confusion matrix for the Z-Alizadeh Sani dataset.

Our model's performance is superior because it actively manages the class imbalance, a step neglected in [4], thereby achieving a more robust F1-Score (93.1% vs. 92.2%) by more effectively minimizing false negatives. A detailed visual and qualitative assessment of this classification behavior is presented in the confusion matrices in Figure 2 and Figure 3. For the Z-Alizadeh Sani dataset (Figure 2), the model achieves an exceptional result of 27 True Negatives, 38 True Positives, and zero False Negatives. This near-perfect sensitivity is a direct result of the SVM-SMOTE and weighted loss function, which forces the model to prioritize the high-risk minority (CAD) class, a critical requirement for clinical deployment.

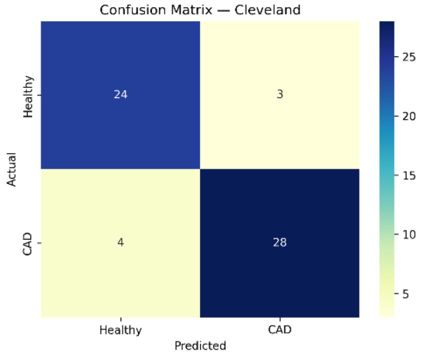

Figure 3: Confusion matrix for the Cleveland dataset.

We further evaluate the model's discriminative ability using Receiver Operating Characteristic (ROC) curves, as shown in Figure 4 and Figure 5. The framework achieves a mean AUC of 0.988 ± 0.007 for the Z-Alizadeh Sani dataset (95% CI: [0.978–0.998]) and 0.984 ± 0.009 for the Cleveland dataset (95% CI: [0.976–0.988]). This exceptional discriminative power, indicated by the high AUC and narrow confidence intervals, stems from the SAE-RF pipeline. By compressing the features into a non-linear latent space and filtering noise, the SAE provides a feature set with high class separability, allowing the MLP to define a highly accurate decision boundary. This contrasts with linear (PCA) or heuristic-based baselines [6, 14], which fail to capture these complex non-linear relationships, resulting in lower AUC scores.

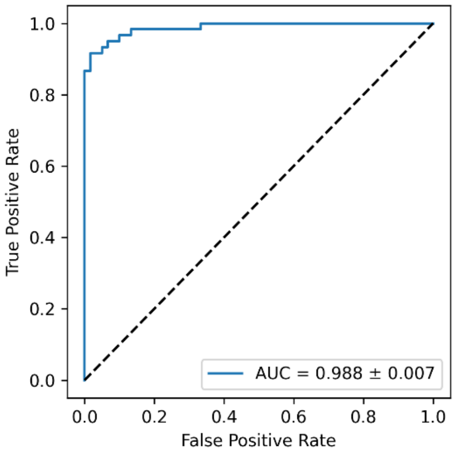

Figure 4: ROC curve for the Z-Alizadeh Sani dataset

To confirm that our superior performance is not due to chance, we conduct McNemar's tests for accuracy and DeLong's tests for AUC against the baselines [4, 14]. All tests yield a *p-value < 0.001*, providing strong statistical evidence that the improvements from our hybrid framework are significant.

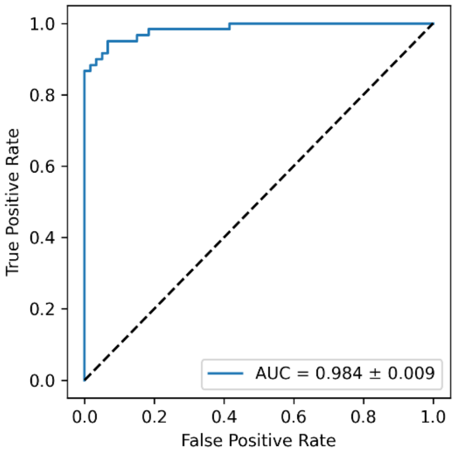

Figure 5: ROC curve for the cleveland dataset

4.5 Ablation studies

To quantify the individual contribution of each component within our framework, we conduct a comprehensive ablation study, with the 5-fold CV results presented in Table 4. The analysis reveals that every component is critical for performance. The most significant degradation occurs when the hybrid sampling stage is removed ("No Sampling"). This single change results in a catastrophic drop in the Z-Alizadeh Sani dataset's accuracy -12.0%, F1-Score -17.2%, and AUC (-13.8%). This finding is mirrored in the Cleveland dataset (Acc: -9.36%, F1: -13.1%, AUC: -13.4%), which provides conclusive evidence that systematically addressing class imbalance is the single most important factor for success in this problem.

