IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 August 2024, accepted 26 September 2024, date of publication 30 September 2024, date of current version 9 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3470537

RESEARCH ARTICLE

Accurate Cardiovascular Disease Prediction: Leveraging Opt_hpLGBM With Dual-Tier Feature Selection

J. JASMINE GABRIEL AND L. JANI ANBARASI
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu 600127, India
Corresponding author: L. Jani Anbarasi (janianbarasi.l@vit.ac.in)

This work was supported by the Vellore Institute of Technology, Chennai, India.

ABSTRACT Reliable forecasting of cardiovascular disease (CVD) outcomes is crucial for efficient patient management. While machine learning (ML) holds promise for disease prediction, challenges arise, particularly with smaller clinical datasets. Feature engineering is essential in this context, as it involves analyzing missing values, managing outliers, and addressing multicollinearity. This process is key to identifying and eliminating unnecessary features from the dataset. To tackle this, a scalable ML based Dual-Tier feature selection framework called ANOVA Chi-Squared ($AnoX^2$) is proposed, utilizing a hybrid statistical method. The framework integrates validation using five different ML classifiers with the selected features from $AnoX^2$. The proposed model Opt_hpLGBM (Optuna hyperparameter tuned Light Gradient Boost Machine) along with $AnoX^2$ feature selection exhibits outstanding performance across four publicly available datasets, consistently achieving remarkable accuracy. For instance, it achieves 94.87% accuracy in the Cleveland dataset with 8 features, 95.12% in the Statlog dataset with the same number of features, 92.81% accuracy with 7 features in the heart disease dataset, and an impressive 98.85% accuracy in the z-Alizadeh Sani dataset with 12 features. These results exceed current benchmarks, establishing it as an industry leader in terms of the number of features utilized, accuracy, precision, recall, F1 score, and log loss metrics. With its potential for early diagnosis and treatment, this innovative framework can transform healthcare, significantly reducing mortality rates associated with cardiovascular disease.

INDEX TERMS ANOVA, cardiovascular disease, chi-squared, feature selection, hyperparameter tuning, preprocessing.

I. INTRODUCTION

The World Health Organization [1] reported that in 2019, 17.9 million deaths worldwide were attributed to CVD, accounting for 32% of all fatalities. Of these deaths, 85% were caused by strokes and cardiovascular disease. Over 75% of CVD related fatalities occurred in low to moderate income countries. Furthermore, three-quarters of the 17 million premature deaths from non-communicable diseases in 2019 were attributed to CVDs. The escalating prevalence of heart-related ailments presents a significant concern, necessitating meticulous treatment [2]. Contributing factors such as diabetes, hypertension, and elevated cholesterol levels pose challenges in early heart disease detection, underscoring the need for healthcare providers to exercise utmost caution. Addressing heart-related concerns mandates a systematic approach for precise assessment and management. Advances in machine learning algorithms and significant drops in data storage costs have caused a noticeable shift in the medical environment in recent years [3]. This has spurred the widespread integration of various machine learning techniques and data mining technologies in healthcare. Augmenting conventional diagnostic procedures with machine learning algorithms enhances predictive accuracy, aiding in refined risk assessment. The analysis of test results alongside machine learning predictions enables nuanced conclusions regarding heart health, facilitating tailored interventions. These interventions, spanning lifestyle

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

modifications, medication regimes, or specialist referrals, are meticulously calibrated based on the comprehensive assessment, incorporating insights from machine learning predictions. Disease classification and prediction benefit significantly from accurate feature selection, which involves removing irrelevant features and building models with only relevant ones. By eliminating redundant or irrelevant data, feature selection lowers dimensionality and enhances model performance by pinpointing the dataset's most relevant and instructive features. This approach reduces the time and expense associated with illness prediction while concurrently enhancing predictability. Feature selection is crucial for managing data complexity, especially in datasets containing a mix of numerical and categorical features. By focusing on the most relevant features, it simplifies the data, making model building and evaluation more straightforward. Additionally, feature selection aids in identifying specific biomarkers or factors unique to individual patients, which facilitates personalized and targeted treatment approaches. The selection techniques for features vary based on factors such as the dataset, outliers, and the quality of dataset preparation. These techniques encompass the Filter, Wrapper, and Embedded methods. Correct model selection ensures that the chosen algorithm is suitable for the problem at hand, considering factors such as the nature of the data, the complexity of relationships, and the desired outcomes. By pinpointing the most important features related to a disease, feature selection also offers valuable clinical insights into its causes and risk factors, thereby enhancing the understanding and managing health conditions. Achieving a perfect model involves considering preprocessing, feature selection, and model building. Many classifiers, such as Support Vector Machines (SVM) [4], Gradient Boost (GB) [5], eXtreme Gradient Boosting (XGB) [6], Random Forest (RF) [7], Multilayer Perceptron (MLP) [8], Decision trees (DT) [3], logistic regression (LR) [9], and Neural networks (NN) [10], were utilized in recent research on CVD prediction. Additionally, evolutionary based feature selection methods have gained attention from researchers due to their potential to improve model performance. Furthermore, hyperparameter tuning adjusts the model's parameters to maximize performance and enhance the model's capacity for good generalization on unseen data. Together, these steps play a vital role in building robust and accurate machine learning models, enhancing their predictive capabilities, and ensuring reliable results in various applications. The following summarizes the primary contributions of the suggested AnoX^2-Opt_hpLGBM:

1. Efficient preprocessing techniques were employed, ensuring dataset readiness for subsequent analysis.
2. Developed a dual-tier feature selection approach to prevent the oversight of relevant features while excluding irrelevant ones.
3. Introduced an automated mechanism for hyperparameter tuning, optimizing hyperparameter selection without manual intervention.
4. Identified the optimal classifier demonstrating superior performance for CVD prediction tasks through rigorous experimentation on four publicly available CVD datasets.
5. A single model was developed that is suitable for all four publicly available CVD datasets and has also been tested on other diseases, such as Chronic Kidney Disease (CKD) and the Pima Indians Diabetes Database (PIDD).
6. In comparison to contemporary cutting-edge methods for CVD prediction, this comprehensive evaluation highlighted the superiority and efficacy of the suggested AnoX^2- Opt_hpLGBM model.

The analysis begins with a systematic review of related literature and a comparison table in Section II. Section III identifies the research gap in disease prediction, presenting it as both a motivation for the study and a call for further investigation. Section IV introduces the system architecture, followed by Section V, which provides a comprehensive explanation of the proposed approach, including model design, feature selection, data preparation, and hyperparameter tuning. Section VI describes the experimental approach and dataset in detail, offering a thorough exploration of the experimentation process and validation of findings. Section VII presents the results obtained in each module, along with the performance metrics. Finally, Section VIII concludes the experiment with a concise summary of key findings, their implications, and potential directions for future research.

II. RELATED WORKS

Alharbi [11] present the Artificial Rabbits Optimizer (ARO), integrated with machine learning methodologies to monitor real-time activity in emergency departments. This system tracks the length of stay (LOS), types of treatment, and patient visit data, and classifies medical data across hospitals in Saudi Arabia. The ARO technique selects a feature subset, which is then classified using a class-specific cost-regulated extreme learning machine classifier. The parameters of this classifier are fine-tuned using the Grasshopper Optimization Algorithm (GOA).

Abdellatif et al. utilized the Cleveland and Statlog datasets to predict clinical heart disease, employing a novel method called Infinite Feature Selection combined with Improved Weighted Random Forests (IWRF). They identified the most important features using Infinite Feature Selection and fine-tuned the hyperparameters of IWRF using Bayesian optimization approaches [12]. The authors compared their model with other traditional machine learning algorithms, including SVM, k-Nearest Neighbors (kNN), and XGB.

An Internet of Medical Things (IoMT) platform was developed by Khan and Algarni [13] aimed to forecast heart disease (HD) using a novel approach. They utilized Modified Slap Swarm Optimization (MSSO) to optimize the parameters of the Adaptive Neuro-Fuzzy Inference System (ANFIS), thereby enhancing the predictive capacity and precision of the

IoMT framework. Additionally, they incorporated the Levy Flight mechanism to enhance efficiency and local optimization capabilities, resulting in the Levy based Crow Search Algorithm (LCSA) for feature selection.

Temidayo et.al proposed a LGBM model optimized with a Tree-structured Parzen Estimator (TPE) for breast cancer diagnosis, achieving a 99% accuracy on the Wisconsin Diagnostic Breast Cancer dataset (WDBC). The study employed the Borderline Synthetic Minority Oversampling Technique (SMOTE) to focus on borderline minority samples, ensuring a balanced dataset. Additionally, the LGBM algorithm was tuned using TPE optimization and validated with 10-fold cross-validation [14].

Rufo et al. developed a prediction model for an Ethiopian based type 2 diabetes dataset. Features with low correlation to the target variable were dropped, missing values were replaced with mean values, and min-max normalization was applied to the dataset. The author utilized 10-fold cross-validation and grid search for parameter tuning of the LGBM classifier. The model was compared with, SVM, Naive Bayes (NB), RF, bagging, and XGB, with LGBM emerging as the best-performing model [15].

The Enhanced Sparrow Search Algorithm (E-SSA) is an optimal feature selection technique presented by Sonam et al. This process was used to pick features, which were then combined with the Aquila Optimization Algorithm (AOA) to create a model. A comprehensive cardiac dataset made up of integrated datasets from Cleveland, Statlog, Hungary, Cleveland, Switzerland, and Long Beach VA was subjected to the suggested model's application [16].

Sonam et al. introduced a method called Distributed-t-Stochastic Neighborhood Embedding (D-t-SNE) to address overfitting concerns and eliminate redundant data, thereby enhancing classifier performance in predicting heart disease. Additionally, they employed a hyper parameter tuned Multi-Layer Perceptron (H-MLP) to efficiently handle classification challenges. Here the feature extraction and feature level fusion were done by the deep Convolution neural network (CNN) model [17].

The objective of Patro et al. [18] was to use important risk factors to create a prediction framework for the diagnosis of heart disease. NB, Bayesian optimized support vector machine (BO-SVM), kNN, and Salp swarm optimized neural network (SSA-NN) are among the classifier methods used by the system. The work tries to improve the efficacy and accuracy of heart illness diagnosis using a dataset on heart disease from the UCI Machine Repository.

Yang et al. [19] introduced HY_OptGBM, an enhanced prediction model for coronary heart disease (CHD). This model makes use of an enhanced version of LGBM, with Optuna, an advanced hyperparameter optimization framework, being used to fine-tune the hyperparameters. To enhance model performance, the authors also used a unique loss function known as Focal Loss (FL). Model construction and evaluation were conducted using the Framingham dataset, which demonstrated the efficacy and resilience of the suggested method in predicting CHD.

Akella et al. [20] presented a method to forecast coronary artery disease (CAD) in 2021. Using the Cleveland Dataset, they ran trials to estimate CAD using six different machine-learning methods and neural network algorithms. The objective was to create a useful clinical tool for CAD detection by combining the strengths of innovative neural network techniques and conventional machine learning approaches.

Using the XGBoost classifier and hyperparameter adjustment via Optuna, Srinivas and Katarya [21] introduced the hyoPTXg model, which is intended to predict heart disease. The authors employed datasets from the Cleveland dataset and with two distinct heart datasets obtained from Kaggle. The research involves removing missing value entities, applying min-max normalization, and using an Optuna tuned XGB classifier. Ali et al. [22] delve into two primary objectives: refining features and addressing challenges posed by predictive modeling, namely underfitting and overfitting. To tackle the issue of irrelevant features, we advocate for the utilization of the $\chi 2$ statistical model. Simultaneously, he employs an exhaustive search strategy to identify the optimal configuration for a deep neural network (DNN). The efficacy of our hybrid model, termed $\chi 2$-DNN, is assessed by benchmarking its performance against traditional artificial neural network (ANN) and DNN models. To predict heart disease, he makes use of the Cleveland dataset.

A Machine Intelligence Framework for Heart Disease Diagnosis (MIFH) was introduced by Gupta et al. [23]. MIFH employs Factor Analysis of Mixed Data (FAMD) to extract and generate features from the Cleveland dataset. The preprocessing steps included missing value imputation, Z-score normalization, and data stratification. Feature extraction resulted in 2 to 28 features, validated using a hold-out method. These extracted features were then used to train machine-learning models for heart disease diagnosis. The author experimented with classification techniques such as RF, kNN, SVM, DT, and LR. Among these, the combination of FAMD and RF demonstrated the highest accuracy and was identified as the best-performing model.

Utilizing the Cleveland dataset, Mohan et al. [24] developed a novel prediction model for heart disease risk known as the Hybrid Random Forest and Linear Model (HRFLM). To improve its ability to predict cardiac illness, this model combines two well-known machine learning approaches: random forest and linear model. The original Cleveland dataset was further divided into 8 datasets based on specific classification rules. The classification was conducted using R Studio. The dataset was trained using three popular classification algorithms: DT, RF, and LR. Among these, the hybrid model combining RF and LM performed the best.

