International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Novel Deep Learning Model for Cardiovascular Disease Prediction (MarCDP)

Marwa Abdulrahman Al-Hadi[1]*, and Ghaleb Hamoud Al-Gaphari[2]

Submitted: 02/12/2024 **Revised:** 22/01/2025 **Accepted:** 30/01/2025

Abstract: Cardiovascular disease (CVD) is one of the leading causes of death worldwide, highlighting the critical need for efficient early identification. Early and accurate prediction is crucial for effective treatment. Despite various proposed solutions, a gap in prediction accuracy still persists. Therefore, this study introduces a novel deep learning model designed to enhance CVD prediction by utilization several advanced techniques. The model of Marwa Cardiovascular disease predication (MarCDP) leverages Recurrent Neural Networks (RNN) to capture patterns, employs Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection, and utilizes the Synthetic Minority Over-sampling Technique (SMOTE) to address data imbalance. A range of optimization approaches is applied to fine-tune the model's parameters, resulting in improved accuracy. The model was developed and evaluated using four benchmark datasets: Cleveland, Hungary, Switzerland, and Long Beach V. The proposed model achieved an accuracy of 98.05%, surpassing the performance of existing deep learning models. This novel approach offers a promising product for early CVD detection.

Keywords: *Recurrent Neural Networks (RNN), Feature selection (LASSO), Synthetic Minority Over-sampling Technique (SMOTE), prediction accuracy.*

1. Introduction

Cardiovascular disease (CVD) remains the primary cause of death worldwide, as reported by the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) [1]. In Yemen, CVD accounts for approximately 27,848 deaths per 100,000 people, while globally, it contributes to around 19.39% of all deaths. CVD was ranked as the leading cause of death in the 2020 WHO report. As the risk of CVD continues to increase, researchers are actively exploring methods to reduce associated mortality rates. [2, 3]. Although numerous diseases impact human health, cardiovascular conditions are among the most widespread [4].

The major risk factors for CVD include high blood pressure, high cholesterol, diabetes, smoking, and a family history of the disease [5]. Early detection and preventive measures are essential to mitigate CVD-related issues.

Traditional methods for detecting CVD, such as electrocardiography (ECG) and stress testing, often exhibit limitations in sensitivity and specificity. In contrast, deep learning has emerged as a promising approach for CVD detection and prevention, due to its ability to identify complex patterns within large datasets [6].

Numerous studies have demonstrated the utility of artificial intelligence (AI), machine learning, and deep learning in enhancing forecasting and decision-making capabilities. AI refers to any machine or system capable of exhibiting intelligent behavior, which includes mimicking human cognitive functions such as learning, problem-solving, and decision-making [7]. Machine learning allows systems to enhance their performance over time through exposure to data, without the need for explicit programming. This field encompasses various methodologies, including supervised learning, unsupervised learning, and deep learning (DL) [8]. Supervised learning involves training models on labelled data, where the desired outcomes are known, such as an email spam filter that learns from emails classified as spam. In contrast, unsupervised learning identifies patterns in unlabeled data and uncovers hidden structures, such as grouping customers based on their purchase history. Deep learning, which is modelled after the structure and function of the human brain, employs artificial neural networks with multiple layers to process information [9]. This approach is particularly effective for handling complex tasks. Artificial intelligence (AI) is a broad field of research, with machine learning serving as a specific subset that focuses on knowledge acquisition from data. Deep learning, an advanced technique within machine learning, excels at addressing complex tasks[9].

Challenges in AI, such as data dependency, complexity, and accuracy, continue to be addressed across various fields, including healthcare, architecture, smart homes, industrial automation, environmental prediction, and energy management [10]. Deep learning, which processes vast amounts of data, plays a critical role in addressing these challenges through the use of neural networks[11].

According to [12], deep learning algorithms are a powerful subset of machine learning, inspired by the structure and function of the human brain. These algorithms have revolutionized AI by enabling machines to perform complex tasks once reserved for humans. Deep learning relies on artificial neural networks, which are composed of multiple layers of interconnected nodes that process and transform information [13]. These networks are trained on large datasets, adjusting the connections between nodes based on performance, to achieve the desired accuracy [14, 15].

Various deep learning algorithms, such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks, each possess specific strengths and applications. CNNs are particularly effective in image and video recognition, excelling at capturing

1Department of Computer Science, Faculty of Computer Science and IT, Sana'a University, Sana'a, Yemen.
2Department of Computer Science, Faculty of Computer Science and IT, Sana'a University, Sana'a, Yemen.
** Corresponding Author Email: marwa.alhadi@su.edu.ye*

spatial relationships. RNNs are designed to process sequential data, making them ideal for tasks such as language translation and time series forecasting. LSTM networks, a variant of RNNs, overcome the vanishing gradient problem, which makes them suitable for applications like handwriting recognition and anomaly detection [16-18]. As deep learning continues to evolve, it holds tremendous potential for transformative applications, including personalized medicine, where it can analyse medical data to predict disease risk, customize treatment plans, and accelerate drug discovery [19]. The selection of an optimal deep learning algorithm depends on the specific task, data type, and other factors. Therefore, the main contribution of this work are as follows:

1. Create a novel algorithm that enhanced CVD prediction.
2. Achieve high accuracy surpassing traditional methods and other deep learning approaches.
3. Enhance performance metrics by excelling in precision, recall, and AU-ROC, which highlights the model's ability to minimize false positives and false negatives and ensures reliability for clinical applications.
4. Reduce training time and improve interpretability by applying LASSO for feature selection, which effectively reduces dataset dimensionality while preserving key features relevant to CVD prediction, streamlining the training process and enhancing interpretability.
5. Optimize convergence and performance through the use of a diverse set of optimizers (Adam, RMSprop, Adagrad, Adadelta, Adamax, and Nadam).

2. Related Work

This study aims to present a novel approach for predicting cardiovascular disease by utilizing benchmark datasets and advanced deep learning techniques. The methodology outlines the systematic process, encompassing data collection, preprocessing, model design, and performance evaluation.

Several deep learning models have been proposed for cardiovascular disease (CVD) detection, including CNNs, RNNs, and LSTM networks. RNNs are particularly suited for tasks involving sequential data, such as ECG recordings, which can be beneficial for CVD prediction [20]. A synthesis of the referenced studies is presented in Table 1. The combination of different optimization algorithms has been shown to enhance both convergence speed and accuracy in deep learning models compared to traditional optimizers [21].

In [22], an artificial neural network (ANN) with multiple levels of perceptron's and weights was compared with Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees (DT). The findings demonstrate that ANN surpasses these machine learning algorithms in prediction accuracy, although it demands more time for generating predictions.

In [23], the Enhanced Deep Learning-assisted (EDCNN) was proposed to improve heart disease diagnosis by analyzing patient clinical test data. This system, deployed on the Internet of Medical Things (IoMT) platform, employs a deeper architecture and regularization learning approaches to enhance efficiency and assess heart disease risk levels more effectively than conventional methods.

In [24], involved feature grouping using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and the TSA-EDL (Tunicate Swarm Algorithm and Ensemble Deep Learning). The TSA-EDL algorithm, utilized on datasets from the University of California, Irvine (UCI) and various CVD datasets, showed enhancements in performance metrics, including accuracy, recall, specificity, precision, and the likelihood of misclassification error.

In [25], the Cuckoo Search-Based Conv LSTM Classifier was applied to the PIMA dataset to address gaps in manual heart disease prediction and reduce errors. The method achieved approximately 97% accuracy with nine features. It emphasizes the importance of studying feature relationships to ensure accurate predictions.

According to [26] an ensemble model combining RNNs and LSTMs with optimization techniques like Artificial Flora Optimization (AFO) and Modified Artificial Flora Optimization (MAFO) demonstrated high prediction accuracy. This approach addressed challenges in smart frameworks and predictive models, although it may affect prediction time.

While in [27], utilized 918 samples from five independent centres' (Cleveland, Hungarian, Switzerland, Long Beach, and Stalog) with 12 clinical features to develop a model using Sparse Autoencoder and Convolutional Classifier. The model showed that prediction could be impacted by human error and delays, despite its effectiveness.

In [28] , data collected from five different sources was used to build an Oversampled Quinary Feed Forward Network (OQFFN) with 11 medically relevant parameters. This network provided real-time and highly accurate predictions of heart disease probability, though further efforts are needed to enhance intelligent resource allocation.