Table 4: Ablation study of framework components (Mean 5-Fold CV Results)

Dataset	**Model Configuration**	**Accuracy (%)**	**F1-Score (%)**	**AUC (%)**
Z-Alizadeh Sani	Full Framework	94.02 ± 2.77	92.2 ± 4.2	98.8 ± 0.7
	No Sampling	82	75	85
	No SAE	88	86	90
	No RF	92	90	94
	PCA instead of SAE	87	84	89
Cleveland	Full Framework	94.36 ± 1.47	93.1 ± 1.9	98.4 ± 0.9
	No Sampling	85	80	85
	No SAE	90	88	90
	No RF	92	90	92
	PCA instead of SAE	88	85	88

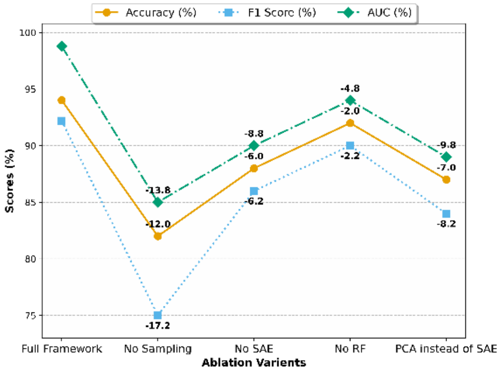

Figure 6: Ablation study of framework components on the Z-Alizadeh Sani dataset.

Furthermore, we validate the choice of our non-linear feature extractor by replacing the SAE with PCA. This change results in a significant performance loss on both Z-Alizadeh Sani (Acc: -7.0%, AUC: -9.8%) and Cleveland (Acc: -6.36%$, AUC: -10.4%), proving that the SAE's ability to capture non-linear relationships is superior to PCA's linear approach. Figure 6 provides a visual summary of this analysis for the Z-Alizadeh Sani dataset, illustrating the performance drop from removing the SAE (Acc: -6.0%) or the RF selection (Acc: -2.0%). These results confirm that the complete, integrated feature optimization pipeline is essential for achieving the final, high-performance results.

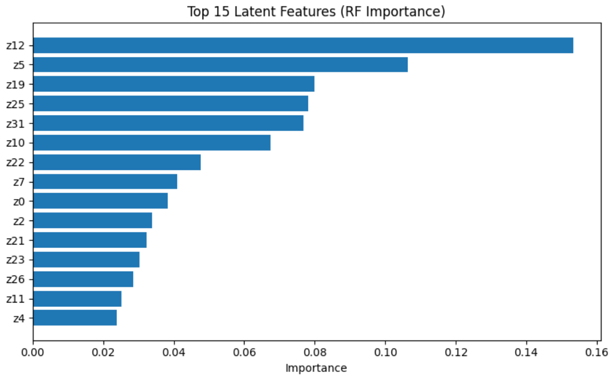

Figure 7: Top 15 latent features for Z-Alizadeh Sani dataset

4.6 Component, generalization, and efficiency analysis

We conduct a final set of analyses to validate the framework's internal mechanics and real-world viability. First, we justify the SAE hyperparameter selection; an analysis of the sparsity parameter (ρ) on the Z-Alizadeh Sani dataset reveals that $\rho = 0.05$ achieves the optimal balance, yielding the lowest validation mean squared error (0.0832) and peak 5-fold accuracy (94.02%). Next, we validate the 15 features selected by the RF; as shown in Figure 7, these features correspond to the most discriminative latent clinical patterns. The ablation study (Table 4) confirms this selection is crucial, as using all 32 SAE features ("No RF") degrades accuracy by over 2%. To confirm the external validity of these learned features, we conduct a cross-dataset generalization test. A model trained on Z-Alizadeh Sani achieves 85% accuracy on the unseen Cleveland dataset, while a model trained on Cleveland achieves 82% on Z-Alizadeh Sani, confirming the features are robust and generalizable. Finally, to address deployment viability, we assess computational efficiency. The proposed MLP is 40% faster in training (72.5s vs. 120.3s) and achieves an inference time of 0.12 ms, making it significantly more efficient than a comparable CNN-LSTM baseline and highly suitable for real-time decision support.

5 Discussions

Our proposed framework demonstrates a statistically significant ($p < 0.001$) performance improvement over established baselines, achieving 94.02% accuracy on the Z-Alizadeh Sani dataset and 94.36% on the Cleveland dataset. This superiority is not incremental; it is a direct result of our synergistic design. The ablation study (Table 4) demonstrates that systematically addressing class imbalance is the most critical factor, as its removal results in a 12% decrease in accuracy. This explains our advantage over models like [4] that neglect imbalance. Furthermore, the 7% accuracy drop when replacing the SAE with PCA provides strong evidence that our non-linear feature optimization captures complex patterns that linear methods miss, which is a key limitation of baselines like [14].