Ali et al. [25] devised a heart disease prediction model by merging two SVM models. In the application of L1 regularization and a linear method, the first SVM model

can eliminate superfluous features by setting their coefficients to zero. In contrast, the predictive portion of the framework is the second SVM model, which employs L2 regularization. Utilizing a hybrid grid search algorithm (HGSA), the two models could be optimized simultaneously, offering quick fine-tuning of both. For this investigation, the Cleveland dataset was used. An ML based technique for HD prediction was developed by Kavitha et al. [26] using regression and classification techniques with RF, DT, and a hybrid of RF and DT approaches. The model, utilizing the Cleveland dataset, achieved an accuracy of 88.7% with the proposed hybrid algorithm combining RF and DT. Using an ANN, Almazroi et al. [27] created a Deep Learning (DL) based framework for HD prediction, with a maximum accuracy of 83% across four datasets. The proposed model involves data acquisition and preprocessing, followed by the application of a classification model. The work incorporates 9 different classification models and one DL model. Validation of the DL model was conducted using various configurations of the dense neural network's hidden layers, ranging from three to nine layers, with each hidden layer containing 100 neurons utilizing the ReLU activation function. Ultimately, the DL model demonstrated superior predictive performance compared to the machine learning models.

A system for predicting cardiac illness was created by Bharti et al. [28] by combining DL and ML techniques. They conducted their analysis using the Cleveland dataset, and the DT approach they used produced encouraging results. An embedded Lasso algorithm was used for feature selection in the study. The author evaluated three approaches for the classification model: the first with no feature selection and no outlier detection, the second with feature selection but no outlier detection, and the third with both feature selection and outlier detection. Ultimately, the third method, using an ANN classifier, showed the best predictive performance on the Cleveland heart dataset.

Chakraborty [29] introduced a novel method called ECR-CNN (Extraction of Classification Rules from CNN) to improve the accuracy of heart disease prediction. This algorithm extracts classification rules, helping healthcare professionals make informed predictions and provide appropriate medication.

Omotehinwa et al. [30] proposed using Multiple Imputation by Chained Equations (MICE) with 10-fold cross-validation for feature selection on the Framingham dataset. Bayesian optimization was applied to tune the LGBM classification algorithm, and Borderline SMOTE was used for cross-validation. The Gradient based One-Side Sampling (GOSS) boosting method was identified as the optimal choice for this model. The study compared three approaches: the first used a baseline LGBM with missing values dropped, the second used a baseline LGBM with MICE-imputed data, and the third combined baseline LGBM with MICE imputation, TPE optimization, and Borderline SMOTE. The third approach achieved the highest accuracy while effectively addressing the class imbalance problem.

Sai et al. [31] proposed an ensemble model combining LGBM and adaptive prediction for the Pima Indians Diabetes dataset, which contains 8 features with no feature selection performed. The ensemble method was evaluated using five-fold cross-validation, with soft voting employed as the meta-classifier. The ensemble model included LGBM, kNN, and AdaBoost classifiers, achieving an accuracy of 90% using all features.

Subramani et al. [32] proposed a disease prediction model using a heart dataset with 11 features. They selected features using the Gradient Boosted Decision Trees (GBDT) combined with the SHAP method. The model was validated through five-fold cross-validation using a stacking approach, which included both base learners and a meta-learner. The machine learning algorithms incorporated in the model were SVM, kNN, LR, Extra Trees (ET), and LGBM.

Kodri and Hadianti [33] proposed a prediction model that was tested in both Python and RapidMiner environments. The model utilized the Cleveland dataset and involved exploratory data analysis through Pearson correlation and multicollinearity assessments. Min-max normalization was applied, and the dataset was upsampling using the SMOTE technique. The author employed classifiers such as RF, kNN, and LR, with 10-fold cross-validation. Grid search was used for hyperparameter tuning in RapidMiner, while Optuna was employed for the same purpose in Python. Chest pain (cp) was identified as the most significant feature, as determined by feature importance analysis using the Skor method.

Table 1 provides a simplified comparison of all the literature surveyed, highlighting the primary details that reveal the research gap for the proposed work. The table includes essential information such as feature selection methods, classification techniques, optimization algorithms, and datasets used, along with their performance metrics. It can be observed that not all authors follow the same process some researchers skip feature selection, while others omit parameter tuning. This inconsistency has opened the door to developing a single model that performs well across various classification problems, particularly in the healthcare sector. In this proposed work, we developed a new feature selection algorithm and classification model designed to handle both categorical and numerical features with categorical targets. While ANOVA is effective for datasets with numerical features and categorical targets and supports assumption testing for validating results, the Chi-Square (χ^2) test is more suitable for datasets with categorical features and categorical targets. We utilized the χ^2 test in our proposed work to select the most relevant features by evaluating their relationships with the target variable. Additionally, LGBM's classification algorithm is known for its fast and lightweight nature. Its efficiency, combined with its capability to handle categorical features, is due to its histogram-based algorithm for finding the best split. This results in improved performance and outcomes.

TABLE 1. A comparative analysis of the existing literature.

Ref. no	Feature selection	Method	Hyper parameter tuning	Dataset	Accuracy (%)
[11]	ARO	AROML-HDC	GOA	Cleveland Statlog	93.22 94.05
[12]	Infinite Feature Selection	If-FSs+BO+IWRF	BO	Statlog Heart disease clinical record	98.3 97.2
[13]	LCSA	MSSO-ANFIS	-	UCI dataset	99.45
[14]	-	Borderline-SMOTE+LGBM+TPE	TPE	WDBC	99.12
[15]	-	LGBM	-	Ethiopian based type 2 diabetes	98.1
[16]	E-SSA	Deep-DenseAquilaNet	AOA	Heart Disease	99.57
[17]	Deep CNN	D-t-SNE	H-MLP	Heart Disease	96.68
[18]	SSA	BO-SVM	BO	Cleveland	93.3
[19]	-	HY_OptGBM	Optuna	Framingham	93.0
[20]	-	ANN	-	Cleveland	93.03
[21]	-	HyOPTXg	Optuna	Cleveland Heart failure Kaggle Heart disease UCI Kaggle	94.71 89.34 85.50
[22]	X^2 model	X^2 Statistical model + DNN	Grid search	Cleveland	93.33
[23]	FAMD	FAMD+ RF	-	Cleveland	93.44
[24]	-	HRFLM	-	Cleveland	88.7
[25]	-	L1Linear SVM + L2 Linear &RBF SVM	HGSA	Cleveland	92.22
[26]	-	Hybrid of RF + DT	-	Cleveland	88.70
[27]	-	DL	-	Heart	83
[28]	Isolation forest	ANN	Grid Search	Cleveland	94.2
[29]	Feature extraction	ECRCNN	-	Statlog	85.19
[30]	-	Borderline-SMOTE + TPE + LGBM	TPE	Framingham	98.82
[31]	-	LGBM+kNN+AdaBoost	-	Pima Indian Diabetes dataset	90.76
[32]	GBT+SHAP	XGB	-	Heart dataset	96
[33]	X^2	RF	Optuna	Cleveland	93.15

III. RESEARCH GAP

Despite advancements in outlier detection and handling overfitting-underfitting, there remains a gap in exploring techniques tailored for datasets with numerous categorical features that require proper encoding. Additionally, many models include irrelevant or redundant features, reducing accuracy and robustness.

This highlights the necessity for feature selection methods that not only enhance model performance but also improve interpretability, leading to more reliable and comprehensible predictions. Moreover, the efficient utilization of Optuna hyperparameter tuning for optimizing model performance is yet to be fully explored. Furthermore, existing studies have not addressed the need for a unified model capable of effectively handling different heart disease datasets. Therefore, this research aims to fill this gap by introducing AnoX2-Opt_hpLGBM and comparing its effectiveness to other techniques in heart disease prediction. We forecast that our suggested model would perform better than the most advanced models and earlier research findings, proving its superiority in classification accuracy.

IV. SYSTEM ARCHITECTURE

The proposed architecture, illustrated in Figure 1, aims to accurately predict patient survival in cardiovascular diseases and the presence of heart disease. It includes data collection from four different datasets, preprocessing, feature selection using a dual-tier method with ANOVA and chi-squared tests, predictive modeling, and automated hyperparameter tuning with Optuna. The process concludes with six different performance evaluations, ultimately predicting the presence or absence of cardiovascular disease.

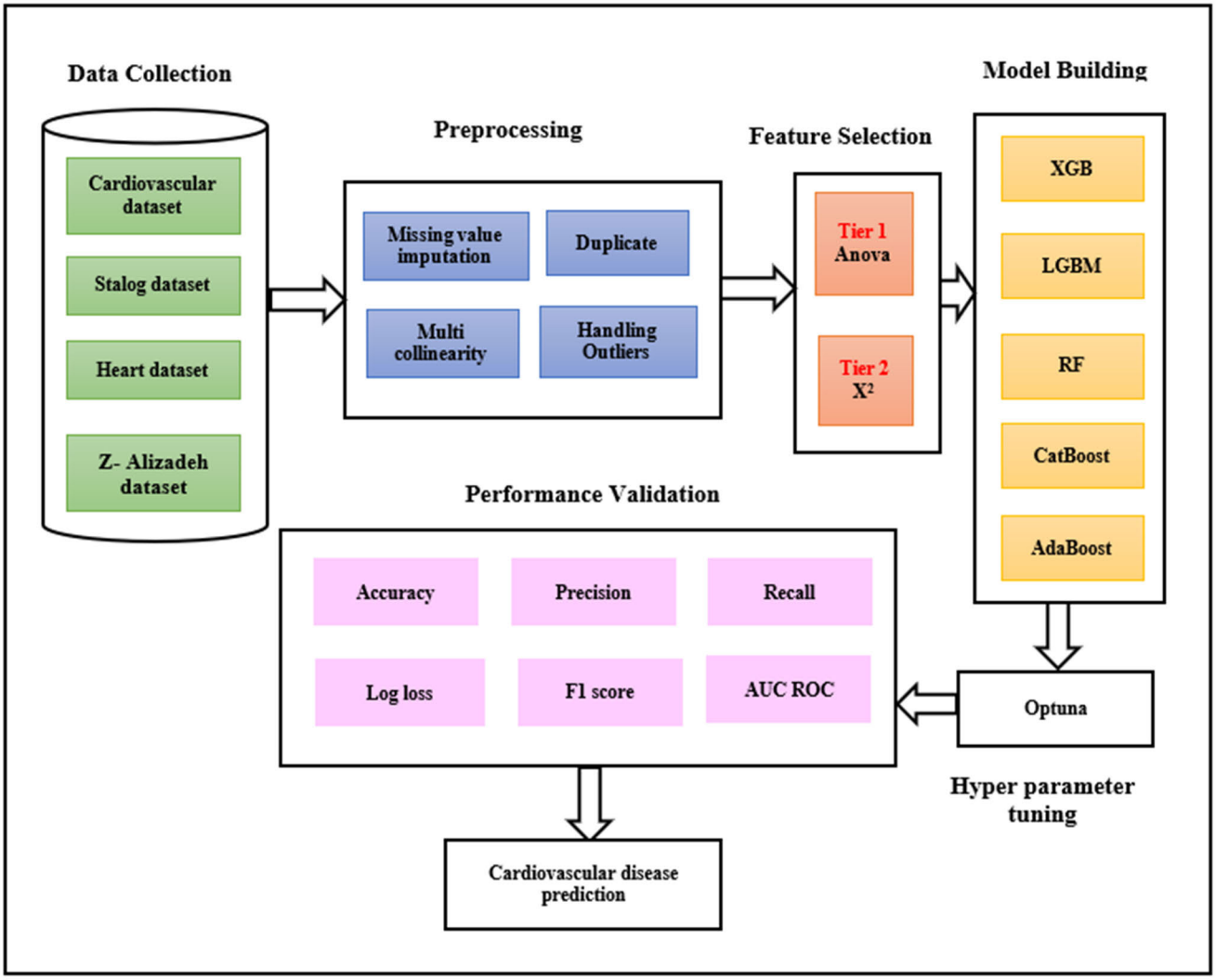

FIGURE 1. System architecture of the proposed model.

V. PROPOSED ML BASED PREDICTION FOR CVD

A. DATA PREPROCESSING

Preprocessing is indispensable to ensure the reliability and quality of data before analysis or modeling. It enhances predictive performance and reduces computational costs by identifying and handling missing values, outliers, and errors.

Feature selection aims to identify the most relevant information for model training, thereby improving predictive performance and reducing computational complexity. Optimizing model parameters, selecting the best algorithm, and evaluating performance metrics are critical steps in model prediction and hyperparameter tuning, all aimed at enhancing prediction accuracy and generalization.

B. DUAL-TIER ANOX^2 FEATURE SELECTION

Feature selection is essential to address the curse of dimensionality, preventing sparsity and computational inefficiency. It mitigates overfitting by excluding irrelevant or redundant features and enhances model generalization. It improves model performance and interpretability by concentrating on the most informative features. Additionally, it facilitates faster training and inference, optimizing computational resources. Here we adopt a dual-tier feature selection process named AnoX^2.