The research presented in [29] developed a deep learning-based ensemble classifier, SWCDTO, for early heart disease detection. This method, which combines pre-processed data and feature fusion, outperforms other heart disease prediction algorithms in specificity, accuracy, and sensitivity while reducing computational time.

Bootstrapping ensemble strategies improve prediction outcomes and are recommended for information fusion and medical drug recommendations. Advances in AI and IoT have facilitated early detection and treatment through AI-assisted diagnostic models, offering benefits like cost reduction, infection control, and reduced mortality [30].

In [31], an enhanced ensemble learning approach for heart disease prediction was proposed using boosting algorithms such as Gradient Boost, XGBoost, and AdaBoost. Data preprocessing techniques, such as outlier detection and missing value imputation, substantially enhanced model performance, with Gradient Boost achieving the highest accuracy of 92.20%.

In [32], utilized self-attention mechanisms and transformer networks to predict CVD risk, incorporating multiple layers and modified attention processes. While this model shows promise, it still requires performance enhancement and handling of additional data.

Research in [33] introduced the Gradient Squirrel Search Algorithm-Deep Maxout Network (GSSA-DMN) for heart disease detection. This approach, involving data preprocessing, feature selection using Relief, and training with the Gradient Squirrel Search Algorithm, achieved high accuracy, specificity and sensitivity.

The new Set of Convolutional Neural Network (HCNN) design for heart disease prediction demonstrated its ability to extract detailed features and identify minor trends in cardiovascular health datasets. The HCNN's high accuracy and predictive power suggest significant potential for improving patient outcomes and healthcare decisions [34].

The DeepVAQ model, designed to predict vascular access quality from Photoplethysmography (PPG) sensors, achieved an accuracy of 0.9213 and a precision of 0.9614, surpassing traditional models such as Decision Tree, Naive Bayes, and K-Nearest Neighbours (KNN). This progress in non-invasive diagnostics has the potential to enhance patient outcomes and reduce mortality rates[35].

This study aims to advance CVD prediction by introducing a robust and effective solution with a focus on feature selection, data balancing, and prediction accuracy. The proposed model presents an innovative strategy for disease prediction by dealing with Lasso feature selection techniques, SMOTE data balancing,

Table 1. Cardiovascular Disease Prediction Using Deep Learning

Ref.	Dataset	Contribution	Pros.	Limitations	Feature taken	Model accuracy
[22]	Set of datasets collected from Kaggle	ANN which has various levels and each level has various perceptions' output depends on the weight	Comparing machine learning and ANN. ANN is more accurate than machine learning.	It may take more time for prediction.	14	85.24
[23]	Heart Disease (Hungary, Cleveland, Switzerland, and Long Beach.)	Enhanced Deep learning assisted Convolutional Neural Network (EDCNN)	EDCNN, an IoMT platform, can predict heart disease risk with high accuracy, improving efficiency and reducing manual feature engineering. It also enables remote prediction from wearable devices and potential scalability for large patient populations.	The EDCNN-IoMT platform faces challenges in data security, interpretability, training data bias, limited availability, and technical expertise. Its accuracy depends on the quality and diversity of training data, and its implementation may require specialized expertise in certain healthcare settings.	14	93.2
[24]	Two datasets (Cleveland University of California Irvine (UCI) and CVD)	Set of TSA-EDL (Set of Tunicate Swarm Algorithm and Ensemble Deep Learning)	HEDTSA enhances heart disease predictions by combining deep learning's pattern learning with TSA's optimization, reducing noise and feature selection, and reducing the risk of overfitting to a specific dataset.	HEDTSA is a complex, computationally expensive, data-dependent, and limited explainable deep learning model that faces challenges in implementation, interpretation, and explainability, particularly in the medical field.	14	(97.5%) in UCI.
[25]	PIMA dataset	Cuckoo Search-Based Conv LSTM Classifier	Solve the problem of the manual detection process is found. This process can be time-intensive and may lead to detection errors that impact diagnostic accuracy.	Number of features taken may affect on the final prediction while adding studying the relationship between features could increase the assurance of this method more than before.	9	97.591%,95.874%, and 97.094% of accuracy
[26]	Three-datasets (Cleveland , Hungarian and Switzerland)	Ensemble RNN and LSTM. Then using of optimization techniques like AFO and MAFO	Handle the lack of smart framework, such systems are unable to manage high-dimensional datasets derived from various data sources during the prediction of heart disease.	Examine the weights of RNN to utilize this model for diagnosing another biomedical research context.	14	97.3
[27]	Sample from (Cleveland, Hungarian, Switzerland, Long Beach, stalog)	Multitasking classifier with CNN (combines the Sparse Autoencoder and the Convolutional Classifier)	Diagnoses may not be fully objective and are prone to human error.	May it take more time to predicate the result.	12	90.09
[28]	Heart Failure Prediction Dataset	Oversampled Quinary Feed Forward Network (OQFFN)	Delivering highly accurate and real-time predictions of heart disease probability.	Greater emphasis is required on the development of intelligent resource allocation strategies for the entire model, both at the edge and on the server.	12	89.25
[29]	Heart Disease (Cleveland)	Feature fusion results are then utilized for heart disease prediction classification through the proposed Social Water Cycle Driving Training Optimization (SWCDTO) ensemble classifier, which combines the driver training-based optimization algorithm with the social water cycle algorithm.	Deep learning enhances heart disease predictions by learning complex data patterns. Ensemble learning reduces reliance on a single model, while feature fusion enhances performance. SWCDTO optimizes ensemble classifier training, resulting in faster convergence and better performance.	Complexity: The approach involves several components (deep learning, ensemble methods, SWCDTO), making it more complex. Computational Cost: Deep learning models frequently necessitate substantial computational resources during training, particularly when handling large datasets. Data Dependency: The effectiveness of the model heavily relies on the quality and size of the training data.	14	95.84%, 94.80%, and 95.36%
[30]	Medical Dataset	Bootstrapping ensemble strategy creates multiple subsets of a single dataset. Voting enhances prediction outcomes. An ensemble learning model is suggested for applications involving information fusion and drug recommendations within the medical field.	Deep learning models can enhance accuracy, discover anomalies earlier, save money, and provide individualized treatment recommendations. They can evaluate enormous volumes of medical data, lowering risk and avoiding unneeded procedures. This technology can also evaluate massive volumes of data to improve diagnosis.	Deep learning models are sophisticated, data-dependent, and difficult to comprehend, causing problems for clinicians. They can also be computationally costly, which raises issues about data security and privacy, particularly when using IoT devices.	11	94.21

Ref.	Dataset	Contribution	Pros.	Limitations	Feature taken	Model accuracy
[31]	cardiovascular diseases (Cleveland, Hungary, Switzerland, and Long Beach.)	ensemble learning approach for heart disease prediction, highlighting the effectiveness of gradient boosting algorithms.	Comprehensive data preprocessing techniques, including outlier detection, data imputation, and normalization, were utilized, significantly enhancing the model's robust performance. The use of ensemble learning methods, improved the predictive performance over individual classifiers.	limit the generalizability of the model to larger and more diverse populations. There is still a potential risk, especially if the model is not properly regularized or if hyperparameters are not well-tuned. Lack of Real-time Application, which is critical for practical healthcare applications.	14	92.20
[32]	Cleveland dataset	Self-attention model with CNN	developed a novel attention-based transformer model which improve accuracy	There is a need to improve performance, particularly in managing situations with limited labeled data.	14	96.51
[7]	cardiovascular diseases, clinical datasets	An ensemble extra tree classifier feature selection based (SVM, KNN , EX, NB, LDA , MLP, and LR)	The study demonstrates that extra tree feature selection improves accuracy in classifiers, reduces feature space, and focuses on important features for heart disease diagnosis, potentially leading to faster training times and better understanding of the disease.	The effectiveness of feature selection can vary depending on the chosen classifier. Some classifiers in the study achieved similar performance with or without feature selection. Removing features, even if they seem irrelevant, might lead to discarding some informative data. The study only investigates the extra tree approach. Other feature selection techniques might be even more effective for certain datasets or classifiers.	5	97
[33]	contains of four databases, such as Cleveland, Hungary, Switzerland, and Long Beach.	Heart disease identification was carried out using the Gradient Squirrel Search Algorithm-Deep Maxout Network (GSSA-DMN).	The GSSA-DMN methodology, which integrates Gradient Descent Optimization with the Squirrel Search Algorithm, demonstrates significant effectiveness in heart disease detection. This approach may enhance training efficiency and accuracy by leveraging data preprocessing and feature selection techniques.	The study on GSSA-DMN lacks transparency and external validation, requiring further research for its effectiveness.	14	93.2
[34]	Heart Disease Dataset (Cleveland)	Set of Convolutional Neural Network (HCNN)	Strong performance across accuracy, highly effective in predicting heart disease. This model could lead to more robust predictions. Flexibility and Generalizability	inner workings difficult to interpret and it may take time.	14	92
[35]	Clinical dataset	The DeepVAQ model utilizes a CNN to analyze data from PPG sensors, identifying particular frequencies and patterns that signify the quality of ventricular arrhythmias (VA).	Deep VAQ, a non-invasive, cost-effective tool, uses Photoplethysmography sensors for predicting VA quality, demonstrating high accuracy and reliability compared to the expensive UDT technique.	Limited generalizability, and specific information on PPG sensor features for VA quality prediction, necessitating further validation on external datasets	6	92.13

and a recurrent neural network (RNN) prediction model, along with a comprehensive set of optimizers.