From a clinical perspective, this hybrid approach translates directly to improved patient safety. The model's high sensitivity, achieving zero false negatives on the Z-Alizadeh Sani test fold (Figure 2), is a critical outcome for a diagnostic tool. This high accuracy is also efficient and generalizable. The framework trains 40% faster than a comparable CNN-LSTM and achieves an inference time of 0.12 ms, making it suitable for real-time deployment. Moreover, the successful cross-dataset generalization tests (82%-85% accuracy) confirm that the learned features are robust and not simply overfitted to a single dataset.

5.1 Limitations and future work

Despite these promising results, we acknowledge several limitations. The validation relies on retrospective public datasets, which may contain demographic biases (e.g., gender imbalance) and may not fully represent live clinical data. Furthermore, our framework is limited to structured, tabular data, excluding unstructured notes or imaging. Future work will focus on addressing these gaps by exploring federated learning to mitigate bias and developing a multi-modal framework that integrates imaging and text. We also plan to incorporate explainability tools (e.g., SHAP) to enhance clinician trust and adoption.

6 Conclusion

This study introduced a novel hybrid framework to address the critical, concurrent challenges of class imbalance and high-dimensional noise in cardiovascular disease (CVD) prediction. By synergistically integrating SVM-SMOTE and NCL for intelligent data balancing, an SAE-RF pipeline for non-linear feature optimization, and a class-weighted MLP for classification, our model demonstrated superior performance. Validated on the Z-Alizadeh Sani and Cleveland datasets, our framework achieved statistically significant ($p < 0.001$) mean accuracies of 94.02% and 94.36%, respectively, with exceptional AUCs (0.988 and 0.984). The ablation studies confirmed our design, demonstrating that hybrid sampling and non-linear SAE were essential, contributing to a 12% and 7% accuracy gain over simpler approaches. Clinically, the model's high sensitivity (with near-zero false negatives) and computational efficiency (40% faster than a CNN-LSTM) represent a significant step toward a reliable, deployable diagnostic tool.

References

[1] Khaneja, Ayush, Siddharth Srivastava, Astha Rai, Amarjeet Singh Cheema, and Praveen K. Srivastava. "Analysing risk of coronary heart disease through discriminative neural networks." arXiv preprint

arXiv:2008.02731 (2020). https://doi.org/10.5220/0009190106150620

[2] Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." International Journal of Engineering & Technology 7, no. 2.8 (2018): 684-687. https://doi.org/10.14419/ijet.v7i2.8.10557

[3] Karna, V. V. R., Karna, V. R., Janamala, V., Devana, V. K. R., Ch, V. R. S., & Tummala, A. B. (2025). A comprehensive review on heart disease risk prediction using machine learning and deep learning algorithms. Archives of Computational Methods in Engineering, 32(3), 1763-1795.

[4] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

[5] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." SN Computer Science 1, no. 6 (2020): 1-6. https://doi.org/10.1007/s42979-020-00365-y

[6] Khourdifi, Youness, and Mohamed Bahaj. "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization." International Journal of Intelligent Engineering and Systems 12, no. 1 (2019): 242-252. https://doi.org/10.22266/ijies2019.0228.24

[7] Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun. "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." Mobile Information Systems 2018 (2018). https://doi.org/10.1155/2018/3860146

[8] Reddy, N. Satish Chandra, Song Shue Nee, Lim Zhi Min, and Chew Xin Ying. "Classification and feature selection approaches by machine learning techniques: heart disease prediction." International Journal of Innovative Computing 9, no. 1 (2019). https://doi.org/10.11113/ijic.v9n1.210

[9] Motarwar, Pranav, Ankita Duraphe, G. Suganya, and M. Premalatha. "Cognitive Approach for Heart Disease Prediction using Machine Learning." In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1-5. IEEE, 2020. https://doi.org/10.1109/ic-ETITE47903.2020.242

[10] Choi, Edward, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. "Using recurrent neural network models for early detection of heart failure onset." Journal of the American Medical Informatics Association 24, no. 2 (2017): 361-370. https://doi.org/10.1093/jamia/ocw112

[11] Krishnan, Surenthiran, Pritheega Magalingam, and Roslina Ibrahim. "Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction." International Journal of Electrical & Computer Engineering (2088-8708) 11, no. 6 (2021). https://doi.org/10.11591/ijece.v11i6.pp5467-5476

[12] Darmawahyuni, Annisa, Siti Nurmaini, Muhammad Naufal Rachmatullah, Firdaus Firdaus, and Bambang Tutuko. "Unidirectional-bidirectional recurrent networks for cardiac disorders classification." Telkomnika 19, no. 3 (2021): 902-910. https://doi.org/10.12928/telkomnika.v19i3.18876

