**Journal of Industrial Information Technology and Application** *ISSN(Online): 2586-0852*
Vol.9 No. 4

# Ensemble Learning with Heuristic Feature Enhancement for Deep Learning Models in Coronary Artery Disease Prediction and Drug Recommendation

*Sang Suh*[*)], *Lakshmi Kiranmai Reddy Voggu, Venkata Sai Jaswanth Kumar Vellanki, Bhavya Muthineni, and Ravin Timalsina*

Department of Computer Science East Texas A&M University, U.S.A

**Abstract**. The research involves utilizing the design with cardiovascular disease (CVD) and improvising the algorithms on autoencoders and LightGBM models to configure layers in three steps. Phase 1 (32K samples) involves model initialization, basic feature engineering, and cross-validation to create a baseline prediction model. In this phase, the LightGBM model provides preliminary classification results, and an autoencoder performs unsupervised feature representation learning. Phase 2 (80K samples) uses distributed learning, advanced cross-validation, and feature engineering to improve the model's resilience and effectiveness. This phase introduces multilabel classification to include more CVD risk variables. In Phase 3 (160K samples), LightGBM's distributed learning capabilities are fully used, including hyperparameter tweaking and thorough testing to improve model prediction accuracy. This step ensures that the model can handle large datasets while maintaining performance, supporting the practical approach. Autoencoder and LightGBM can reliably predict cardiovascular disease risk in binary and multilabel classification situations with flexibility and scalability. The method accurately predicts cardiovascular disease risk in binary and multilabel classification situations without overfitting at 99%.

**Key word;** Coronary Artery Disease(CAD) Prediction, Deep Learning, Ensemble Approach, LightGBM

Cite this paper as : Sang Suh, Lakshmi Kiranmai Reddy Voggu, Venkata Sai Jaswanth Kumar Vellanki, Bhavya Muthineni, and Ravin Timalsina (2025) “Ensemble Approach Towards Heuristic Features on Deep Learning Algorithms for Coronary Artery Disease Prediction and Drug Recommendations”, Journal of Industrial Information Technology and Application, Vol. 9. No. 4, pp. 1168-1187

[*] Corresponding author: Sang.Suh@etamu.edu

Received: Aug. 8. 2025 Accepted: Nov. 11. 2025 Published: Dec. 31. 2025

# 1. Introduction

### *A.* ***Privacy and Data Security in CAD Prediction***

The development of studies on the detection of coronary artery disease (CAD) has shifted its focus to data security and privacy. Shao et al. (2022) [1] concentrate on cloud-based systems with hyperplane decision-based classifiers, which guarantee the security of patient information, but they do not deal with the issues of enhancing the accuracy of CAD predictions. Likewise, Jamthikar et al. (2022) [2] apply ensemble machine learning to carotid ultrasound data, where the predictions are improved by multiple models together, however, the model does not include contemporary feature extraction methods. The two papers are informative on the aspects of preserving patient privacy and achieving better prediction results, but they nonetheless encounter the issue of enhancing model accuracy.

### *B.* ***Advancements in CAD Detection Techniques***

The recent developments in CAD detection use deep learning and novel machine learning algorithms to enhance the accuracy of the diagnosis. The study by Hasan et al. (2020) [8] applies deep learning to cardiac CT angiography and represents the use of neural networks to select patients undergoing invasive treatment. Jiménez-Partiten et al. (2024) [11] then go a step further and use deep learning to classify CAD according to the severity of the lesion with the aim of achieving more precise diagnostics. Further, Pathak et al. (2022) [12] and Yin et al. (2024) [13] used phonocardiogram data with large multilayered perceptron neural network models indicating the increasing importance of multimodal data and transfer learning in enhancing CAD detection.

### *C.* ***Non-invasive and Multimodal Approaches for CAD Diagnosis***

Non-invasive technologies like phonocardiogram and facial video analysis are getting popular in detecting CAD. Pathak et al. (2022) [12] apply phonocardiogram with transfer learning to diagnose atherosclerotic CAD and Yin et al. (2024) [13] further improves it by using a multiscale attention convolutional network. On the same note, Yao et al. (2020) [16] and Liu et al. (2023) [17] use ECG characteristics and facial videos to predict CAD, showing how multimodal data can be combined to increase the accuracy of the diagnosis. These advancements correspond to the trend of less invasive, easier to access diagnostic methods in CAD studies.

The Proposed research implicates to apply an ensemble of autoencoders (EAs) with improved autoencoder architectures to improve classification in predictive analysis and detection algorithms, especially on medical data. The common challenges encountered with these datasets include high dimensionality, class imbalance and privacy issues. The main objective is to develop a powerful model that combines the feature extracting ability of autoencoders and predictive ability of ensemble learning which would enhance the accuracy of diagnosis and be able to process large complex datasets. The solution is meant to fit the needs of the healthcare data analysis which are expected to change over time, offering a framework that offers a compromise between feature extraction and predictive performance across a range of medical tasks. The research will be concerned with the investigation of various autoencoder models, such as deep autoencoders and variational autoencoders, in order to establish the most efficient strategies of converting raw data of high dimension into meaningful representations that can be used in subsequent classification procedures.

Further to improve the performance of predictions, the research will use ensemble learning techniques including bagging, boosting, and stacking, to integrate the outputs of many classifiers that will be trained on the smaller set of features. The proposed apporach behind this two-level design is to minimize the biases of the individual models, at the same time as increasing the predictive accuracy. Also, ensemble methods will be adjusted to the problem of classes imbalance, where the minority classes will be represented properly in the training procedure. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) will be introduced to increase the ability of the model to make realistic predictions on different classes and therefore the approach will be more useful in practical medical cases. Evaluation and validation of the EA model will be substantially performed utilizing various medical datasets to determine the flexibility and efficacy in various settings. The Functionalities of the model will be thoroughly examined with the help of such performance measures as accuracy, precision, recall, and F1-score. This study can contribute to the sphere of medical diagnostics, namely the prediction of coronary artery disease (CAD), as it combines the benefits of autoencoders and ensemble learning and will result in the improved care and condition of patients

## 2. Literature Survey

### Privacy-Preserving Disease Diagnosis vs. Traditional Disease Prediction Models

**Shao et al. (2022)** [1] propose a privacy-preserving cloud-assisted framework for disease diagnosis, utilizing a hyperplane decision-based classifier. This approach focuses on preserving data security while ensuring accuracy in medical diagnoses. In comparison, **Jamthikar et al. (2022)** [2] employ ensemble machine learning techniques to predict coronary artery disease (CAD) and acute coronary syndrome (ACS) using data from carotid ultrasound. While **Shao et al.** emphasize privacy protection in cloud-based healthcare systems, **Jamthikar et al.** focus on model robustness and prediction accuracy. Both models enhance disease detection but address distinct aspects of healthcare challenges: data security and model performance.