By utilizing advanced feature selection techniques, the model emphasizes the significance of pertinent features while mitigating the influence of noise. The use of optimization techniques enhances the model's performance across varied datasets. Adopting this model leverages the strengths of various methods and addresses the limitations identified in existing studies, presenting a promising solution for CVD prediction.

3. Materials and Methods

This study aims to present a novel approach for predicting cardiovascular disease by utilizing benchmark datasets and advanced deep learning techniques. The methodology outlines the systematic process followed, from data collection and preprocessing to model design and performance evaluation.

The approach consists of several stages: acquiring cardiovascular datasets, preprocessing to ensure data consistency, applying LASSO for feature selection,

balancing the data with SMOTE, then training and prediction a recurrent RNN optimized with cutting-edge algorithms. To enhance the accuracy and reliability of the novel deep learning model for cardiovascular disease prediction, mathematical proofs are employed to validate the efficacy of this work starting from feature selection and preprocessing methods until predication. It also provided for the effectiveness of the standard scaler in normalizing features, which is crucial for maintaining model performance. The methodology is thoroughly designed and detailed as follows:

3.1. Data Collection and Data Preprocessing

Data were sourced and collected from four widely recognized benchmark datasets: Hungary, Cleveland, Long Beach V, and

Switzerland. These datasets are pivotal to cardiovascular disease (CVD) research, offering diverse patient records with various features related to CVD health. They were aggregated to form a comprehensive dataset, establishing a robust foundation for training and evaluating of the proposed model. The cardiovascular disease dataset was obtained from Kaggle's web platform (https://www.kaggle.com/datasets/Johnsmith88/heart-disease-dataset). It comprises 1,025 instances and 14 attributes, with data collected in 1988. The "target" field represents the presence or absence of cardiac disease, where 0 signifies no disease and 1 signifies the presence of disease. The features included in the dataset are detailed in Table 2. Following the aggregation of these datasets, data preprocessing becomes the subsequent step, which plays a critical role in refining the raw data for model training. This process is described as follows:

- Handling Missing Values: Missing or incomplete data entries were addressed by removing affected records, thereby ensuring the integrity and accuracy of the dataset for further analysis.
- Normalization: Features were normalized to guarantee that all variables contributed equally toward the model. This step was essential to prevent any single feature from disproportionately influencing the model due to differences in scale.

Table 2. Descriptions Of Dataset's Attributes

No	Feature Code	Feature Description	Value
1.	age	Age	29-77
2.	sex	Gender	1=male 0=female
3.	cp	chest pain types	0=Atypical angina, 1=typical angina, 2=asymptotic, 3=non angina pain
4.	trestbps	Resting blood pressure	94-200
5.	Chol	serum cholesterol in mg/dl	126-564
6.	Fbs	fasting blood sugar , <=120 mg/dl normal >=120 not normal	0=false 1=true
7.	restecg	resting electrocardiographic results (values 0,1,2)	0=normal 1=ST-T wave abnormalities 2= left ventricular Hypertrophy
8.	Thalach	maximum heart rate achieved	71-202
9.	Exang	exercise induced angina	0=no 1=yes
10.	Oldpeak	ST depression induced by exercise related to rest	0.0-6.2
11.	Slope	the slope of the peak exercise ST segment	0= un sloping 1=flat 2=down sloping
12.	Ca	Count of main vessels (0-3) colored by Fluoroscopy	0-3
13.	Thal	Thallium Scan	0 = normal; 1 = fixed defect; 2 = reversable defect
14.	Target	Class Attribute	0=no 1=yes

The dataset comprises 14 numerical attributes, which are summarized in Table 2. The age attribute ranges from 29 to 77 years. Epidemiological studies have consistently shown that the incidence of heart disease is notably low among individuals over 65 years of age [36, 37]. Gender is coded as 1 for male and 0 for female; evidence suggests that males have a higher likelihood of developing cardiovascular disease compared to females. However, female patients with diabetes are at a higher risk of cardiovascular disease than their male counterparts. The dataset includes four types of chest pain: non-anginal pain, atypical angina, asymptomatic and typical angina. Typical angina results from reduced blood circulation to the cardiac muscle due to a shortage of oxygen-rich blood, while atypical angina is triggered by through or psychological stress. Asymptomatic pain does not signify a cardiac condition. The TRESTBPS attribute, measured in mmHg, represents an individual's resting blood pressure. Serum cholesterol denotes the total cholesterol content in the blood, with Low-Density Lipoprotein (LDL) known as "bad cholesterol" that contributes to arterial narrowing, and High-Density Lipoprotein (HDL), or "good cholesterol," which decreases the likelihood of heart attacks. The Fasting Blood Sugar (FBS) attribute indicates an individual's blood sugar level, with a value of 0 if FBS is less than 120 mg/dl and 1 if it exceeds 120 mg/dl. Elevated blood sugar levels due to improper insulin response are correlated with a significant risk of cardiovascular disease. The Resting Electrocardiogram (RESTECG) results are recorded as 0 for normal, 1 for abnormal ST-T wave, and 2 for left ventricular hypertrophy. The Maximum Heart Rate Achieved (MHR) reflects the highest heart rate attained. Exercise-Induced Angina (EIA) is recorded as 0 if absent and 1 if present, with angina typically manifesting as pain in the chest or shoulders. An Oldpeak value represents the ST-segment depression during an exercise stress test compared to the resting state. A value of 0.0 mm is considered normal, indicating no significant ischemia. However, any depression greater than 0 mm, typically ranging from 0.1 to 6.2 mm, is generally considered abnormal. The slope of the peak exercise ST segment during an exercise stress test provides valuable insights into heart function under physical exertion. An upsloping ST segment (value of 0) is generally considered normal and less indicative of ischemia, though further evaluation may sometimes be needed. A flat ST segment (value of 1) is typically considered abnormal, suggesting possible ischemia or inadequate blood flow to the heart. A down-sloping ST segment (value of 2) is clearly abnormal and is often associated with significant ischemia, indicating a higher likelihood of CVD. Ca (count of major vessels), assessed by fluoroscopy, measures the number of major coronary arteries (ranging from 0 to 3) that are visible or "colored," helping to detect blockages or narrowing. A count of 0 is considered normal, indicating no significant obstruction in the major vessels. However, a count between 1 and 3 is typically abnormal, suggesting that one or more coronary arteries are blocked or narrowed, which may indicate the presence of CVD. he Thal (Thallium Scan) is used to assess blood flow to the heart muscles and detect areas of damage or reduced perfusion. A value of 0 indicates a normal result, meaning there is no significant defect in blood flow. A value of 1 represents a fixed defect, which suggests permanent damage to the heart tissue, often from a previous heart attack. A value of 2 indicates a reversible defect, meaning blood flow is temporarily reduced during stress but returns to normal at rest, which may signal ischemia or coronary artery disease. Both values 1 and 2 are considered abnormal and indicative of potential heart issues. Finally, the "Target" attribute is the class label, with a value of 1 representing the existence of cardiovascular disease and 0 representing its absence.