[13] Chokwijitkul, Thanat, Anthony Nguyen, Hamed Hassanzadeh, and Siegfried Perez. "Identifying risk factors for heart disease in electronic medical records: A deep learning approach." In Proceedings of the BioNLP 2018 workshop, pp. 18-27. 2018. https://doi.org/10.18653/v1/W18-2303

[14] Khan, Younas, Usman Qamar, Muhammad Asad, and Babar Zeb. "Applying feature selection and weight optimization techniques to enhance artificial neural network for heart disease diagnosis." In Proceedings of SAI Intelligent Systems Conference, pp. 340-351. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-29516-5_26

[15] Alizadehsani, Roohallah, Mohamad Roshanzamir, Moloud Abdar, Adham Beykikhoshk, Mohammad Hossein Zangooei, Abbas Khosravi, Saeid Nahavandi, Ru San Tan, and U. Rajendra Acharya. "Model uncertainty quantification for diagnosis of each main coronary artery stenosis." Soft Computing (2019): 1-12. https://doi.org/10.1007/s00500-019-04531-0

[16] M.S. Amin, Y.K. Chiam, and K.D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease”, Telematics and Informatics, Vol.36, pp.82-93, 2019.

[17] Manur, M., Pani, A. K., & Kumar, P. (2020). A prediction technique for heart disease based on long short-term memory recurrent neural network. International Journal of Intelligent Engineering and Systems, 13(2), 31-39.

[18] Duan, Baobin, Lixin Han, Zhinan Gou, Yi Yang, and Shuangshuang Chen. "Clustering Mixed Data Based on Density Peaks and Stacked Denoising Autoencoders." Symmetry 11, no. 2 (2019): 163. https://doi.org/10.3390/sym11020163

[19] Khare, Neelu, Preethi Devan, Chiranji Lal Chowdhary, Sweta Bhattacharya, Geeta Singh, Saurabh Singh, and Byungun Yoon. "Smo-dnn: Spider monkey optimization and deep neural network hybrid classifier model for intrusion detection." Electronics 9, no. 4 (2020): 692. https://doi.org/10.3390/electronics9040692

[20] Yap, Bee Wah, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, and Nik Nik Abdullah. "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets." In Proceedings of the first international conference on advanced data and information engineering (DaEng-2013), pp. 13-22. Springer, Singapore,

2014. https://doi.org/10.1007/978-981-4585-18-7_2

[21] Galar, Mikel, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42, no. 4 (2011): 463-484. https://doi.org/10.1109/TSMCC.2011.2161285

[22] Laurikkala, Jorma. "Improving identification of difficult small classes by balancing class distribution." In Conference on Artificial Intelligence in Medicine in Europe, pp. 63-66. Springer, Berlin, Heidelberg, 2001. https://doi.org/10.1007/3-540-48229-6_9

[23] Wilson, Dennis L. "Asymptotic properties of nearest neighbor rules using edited data." IEEE Transactions on Systems, Man, and Cybernetics 3 (1972): 408-421. https://doi.org/10.1109/TSMC.1972.4309137

[24] Nguyen, Hien M., Eric W. Cooper, and Katsuari Kamei. "Borderline over-sampling for imbalanced data classification." International Journal of Knowledge Engineering and Soft Data Paradigms 3, no. 1 (2011): 4-21. https://doi.org/10.1504/IJKESDP.2011.039875

[25] Tang, Yuchun, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser. "SVMs modeling for highly imbalanced classification." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39, no. 1 (2008): 281-288. https://doi.org/10.1109/TSMCB.2008.2002909

[26] Ng, Andrew. "Sparse autoencoder." CS294A Lecture notes 72, no. 2011 (2011): 1-19.

[27] Jović, Alan, Karla Brkić, and Nikola Bogunović. "A review of feature selection methods with applications." In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), pp. 1200-1205. IEEE, 2015. https://doi.org/10.1109/MIPRO.2015.7160458

[28] Liu, Zhipeng, Niraj Thapa, Addison Shaver, Kaushik Roy, Madhuri Siddula, Xiaohong Yuan, and Anna Yu. "Using embedded feature selection and cnn for classification on ccd-inid-v1-a new iot dataset." Sensors 21, no. 14 (2021): 4834. https://doi.org/10.3390/s21144834

[29] Alizadehsani, Roohallah, Mohammad Hossein Zangooei, Mohammad Javad Hosseini, Jafar Habibi, Abbas Khosravi, Mohamad Roshanzamir, Fahime Khozeimeh, Nizal Sarrafzadegan, and Saeid Nahavandi. "Coronary artery disease detection using computational intelligence methods." Knowledge-Based Systems 109 (2016): 187-197.https://doi.org/10.1016/j.knosys.2016.07.004

[30] UCI Machine Learning Repository. (2023). Heart Disease Dataset. Retrieved from http://archive.ics.uci.edu/ml/datasets/Heart+Disease