1) TIER 1: INITIAL FEATURE SELECTION

An analysis of variance, or ANOVA for short, is a statistical method that compares the means of two or more groups to see if there are any statistically significant differences between them. It is used to perform the initial feature selection. ANOVA evaluates each feature's relevance to the target variable in the context of feature selection by computing the F-value, which is the ratio of inter-group variance compared to intra-group variance [30]. Features with higher F-values are considered more relevant to the target variable. The first tier involves calculating the F-value for each feature in the

dataset [34]. Features with higher F-values are retained for further analysis in the second tier. ANOVA's two statistical metrics used to evaluate the significance of mean differences between two or more groups in a dataset are the p-value and the F-statistic. The variation within groups and the variance between groups are the two variances that make up the F-statistic. It measures how much the group means varies from one another about how much they vary within the groups. Greater variations between group means are indicated by a bigger F-statistic The p-value associated with the F-statistic indicates the probability of receiving the observed F-statistic (or one more extreme) under the null hypothesis, which assumes that the group means are equal.

When a small p-value is typically less than a chosen significance level, such as 0.05 indicates that it is implausible that the observed differences between group averages are the product of random chance, the null hypothesis is rejected.

2) TIER 2: REFINEMENT AND VALIDATION

X^2, a statistical metric used to determine the relationship or independence between two categorical variables, runs the second tier with further refinement [35]. It quantifies the discrepancy between the observed frequencies of the variables and the frequencies that would be expected under the assumption of independence. The second tier involves applying the X^2 method to the selected features from Tier 1. It computes the gap between the frequencies that are observed and those that are predicted under the independence null hypothesis. The test statistic is distributed according to a chi-squared distribution; the greater the chi-squared value, the less likely it is that the link is the result of chance.

The dual-tier feature selection method combines the advantages of both ANOVA and X^2 techniques. When these two techniques are combined, they provide a more comprehensive assessment of feature importance. ANOVA is effective for identifying influential continuous features, while X^2 is useful for highlighting the importance of categorical features. By integrating both methods, the combined approach can handle datasets with mixed types of features more effectively, resulting in a more robust and accurate feature selection process. Algorithm 1 provides a detailed explanation of our model's inner workings during the feature selection process, utilizing a dual-tier hybrid two-filter feature selection method. Given the binary nature of our classification task and the linear separability of the dataset, a tree-based classifier emerged as an ideal choice for our methodology.

Algorithm 1 Proposed Two-Tier feature selection

Step 1	Tier 1: OneWayANOVA (all features): Define Hypotheses H_0: All levels or groups in the feature have equal variance. H_1: At least one group is different. .
Step 2	Calculate the Sum of Squares Total Sum of Squares (SST) = Σ (Yi - $\bar{Y}$) ^{2} Between Sum of Squares (SSB) = Σni ($\bar{Y}$i - $\bar{Y}$) ^{2} Within Sum of Squares (SSW) = $\Sigma\Sigma$ (Yij - $\bar{Y}$i) ^{2}
Step 3	Determine Degrees of Freedom (df) df1 = k - 1 df2 = n – k
Step 4	Calculate F-value F-value (F) = SSB / SSW $*$ (df2 / df1)
Step 5	Accept or Reject the Null Hypothesis If F-value > Critical F-value: Reject Null Hypothesis (drop the feature) Else: Fail to Reject Null Hypothesis (accept the feature)
Step 6	Tier 2: Chi-squared (χ^2) test Define Hypotheses: H0: There is no relationship between the feature and the target variable. H1: There is a relationship between the feature and the target variable.
Step 7	Calculate the Chi-squared Statistic: Use the following formula to compute the chi-squared statistic: $X^2 = \sum (O_i - E_{i)})^2 / E_i$
Step 8	Selected feature A significant chi-squared statistic suggests that the feature is associated with the target variable.

C. LGBM

Light Gradient Boosting Machine is a robust gradient-boosting framework developed by Microsoft acclaimed for its exceptional accuracy and speed. Engineered for the streamlined training of vast datasets, LGBM boasts unique features such as Radiant Boosting, where each tree rectifies errors from its predecessors, ensuring a powerful ensemble for predictive tasks. Its leaf-wise strategy prioritizes splits based on maximum gain, enhancing training efficiency. Notably, LGBM offers seamless support for categorical features, eliminating the need for preprocessing, which is advantageous for datasets like our CVD dataset. Widely applied across classification, regression, and ranking tasks, LGBM is renowned for its efficiency, scalability, and outstanding performance [36]. The Eq. 1 gives LGBM prediction for a single instance i is

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

$\hat{y_i}$- the predicted value for the i^{th} instance, f_k (x_i)- prediction made by the k^{th} tree for the i^{th} instance x_i, K - total number of trees.

D. OPTUNA HYPERPARAMETER TUNING

Optuna emerged from the collaborative efforts of Takuya Akiba and his team at Preferred Networks, an influential player in the field of artificial intelligence based in Japan. The project's inception in 2018 was a response to the pressing demand within the machine learning domain for streamlined and effective hyperparameter optimization techniques. Led by Takuya Akiba, the development team poured their expertise into crafting Optuna, a sophisticated open-source software solution. Since its initial release, Optuna has garnered widespread acclaim throughout the machine learning community, distinguishing itself for its intuitive interface, adaptable functionality, and cutting-edge

optimization algorithms. Optuna is an open-source hyperparameter optimization framework for ML and DL models. The search for ideal hyperparameters, or the variables that regulate the training process of machine learning and deep learning models, such as learning rate, number of layers, and dropout rate, may be automated with Optuna's adaptable and effective interface To effectively search the hyperparameter space and identify the set of hyperparameters that maximizes or minimizes a particular objective function, Optuna uses a variety of optimization algorithms, including evolutionary algorithms and Bayesian optimization. This allows users to quickly and effectively tune their models for improved performance without the need for manual trial and error [21].

E. ANOX^2 - OPT_HPLGBM

ANOVA is particularly useful for assessing the statistical significance of numerical features in relation to categorical target variables, such as the presence or absence of heart disease. Conversely, the Chi-Square test is ideal for evaluating the independence of categorical features. Together, these dual feature selection methods ensure that only the most relevant features are included in the model. ANOVA is particularly effective for evaluating the statistical significance of numerical features in relation to categorical target variables, such as the presence or absence of heart disease. In contrast, the Chi-Square test is well-suited for assessing the independence of categorical features. Combining these dual feature selection methods ensures that only the most relevant features are included in the model. LGBM, a gradient-boosting framework known for its efficiency, excels in handling both numerical and categorical data through its histogram-based algorithm. Its innovative leaf-wise growth strategy, which focuses on growing the leaf with the maximum delta loss, allows LGBM to achieve higher accuracy with fewer iterations compared to the level-wise growth strategy used by algorithms like XGB. LGBM's efficiency and speed are significantly enhanced by its leaf-wise growth strategy and histogram-based algorithm. Optuna's hyperparameter optimization further refines LGBM's performance by adjusting key parameters like learning rate, number of leaves, and feature fraction. This fine-tuning is crucial for maximizing the model's accuracy and computational efficiency, leading to precise predictions of heart disease. The integration of effective feature selection, advanced boosting techniques, and optimized hyperparameters ensures that the model delivers both accuracy and efficiency, which is essential for real-world medical applications where rapid and reliable predictions are critical.

VI. EXPERIMENTAL SETUP

A. SETUP

The proposed work was executed using Jupyter Notebook, and the system configuration is provided in Table 2.

B. DATA COLLECTION

In our work, we leverage four distinct types of heart datasets. These include the Cleveland dataset, the Statlog dataset, the heart dataset (a comprehensive compilation of Cleveland, Statlog, Hungarian, and Long Beach VA datasets), and the Alizadeh dataset. The research work incorporated four distinct datasets, each with its unique characteristics.

TABLE 2. System configuration.

Name	Configuration
OS	Windows 11
processor	11th-generation Intel i3 processor
Ram	64GB
Hard Disk	1TB
Python	3.11.5 (64 bit)
Libraries	matplotlib (3.7.2) pandas (2.0.3) NumPy (1.24.3), scikit-learn (1.3.0) seaborn (0.12.2) plotly (5.9.0) Optuna (3.6.0).

1) CLEVELAND DATASET

The Cleveland Dataset is renowned in machine learning research and encompasses 76 attributes. It comprises 303 instances with 14 features. In this dataset, the target field represents the existence of cardiac disease as an integer value between 0 and 4. Heart disease is not present with a score of 0, and its severity is represented by the values 1, 2, 3, and 4.

2) STATLOG DATASET

The UCI Repository Statlog Heart Disease has 270 subjects in the original dataset, with 13 attributes and one output class. In particular, cardiac disease is classified as present in 120 participants (positive class) and absent in 150 subjects (negative class) in this study. Notably, dataset I contains no missing values. The Statlog dataset shares identical data descriptions with the Cleveland dataset with all 14 features.

3) HEART DISEASE DATASET

The heart disease dataset consolidates four prominent datasets: Cleveland, Hungarian, Switzerland, and Long Beach VA. During preprocessing, features such as 'ca' and 'thal' were removed, resulting in a combined dataset containing 11 features and 1 target variable. With a total of 1190 instances, this amalgamated dataset is intended to advance the development of CAD related machine learning and data mining algorithms.

4) Z ALIZADEH DATASET

There are 303 patients' records in the Z-Alizadeh Sani dataset, and each patient has 54 features. The ECG, symptom and examination, demographic, laboratory, and echo aspects comprise the four groups into which these features

are divided. Patients are divided into two groups: Normal and CAD. If a patient's diameter narrowing is 50% or more, they are classified as having CAD; if not, they are classified as normal. Table 3 presents the positive and negative target values along with information regarding missing values.

VII. EXPERIMENTAL RESULT AND DISCUSSION

A. DATA PREPROCESSING FOR THE PROPOSED MODEL

For the Cleveland dataset, missing values are encountered in categorical features, and they are filled using the fillna method. No duplicate values are found. Outliers are addressed using the z-score method, resulting in a fully preprocessed dataset. In the Statlog dataset, there are no missing or duplicate values. Additionally, no high correlation between features is observed. Outliers are managed using the z-score method. The heart disease dataset is comprehensive and does not contain missing values, but duplicate values are present and are consequently dropped. Similarly, outliers are managed using the z-score transformation method. There are no highly correlated features in this dataset. The three datasets were tested for feature correlation, but no highly correlated features were found. In the Z-Alizadeh Sani dataset, although there are no missing or duplicate values, certain features, such as 'lymph' (a type of white blood cell produced in the bone marrow) and 'neut' (another type of white blood cell), show a high degree of correlation, as highlighted by the red dot in the heatmap in Figure 2. Since both features represent types of white blood cells, they were found to be highly correlated. Since highly correlated features provide redundant information to the model, including both can introduce unnecessary complexity without enhancing predictive power. Therefore, the 'lymph' feature has been removed. Additionally, the dataset has been identified as imbalanced. Prompting the implementation of the SMOTE method to address this imbalance in the target classes. Outliers are handled using the z-score method. The Z-score in this method calculates the number of standard deviations an observation deviates from the dataset mean standard deviations. Potential outliers are observations whose z-scores are greater than a

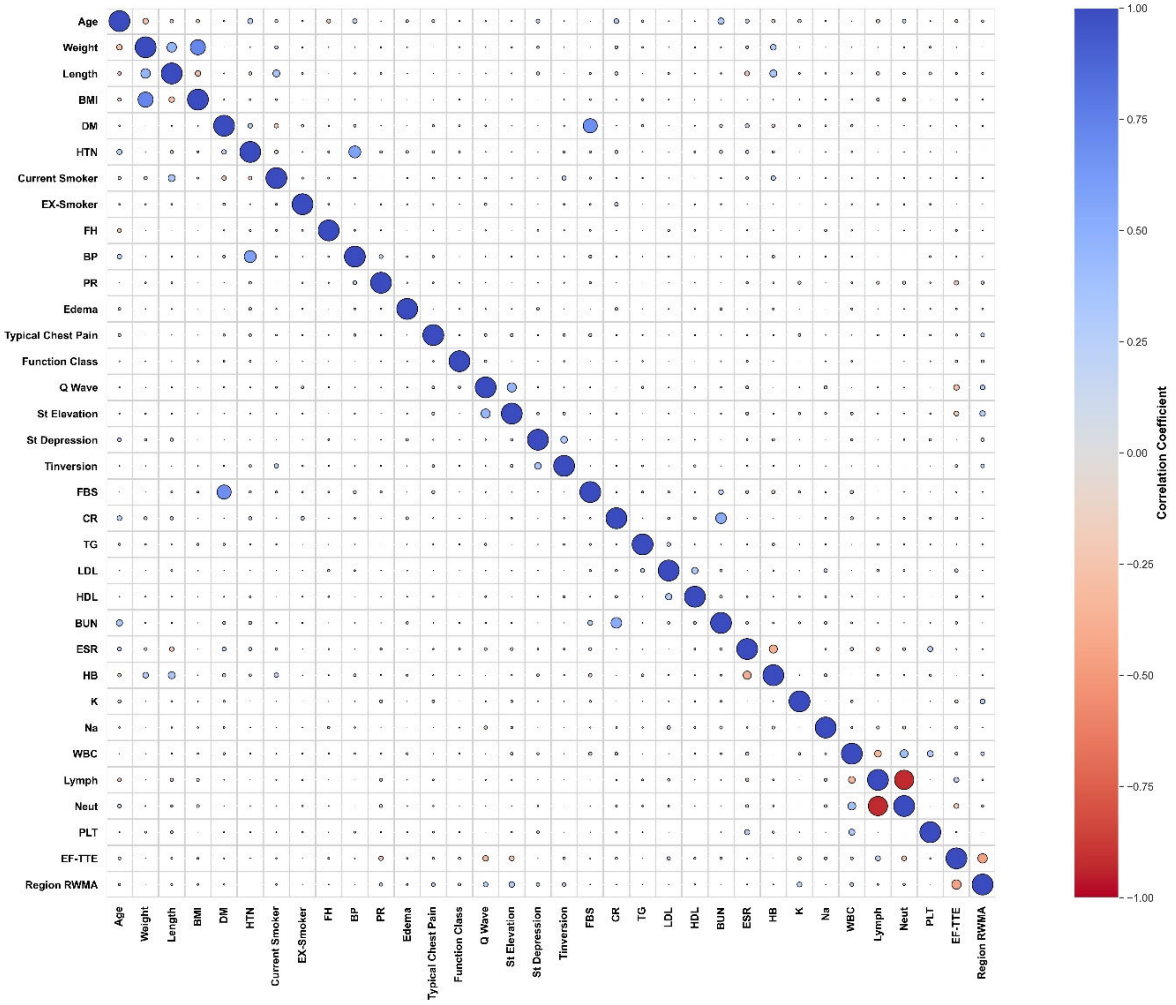

FIGURE 2. Heat map of Z-Alizadeh Sani.