To validate the integrity of the combined dataset, it is crucial to verify that the aggregation process maintains the statistical properties of the original datasets. Specifically, the mathematical proof for data aggregation is detailed below:

- Let X_{All} denote the integrated with n samples and m features.
- Each sample x_i where i a range from 1 to n as $x1, x2, x3 \ldots$

- Each feature represented as $x_{ij}$where j range from 1 to m as x_{11}, x_{12}
- Let M denote a missing value, if x_{ij} is missing value then $(i,j) \in M$
- While the cleaned dataset mentioned as X therefore:

$$X = X_{All} \setminus \{rows\ where\ (i,j) \in M\} \quad (1)$$

$$X = x_{ij}\ \ where\ \forall j\ (i,j) \notin M \quad (2)$$

The aggregation and preprocessing of the datasets were rigorously validated. These proofs will confirm that the combined dataset retains the statistical properties of the original datasets and that the data cleaning methods effectively address missing values, thereby maintaining the integrity of the dataset.

3.2. Dataset Splitting

The dataset was divided into training and testing subsets, with a split of 80:20. The training set was utilized for model fitting, while the testing set provided an unbiased evaluation of model performance. Common splits include 80/20 or 70/30, with 80/20 cross-validation applied here for more robust estimation[38, 39]. Of the 1025 samples, total of 820 instances were designated for training the model, while the remaining 205 instances were set aside for testing to evaluate the model work.

- Let the overall number of samples in the dataset be denoted by n,where n=1025.
- **Training set size:**

$$n_{train} = 0.80 * n = 0.80 * 1025 = 820\ instances \quad (3)$$

- **Testing set size:**

$$n_{Test} = 0.20 * n = 0.20 * 1025 = 205\ instances \quad (4)$$

Thus, out of 1025 total samples, 820 are used for training, and 205 are used for testing. This division ensures the model is trained on the common of data while maintaining a portion for unbiased evaluation.

3.3. Feature Selection

To improve model performance and decrease computational complexity, feature selection was conducted using the Least Absolute Shrinkage and Selection Operator (LASSO) regression [40]. LASSO is particularly effective in selecting a subset of relevant features by penalizing the coefficients of less important variables, thus reducing the number of dimensions in the dataset while retaining the most critical features for cardiovascular disease prediction [41].Before feeding the dataset into the RNN model, LASSO feature selection was applied to determine the most significant features, enhancing both the model's interpretability and performance.

LASSO regression is a technique used for regularization, where the alpha parameter controls the strength of the penalty applied to the coefficients. In this study, the alpha parameter was set to 0.1. A higher alpha value forces more coefficients to shrink to zero, leading to the exclusion of less important features. The model fitting was performed on the training data n_{train} (features) and n_{Test} (target labels). During training, LASSO estimates the weights (coefficients) of the features, with non-zero weights indicating important features. Features with zero coefficients are excluded from the model. This approach mitigates overfitting by focusing on the most relevant features, thereby improving model performance[42]. Knowing that lasso is used for minimize the cost and it clarified as below:

$$selected_{feautres} = minimize_{\beta}\left(\frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 + a\sum_{j=1}^{m}|\beta_j|\right) \quad (5)$$

Where:

- $n\ is\ the\ number\ of\ samples,$
- $m\ is\ he\ number\ of\ features,$
- $x_i\ is\ the\ feature\ vector\ for\ sample\ i,$
- $y_i\ is\ the\ target\ for\ smaple\ i$
- $\beta_j\ is\ the\ coeffecinet\ for\ feature\ j$
- $a\ is\ the\ regularitzation\ paprpmenter\ based\ on\ the\ toatl\ number\ of\ features$

While $a\sum_{j=1}^{m}|\beta_j|$ which shrinking some coefficients βj towards zero, effectively excluding the less relevant features. When a coefficient $\beta j = 0$, the corresponding feature is excluded from the model, means lasso has identified it excluding the less relevant features.

Therefore, the overall extracted features from data are represented as $selected_{feautres}$ which represents the optimal value of the coefficient vector.

3.4. Data Balancing

While the selected features were identified, the process of prediction still faces the issue of data imbalance, which can lead to inaccurate predictions. To address this challenge of imbalance in the cardiovascular disease dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied following feature selection using LASSO. SMOTE is specifically employed to mitigate class imbalance in datasets, particularly in medical data where there is often a significant disparity between the number of positive (disease-present) and negative (disease-absent) instances. This imbalance can bias the model toward the majority class, leading to skewed predictions [43]. SMOTE is effective in generating synthetic samples for the minority class (e.g., disease-positive cases), thus preventing the model from overfitting to the majority class during training. The technique works by selecting instances from the minority class, identifying their k-nearest neighbors, and generating synthetic data points along the line segments connecting them. This method increases the representation of the minority class without duplicating data, ensuring a more balanced and robust training process [44]. In this work, SMOTE was applied before model training to create a more balanced dataset, enhancing the model's ability to predict both classes accurately and reducing the risk of overfitting. The mathematical formulation of SMOTE is detailed as follows:

Let $X_{train} \in X^{n*m}$represent the training data matrix after applying lasso

- Where n is the number of samples and m is the number of features.

- Let $y_{train} \in \{0,1\}^n$be the corresponding class labels
  - Where:
    - $y = 1$ represents the minority class
    - $y = 0$ represents the majority class

- Given the imbalanced dataset (X_{train}, y_{train}), applying SMOTE to produce synthetic sample for the minority class to balance the dataset.
- Let minority class denoted as
  - ($S_{minority} = \{x_i \in X_{train} : y_i = 1\}$ and
- Let majority class denoted as
  - ($S_{majority} = \{x_i \in X_{train} : y_i = 0\}$

Knowing that, the goal of SMOTE is to generate synthetic samples ($x_{new} \in X^{n*m}$ such that the number of samples in the minority class is equalized to the number of samples in the majority class. This is achieved by interpolating between the minority class samples and their nearest neighbors and nearest neighbor in minority class is clarified as:

- For each sample
  - $x_i \in S_{minority}$ SMOTE
  - selects a random nearest neighbor $x_{nn} \in S_{minority}$
  - Where the distance between x_i and x_{nn} is calculated using a distance matric such as Euclidean distance: $d(x_i, x_{nn}) \sqrt{\sum_{j=1}^{m}(x_{i,j} - x_{nn,j})^2}$ **(6)**

Where $x_{i,j}$ and $x_{nn,j}$ represents the j^{th} the features of x_i and x_{nn},respectively. While synthetic sample generations and work as:

- For each sample $x_i \in S_{minority}$ a new synthetics sample x_{new} is produced by linearly interpolating between x_i and one of its randomly chosen nearest neighbors x_{nn}. The interpolation is given by:
  - $x_{new} = x_i + \lambda.(x_{nn} - x_i)$ **(7)**

Where $\lambda \in [0,1]$ is a random number generated uniformly within the interval $[0,1]$.
It repeated until the number of synthetic samples generated equal the difference in size between majority and minority classes therefore Data Balancing will be:

- ($X_{balanced\ train} = X_{train} \cup \{x_{new}\}$
- $y_{balanced\ train} = y_{train} \cup \{1,0,\dots 1\}$

$(X_{balanced\ train}, y_{balanced\ train}) \leftarrow SMOTE(X_{train}, y_{train})$

3.5. Recurrent Neural Network (RNN)

Several deep learning models have been applied for CVD detection, including CNNs, RNNs, and LSTMs. While RNNs are highly suitable for tasks that involve sequential data, making them effective for processing and analyzing such data [20]. This research utilizes an RNN to model the relationships present in cardiovascular data. After feature selection using LASSO and data balancing with SMOTE, the selected balanced features are fed into the RNN. The choice of RNN is driven by its ability to effectively process sequential data, which is crucial for accurate prediction[20].
The RNN architecture comprises multiple layers. The model is trained using a hybrid optimization approach that utilizes multiple optimization algorithms, ensuring efficient convergence and minimizing the risk of overfitting.
The RNN generates predictions based on historical data, enhancing its ability to forecast cardiovascular disease with high accuracy. This approach leverages temporal patterns, which are essential for predicting outcomes in time-sensitive medical data. The mathematical formulation used for RNN model with optimizer utilization algorithms to predict cardiovascular disease is as follows:

- Let $X_{balanced\ train} \in X^{n*m}$ represent the training selected balanced data
- Let the model output is as $y_{balanced\ train} \in \{0,1\}^n$

While the architecture of RNN model can be described as follows:

1. **Input layer:** the input data at time step t is $x_i \in X^{n*m}$.