**Ensemble Learning vs. Neutral Network approaches for Disease Prediction**

**Nadeem et al. (2024)** [3] hybridize artificial neural networks (ANNs) with traditional machine learning algorithms for improving colon cancer prediction. This hybrid approach contrasts with **Das et al. (2023)** [4], who employ a computational framework combining binary relevance and MLSMOTE to predict drug-target interactions. While **Nadeem et al.** focus on enhancing prediction accuracy and generalization in biological datasets using neural networks, **Das et al.** address class imbalance through oversampling techniques to improve drug function predictions. Both studies tackle complex prediction problems but differ in their approach: neural networks for enhanced accuracy vs. balancing techniques for more reliable model performance.

**Feature Reduction vs. Oversampling Techniques for Class Imbalance**

**Pasha and Mohamed (2020)** [5] propose a feature reduction model for effective disease risk prediction, which aims to reduce the complexity of high-dimensional healthcare datasets while maintaining predictive accuracy. On the other hand, **Noor et al. (2023)** [6] utilize a stacking model combined with balancing techniques and dimensionality reduction to predict heart disease. **Pasha and Mohamed** aim to reduce input data dimensions to enhance model interpretability, while **Noor et al.** deal with class imbalances in heart disease datasets by incorporating balancing techniques alongside dimensionality reduction. Both studies aim to improve prediction performance but utilize different strategies feature reduction vs. class balancing.

**Deep Learning for Medical Imaging: Segmentation vs. Classification**

**Zreik et al. (2020)** [10] apply deep learning for segmentation tasks in cardiac CT angiography, aiming to detect CAD early through accurate coronary artery segmentation. This contrasts with **Jiménez-Partinen et al. (2024)** [11], who use convolutional neural networks (CNNs) for classifying coronary artery disease based on different lesion degrees. The primary difference lies in the objective: **Zreik et al.** aim to segment coronary arteries for early disease detection, while **Jiménez-Partinen et al.** classify CAD severity to guide treatment planning. Both approaches enhance disease detection through deep learning but differ in their focus on segmentation versus classification.

### Non-invasive Disease Detection: Phonocardiogram vs. ECG vs. Video-based Approaches

Pathak et al. (2022) [12] use phonocardiogram data combined with transfer learning and multiple kernel learning to detect atherosclerotic coronary artery disease (CAD). Yin et al. (2024) [13] also explore phonocardiogram data but apply multiscale attention convolutional compression networks for CAD detection, focusing on the most relevant features in the signal. Both studies aim to improve CAD detection through phonocardiograms, but their methods differ in algorithmic techniques—Pathak et al. employ transfer learning and kernel learning, while Yin et al. use deep learning with attention mechanisms. In contrast, Yao et al. (2020) [16] focus on ECG-based detection, using QT interval time series and ST-T waveform features to enhance the accuracy of CAD diagnosis. Lastly, Liu et al. (2023) [17] introduce a novel approach that utilizes facial video data to predict CAD, leveraging facial features like texture and movement. This unique method provides a non-invasive and easily accessible diagnostic tool, making it stand out in the realm of non-invasive disease detection. While all studies aim to detect CAD using non-invasive techniques, the methodologies differ, with Liu et al.'s facial video-based approach being particularly innovative in terms of accessibility and simplicity.

### Advanced Predictive Models for Disease Risk

Kapila et al. (2023) [7] introduce a Quine McCluskey Binary Classifier (QMBC) for heart disease prediction, which simplifies the classification process using Boolean algebra, making it computationally efficient. In contrast, Hasan et al. (2020) [8] use an ensemble of various machine learning classifiers to predict diabetes, enhancing prediction accuracy by combining models like decision trees and support vector machines. While both focus on disease prediction, Kapila et al. offer a simple yet effective binary classification approach, whereas Hasan et al. rely on multiple classifiers to boost prediction accuracy for diabetes. Kapila et al.'s approach prioritizes computational efficiency and interpretability, while Hasan et al. emphasize robust prediction through model diversity.

### Risk Management in Healthcare Using Data Envelopment Analysis (DEA)

Jomthanachai et al. (2021) [9] present a unique integration of Data Envelopment Analysis (DEA) and machine learning for risk management in healthcare. By evaluating the efficiency of healthcare processes using DEA and applying machine learning models for predicting outcomes, this study targets the optimization of healthcare systems. This approach is distinct from Mahmood et al. (2024) [19], who combine hybrid deep learning models with neural fuzzy inference systems for improving coronary artery disease prognosis. While Jomthanachai et al. focus on the broader management of healthcare risks, Mahmood et al. concentrate on refining disease prognosis, combining deep learning with fuzzy systems for optimized decision-making.

### Hybrid Deep Learning and Expert Knowledge for Disease Classification

Ramesh and Lakshmanna (2024) [20] combine hybrid deep learning with a neural fuzzy inference system for the early detection and prevention of coronary heart disease (CHD). This model is aimed at enhancing both prediction accuracy and interpretability. In contrast, Jiang et al. (2024) [18] integrate expert ECG features with machine learning to automatically classify coronary microvascular dysfunction, a subtle yet challenging condition to diagnose. Both studies integrate deep learning with expert knowledge, but Ramesh and Lakshmanna focus on disease prevention through early detection and fuzzy logic, while Jiang et al. address the complexity of classifying microvascular dysfunction using a hybrid approach.

## 3. Proposed Work

### Concept

Stroke and coronary artery disease (CAD) are major cardiovascular diseases that present severe health hazards in the world. CAD is a condition in which coronary arteries get narrowed or blocked by deposits known as plaque, resulting in the decreased supply of blood to the heart, causing angina or heart attacks. The risk factors are age, gender, hypertension, high cholesterol, diabetes, smoking, and a sedentary lifestyle. Likewise, a stroke, especially an ischemic stroke, is the result of a failure in blood supply to the brain usually due to clots. Stroke risk factors substantially coincide with CAD risk factors, which is why preventing and identifying it at an early age is crucial. In order to successfully categorize those illnesses, datasets including demographic factors, clinical measures such as blood pressure and cholesterol levels, and lifestyle choices are used. Such information is essential in the modeling of models capable of predicting the existence of such diseases. To discuss the issue of CAD and stroke classification, binary and multi-label learning is applicable. Binary classification aims at identifying the presence of either CAD or stroke in a patient and therefore the prediction is a basic presence or absence. Conversely, multi-label classification enables the prediction of many conditions simultaneously, which is sensible given that a patient may be in danger of developing both diseases.