TABLE 3. Dataset description of all four datasets.

Dataset Reference	Dataset	No of features no. of instance	Missing value	No. of positive class (1)	No. of negative class (0)
[37]	Cleveland	(14,303)	Yes	139	164
[38]	Statlog	(14,270)	No	120	150
[39]	Heart	(11,1190)	No	629	591
[40]	Z-Alizadeh Sani	(56,303)	No	216	87

TABLE 4. F-Statistics and p-value of a) Cleveland b) Statlog and c) heart dataset.

Feature of Cleveland	F- Statistic	P-Value	Accepted /Rejected	Feature of Statlog	F- Statistic	P-Value	Accepted /Rejected	Feature of heart	F- Statistic	P-Value	Accepted/Rejected
ca	39.3166	2.30976e-e9	A	ca	50.9997	1.616627e-11	A	slope	406.806	5.811e-75	A
thal	38.1443	3.80844e-e9	A	cp	42.5338	5.41154e-10	A	exang	299.357	3.89488e-58	A
cp	35.1883	1.36161e-e8	A	oldpeak	36.8033	6.29796e-09	A	oldpeak	212.89	1.89801e-43	A
exang	34.3573	1.95476e-e8	A	thal	36.4457	7.35662e-09	A	thalach	169.8	1.09281e-35	A
oldpeak	29.6868	1.53633e-07	A	exang	36.3054	7.8194e-09	A	cp	158.663	1.24958e-33	A
thalach	28.4116	2.72198e-07	A	thalach	29.028	1.96688e-07	A	sex	92.9604	5.2742e-21	A
sex	21.0659	7.98129e-06	A	sex	24.414	1.84036e-06	A	age	76.6704	9.83656e-18	A
slope	14.5004	0.00018817	A	slope	12.7296	0.000449049	A	fbs	62.8852	6.459e-15	A
age	8.9969	0.0036a86	A	restecg	7.56117	0.0065023	A	chol	45.309	2.9901e-11	A
restecg	6.15248	0.0139788	A	age	7.49496	0.0067367	A	trestbps	12.7714	0.000370621	A
trestbps	0.81435	0.367961	R	chol	2.00138	0.158688	R				
chol	0.42264	0.516394	R	trestbps	0.283905	0.594736	R	restecg	11.2111	0.000846759	A
				fbs	0.013372	0.908052	R				

*A-Accepted, R-Rejected

TABLE 5. F-Statistics and p-value of Z-Alizadeh Sani dataset.

Feature of Z-Alizadeh Sani	F- Statistic	P-Value	Accepted /Rejected	Feature of Z-Alizadeh Sani	F- Statistic	P-Value	Accepted /Rejected
Typical chest pain	125.837	1.24157e-24	A	Na	2.8156	0.0943897	R
Atypical	62.9622	4.1877e-14	A	Airway disease	2.14814	0.143787	R
Age	44.0354	1.49903e-10	A	BUN	2.07466	0.150803	R
eF-TTe	37.744	2.55043e-09	A	Current Smoker	1.63507	0.201988	R
Region RWMA	33.3935	1.87921e-08	A	CR	1.39577	0.238366	R
HTN	27.175	3.45946e-07	A	Sex	1.35896	0.244641	R
Nonanginal	24.4676	1.25883e-06	A	VHD	1.30412	0.254369	R
DM	20.5664	8.32655e-06	A	PLT	1.23608	0.267115	R
FBS	18.8664	1.91856e-05	A	HDL	1.00653	0.316541	R
Tinversion	17.9023	3.09041e-05	A	Edema	0.882549	0.348258	R
BP	16.3188	6.80062e-05	A	LVH	0.791015	0.374504	R
ESR	12.4008	0.000495634	A	WBC	0.739521	0.390499	R
TG	10.5126	0.00131879	A	Thyroid Disease	0.697351	0.404338	R
K	8.96213	0.00298489	A	Lung rales	0.61563	0.433293	R
Q Wave	6.91406	0.00899146	A	BBB	0.568876	0.451295	R
Diastolic Murmur	6.6274	0.0105201	A	BMI	0.56289	0.453684	R
ST Depression	6.41226	0.0118418	A	HB	0.542592	0.461934	R
PR	6.30663	0.0125524	A	LDL	0.456528	0.49977	R
ST Elevation	5.9899	0.0149593	A	FH	0.382077	0.536961	R
Dyspnea	4.79417	0.0293237	A	Ex- Smoker	0.381491	0.537274	R
Poor R Progression	3.75764	0.0535014	R	Obesity	0.151935	0.696969	R
Lymph	3.31846	0.0694984	R	DLP	0.0486959	0.825498	R
Function class	2.86421	0.0916051	R	Systolic Murmur	0.00709932	0.932908	R

predetermined threshold, which is typically set at ±2 or ±3. The fully preprocessed dataset serves as input for the subsequent feature selection phase utilizing the proposed AnoX^2- Opt_hpLGBM method.

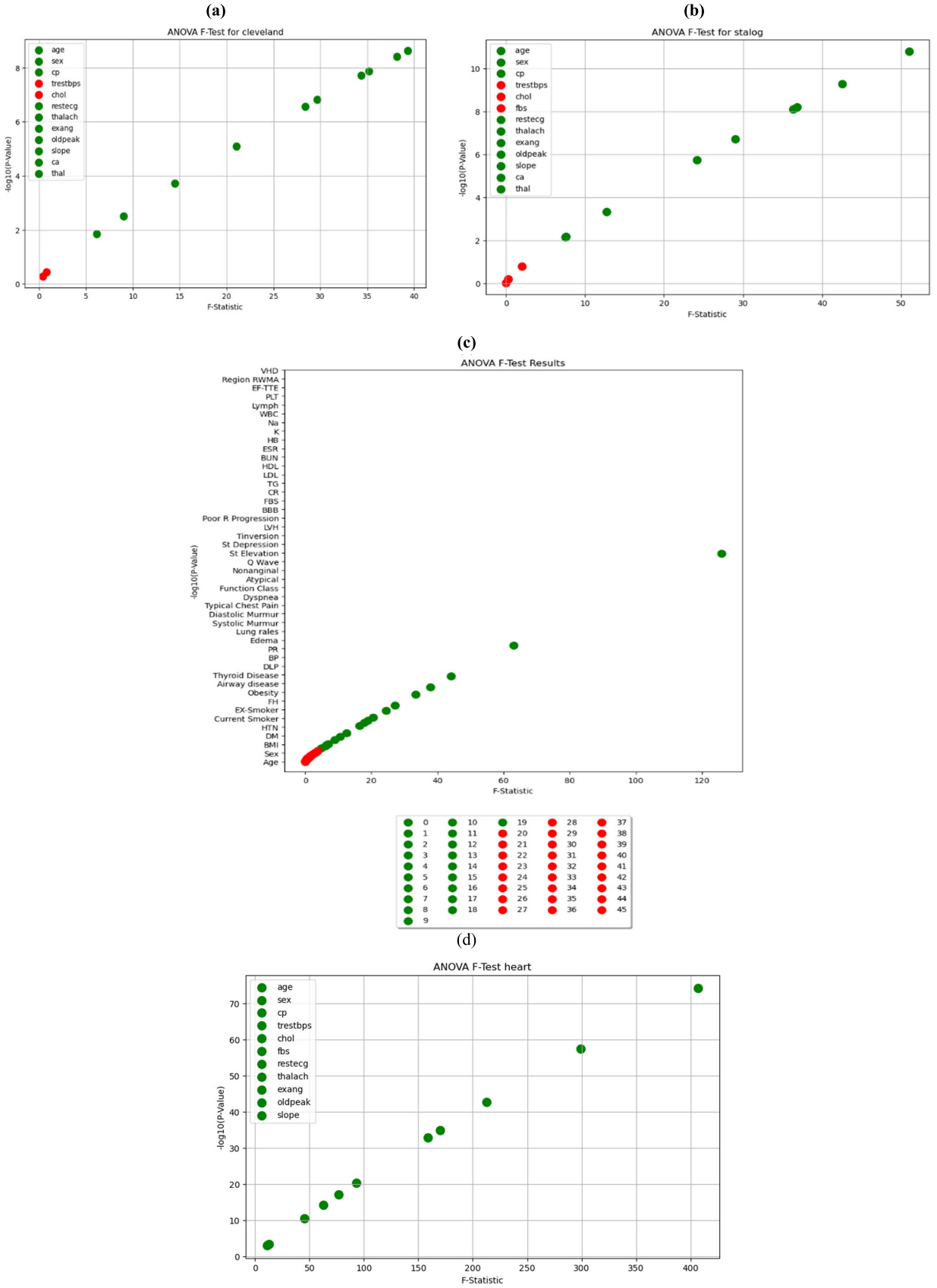

FIGURE 3. Feature selected by ANOVA (Tier 1) a) Cleveland b) Statlog c) Z-Alizadeh Sani d) heart disease.

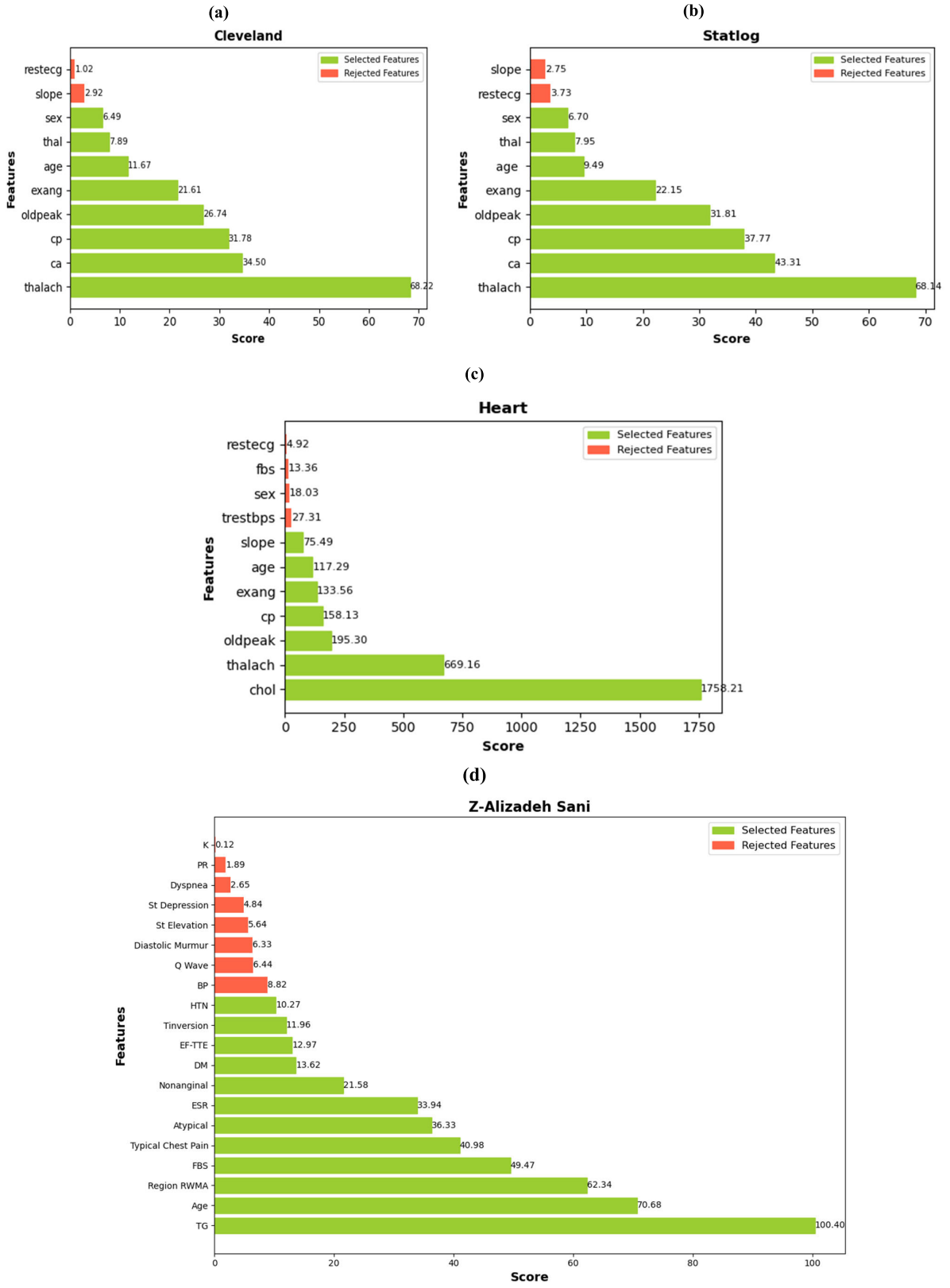

FIGURE 4. Feature selected by X^2 (Tier 2) a) Cleveland b) Statlog c) heart disease d) Z-Alizadeh Sani.