   Where:
   - n is the number of samples and
   - m is the number of features after feature selection
2. **Simple recurrent neural network layer**: the hidden state h_t at time t in the simple RNN layer with k hidden units is calculated as:
   - $h_t = \sigma(W.x_t + U.h_{t-1} + b)$ **(8)**

   Where:
   - K refers to the number of neurons (or units) in the hidden layer
   - $W \in X^{k*m}$ is the weight matrix for input,
   - $U \in X^{k*k}$ is the weight matrix for the recurrent connection,
   - $b \in X^k$ is the bias term,
   - σ is the Sigmoid activation function
   - h_{t-1} is the hidden state from the previous time step.

The output of this layer is the hidden state vector at the last time step.

3. **Dense output layer:** the final output $\hat{y}$ is calculated using a dense layer with a sigmoid activation function for binary classification:
   $\hat{y} = sigmoid(W_{out}.h_T + b_{out})$ **(9)**
   Where:
   - $W_{out} \in X^{1*k}$ is the weight matrix for the dense layer,
   - $b_{out} \in X$ is the total number of time steps.

While the sigmoid function is defined as:

$$sigmoid(z) = \frac{1}{1+e^{-z}} \textbf{ (10)}$$

This output $\hat{y}$ represents the predicted probabilities of the samples belonging to class 0 or class 1, and to enhance model performance while minimizing potential loss functions, an optimizer is utilized.

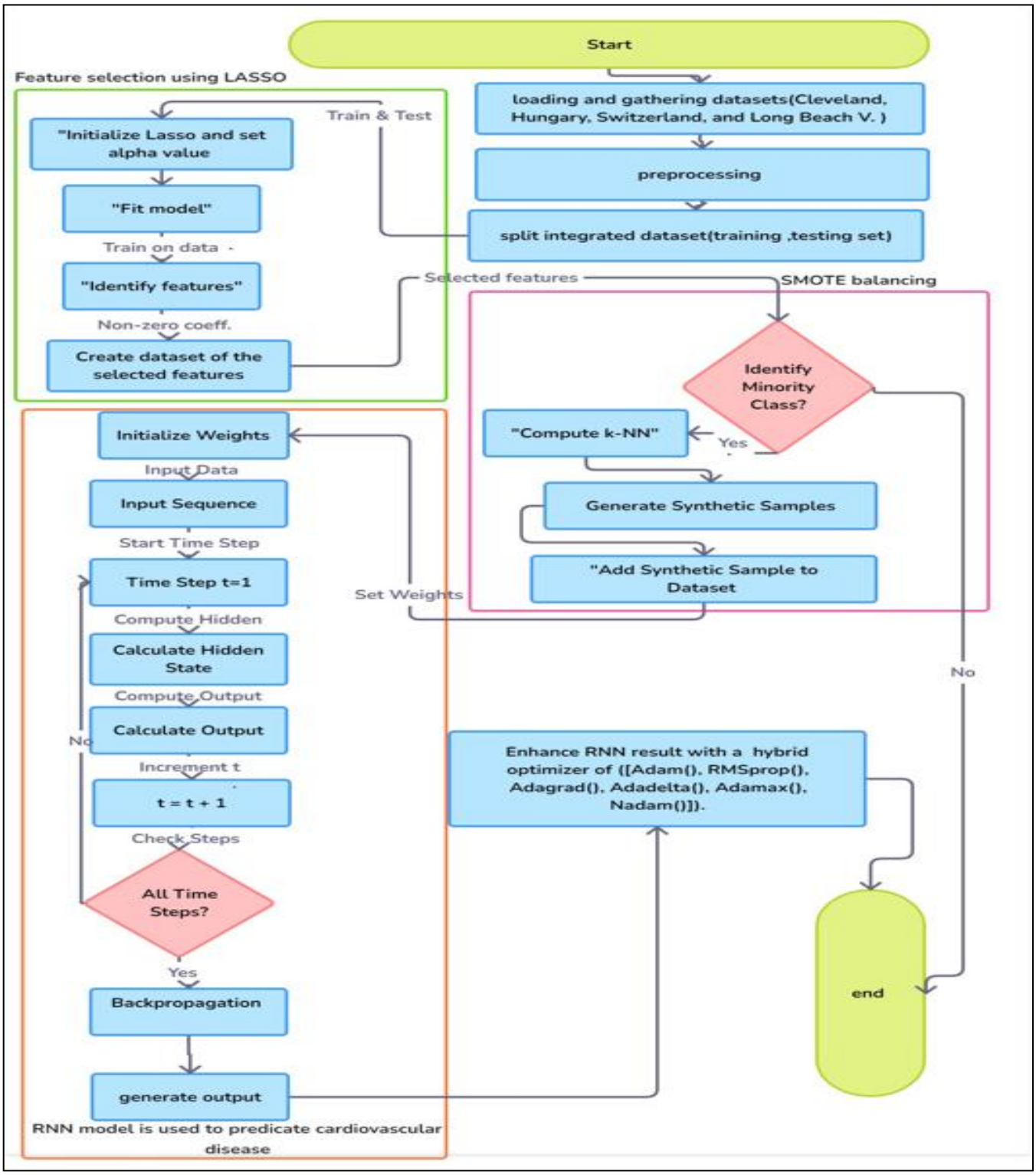


Fig. 1. MarCDP Deep Learning Model for Cardiovascular Disease Prediction

Therefore, the model is compiled and trained using multiple optimizers to evaluate their performance via:

Let the set of optimizers be denoted by $O = \{Adam, RMprop, Adagrad, Adadleta, Adamax, Nadam\}$, while the optimization problem aims to minimize the loss function $L(y, y^{\wedge})$.

$$L_{(y,y^{\wedge})} = -\frac{1}{n}\sum_{i=1}^{n}[(y_i \log(y_i^{\wedge})) + (1 - y_i)\log(1 - y_i^{\wedge})] \quad \textbf{(11)}$$

Where:

- y_i is the ground truth for the i[th] sample
- $y_i^{\wedge}$ is the predicated probability for sample i,
- n is the total number of samples.

For each optimizer $o \epsilon O$, the model is trained by updating the weights W and biases b using gradient-based methods.

Mathematically, the weight update rule for a given optimizer is:

$$W^{(t+1)} = W^{(t)} - \iota \nabla w L(y, y^{\hat{}}) \quad (12)$$

where:

- ι is the learning rate(optimizer-specific),
- $\nabla w L(y, y^{\hat{}})$ is the gradient of the loss function with respect to the weights.

The difference among optimizers lies in how ι and the gradient updated are computed for each optimizer.

3.6. MarCDP Deep Learning Model for Cardiovascular Disease Prediction

Marwa Cardiovascular disease predication (MarCDP) Deep Learning Model for Cardiovascular Disease Prediction is designed to enhance predictive accuracy and model performance through a systematic and scientifically grounded approach. This model incorporates several key components, each playing a vital role in improving effectiveness in identifying cardiovascular disease cases.

The model begins with LASSO for feature selection, with an alpha parameter set to 0.1. By performing regularization and dimensionality reduction, LASSO removes irrelevant or redundant features, ensuring that only the most significant variables are included. This reduces overfitting and improves the generalization of the model, which is essential for handling the complexity of real-world cardiovascular disease data. By streamlining the dataset, LASSO helps focus the model's attention on critical features, thus enhancing prediction accuracy.

Following feature selection, the SMOTE is applied to tackle class imbalance in the dataset. Cardiovascular disease datasets often exhibit a higher number of negative (disease-free) cases compared to positive (disease-present) ones. SMOTE generates synthetic samples for the minority class (positive cases) by interpolating between existing data points. This improves the balance between classes and ensures that the model receives sufficient training on both positive and negative cases. By addressing class imbalance, SMOTE reduces bias towards the majority class, thereby enhancing the model's ability to accurately predict minority (disease-present) instances.

The Recurrent Neural Network (RNN) at the core of the MarCDP model is well-suited for handling sequential data, which is particularly beneficial in analyzing time-dependent cardiovascular data (e.g., patient history, diagnostic metrics over time). The RNN architecture can capture patterns, leading to more nuanced predictions. By learning dependencies in the input sequence, the RNN enhances the model's capacity to understand complex relationships between features, thus improving overall prediction performance.