In order to further improve the predictive performance of such classifications, ensemble learning and autoencoder models are a very effective combination. Autoencoders are applied in dimensionality reduction and feature extraction which helps the model to concentrate on fundamental patterns and reduces noise in the data. The process is useful in detecting the abnormalities in patient profiles, which may represent potential health risks and require additional research.

A combination of ensemble techniques and autoencoders makes a powerful framework of disease prediction. Ensemble learning combines a number of classifiers which include decision trees and neural networks to enhance accuracy and reliability. This approach uses more of the patterns present in the data by utilizing different models. Ensembles also provide voting mechanism on the final classification which makes it robust, be it in binary or multi-label cases. Practically, the steps of the workflow are pre-processing of the data, training the autoencoder to produce low-dimensional representations, and classification using the ensemble model. This integrated framework could offer real-time predictions on novel patient data that would be vital in early diagnosis and intervention of CAD and stroke, which would eventually translate to improved patient outcomes and informed clinical practice.

**Block Diagram**

The block diagram shown in figure 1 below starts with Data Collection and Preprocessing. The Data.gov dataset used is the U.S. mortality data concerning coronary heart disease and stroke. The initial stage is data loading, cleaning (dealing with missing values) and preparation to train the model. Such preprocessing step is essential to the quality and consistency of data that is commonly done by normalization (scaling numerical values) and encoding categorical variables to machine learning models. The preprocessing of the data may also include the division of data into training and testing data and possibly the augmentation of the dataset to address any class imbalances which is prevalent in healthcare datasets.

Consequently, Feature Extraction is implemented with the help of Autoencoder, a deep learning-based technique of decreasing the dimensionality of the data set, retaining important features. Autoencoder is a neural network, which comprises two primary components: encoder and decoder. The encoder reduces the input data to a lower-dimensional latent space, where internal node represents the most important characteristics. The task of the decoder to generate the original input based on this latent space. Hence, perfect reconstruction per se that is desired, but the latent space serves as a compact and informative representation of the data. This step of feature extraction aids in removing noise and makes the machine learning models more efficient and accurate as they would consider the important patterns in the data rather than all the features. The latent space is subsequently fed into the next step which is the model training.

Lastly, the proposed system uses Model Training Ensemble Learning that is, training several models and then combining them to get a better prediction. Within this framework, a number of machine learning models, including Decision Trees, Random Forests, Support Vector Machines, and Neural Networks are trained in parallel on the resulting features. The models predict independently and the predictions are combined using a voting mechanism to make the final output. This combination method is useful in minimising the bias of the individual models resulting in stronger predictions. Once the models have been trained, they are tested with accuracy, precision and F1 score and predictions are made in real-time on unseen data. The results are reported using visualization tools to monitor the performance of the model.

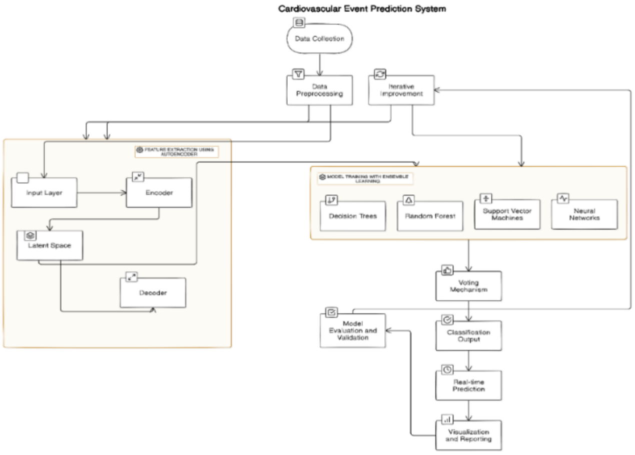


Figure1. Block Diagram of the Cardiovascular Event Prediction System using Ensemble and Autoencoder

## Methods

*Formulations:*

Encoder Function: The encoder reduces the input data X into a latent space representation ZZZ through a series of hidden layers. This can be mathematically represented as:

$$Z = f(X) = \sigma(W_e \cdot X + b_e) \qquad (1)$$

where $W_e$ and $b_e$ are the weights and biases of the encoder layers, and σ is the activation function (ReLU or Sigmoid).

Latent Space: The latent space Z represents a compressed version of the input data, capturing the most essential features for reconstruction.

Decoder Function: The decoder reconstructs the original input data X′ from the latent space Z, represented as:

$$X' = g(Z) = \sigma(W_d \cdot Z + b_d) \qquad (2)$$

where $W_d$ and $b_d$ are the weights and biases of the decoder layers. The objective is to minimize the reconstruction loss LLL, typically using mean squared error (MSE):

$$L = \frac{1}{n}\sum_{i=1}^{n}(Xi - Xi')^2 \qquad (3)$$

Formulation for Ensemble Learning:

1. Individual Models: The ensemble consists of multiple base models $M_1, M_2, \ldots\ldots, M_n$, each trained on the same dataset. Each model $M_i$ generates a prediction$\hat{y}$ for a given input X.
2. Voting Mechanism: The predictions from all models are combined using either majority voting (for classification) or averaging (for regression). In the case of majority voting for classification:

$$\hat{y} = mode(\widehat{y^1}, \widehat{y^2}, \ldots, \widehat{y^n})$$

where$\hat{y}$ is the final predicted class indicated above formulations.

3. Weighted Voting: Optionally, models can be assigned weights based on their performance, and the final prediction is calculated as a weighted sum:

$$\hat{y} = argmax \sum_{i=1}^{N} w_i * \hat{y} \qquad (4)$$

where $w_i$ is the weight assigned to model $M_i$, This allows better-performing models to have more influence on the final prediction.

In an autoencoder, the encoder map in an autoencoder is a compression map that maps input data X to a latent space representation Z, which is a transformation of weights, biases and an activation function as in (1). The latent space Z is a compressed form of X, and the key features that allow reconstruction. At the decoder, a reverse transformation is used to recover the original data X 1 from Z, as in (2). The goal is to minimize the reconstruction loss which is usually mean squared error (MSE) in (3). In ensemble learning, a number of models are trained on identical data, and the predictions of the models are combined through a voting system. In classification problems, majority voting is used to make the final prediction as described in (4). In another variation, weighted voting may be used, in which a weight is assigned to each model based on its prediction, and the overall prediction is a weighted average, also defined in (4).