TABLE 6. Optuna tuned light gradient boosting machine hyperparameters.

Dataset/parameter	Cleveland	Statlog	Heart Disease	Z-Alizadeh Sani	CKD	Pima
n_estimators	93	124	150	109	156	50
max_depth	4	10	10	4	6	4
learning_rate	0.0920	0.0067	0.0026	0.0687	0.0355	0.0108
mn_child_samples	33	7	40	56	10	8
subsample	0.3012	0.1488	0.4577	0.3453	0.3586	0.3702
colsample_bytree	0.5869	0.9178	0.2172	0.3975	0.2557	0.3412

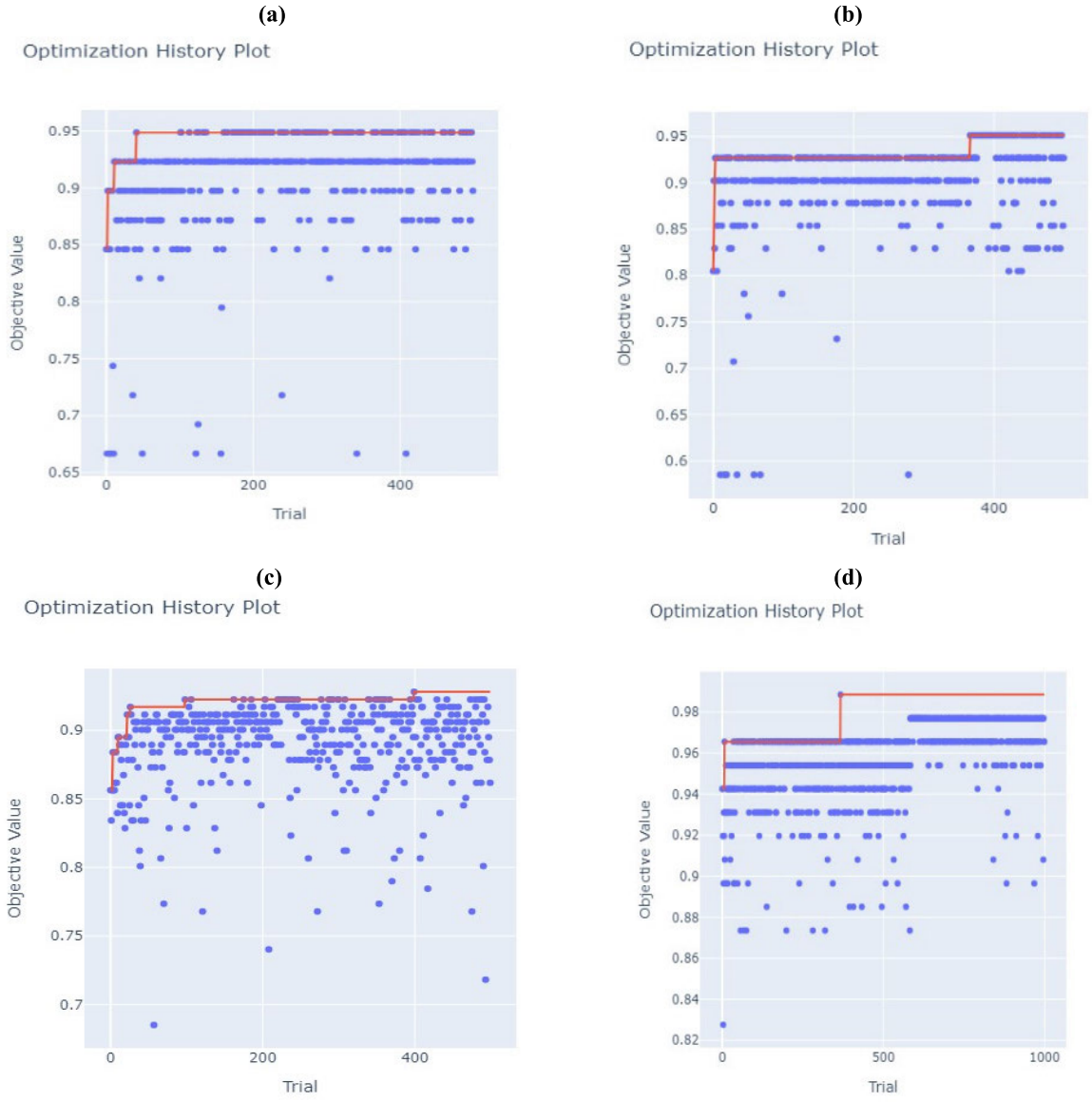

FIGURE 5. Optimization history comparison for the proposed model a) Cleveland b) Statlog c) heart disease d) Z-Alizadeh Sani.

B. DUAL-TIER FEATURE SELECTION

For each of the four datasets, the F-statistics and associated p-values are shown in Table 4 and Table 5. Here, we set the significance level (α) to 0.05. If the p-value is less than or equal to the threshold, the feature is rejected; otherwise, it is accepted. In Fig. 3, the red dots represent features

TABLE 7. Proposed model evaluation for cleveland and statlog datasets.

Dataset	Cleveland					Stalog				
Model	RF	Ada	XGB	Cat	**LGBM**	RF	Ada	XGB	Cat	**LGBM**
Training Accuracy (%)	83.97	87.5	88.46	93.58	**93.58**	85.36	84.75	88.41	100	**86.58**
Testing Accuracy (%)	89.74	92.3	92.3	87.17	**94.87**	90.24	92.68	92.68	90.24	**95.12**
Precision (%)	84.61	91.66	85.71	78.57	**86.66**	88.46	92	92	91.66	**95.83**
Recall (%)	84.61	84.61	92.3	84.61	**100**	95.83	95.83	95.83	91.66	**95.83**
F1 (%)	84.61	84.61	88.88	81.48	**92.85**	92	93.87	93.87	91.66	**95.83**
Specificity (%)	92.3	84.61	92.3	88.46	**92.3**	82.35	88.23	88.23	88.23	**94.11**
Log Loss (%)	3.69	3.69	2.77	4.62	**1.84**	3.51	2.63	2.63	3.51	**1.75**

TABLE 8. Proposed model evaluation for Heart and z- Alizadeh CVD datasets.

Dataset	Heart					Z-Alizadeh Sani				
Model	RF	Ada	XGB	Cat	**LGBM**	RF	Ada	XGB	Cat	**LGBM**
Training Accuracy (%)	85.61	85.2	84.23	88.1	**82.15**	92.46	92.75	94.49	97.68	**94.2**
Testing Accuracy (%)	87.29	86.18	92.81	85.63	**92.81**	93.1	95.4	97.7	96.55	**98.85**
Precision (%)	93.06	93.81	92.17	92	**92.17**	93.18	93.47	97.72	95.55	**97.77**
Recall (%)	85.45	82.72	96.36	83.63	**96.36**	93.18	97.72	97.72	97.72	**100**
F1 (%)	89.09	87.92	94.22	87.61	**94.22**	93.18	95.55	97.72	96.62	**98.87**
Specificity (%)	90.14	91.54	87.32	88.73	**87.32**	93.02	93.02	97.67	95.34	**97.67**
Log Loss (%)	4.58	4.97	2.58	5.17	**2.58**	2.48	1.65	0.82	1.24	**0.41**

rejected by the ANOVA method, based on metrics like the F-value and p-test, while the green dots signify accepted features. Among these datasets, the Cleveland dataset rejects 2 features (trestbps, chol), Statlog rejects 3 features (chol, trestbps, fbs), and the heart disease dataset rejects no features. The Z-Alizadeh Sani dataset rejects 23 features (Sex, BMI, Current Smoker, EX-Smoker, FH, Obesity, Airway disease, 'Thyroid Disease, DLP, Edema, Lung rales, Systolic Murmur, Function Class, LVH, Poor R Progression, BBB, CR, LDL, HDL, BUN, HB, Na, WBC, Lymph, PLT, VHD). As a result, the selected green features are utilized as input for the next stage of feature selection. In Fig. 4, the selected features identified by chi-square are depicted. Green bars represent the selected features, while red bars indicate the rejected features. Making sure that all relevant features are taken into account throughout the decision-making process, improves the overall performance of machine learning models. From the proposed method it is understood that resting systolic blood pressure (trestbps), Fasting blood sugar (fbs), and resting electrocardiographic findings (restecg) were deemed irrelevant and rejected in the Cleveland, Stalog, and Heart datasets.

C. PROPOSED ANOX^2- OPT_HPLGBM FOR CVD DATASET

The proposed model was constructed based on the selected features obtained from the dual-tier feature selection process. Each of the four datasets underwent feature selection using the AnoX^2 methodology, followed by optimization with Optuna across five different classifiers: Random Forest, XGBoost, LGBM, AdaBoost, and CatBoost. In Fig. 5 the y-axis represents the Objective Value, which corresponds to the accuracy of the model. The plot suggests that conducting 500 trials might have been excessive for the relatively straightforward hyperparameters of the FM (Factorization Machines) model. Table 6 presents the optimal hyperparameters identified via the Optuna optimization process to attain peak performance for the Light Gradient Boosting Machine model. Figure 5 depicts the results of hyperparameter tuning using FM models belonging to a class of machine-learning models adept at handling sparse data and high-dimensional feature spaces. They excel in capturing interactions between features, particularly in scenarios with numerous features but limited training data. The pronounced angle of the best value line indicates that Optuna efficiently identified optimal hyperparameters in fewer than twenty trials. Further tuning beyond this threshold did not notably enhance model performance, indicating that Optuna rapidly approached near-optimal results. The blue dots are the objective values whereas the red line is the best value obtained by Optuna hyperparameter tuning.

Table 7 and Table 8 present a comparative analysis of each dataset with each model, showcasing their performance metrics. The evaluation includes training, testing accuracies, precision, recall f1 score, specificity, and log loss.

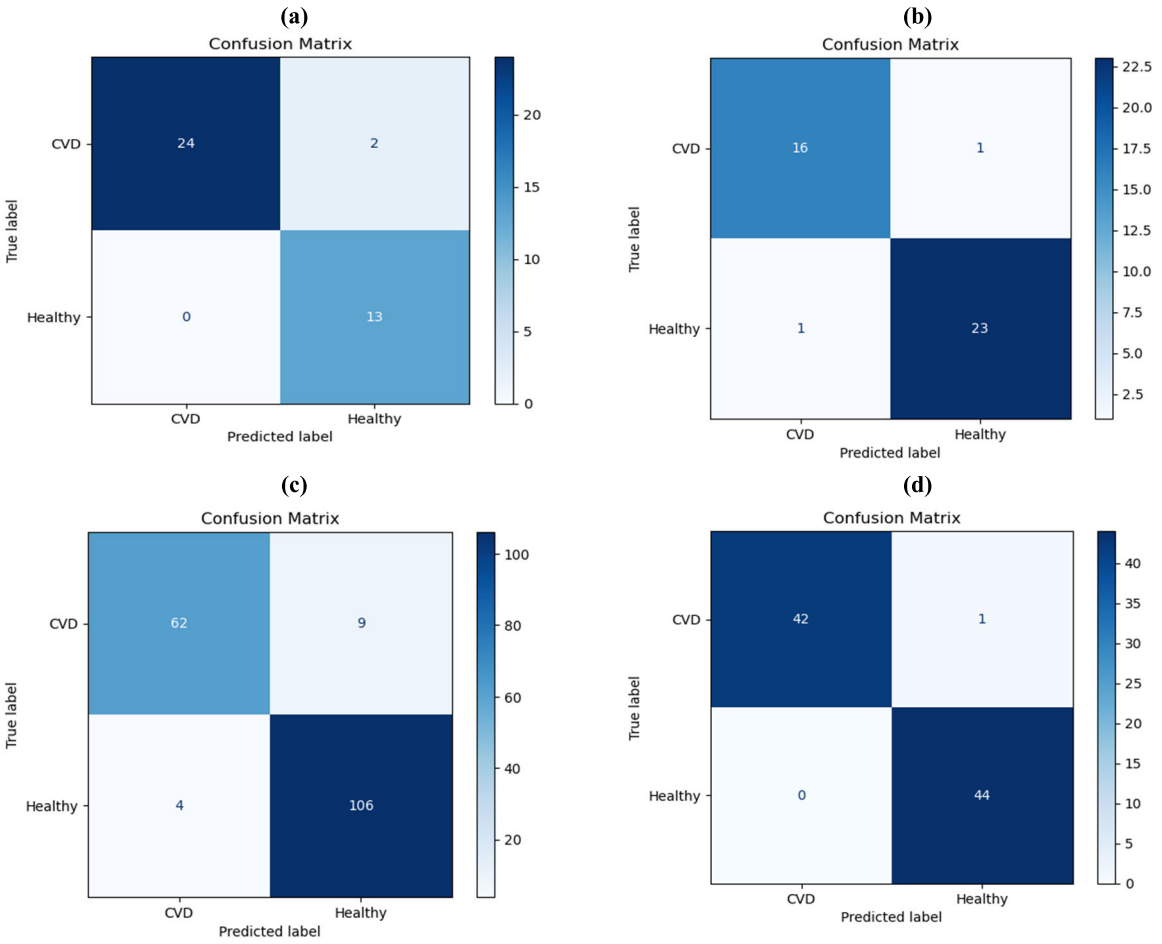

FIGURE 6. Confusion Matrix Comparison for the Proposed Model Across Different Datasets a) Cleveland b) Statlog c) heart disease d) Z-Alizadeh Sani Datasets.