The utilization of optimizers further boosts the performance of the MarCDP model by improving the efficiency of the training process. These optimizers combine the strengths of multiple optimization techniques, ensuring faster convergence, reduced error, and better fine-tuning of the model parameters. As a result, the model can achieve a higher level of precision and recall, which are critical metrics in disease prediction tasks.

MarCDP Deep Learning Model leads to a more robust model that generalizes well to unseen data, providing more accurate, reliable, and timely predictions for cardiovascular disease. By addressing key challenges like feature redundancy, class imbalance, and complex data patterns, the MarCDP model offers a significant improvement over traditional models, making it a valuable tool for medical diagnosis and risk assessment. The proposed model is presented as algorithm in Table 3 and Figure 1.

4. Evaluation Criteria's

The model was trained on the identified, selected, pre-processed and balanced dataset. The training process involved adjusting the weights of the RNN using the set of optimizer strategy. To assess the model's performance, a range of evaluation metrics was employed, including accuracy, precision, recall, and Receiver Operating Characteristic Area Under Curve(ROC-AUC) [45]. It will be employed to assess the performance of the proposed model. The proposed model employs the test-holdout approach for validation, utilizing eighty percent of the data for training and reserving twenty percent for testing, in accordance with the 80-20 train-test validation method. The sklearn library is used for evaluation process as it clarified in Table 4.

Table 3. Pseudo Code For MarCDP Deep Learning Model For Cardiovascular Disease Prediction

```
Start
  //1.data collection and data preprocessing as in (equation (1,2))
loading and gathering datasets
preprocessing
//2.data splitting (equation (3,4))
Split integrated dataset (training, testing set)
//3. feature selection as in (equation (5))
Initialize lasso and set alpha value
Fit model
Identify features
Create dataset of the selected features
//4.data balancing as in (equation (6,7))
Identify minority class
if it's a minority class
   Compute k-NN
   Generate synthetic samples
   Add synthetic sample to dataset
//5.RNN as in (equation (8,9))
  Initialize weight
  Input sequence
  Start time step t=1
  Calculate hidden state
  Calculate Output
   t=t+1
   If all time steps
      Back-propagations
      Generate output
//6. Enhance RNN as in (equation (10,11,12))
      Enhance RNN result with optimizer
      Calculate activation function
      Calculate loss function
End
```

Table 4. Metrics Used to Evaluate the Performance of the Proposed Model

Metric	Equation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FN}$
Recall	$\frac{TP}{TP + FN}$
AU-ROC	$\frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FN}\right)$

5. Results And Discussion

The proposed healthcare prediction model, which utilizes RNN with LASSO feature selection, SMOTE data balancing, and a set of optimizer strategy, was evaluated using four benchmark datasets: Cleveland, Hungary, Switzerland, and Long Beach V. Feature selection is a crucial step in improving the performance and interpretability of deep learning models. In this study, the LASSO was utilized to identify the most relevant features for cardiovascular disease prediction. LASSO not only helps in reducing the dimensionality of the dataset by eliminating irrelevant or redundant features, it also enhances model performance by focusing on the most significant predictors. Table 5 provides a detailed summary of the features selected by LASSO versus those that were excluded. This information highlights the key features used in the model and offers insights into which variables contribute most to predicting cardiovascular disease. Knowing that, understanding the relative importance of different features is critical for interpreting the performance and results of predictive models. In this study, using of LASSO regression technique to identify and evaluate the significance of each feature in predicting cardiovascular disease is a critical point of success. LASSO helps in feature selection by applying the coefficients of less important features, effectively shrinking them towards zero and thereby highlighting the most impactful features.

In Figure 2, it illustrates the importance of features based on the coefficients obtained from the LASSO model. Each bar represents a feature, with its length corresponding to the magnitude of its coefficient. Features with larger coefficients are deemed more significant in the prediction process. This visualization provides intuitive way to assess which features are most influential, offering valuable insights into the underlying factors contributing to cardiovascular disease predictions. It visualizes the importance of different features in a deep learning model, as determined by their LASSO coefficients. Understanding that, LASSO is a regularization technique that tends to shrink less important feature coefficients towards zero. The features cp and old peak have the highest absolute coefficient values, indicating their significant influence on the model's predictions. The features thalach, thal, exang, and ca have moderate coefficient values. The feature sex has the smallest coefficient value, indicating its minimal impact on the model's predictions. Based on the LASSO coefficients, the model suggests that cp and oldpeak are the most critical factors influencing the target variable. These features likely contribute significantly to the model's ability to make accurate predictions.

Table 5. Feature Selection Results

No	Feature Code	Selected
1.	age	No
2.	sex	Yes
3.	cp	Yes
4.	trestbps	No
5.	Chol	No
6.	Fbs	No
7.	restecg	No
8.	Thalach	Yes
9.	Exang	Yes
10.	Oldpeak	Yes
11.	Slope	No
12.	Ca	Yes
13.	Thal	Yes

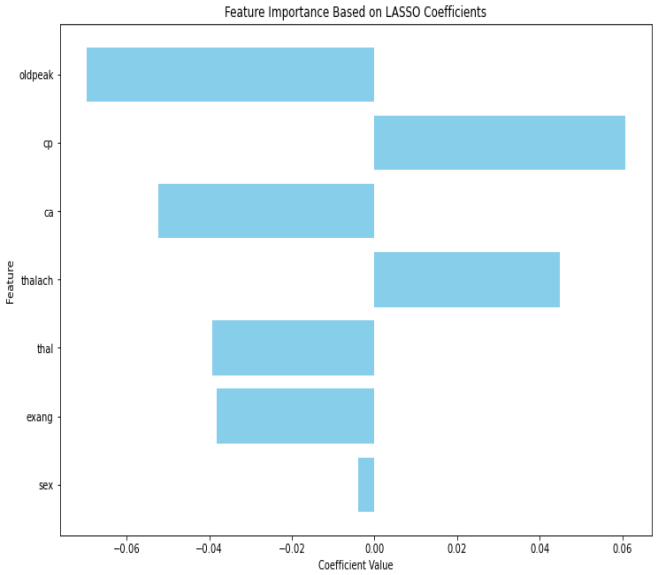


Fig. 2. Coefficient value for each feature selected

To address the class imbalance in the cardiovascular disease datasets, the SMOTE was applied. Class imbalance can significantly impact the performance of deep learning models, leading to biased predictions where the minority class may be underrepresented. SMOTE works by generating synthetic samples for the minority class, thus balancing the dataset and improving the model's ability to generalize. Table 6 illustrates the distribution of classes before and after the application of SMOTE. This balancing act ensures that the model is not biased toward the majority class and provides a more accurate evaluation of its performance across different classes as clarified in Table VI. Figure 3 presents comparing the distribution of classes before and after applying the SMOTE algorithm. The Class Distribution Before SMOTE shows an imbalance in the dataset, with a significantly higher number of instances in one class (likely class 0) compared to the other class (class 1). The Class Distribution After SMOTE demonstrates that the SMOTE technique has successfully balanced the dataset by increasing the number of instances in the minority class (class 1).

Table 6. SMOTE Balancing Results

CLASS	BEFORE SMOTE	AFTER SMOTE
CLASS 0	800	800
CLASS 1	200	800

The application of SMOTE has effectively addressed the class imbalance issue in the dataset. This is generally considered beneficial for deep learning models, as imbalanced datasets can lead to biased models that perform poorly on the minority class.

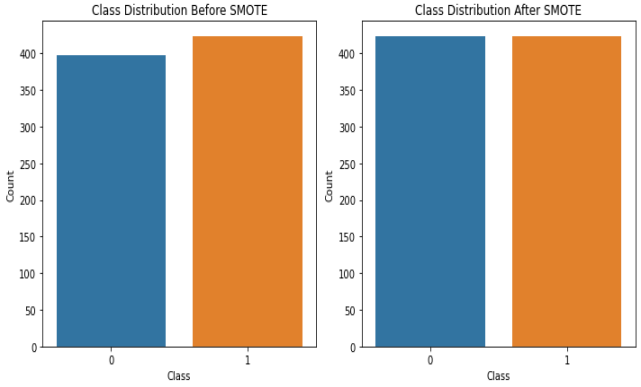


Fig. 3. Class distribution using SMOTE

The training and validation loss/accuracy curves is essential for understanding how well a model is learning and generalizing over time. These curves provide insights into the model's performance during training and help identify potential issues such as overfitting or underfitting. Training Loss/Accuracy Curves plots show the progression of the model's performance on the training dataset over epochs. The training loss curve reflects how well the model is minimizing the error on the training data, while the training accuracy curve indicates how accurately the model classifies the training samples. Validation Loss/Accuracy Curves track the model's performance on a separate validation dataset, providing an estimate of how well the model will perform on data. A decreasing validation loss and increasing validation accuracy proposed that the model is learning to generalize well. Figure 4 illustrates the loss and accuracy curves for different optimizers used in training RNN. These visualizations help assess the effectiveness of each optimizer in improving model performance and ensuring robust learning. Training Loss decreases steadily over epochs, indicating the model is learning from the training data.