## 4. Experimental Setup

The Cardiovascular Event Prediction System Experimental Setup was carried on a DELL Inspiron 15R Laptop. Minimum specifications capable of running machine learning workloads on the system include a multi-core processor, 8GB or more of RAM, and an integrated GPU. Since the feature extraction model based on the Autoencoder model is relatively complex and the ensemble of machine learning models, the resources of the laptop were optimized through experimentation in batches, using Python libraries, such as TensorFlow, Keras, and Scikit-learn. The Autoencoder was trained in a manner that it compressed the input data into latent space and the outputs were stored and subsequently used as inputs to the ensemble learning models. All the training was done on a GPU to speed up the process, particularly the deep learning-based autoencoder, whereas the ensemble models (Decision Trees, Random Forests, SVMs, Neural Networks) were effectively trained on the CPU.

Real-time data processing, such as preprocessing, model training, and evaluation was also supported on the Dell Inspiron 15R laptop. The development and testing were done in python-based environments such as Jupyter Notebook. The ensemble learning models were optimized to work with parallel processing in spite of the memory and computational resources being limited relative to a high-end workstation machine, where multi-threaded processor capabilities were utilized. The training set was split into batches and the models were fitted on the extracted features of the Autoencoder. The evaluation metrics such as accuracy, precision, recall, and F1 score were calculated, and real-time predictions were produced on unseen data of the test set. Performance tracking of the results was done through Python libraries Matplotlib and Seaborn to visualize the results, and the final outputs were saved locally on the system to be used in further analysis.

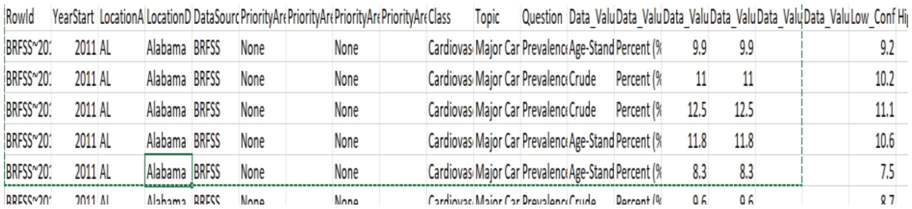
Implementation Analysis

Phase 1 starts with 32,000 samples to emphasize the necessity to set a baseline of model performance by performing data preprocessing and feature engineering. The preprocessing step consists of missing value replacement, noise removal, and Encoding of categorical variables where Label Encoding is used when dealing with classification problems. Numerical attributes are normalised or standardised to make them consistent throughout the data. Heuristic feature engineering is used to generate new features, e.g.

moving averages or polynomial transformations and feature importance analysis is conducted to determine the most significant variables. As a model, LightGBM will be selected based on its effectiveness in using computational resources (low cost) on large datasets. Model is first trained with default parameters and the performance of model is evaluated with K-Fold cross-validation (5 or 10 folds usually) to avoid overfitting and underfitting. Model performance is evaluated using accuracy, precision, recall, and F1-score and then hyperparameter optimization is done to maximize accuracy using cross-validation scores.

Phase 2 In this phase using a larger dataset of 80,000 samples the preprocessing is scaled to the larger dataset and more complex feature engineering is employed, such as interaction terms and non-linear transformations. Recursive Feature Elimination (RFE): This is applied to decrease the number of features to the most interpretive variables. Since all of this is bigger, the distributed learning and parallelization capabilities of LightGBM are used to accelerate training, and early stopping is used to avoid overfitting. The data is divided into 80 percent training, 10 percent validation and 10 percent testing and 10-fold cross-validation is employed to make sure that the model is a good generalizer. The process of hyperparameter tuning is performed with the help of Grid Search or Random Search, and the performance of a model is measured with the help of precision, recall, F1-score, and AUC measures.

The scalability of the model is further evaluated in Phase 3 with 160,000 sample dataset. The preprocessing of the data is made more efficient and is frequently automated via pipelines and the batch processing is employed to address the limitations of memory. The distributed learning mode provided by LightGBM is critical to training with the huge dataset, and stratified sampling makes sure that all classes are represented. Training of the model is done with 10-fold cross-validation and further sophisticated hyperparameter optimization is performed using Grid Search and Bayesian optimization. More complicated parameters, like learning rate decay and max bin values are tweaked to get the best performance. Lastly, the model is tested on the independent test set with precision, recall, F1-score, and AUC, the model is confirmed not to be overfitted and is capable of delivering good results even in the case of large datasets.



| RowId | YearStart | LocationA | LocationD | DataSourc | PriorityAre | PriorityAre | PriorityAre | PriorityAre | Class | Topic | Question | Data_Valu | Data_Valu | Data_Valu | Data_Valu | Data_Valu | Data_Valu | Low_Conf | Hi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BRFSS~20: | 2011 | AL | Alabama | BRFSS | None | | None | | Cardiovas | Major Car | Prevalenc | Age-Stand | Percent (% | 9.9 | 9.9 | | | 9.2 | |
| BRFSS~20: | 2011 | AL | Alabama | BRFSS | None | | None | | Cardiovas | Major Car | Prevalenc | Crude | Percent (% | 11 | 11 | | | 10.2 | |
| BRFSS~20: | 2011 | AL | Alabama | BRFSS | None | | None | | Cardiovas | Major Car | Prevalenc | Crude | Percent (% | 12.5 | 12.5 | | | 11.1 | |
| BRFSS~20: | 2011 | AL | Alabama | BRFSS | None | | None | | Cardiovas | Major Car | Prevalenc | Age-Stand | Percent (% | 11.8 | 11.8 | | | 10.6 | |
| BRFSS~20: | 2011 | AL | Alabama | BRFSS | None | | None | | Cardiovas | Major Car | Prevalenc | Age-Stand | Percent (% | 8.3 | 8.3 | | | 7.5 | |
| BRFSS~20: | 2011 | AL | Alabama | BRFSS | None | | None | | Cardiovas | Major Car | Prevalenc | Crude | Percent (% | 9.6 | 9.6 | | | 8.7 | |