TABLE 9. Proposed model evaluation for Kidney and Pima datasets.

Dataset	CKD					PIDD				
Model	RF	Ada	XGB	Cat	**LGBM**	RF	Ada	XGB	Cat	**LGBM**
Training Accuracy (%)	98.61	98.05	100	98.23	**96.25**	79.5	78.5	83.12	81.75	**82**
Testing Accuracy (%)	98.06	97.50	98.20	96.54	**98.75**	81.5	82	84.5	81.5	**85**
Precision (%)	97.63	96.49	98.23	94.23	**96.55**	82.69	83.49	96.25	84	**88.65**
Recall (%)	98.21	98.21	94.40	95.32	**100**	81.90	81.90	73.33	80	**81.90**
F1 (%)	97.92	97.34	98.89	95.75	**98.24**	82.29	82.69	83.24	81.95	**85.14**
Specificity (%)	97.92	96.89	98.44	96.25	**98.07**	81.05	82.10	96.84	83.15	**88.4**
Log Loss (%)	0.69	0.85	.49	0.78	**0.45**	81.47	6.48	55.58	6.66	**5.40**

Notably, Opt_hpLGBM consistently demonstrates promising results across all datasets. Specifically, the Cleveland dataset achieves the highest accuracy of 94.87%, followed by Statlog with 95.12%, heart disease with 92.81%, and Z- Alizadeh Sani with 98.85%. These accuracies surpass those obtained using existing methods, indicating the effectiveness of the proposed approach. Figure 6 shows the confusion matrices for the AnoX^2-Opt_hpLGBM model across all datasets. Both Type I and Type II errors are minimal, contributing to a higher overall accuracy. Additionally, the training accuracy, as well as the testing accuracy of the suggested model AnoX^2-Opt_hpLGBM was shown in Fig. 7. Each dataset attains more than 90% accuracy. In summary, Figure. 8 presents the code for plotting the ROC curve for multiple models, with the inclusion of the AUC score in the legend for each model's ROC curve. Among well-known ML techniques,

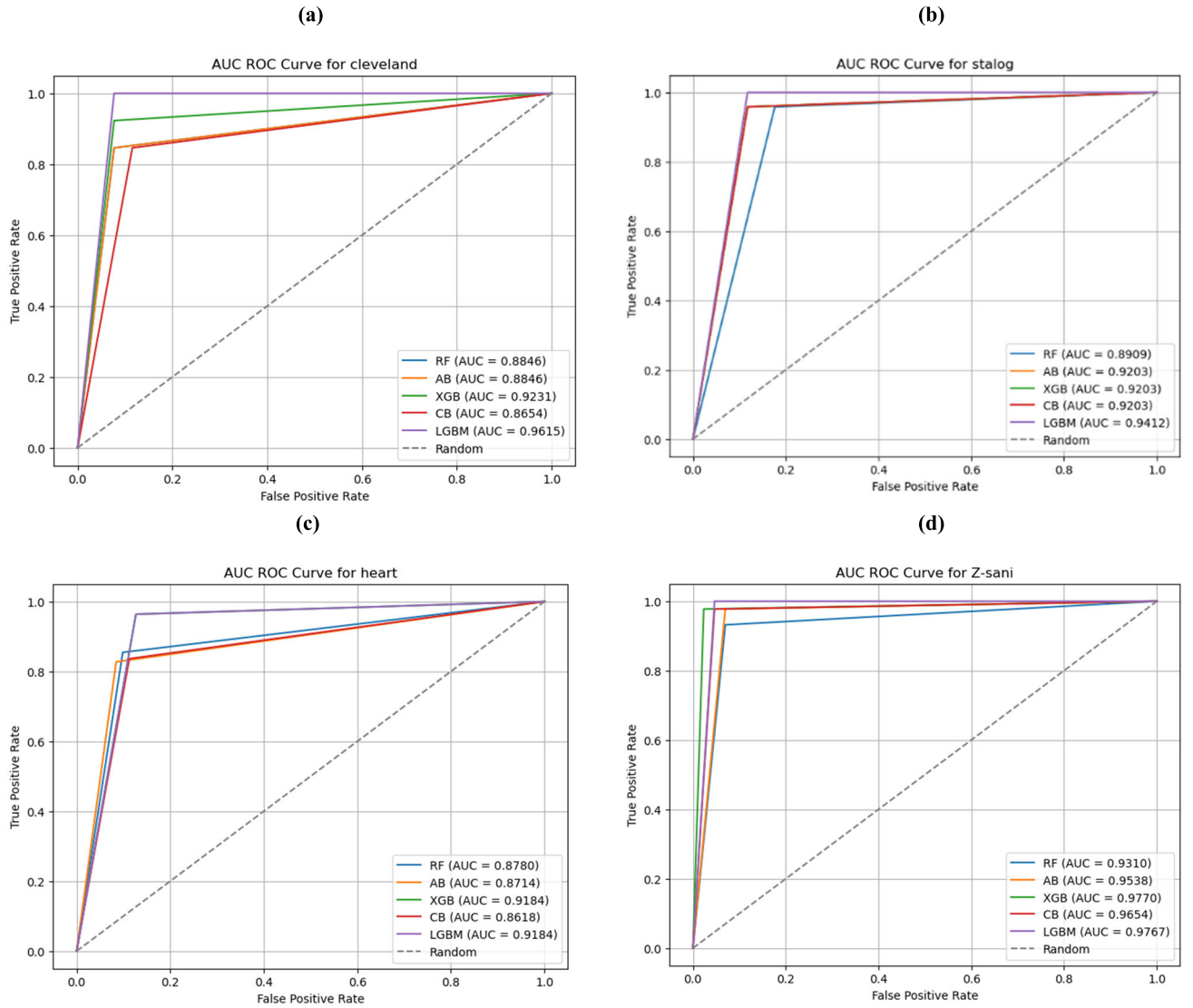

FIGURE 7. AUC ROC of the proposed model for) a) Cleveland b) Statlog c) heart disease d) Z-Alizadeh Sani.

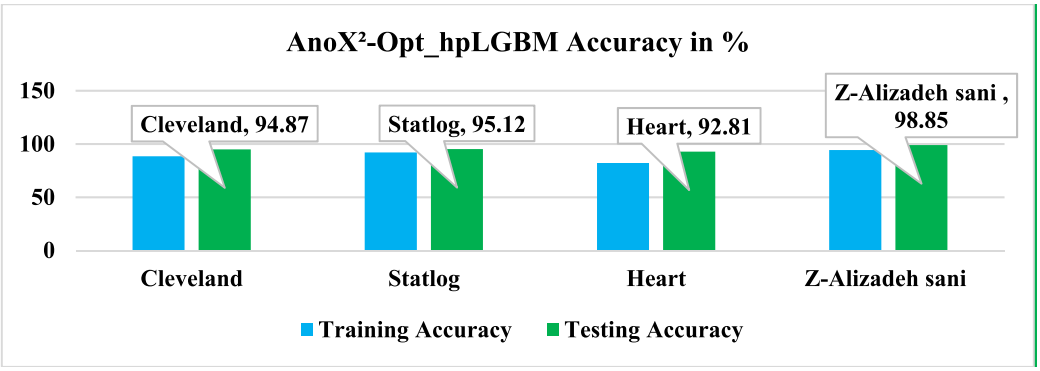

FIGURE 8. Accuracy of all 4 datasets of AnoX^2-Opt_hpLGBM.

including SVM, DT, RF, K-NN, XGB, and even some deep learning approaches, our proposed AnoX^2-Opt_hpLGBM model emerged as the best performer. Our detailed analysis, focused on the accuracy of the AUC metric, clearly

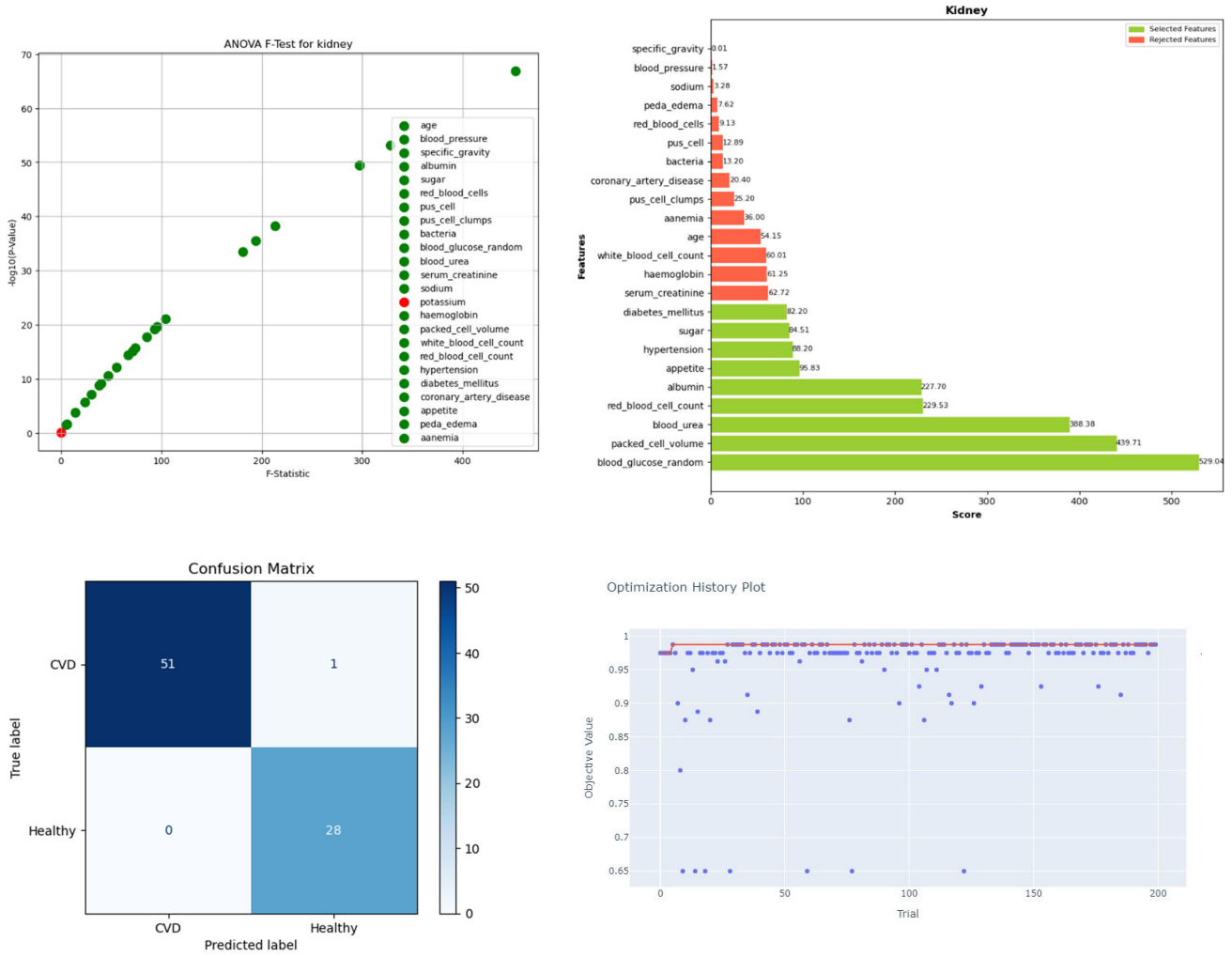

FIGURE 9. Dual tier feature selection with confusion matrix and optimization history of CKD dataset.

TABLE 10. Optuna tuned light gradient boosting machine hyperparameters for other disease.