Table 7. Performance Metrics

Metric	Proposed Model
Accuracy	98.05%
Precision	99.01%
Recall	97.09%
ROC AUC	99.83%

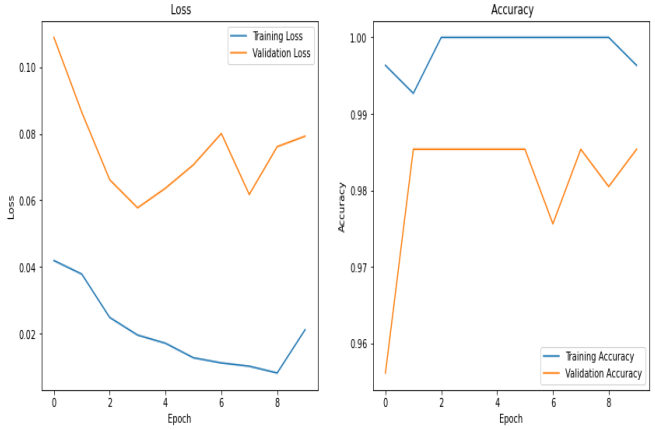


Fig. 4. coefficient Training loss, validation loss, training accuracy and validation accuracy

Validation Loss shows a similar trend to training loss, though with slight fluctuations. This suggests that the model is generalizing well to data. No Overfitting and no gap between training and validation loss is relatively small, indicating a low risk of overfitting. Training Accuracy steadily increases, reaching a around 0.99, suggesting the model is learning to classify training samples accurately. Validation Accuracy is Also increases and stabilizes, with a slight gap compared to training accuracy. This indicates good generalization performance.

Confusion matrix is also used to visualize the performance of this model in classifying different classes. Visualizing the results enables a deeper understanding of the model's strengths and limitations in predicting cardiovascular disease cases. The following heatmap displays the confusion matrix for each optimizer used, highlighting how the model's predictions compare to the actual outcomes as it described in Figure 5. The model correctly classified 102 + 100 = 202 out of 205 instances, resulting in an accuracy of 98.5%. The dataset appears to be balanced, with approximately equal numbers of instances in Class 0 and Class 1. While Misclassifications is represented as 0 instances of Class 0 were incorrectly predicted as Class 1 (false positives), 3 instances of Class 1 were incorrectly predicted as Class 0 (false negatives). The model correctly identified 102 out of 102 true negative instances (Class 0), resulting in a specificity of 100%. The model correctly identified 100 out of 103 true positive instances (Class 1), resulting in a sensitivity of 97.09%. Overall, the model demonstrates high accuracy, specificity, and sensitivity, suggesting it performs well in classifying both Class 0 and Class 1 instances. The evaluation metrics are clarified in Table 7. It demonstrates the model's strong performance in predicting cardiovascular diseases.

As clarified in Table VII, the model achieved an accuracy of 98.05%, indicating that nearly all predictions made by the model were correct. This high level of accuracy highlights the effectiveness of the model in distinguishing between patients with and without cardiovascular disease, making it a reliable tool for clinical decision-making. While with a precision score of 99.01%, the model showed an exceptional ability to correctly identify positive instances of cardiovascular disease while minimizing false positives. This is critical in a healthcare setting where false can lead to unnecessary interventions and increased healthcare costs. Whereas the recall score of 97.09% reflects the model's ability to correctly identify the majority of actual cases of cardiovascular disease as it clarified in Figure 6. A high recall is crucial for ensuring that patients with the disease are accurately diagnosed, thus reducing the likelihood of missed diagnoses that could lead to adverse outcomes. In addition, using of ROC AUC underscores its superior discriminatory power in distinguishing between healthy and diseased patients. The model's ROC AUC score of 99.83% as it clarified in Figure 7. This near-perfect score indicates that the model performs exceptionally well across various thresholds, making it highly effective in clinical scenarios where both sensitivity and specificity are vital.

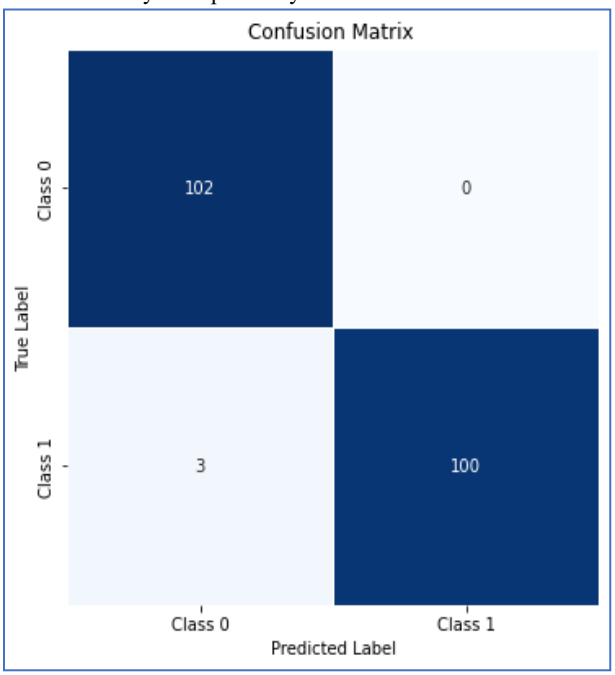
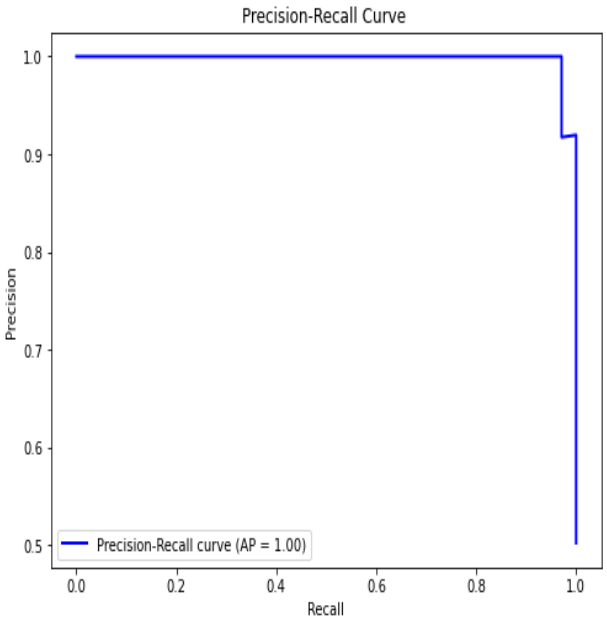


Fig. 5. confusion matrix of model performance



Fi

Fig.6. precision and recall accuracy curve

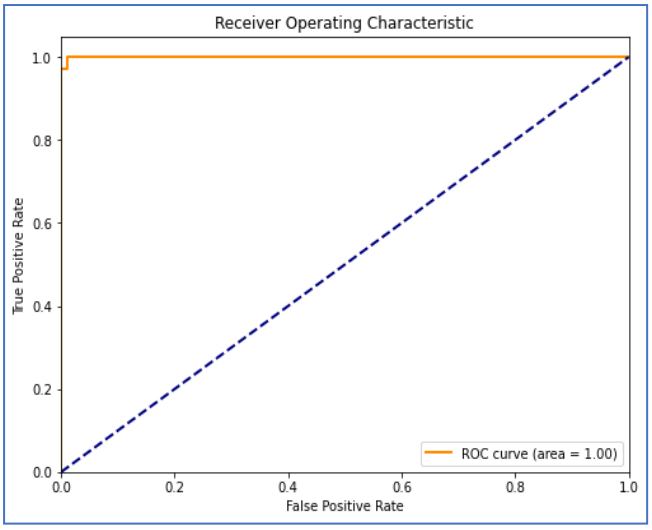


Fig7. Roc Model performance

Evaluated the Precision-Recall (PR) curve provides more nuanced view of a model's performance, particularly in the context of class imbalance. Precision represents the proportion of true positive predictions among all positive predictions made by the model, while recall (or sensitivity) represents the proportion of true positives among all actual positive cases. By plotting these metrics against each other at various thresholds, the PR curve offers insights into the trade-off between precision and recall across different decision boundaries. This is especially valuable in medical applications where both false positives and false negatives have significant implications.