Figure 1. Representing the Dataset for the CVD

The CHD data provided by data.gov is comprehensively preprocessed through heuristic rules to assure data quality and relevancy. The missing values are filled with regional or topic means, and the categorical variables such as LocationAbbr and PriorityArea1 are one-hot encoded, and the continuous variables such as Data_Value are standardized. Domain knowledge will focus on such important aspects as access to healthcare and geography. The Ensemble with Heuristic Model LGBM incorporates k-fold newton cross-validation and differential equations to capture the time series health patterns and optimizes feature importance based on heuristics that gives more weight to influential variables. This method trades off completeness, interpretability and solidness in forecasting CHD prevalence. The heuristics are developed as mathematical formulations as

$$AFW = f + w_1 \times DAT$$

Where AFT (Adjusted Feature Weight)

F = Feature weight

DAT = Domain Adjusted Factor

This feature weights are used in model for training, especially in cross-validation, which improves generalization and the accuracy of prediction. Grid Search, Random Search, and Bayesian optimization are used in hyperparameter tuning, guided by heuristics to improve the AFW. The performance is further improved by using ensemble methods such as bagging, boosting and stacking. Trained on 10,000 cases and cross-validated 5 times, the model is highly reliable in terms of accuracy, F1 score, AUC, log loss, MSE, and RMSE.

Table 1. Representing each fold provides several crucial evaluation metrics

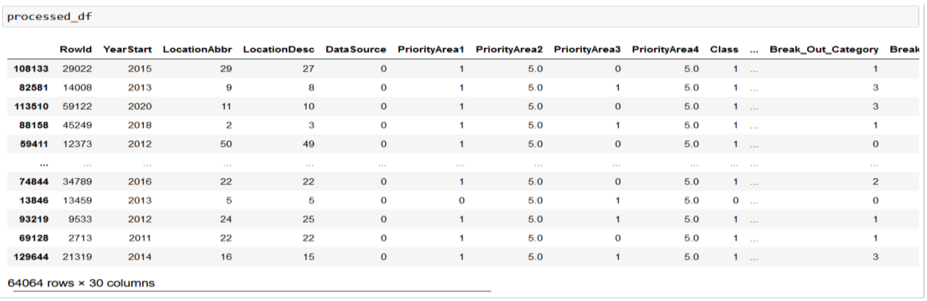
| fold | train_accuracy | test_accuracy | train_f1_score | test_f1_score | train_confusion_matrix | test_confusion_matrix | train_log_loss |
|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 1 | 1 | 1 | [[3987, 0], [0, 4013]] | [[1013, 0], [0, 987]] |
| **1** | 1 | 1 | 1 | 1 | 1 | [[3972, 0], [0, 4028]] | [[1028, 0], [0, 972]] |
| **2** | 2 | 1 | 1 | 1 | 1 | [[4000, 0], [0, 4000]] | [[1000, 0], [0, 1000]] |
| **3** | 3 | 1 | 1 | 1 | 1 | [[3988, 0], [0, 4012]] | [[1012, 0], [0, 988]] |
| **4** | 4 | 1 | 1 | 1 | 1 | [[4053, 0], [0, 3947]] | [[947, 0], [0, 1053]] |

Table 2. Representing the overall accuracy and Loss for Light GBM model

| TEST_LOG_LOSS | TRAIN_RECALL | TEST_RECALL | TRAIN_AUC | TEST_AUC | TRAIN_MSE | TEST_MSE | TRAIN_RMSE | TEST_RMSE |
|---|---|---|---|---|---|---|---|---|
| **2.2E-05** | 2.2E-05 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| **2.2E-05** | 2.2E-05 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| **2.2E-05** | 2.2E-05 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| **2.2E-05** | 2.2E-05 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| **2.2E-05** | 2.2E-05 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |

As illustrated in Table-I and Table-II, the proposed model demonstrates excellent performance in terms of various evaluation measures. The accuracies on train and test are close to 100% and F1 scores are 1.0, which means the classification is precise and balanced. This is confirmed by the confusion matrices that indicate no false positives or false negatives. The model confidence and discriminative ability are emphasized by low log loss and an ideal AUC of 1.0. Metric measures of error such as MSE and RMSE are very close to zero indicating a low classification error which is important in high stakes areas. The results of cross-validation are also stably high, which proves the ability of the model to generalize. These results confirm the high accuracy, stability of the LightGBM-ensemble method and its suitability to be used in the real world.

## Phase-II: Autoencoder Design Analysis



```
processed_df
```

| | RowId | YearStart | LocationAbbr | LocationDesc | DataSource | PriorityArea1 | PriorityArea2 | PriorityArea3 | PriorityArea4 | Class | ... | Break_Out_Category | Break |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **108133** | 29022 | 2015 | 29 | 27 | 0 | 1 | 5.0 | 0 | 5.0 | 1 | ... | 1 | |
| **82581** | 14008 | 2013 | 9 | 8 | 0 | 1 | 5.0 | 1 | 5.0 | 1 | ... | 3 | |
| **113510** | 59122 | 2020 | 11 | 10 | 0 | 1 | 5.0 | 0 | 5.0 | 1 | ... | 3 | |
| **88158** | 45249 | 2018 | 2 | 3 | 0 | 1 | 5.0 | 1 | 5.0 | 1 | ... | 1 | |
| **59411** | 12373 | 2012 | 50 | 49 | 0 | 1 | 5.0 | 0 | 5.0 | 1 | ... | 0 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **74844** | 34789 | 2016 | 22 | 22 | 0 | 1 | 5.0 | 0 | 5.0 | 1 | ... | 2 | |
| **13846** | 13459 | 2013 | 5 | 5 | 0 | 0 | 5.0 | 1 | 5.0 | 0 | ... | 0 | |
| **93219** | 9533 | 2012 | 24 | 25 | 0 | 1 | 5.0 | 1 | 5.0 | 1 | ... | 1 | |
| **69128** | 2713 | 2011 | 22 | 22 | 0 | 1 | 5.0 | 0 | 5.0 | 1 | ... | 1 | |
| **129644** | 21319 | 2014 | 16 | 15 | 0 | 1 | 5.0 | 1 | 5.0 | 1 | ... | 3 | |