Dataset/parameter	CKD	PIDD
n_estimators	156	50
max_depth	6	4
learning_rate	0.0355	0.0108
mn_child_samples	10	8
subsample	0.3586	0.3702
colsample_bytree	0.2557	0.3412

demonstrated that AnoX2-Opt_hpLGBM outperformed its competitors with accuracy rates of 96.15% for the Cleveland dataset and 94.12% for Stalog. Whereas the proposed model achieved 91.84% for the Heart dataset and 97.67% for Z Alizadeh Sani. The exceptional predictive ability of AnoX2-Opt_hpLGBM not only improved accuracy but also reduced the number of features and achieved lower log loss rates of 1.84% for Cleveland, 1.75% for Stalog, 2.58% for the Heart dataset, and 0.41% for Z Alizadeh, marking a significant achievement. The proposed model identifies 8 relevant features in the Cleveland and Statlog datasets, 7 relevant features in the Heart dataset, and 12 dominant features in the Z-Alizadeh Sani dataset, leading to improved accuracy and reduced overfitting. The proposed dual-tier algorithm has consistently identified ''thalach,'' representing the maximum heart rate achieved during exercise or physical activity, as the most significant feature in both the Cleveland and Statlog datasets. This feature is essential for assessing how well the heart functions under stress. ''Oldpeak'' (ST depression induced by exercise) was identified as the most significant feature in the Heart dataset. This measurement reflects the heart's electrical activity under stress compared to rest, which is critical in assessing the severity of reduced blood flow to the heart. In the Z-Alizadeh Sani dataset, Triglycerides (TG) have been identified as a significant feature. Elevated levels of TG are a marker of metabolic health and can indicate an increased risk of cardiovascular issues, making it a crucial factor in the dataset's analysis.

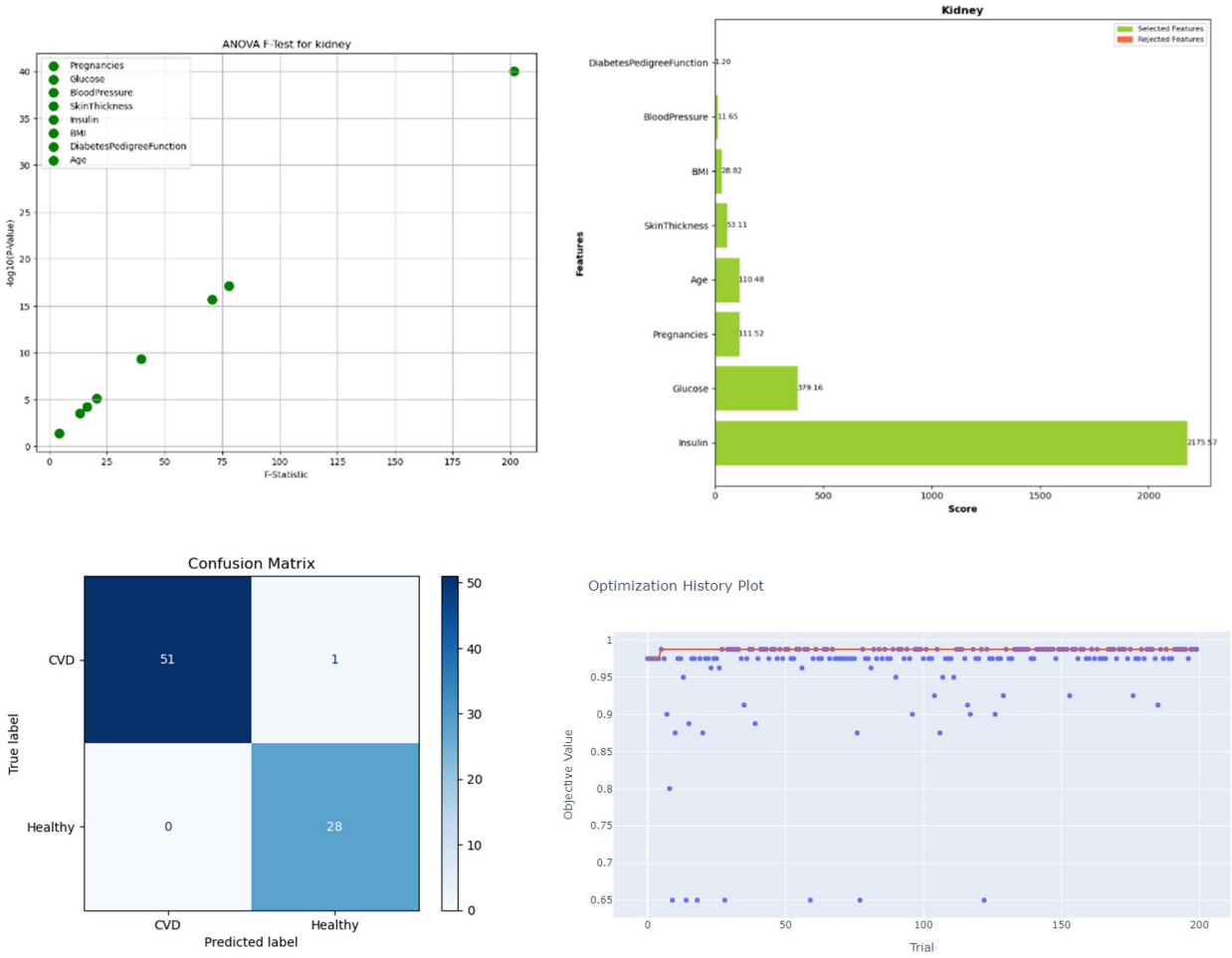

FIGURE 10. Dual tier feature selection with confusion matrix and optimization history of Pima dataset.

TABLE 11. Selected features using Dual tier feature selection for all dataset.

Dataset	Selected features	No. of selected features	Significant feature
Cleveland	Age, sex, cp, thalach, exang, oldpeak, ca, thali	8/14	Thal
Statlog	Age, sex, cp, thalach, exang, oldpeak, ca, thal	8/14	Thal
Heart	Age, cp, chol, thalach, exang, oldpeak, slope	7/12	Oldpeak
z-Alizadeh Sani	Typical Chest Pain, Atypical, ESR, Nonanginal, DM, EF-TTE, Tinversion, HTN, TG, Age, Region RWMA, FBS	12/56	TG (Triglyceridein)
Chronic Kidney Disease	albumin, sugar, blood_glucose_random, blood_urea, packed_cell_volume, red_blood_cell_count, hypertension, diabetes_mellitus, appetite,	9/26	blood_glucose_random
Pima	Pregnancies, Glucose, SkinThickness, Insulin, BMI, Age	6/9	Insulin

D. PROPOSED ANOX²-OPT_HPLGBM FOR OTHER DISEASE

The proposed model was also evaluated on two distinct datasets related to different diseases. The first dataset focused on CKD and comprised 400 instances with 24 features. The goal was to predict the presence or absence of CKD using binary classification [42]. Figure 9 indicated that the proposed model outperformed other models commonly used for kidney disease prediction, achieving an accuracy of 98.75% on the CKD dataset. Similarly, the second dataset pertained to

TABLE 12. Systematic evaluation and comparison of previous findings for all dataset.

Dataset	Work	Year	Techniques	Accuracy (%)
Cleveland	[44]	2020	NB	84.51
	[23]	2020	FAMD+RF	93.44
	[26]	2021	RF+DT	88.70
	[18]	2021	BO-SVM	93.3
	[28]	2021	DL	94.2
	[21]	2022	XGBoost	94.7
	[35]	2022	X^2 – SVM	89.47
	[11]	2023	Artificial rabbit optimizer	93.22
	[45]	2023	LoFS-ANN	90.50
		Proposed	AnoX^2 - Opt_hpLGBM	94.87
Stalog	[46]	2018	PSO+NB	87.91
	[11]	2023	Artificial rabbit optimizer	94.05
		Proposed	AnoX^2 - Opt_hpLGBM	95.12
Heart	[47]	2022	Stacked ensemble classifier with ExtraTrees	92.34
	[48]	2022	Stacking	89.86
		Proposed	AnoX^2 - Opt_hpLGBM	92.81
Z-Alizadeh Sani	[49]	2020	Heterogeneous hybrid feature selection algorithm + SMOTE + XGB	92.58
	[50]	2020	Random trees	91.47
	[51]	2021	Hybrid PSO	84.25
	[52]	2022	XGB + Feature construction + SMOTE	94.70
	[53]	2022	Fixed analysis of mixed data + Binary Bat Algorithm	97.37
	[54]	2023	Mixed SMOTE-NORM-CNN	98.57
		Proposed	AnoX^2 - Opt_hpLGBM	98.85
Chronic Kidney Disease	[55]	2017	AdaBoost+CART	66.2
	[56]	2017	SVM+ FilterSubsetEval with Best First	98.5
	[57]	2019	Density-based feature selection with Ant colony optimization (D-ACO)	95.0
	[58]	2020	Sparse Autoencoder (SAE)+SoftMax Regression	98
	[59]	2022	LR model + Chi-Square feature selection	97.54
		Proposed	AnoX^2 - Opt_hpLGBM	98.75
Pima Indians Diabetes Database	[60]	2015	Genetic Algorithm naïve Bayes (GA_NBs)	78.69
	[61]	2018	Naive bayes	76.30
	[62]	2021	kNN with Standard Deviation (SDKNN)	83.76
	[63]	2021	Ensemble soft voting	79.08
	[64]	2021	kNN	78.68
	[31]	2022	LGBM+kNN+AdaBoost	90.76
	[65]	2023	Naïve Bayes	79.13
	[66]	2023	RBF and RBF city block kernel	85.5
		Proposed	AnoX^2 - Opt_hpLGBM	85.0

diabetes prediction, obtained from Kaggle, the Pima Indians Diabetes dataset [43]. This dataset includes 768 instances with 9 features, where the target variable, labeled "Outcome," predicts whether a patient is healthy or diabetic. Figure 10 illustrates the dual-tier feature selection process for the diabetes dataset, along with its confusion matrix displaying Type I and Type II errors.

The Optuna hyperparameter tuning was completed early, within 50 epochs. Table 9 presents the performance of the kidney and diabetes datasets using the five proposed classifiers tuned with Optuna. The proposed model performed moderately well on this dataset, achieving an accuracy of 85% while utilizing only 6 relevant features. A detailed comparison of the model's performance on both datasets is presented in Table 9. Table 10 gives the hyperparameter tuning for CKD and PIDD datasets.

Table 11 presents the selected features from all datasets after undergoing the dual-tier AnoX^2 process. It also

highlights the most significant feature identified by AnoX^2-Opt_hpLGBM. The dual-tier approach simplifies the process by selecting only the most relevant features, ranging from 6 to 12, which leads to better results compared to existing methods. Table 12 provides a comprehensive comparison of various methodologies applied to all six datasets utilized in our study. These findings are methodically arranged according to publication year, spanning from 2015 to 2024. It is evident that our proposed model, AnoX^2-Opt_hpLGBM, demonstrates outstanding predictive performance in forecasting cardiovascular disease compared to other methodologies, with results separately detailed for each of the four heart experimented and two other datasets.

VIII. CONCLUSION AND FUTURE WORK

Our proposed methodology tackles the challenge of predicting CVD by integrating a dual-tier feature selection approach, ANoX^2, across four distinct datasets: Cleveland, Heart, Statlog, and Z-Alizadeh Sani. This approach combines the major advantages of ANOVA, which assesses variation between different groups, and Chi-square, which identifies significant associations between categorical variables, enhancing algorithm performance. Additionally, we leverage LGBM for its speed and efficiency, coupled with regularization techniques like max_depth and min_child_samples to mitigate overfitting. The integration of these features in our AnoX^2-Opt_hpLGBM model results in exceptional performance across all metrics. Our proposed method customizes feature selection, resulting in each dataset selecting an optimal number of features: Cleveland and Statlog datasets opt for 8 features each, the Heart dataset selects 7 features, and the Z-Alizadeh Sani dataset utilizes 12 features. Through extensive experimentation with machine learning algorithms including Random Forest, XGBoost, AdaBoost, CatBoost, and LGBM, and employing hyperparameter tuning via Optuna, we achieve significant accuracy improvements. Specifically, our model achieves notable accuracy rates of 94.87% for Cleveland, 92.81% for Heart, 95.12% for Statlog, and an impressive 98.85% for the Z-Alizadeh Sani dataset. The performance of the proposed work was also tested on other diseases, such as Chronic Kidney Disease (CKD), achieving an accuracy of 98.65%, and on a diabetes dataset, with an accuracy of 85%. These results demonstrate an improvement over existing methods. The Pima dataset's moderate performance with LGBM can be attributed to the model's sensitivity towards the treatment of missing values. The primary drawback of LightGBM is its sensitivity to the preprocessing of categorical features and hyperparameter tuning. However, these limitations can be effectively addressed through advanced preprocessing techniques and an efficient dual-tier feature selection method, which optimally handles categorical features. Additionally, the use of Optuna, an automated hyperparameter tuning framework, mitigates the sensitivity to hyperparameter settings, ensuring improved performance. This performance surpasses existing methodologies, particularly in its ability to generalize across multiple datasets for the same disease. Furthermore, our model demonstrates enhanced AUC ROC values, underscoring its robustness and reliability in CVD prediction. By offering a non-invasive means of predicting CVD, our model eliminates the need for invasive angiography procedures, thereby reducing patient risks, time, and costs. Our experimental validation on benchmark datasets highlights the superiority of the AnoX^2-Opt_hpLGBM technique over recent methods, showcasing its potential applicability in diverse disease domains. Looking ahead, our methodology could be further explored within deep learning frameworks to assess its efficacy and adaptability across various disease datasets.

REFERENCES

[1] Accessed: Jan. 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain, T. Dawson, P. Fergus, and M. Al-Jumaily, "Predicting the likelihood of heart failure with a multi level risk assessment using decision tree," in *Proc. 3rd Int. Conf. Technol. Adv. Electr., Electron. Comput. Eng. (TAEECE)*, Apr. 2015, pp. 101–106.