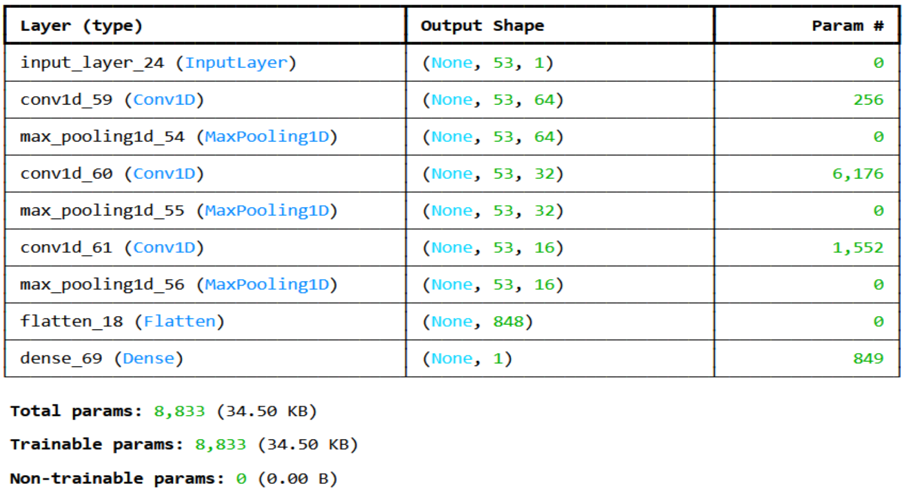
64064 rows × 30 columns

Figure 3. Representing encoded dataset for the CVD case

### *Auto Encoder Design Using Class:*

The suggested model combines Autoencoder, Dense Classifier, and LightGBM to improve the classification. The convolutional autoencoder is constructed using Conv1D and MaxPooling1D layers and acquired compact and meaningful representations of the input data. These encoded characteristics are subsequently fed into dense classifier to make binary predictions. The model has a binary output head when classification is enabled, rendering it to be used in supervised tasks. LightGBM is executed in parallel to do feature selection, and rank the most significant features in order to reduce overfitting and enhance generalization. The advantage of this unified method is that it uses unsupervised learning to optimize representation, supervised learning to optimize prediction and gradient boosting to optimize features, making the classification pipeline highly accurate and robust.

*Results with Prediction plots:*



| Layer (type) | Output Shape | Param # |
|---|---|---|
| input_layer_24 (InputLayer) | (None, 53, 1) | 0 |
| conv1d_59 (Conv1D) | (None, 53, 64) | 256 |
| max_pooling1d_54 (MaxPooling1D) | (None, 53, 64) | 0 |
| conv1d_60 (Conv1D) | (None, 53, 32) | 6,176 |
| max_pooling1d_55 (MaxPooling1D) | (None, 53, 32) | 0 |
| conv1d_61 (Conv1D) | (None, 53, 16) | 1,552 |
| max_pooling1d_56 (MaxPooling1D) | (None, 53, 16) | 0 |
| flatten_18 (Flatten) | (None, 848) | 0 |
| dense_69 (Dense) | (None, 1) | 849 |

**Total params:** 8,833 (34.50 KB)

**Trainable params:** 8,833 (34.50 KB)

**Non-trainable params:** 0 (0.00 B)

Figure 4: Representing the summary of the model design (Autoencoder model)

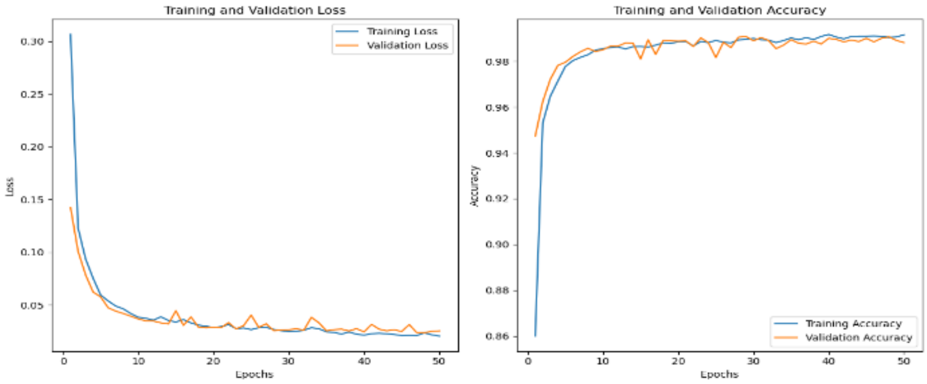


Figure 5. Representing the (Autoencoder model) loss and accuracy plots

The model of Fig.5. starts with an input layer that takes sequences of length 53 and one feature per time step (None, 53, 1). It is converted through a convolutional layer with 64 filters of size 3, and the output has the shape of (None, 53, 64). This is followed by a max-pooling layer, which keeps the sequence size, but improves feature retention to add efficiency.

The data next passes through a second convolutional layer of 32 filters, another max-pooling layer, but with the number of channels now reduced to 32, and more abstract features extracted. The channels are further reduced to 16 with a third convolutional layer having 16 filters and preserving the sequence length. The output after pooling is then flattened into a 1D vector and fed to a fully connected layer to perform binary classification, and the overall number of trainable parameters is 8,833.

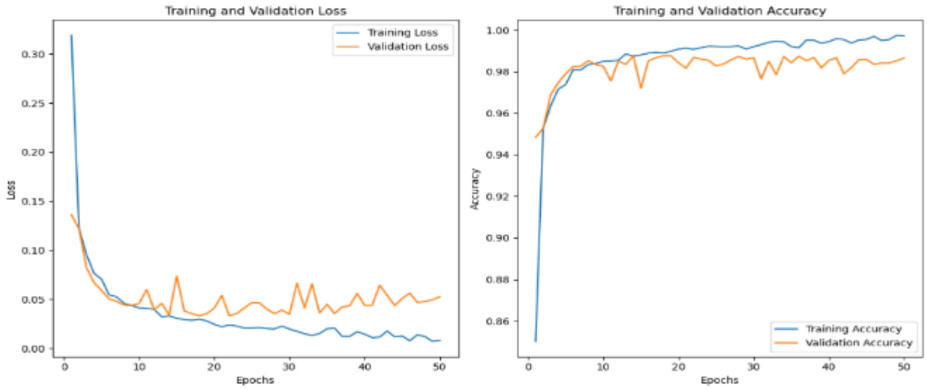


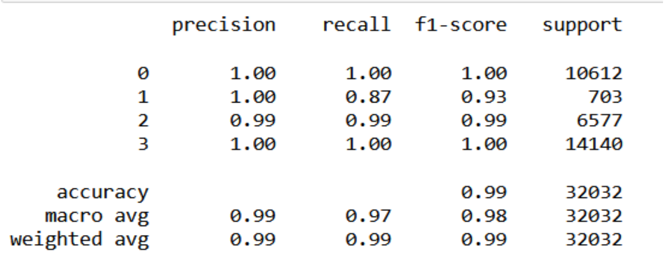
Figure 6. Representing the (Autoencoder-Light-GBM model) loss and accuracy plots

The proposed model combines a convolutional autoencoder with a dense classifier and a custom LightGBM (LGB) feature selection component, demonstrating strong performance over training epochs. Accuracy improves from 75% to 98–99%, and loss drops from 0.48 to 0.0075 by epoch 50, achieving a test accuracy of 98.64%. This highlights the effectiveness of integrating unsupervised feature learning with supervised classification and gradient-boosted feature ranking. In Phase-3, with 60k and 120k samples in binary and multilabel formats, the DataPreprocessor class handles missing values and categorical encoding. It replaces NaNs using placeholders or differential equations for numeric data, and encodes categorical variables using LabelEncoder, preparing data for model training. The MultiLabelGeneration class creates a dynamic multi-label column using a system of linear differential equations defined as

$$\frac{dx}{dt} = A \cdot x + B.$$

For each row, it calculates differential changes from input priority areas, introduces variability using a random factor, and derives differential columns. These are summed to compute an average weight, which is classified into severity levels ('Low' to 'Severe') and stored as multi_labels. This method simulates dynamic system behavior, enabling robust multilabel classification reflective of real-world health data patterns.

*Multi label results:*



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     10612
           1       1.00      0.87      0.93       703
           2       0.99      0.99      0.99      6577
           3       1.00      1.00      1.00     14140