[3] G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. H. Akhter, "Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection," *IEEE Access*, vol. 10, pp. 23808–23828, 2022, doi: 10.1109/ACCESS.2022.3153047.

[4] S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, and A. Ghani, "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," *Knowl. Inf. Syst.*, vol. 58, no. 1, pp. 139–167, Jan. 2019, doi: 10.1007/s10115-018-1185-y.

[5] K. Yongcharoenchaiyasit, S. Arwatchananukul, P. Temdee, and R. Prasad, "Gradient boosting based model for elderly heart failure, aortic stenosis, and dementia classification," *IEEE Access*, vol. 11, pp. 48677–48696, 2023, doi: 10.1109/ACCESS.2023.3276468.

[6] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133034–133050, 2020.

[7] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection," *IEEE Access*, vol. 7, pp. 180235–180243, 2019.

[8] D. Deepika and N. Balaji, "Effective heart disease prediction using novel MLP-EBMDA approach," *Biomed. Signal Process. Control*, vol. 72, Feb. 2022, Art. no. 103318.

[9] G. Ambrish, B. Ganesh, A. Ganesh, C. Srinivas, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," *Global Transitions Proc.*, vol. 3, no. 1, pp. 127–130, Jun. 2022.

[10] P. Ramprakash, R. Sarumathi, R. Mowriya, and S. Nithyavishnupriya, "Heart disease prediction using deep neural network," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Coimbatore, India, Feb. 2020, pp. 666–670.

[11] L. A. Alharbi, "Artificial rabbits optimizer with machine learning based emergency department monitoring and medical data classification at KSA hospitals," *IEEE Access*, vol. 11, pp. 59133–59141, 2023, doi: 10.1109/ACCESS.2023.3284390.

[12] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "Improving the heart disease detection and patients' survival using supervised infinite feature selection and improved weighted random forest," *IEEE Access*, vol. 10, pp. 67363–67372, 2022.

[13] M. A. Khan and F. Algarni, "A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS," *IEEE Access*, vol. 8, pp. 122259–122269, 2020.

[14] T. O. Omotehinwa, D. O. Oyewola, and E. G. Dada, "A light gradient-boosting machine algorithm with tree-structured Parzen estimator for breast cancer diagnosis," *Healthcare Anal.*, vol. 4, Dec. 2023, Art. no. 100218, doi: 10.1016/j.health.2023.100218.

[15] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)," *Diagnostics*, vol. 11, no. 9, p. 1714, Sep. 2021, doi: 10.3390/diagnostics11091714.

[16] S. P. Barfungpa, H. K. Deva Sarma, and L. Samantaray, "An intelligent heart disease prediction system using hybrid deep dense Aquila network," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104742.

[17] S. P. Barfungpa, L. Samantaray, H. Kumar Deva Sarma, R. Panda, and A. Abraham, "D-t-SNE: Predicting heart disease based on hyper parameter tuned MLP," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105129.

[18] S. P. Patro, G. S. Nayak, and N. Padhy, "Heart disease prediction by using novel optimization algorithm: A supervised learning prospective," *Informat. Med. Unlocked*, vol. 26, Jan. 2021, Art. no. 100696, doi: 10.1016/j.imu.2021.100696.

[19] H. Yang, Z. Chen, H. Yang, and M. Tian, "Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison," *IEEE Access*, vol. 11, pp. 23366–23380, 2023, doi: 10.1109/ACCESS.2023.3253885.

[20] A. Akella and S. Akella, "Machine learning algorithms for predicting coronary artery disease: Efforts toward an open source solution," *Future Sci. OA*, vol. 7, no. 6, Jul. 2021, Art. no. FSO698, doi: 10.2144/fsoa-2020-0206.

[21] P. Srinivas and R. Katarya, "HyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost," *Biomed. Signal Process. Control*, vol. 73, Mar. 2022, Art. no. 103456.

[22] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network," *IEEE Access*, vol. 7, pp. 34938–34945, 2019, doi: 10.1109/ACCESS.2019.2904800.

[23] A. Gupta, R. Kumar, H. Singh Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," *IEEE Access*, vol. 8, pp. 14659–14674, 2020.

[24] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[25] L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," *IEEE Access*, vol. 7, pp. 54007–54014, 2019.

[26] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Jan. 2021, pp. 1329–1333, doi: 10.1109/ICICT50816.2021.9358597.

[27] A. A. Almazroi, E. A. Aldhahri, S. Bashir, and S. Ashfaq, "A clinical decision support system for heart disease prediction using deep learning," *IEEE Access*, vol. 11, pp. 61646–61659, 2023.

[28] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, Jan. 2021, Art. no. 8387680, doi: 10.1155/2021/8387680.

[29] M. Chakraborty, "Rule extraction from convolutional neural networks for heart disease prediction," *Biomed. Eng. Lett.*, vol. 14, no. 4, pp. 649–661, Jul. 2024, doi: 10.1007/s13534-024-00358-3.

[30] T. O. Omotehinwa, D. O. Oyewola, and E. G. Moung, "Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease," *Informat. Health*, vol. 1, no. 2, pp. 70–81, Sep. 2024, doi: 10.1016/j.infoh.2024.06.001.

[31] M. J. Sai, P. Chettri, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, "An ensemble of light gradient boosting machine and adaptive boosting for prediction of type-2 diabetes," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, p. 14, Feb. 2023, doi: 10.1007/s44196-023-00184-y.

[32] S. Subramani, N. Varshney, M. V. Anand, M. E. M. Soudagar, L. A. Al-keridis, T. K. Upadhyay, N. Alshammari, M. Saeed, K. Subramanian, K. Anbarasu, and K. Rohini, "Cardiovascular diseases prediction by machine learning incorporation with deep learning," *Frontiers Med.*, vol. 10, Apr. 2023, Art. no. 1150933, doi: 10.3389/fmed.2023.1150933.

[33] S. Hadianti and W. A. G. Kodri, "Optimization of the machine learning approach using Optuna in heart disease prediction," *J. Med. Inform. Technol.*, vol. 3, pp. 59–64, Sep. 2023.

[34] G. Tripathy and A. Sharaff, "AEGA: Enhanced feature selection based on ANOVA and extended genetic algorithm for online customer review analysis," *J. Supercomput.*, vol. 79, no. 12, pp. 13180–13209, Aug. 2023.

[35] R. R. Sarra, A. M. Dinar, M. A. Mohammed, and K. H. Abdulkareem, "Enhanced heart disease prediction based on machine learning and χ^2 statistical optimal feature selection model," *Designs*, vol. 6, no. 5, p. 87, Sep. 2022.

[36] B. S. Ahamed, M. S. Arya, and A. O. V. Nancy, "Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation," *Adv. Hum.-Comput. Interact.*, vol. 2022, Sep. 2022, Art. no. 9220560, doi: 10.1155/2022/9220560.

[37] Accessed: Jan. 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/45/heart+disease

[38] Accessed: Jan. 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/145/statlog+heart

[39] Accessed: Jan. 2024. [Online]. Available: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

[40] Accessed: Jan. 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/412/z+alizadeh+sani

[41] Accessed: Jan. 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease

[42] Accessed: Jan. 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease

[43] [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[44] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," *Health Technol.*, vol. 10, no. 5, pp. 1137–1144, Sep. 2020, doi: 10.1007/s12553-020-00438-1.

[45] S. Goyal, "Predicting the heart disease using machine learning techniques," *ICT Analysis and Applications* (Lecture Notes in Networks and Systems), vol. 517, S. Fong, N. Dey, and A. Joshi, Eds. Singapore: Springer, 2023, pp. 191–199.

[46] U. N. Dulhare, "Prediction system for heart disease using naive Bayes and particle swarm optimization," *Biomed. Res.*, vol. 29, no. 12, pp. 2646–2649, 2018, doi: 10.4066/biomedicalresearch.29-18-620.

[47] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105624, doi: 10.1016/j.compbiomed.2022.105624.

[48] J. Liu, X. Dong, H. Zhao, and Y. Tian, "Predictive classifier for cardiovascular disease based on stacking model fusion," *Processes*, vol. 10, no. 4, p. 749, Apr. 2022, doi: 10.3390/pr10040749.

[49] E. Nasarian, M. Abdar, M. A. Fahami, R. Alizadehsani, S. Hussain, M. E. Basiri, M. Zomorodi-Moghadam, X. Zhou, P. Pławiak, U. R. Acharya, R.-S. Tan, and N. Sarrafzadegan, "Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach," *Pattern Recognit. Lett.*, vol. 133, pp. 33–40, May 2020.

[50] J. H. Joloudari, E. Hassannataj Joloudari, H. Saadatfar, M. Ghasemigol, S. M. Razavi, A. Mosavi, N. Nabipour, S. Shamshirband, and L. Nadai, "Coronary artery disease diagnosis; ranking the significant features using a random trees model," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, p. 731, Jan. 2020.

[51] M. Zomorodi-Moghadam, M. Abdar, Z. Davarzani, X. Zhou, P. Pławiak, and U. R. Acharya, "Hybrid particle swarm optimization for rule discovery in the diagnosis of coronary artery disease," *Expert Syst.*, vol. 38, no. 1, Jan. 2021, Art. no. e12485.

[52] S. Zhang, Y. Yuan, Z. Yao, X. Wang, and Z. Lei, "Improvement of the performance of models for predicting coronary artery disease based on XGBoost algorithm and feature processing technology," *Electronics*, vol. 11, no. 3, p. 315, Jan. 2022.

[53] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "C-CADZ: Computational intelligence system for coronary artery disease detection using Z-Alizadeh Sani dataset," *Appl. Intell.*, vol. 52, pp. 2436–2464, Jun. 2021.

[54] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective class-imbalance learning based on SMOTE and convolutional neural networks," *Appl. Sci.*, vol. 13, no. 6, p. 4006, Mar. 2023.

[55] L.-C. Cheng, Y.-H. Hu, and S.-H. Chiou, "Applying the temporal abstraction technique to the prediction of chronic kidney disease progression," *J. Med. Syst.*, vol. 41, no. 5, p. 85, May 2017.

[56] H. Polat, H. Danaei Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *J. Med. Syst.*, vol. 41, no. 4, p. 55, Apr. 2017.

[57] M. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent diagnostic prediction and classification system for chronic kidney disease," *Sci. Rep.*, vol. 9, no. 1, p. 9583, Jul. 2019.
[58] S. A. Ebiaredoh-Mienye, E. Esenogho, and T. G. Swart, "Integrating enhanced sparse autoencoder-based artificial neural network technique and softmax regression for medical diagnosis," *Electronics*, vol. 9, no. 11, p. 1963, Nov. 2020.
[59] R. C. Poonia, M. K. Gupta, I. Abunadi, A. A. Albraikan, F. N. Al-Wesabi, and M. A. Hamza, "Intelligent diagnostic prediction and classification models for detection of kidney disease," *Healthcare*, vol. 10, no. 2, p. 371, Feb. 2022, doi: 10.3390/healthcare10020371.
[60] D. K. Choubey, S. Paul, S. Kumar, and S. Kumar, "Classification of Pima Indian diabetes dataset using naive Bayes with genetic algorithm as an attribute selection," in *Proc. Int. Conf. Commun. Comput. Syst. (ICCCS)*, 2017, pp. 451–455.
[61] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Proc. Comput. Sci.*, vol. 132, pp. 1578–1585, Jan. 2018.
[62] R. Patra and B. Khuntia, "Analysis and prediction of Pima Indian diabetes dataset using SDKNN classifier technique," in *Proc. IOP Conf., Mater. Sci. Eng.*, vol. 1070, 2021, Art. no. 012059, doi: 10.1088/1757-899X/1070/1/012059.
[63] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cognit. Comput. Eng.*, vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.
[64] R. Saxena, "Role of k-nearest neighbour in detection of diabetes mellitus," *Turk J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 373–376, 2021.
[65] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023.
[66] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset," *Comput. Methods Programs Biomed. Update*, vol. 4, Jan. 2023, Art. no. 100118.

J. JASMINE GABRIEL received the B.E. degree from the Adhiparasakthi Engineering College, Melmaruvathur, India, in 2009, and the M.E. degree in computer science and engineering from the Saveetha Engineering College, Chennai, India, in 2012. She is currently pursuing the Ph.D. degree with the School of Computing Science and Engineering, VIT Chennai, India. She has seven years of teaching experience and has published a book on computer programming. She has published two international conference papers, two national conference papers, and a book chapter. Her research interests include machine learning, data science and analytics, and healthcare. She received best paper awards in the International Conference on Cloud Computing and Services (ICCCS), in December 2011, and the International Conference on Computing and Control Engineering (ICCCE 2012), in April 2012.

L. JANI ANBARASI received the B.E. degree in computer science and engineering from Manonmaniam Sundaranar University, in 2000, and the M.E. and Ph.D. degrees from Anna University, in 2005 and 2015, respectively. She has around 15 years of experience in various institutions and is currently an Associate Professor with the School of Computing Science and Engineering, VIT Chennai. She has published around 96 technical publications in various international journals and conferences. Her research interests include cryptography, image processing, and soft computing techniques. Her Professional membership includes India Society for Technical Education (Life Member) and ACM. She received the Best Paper Award in the International Conference on Knowledge based Computing Technologies, Chennai, in February 2017, and the Best Paper Award in the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), in December 2017.