    accuracy                           0.99     32032
   macro avg       0.99      0.97      0.98     32032
weighted avg       0.99      0.99      0.99     32032
```

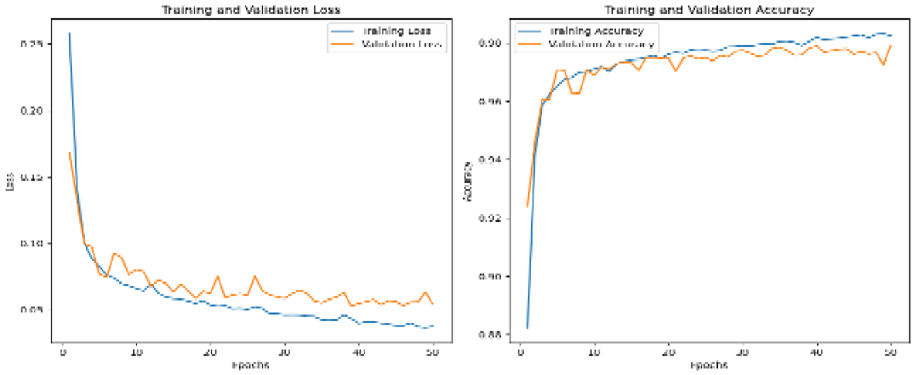


Figure 7. Representing the Accuracy and loss plot for proposed Algorithm (Heuristics Autoencoder Light-GBM and Classification report

Representing the Classification Report and Accuracy-loss plot for Heuristics Autoencoder Light-GBM model.

As shown in Fig. 7, the training log highlights the model's performance over 50 epochs, reflecting the synergy of the autoencoder, LightGBM (LGB) classifier, and dense layers. Accuracy improves rapidly from 82.89% to 97.09% in the first 10 epochs, with loss dropping from 0.3485 to 0.0665, as the autoencoder effectively compresses data into meaningful features. These encoded features are refined by the LGB model for classification, while dense layers map them to final outputs. By epoch 50, the model reaches 98.29% training and 97.93% validation accuracy. Slight fluctuations suggest mild overfitting, but overall performance remains strong on unseen data.

Table 3. Representing existing and proposed model performances

| *SNO* | *Model* | *Average Precision* | *Average Recall* | *Average F1-Score* | *Test Accuracy* | *Train Accuracy* |
|---|---|---|---|---|---|---|
| ***0*** | Random forest | 0.955993 | 0.947734 | 0.951316 | 0.955993 | 1 |
| ***1*** | Gradient boosting | 0.944444 | 0.935369 | 0.93858 | 0.944444 | 0.954886 |
| ***2*** | Adaboost | 0.934457 | 0.926041 | 0.927783 | 0.934457 | 0.943334 |
| ***3*** | Support vector machine | 0.884831 | 0.85823 | 0.869284 | 0.884831 | 0.891508 |
| ***4*** | dl_model_CNN | 0.924469 | 0.913655 | 0.916533 | 0.924469 | 0.938417 |
| ***5*** | dl_model_LSTM | 0.940699 | 0.935542 | 0.93501 | 0.940699 | 0.951062 |
| ***6*** | dl_model_DENSE | 0.887953 | 0.861859 | 0.872888 | 0.887953 | 0.891664 |
| ***7*** | AUTOEND+HDF | 99 | 97 | 99 | 96.5 | 99 |
| ***8*** | AUTOEND+LGB | 99 | 94 | 99 | 96.9 | 99 |
| ***9*** | AUTOEND+HDF+LGB | 99 | 99 | 99 | 97.9 | 99 |

When comparing the results of the current models with the proposed Autoencoder-based ones on identifying cardiovascular disease (CVD) in TABLE-III, it can be seen that the performance is much better. The classic machine learning models such as the Random Forest (RF) obtain good performance with high Test Accuracy of 95.6% and Average Precision and Recall. Nevertheless, deep learning algorithms such as DNN are slightly outperformed by RF and Gradient Boosting (GB), demonstrating a weakness in terms of precision and recall. Conversely, the Autoencoder based models show the best performance over all existing models on all metrics. AUTOEND+HDF model shows a Test Accuracy of 96.5%, whereas AUTOEND+LGB and AUTOEND+HDF+LGB attain even better results, with the latter having a Test Accuracy of 97.9%. These models also have the best Average Precision and Recall, which underlines their better capacity to detect CVD. The models with Autoencoders have good feature extraction and strong classification, which makes them very applicable on complicated medical tasks, such as CVD detection.

## 5. Conclusion

The proposed research contributes to the current knowledge by emphasizing that the models based on Autoencoders (AUTOEND+HDF, AUTOEND+LGB, AUTOEND+HDF+LGB) demonstrate the best results in terms of cardiovascular disease (CVD) detection when compared to the traditional machine learning (ML) and deep learning (DL) approaches using such important metrics as accuracy, precision, recall, and F1-score. Autoencoders coupled with a LightGBM classifier provides feature extraction and classification with high test accuracy of 97.9 percent. Such models as Random Forest and MLPs are promising, but they are behind in terms of complex tasks and need more tuning.

## 6. Scope

To improve in the future, it is possible to include heuristic-based feature engineering, with domain-specific indicators, such as cholesterol or heart rate variability. The multi-task learning is capable of improving the model to predict several heart conditions at once. Robustness and generalization can be enhanced by ensemble approaches, e.g. by stacking different Autoencoder-classifier pairs. Further, more advanced deep learning models such as LSTMs or Transformers are more suitable to learn temporal patterns in health-related data such as ECGs, and finally, model explainability and real-time inference will be necessary to make practical use of such models in healthcare diagnostics.

## References

[1] Shao, Y. et al., ‘Privacy-Preserving and Verifiable Cloud-Aided Disease Diagnosis and Prediction With Hyperplane Decision-Based Classifier’, IEEE Internet of Things Journal, 9(21), pp. 21648-21661. doi: 10.1109/JIOT.2022.3181734. (2022).

[2] Jamthikar, A. D. et al., ‘Ensemble Machine Learning and Its Validation for Prediction of Coronary Artery Disease and Acute Coronary Syndrome Using Focused Carotid Ultrasound’, IEEE Transactions on Instrumentation and Measurement, 71, pp. 1-10. doi: 10.1109/TIM.2021.3139693. (2022).

[3] Nadeem, M. S. A. et al., ‘Hybridizing Artificial Neural Networks Through Feature Selection Based Supervised Weight Initialization and Traditional Machine Learning Algorithms for

Improved Colon Cancer Prediction', IEEE Access, 12, pp. 97099-97114. doi: 10.1109/ACCESS.2024.3422317. (2024).

[4] Das, P. et al., 'BRMCF: Binary Relevance and MLSMOTE Based Computational Framework to Predict Drug Functions from Chemical and Biological Properties of Drugs', IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(3), pp. 1761-1773. doi: 10.1109/TCBB.2022.3215645. (2023).

[5] Pasha, S. J. and Mohamed, E. S., 'Novel Feature Reduction (NFR) Model With Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction', IEEE Access, 8, pp. 184087-184108. doi: 10.1109/ACCESS.2020.3028714. (2020).

[6] Noor, A. et al., 'Heart Disease Prediction Using Stacking Model With Balancing Techniques and Dimensionality Reduction', IEEE Access, 11, pp. 116026-116045. doi: 10.1109/ACCESS.2023.3325681. (2023).

[7] Kapila, R. et al., 'Heart Disease Prediction Using Novel Quine McCluskey Binary Classifier (QMBC)', IEEE Access, 11, pp. 64324-64347. doi: 10.1109/ACCESS.2023.3289584. (2023).

[8] Hasan, M. K. et al., (2020). 'Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers', IEEE Access, 8, pp. 76516-76531. doi: 10.1109/ACCESS.2020.2989857.

[9] Jomthanachai, S. et al., (2021). 'An Application of Data Envelopment Analysis and Machine Learning Approach to Risk Management', IEEE Access, 9, pp. 85978-85994. doi: 10.1109/ACCESS.2021.3087623.

[10] Zreik, M., El-Segaier, M., and Khan, R., (2020). 'Deep Learning Analysis of Coronary Arteries in Cardiac CT Angiography for Detection of Patients Requiring Invasive Coronary Angiography,' IEEE Transactions on Medical Imaging, 39(5), pp. 1545-1557. doi: 10.1109/TMI.2019.2953054.

[11] Jiménez-Partinen, A., Thurnhofer-Hemsi, K., Rodríguez-Capitán, J., Molina-Ramos, A. I., and Palomo, E. J., (2024). 'Coronary Artery Disease Classification With Different Lesion Degree Ranges Based on Deep Learning,' IEEE Access, 12, pp. 69229-69239. doi: 10.1109/ACCESS.2024.3401465.

[12] Pathak, A., Mandana, K., and Saha, G., (2022). 'Ensembled Transfer Learning and Multiple Kernel Learning for Phonocardiogram Based Atherosclerotic Coronary Artery Disease Detection,' IEEE Journal of Biomedical and Health Informatics, 26(6), pp. 2804-2813. doi: 10.1109/JBHI.2022.3140277.

[13] Yin, C., Zheng, Y., Ding, X., Shi, Y., Qin, J., & Guo, X. (2024). Detection of coronary artery disease based on clinical phonocardiogram and multiscale attention convolutional compression network. IEEE Journal of Biomedical and Health Informatics, 28(3), 1353–1362. https://doi.org/10.1109/JBHI.2024.3354832

[14] Zhang, X., Yang, L., Zhang, S., & Liu, H. (2024). An anatomy- and topology-preserving framework for coronary artery segmentation. IEEE Transactions on Medical Imaging, 43(2), 723–733. https://doi.org/10.1109/TMI.2023.3319720

[15] Mahmood, T., Singh, A., & Usman, M. (2024). Enhancing coronary artery disease prognosis: A novel dual-class boosted decision trees strategy for robust optimization. IEEE Access, 12, 107119–107143. https://doi.org/10.1109/ACCESS.2024.3435948

[16] Yao, L., Xu, L., & Zhang, Y. (2020). Enhanced automated diagnosis of coronary artery disease using features extracted from QT interval time series and ST–T waveform. IEEE Access, 8, 129510–129524. https://doi.org/10.1109/ACCESS.2020.3008965

[17] Liu, X., Yang, H., Zhang, X., & Li, H. (2023). VideoCAD: An uncertainty-driven neural network for coronary artery disease screening from facial videos. IEEE Transactions on Instrumentation and Measurement, 72, 1–12. https://doi.org/10.1109/TIM.2022.3229704

[18] Cheung, W. K., Li, Y., & Chan, W. (2021). A computationally efficient approach to segmentation of the aorta and coronary arteries using deep learning. IEEE Access, 9, 108873–108888. https://doi.org/10.1109/ACCESS.2021.3099030

[19] Ramesh, B., & Lakshmanna, K. (2024). A novel early detection and prevention of coronary heart disease framework using hybrid deep learning model and neural fuzzy inference system. IEEE Access, 12, 26683–26695. https://doi.org/10.1109/ACCESS.2024.3366537

[20] Jiang, M., Zhang, X., Wang, J., & Chen, Y. (2024). An automatic coronary microvascular dysfunction classification method based on hybrid ECG features and expert features. IEEE Journal of Biomedical and Health Informatics, 28(9), 5103–5112. https://doi.org/10.1109/JBHIssss