The ROC curve represents, true Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The ideal point for an ROC curve is at the top-left corner (TPR=1, FPR=0), indicating perfect classification. The diagonal line represents the performance of a purely random classifier.

Analysis of the Provided ROC Curve:

- Perfect Classification: The ROC curve is a straight horizontal line at the top (TPR=1) with an area under the curve (AUC) of 1.00. This indicates a perfect classifier.
- No False Positives: The curve lies along the y-axis, meaning there are no false positives (FPR=0) at any threshold setting.
- All True Positives: The curve reaches the top-left corner, indicating that all true positives are correctly identified.

Therefore, This ROC curve demonstrates an exceptional model performance. It perfectly discriminates between the positive and negative classes without any errors. Such a result is often seen in ideal scenarios or with perfectly separable datasets.

Table 8. Comparing Proposed Study to Previous Studies.

Ref.	Dataset	Feature taken	Model accuracy
[22]	Set of datasets collected from Kaggle	14	85.24
[23]	Heart Disease (Cleveland, Hungary, Switzerland, and Long Beach.)	14	93.2
[24]	Two datasets (Cleveland University of California Irvine (UCI) and cardiovascular disease (CVD))	14	(97.5%) in UCI.
[25]	PIMA dataset	9	97.591%,95.874% , and 97.094% of accuracy
[26]	Three-datasets (Cleveland , Hungarian and Switzerland)	14	97.3
[27]	Sample from (Cleveland, Hungarian, Switzerland, Long Beach, stalog)	12	90.09
[28]	Heart Failure Prediction Dataset	12	89.25
[29]	Heart Disease (Cleveland)	14	95.84%, 94.80%, and 95.36%
[30]	Medical Dataset	11	94.21
[31]	cardiovascular diseases (Cleveland, Hungary, Switzerland, and Long Beach.)	14	92.20
[32]	Cleveland dataset	14	96.51
[7]	cardiovascular diseases, clinical datasets	5	97
[33]	comprises of four databases, such as Cleveland, Hungary, Switzerland, and Long Beach.	14	93.2
[34]	Heart Disease Dataset (Cleveland)	14	92
[35]	Clinical dataset	6	92.13
Proposed work(MarCDP)	Four datasets (Cleveland, Hungary, Switzerland, and Long Beach)	14	98.05

Comparing MarCDP to other studies that have the same number of features and same datasets as it represented in Table 8, the proposed work presents a novel work with higher predication result as 98.5 model accuracy. The compared studies as [23], [31], and [33] with the proposed work, MarCDP presents a clear enhancement in accuracy as it presents in Figure 8. Additionally, the model incorporates a set of optimizer strategy to further enhance prediction performance.

This novel approach addresses the gaps in the existing models by enhancing feature selection, improving data balance, and

achieving higher prediction accuracy, thereby offering a more robust and reliable healthcare prediction solution.

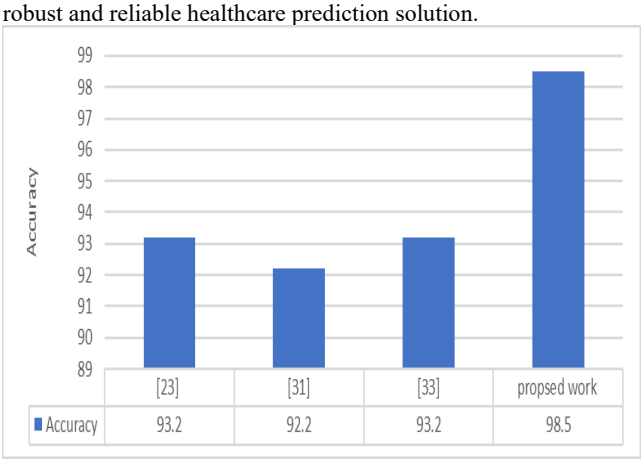


Fig. 8. Comparison of our proposal with the model with the same datasets and features

6. Conclusion And Future Work

This article introduces a novel deep learning model aimed at enhancing the accuracy of cardiovascular disease (CVD) prediction. Through the utilization of Recurrent Neural Networks (RNN) for sequence learning, LASSO for feature selection, and SMOTE for handling data imbalance, the model achieved a significant prediction accuracy of 98.05% on four benchmark datasets and **a precision of 99.01%, a recall of 97.09%, and an ROC AUC of 99.83%**. These results surpass existing methods, demonstrating the model's potential to assist in early detection and prevention of CVD. This work contributes to advancing the field of predictive healthcare, offering a more precise tool for identifying at-risk patients.

Future work will focus on expanding the model's capabilities by incorporating real-time data from wearable IoT devices, allowing for continuous monitoring and dynamic prediction of CVD risk. Additionally, exploring the integration of more sophisticated deep learning architectures and larger, more diverse datasets could further enhance model performance. Another direction involves developing a user-friendly mobile health application that provides real-time CVD risk assessment, making the model more accessible and applicable in clinical practice and incorporate this work with [46, 47]**.**

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] S. S. Martin *et al.*, "2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association," *Circulation,* vol. 149, no. 8, pp. e347-e913, 2024.

[2] W. H. Organization, *World health statistics 2023: monitoring health for the SDGs, sustainable development goals*. World Health Organization, 2023.

[3] G. ZWIELEWSKI, "WORLD HEALTH ORGANIZATION. World health statistics 2022: monitoring health for the," *Gestão de qualidade em saúde: conceitos e ferramentas da qualidade como estratégia de construção e práticas em gestão em saúde,* 2023.

[4] A. Selzer, *Understanding heart disease*. Univ of California Press, 2023.

[5] M. M. Hussain, U. Rafi, A. Imran, M. U. Rehman, and S. K. Abbas, "Risk Factors Associated with Cardiovascular Disorders: Risk Factors Associated with Cardiovascular Disorders," *Pakistan BioMedical Journal,* pp. 03-10, 2024.

[6] P. Branigan *et al.*, "Towards Optimal Cardiovascular Health: A Comprehensive Review of Preventive Strategies," *Cureus,* vol. 16, no. 5, 2024.

[7] H. Abubaker, F. Muchtar, A. R. Khairuddin, A. N. A. Nuar, Z. M. Yunos, and C. Salimun, "Exploring Important Factors in Predicting Heart Disease Based on Ensemble-Extra Feature Selection Approach," *Baghdad Science Journal,* vol. 21, no. 2 (SI), pp. 0812-0812, 2024.

[8] D. Touretzky, C. Gardner-McCune, and D. Seehorn, "Machine learning and the five big ideas in AI," *International Journal of Artificial Intelligence in Education,* vol. 33, no. 2, pp. 233-266, 2023.

[9] M. M. Taye, "Understanding of machine learning with deep learning: architectures, workflow, applications and future directions," *Computers,* vol. 12, no. 5, p. 91, 2023.

[10] L. Alzubaidi *et al.*, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data,* vol. 10, no. 1, p. 46, 2023.

[11] J. Kufel *et al.*, "What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine," *Diagnostics,* vol. 13, no. 15, p. 2582, 2023.

[12] K. Sharifani and M. Amini, "Machine learning and deep learning: A review of methods and applications," *World Information Technology and Engineering Journal,* vol. 10, no. 07, pp. 3897-3904, 2023.

[13] M. M. Taye, "Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions," *Computation,* vol. 11, no. 3, p. 52, 2023.

[14] F. Mehmood, S. Ahmad, and T. K. Whangbo, "An efficient optimization technique for training deep neural networks," *Mathematics,* vol. 11, no. 6, p. 1360, 2023.

[15] L. Jacobs *et al.*, "Enhancing their quality of life: environmental enrichment for poultry," *Poultry science,* vol. 102, no. 1, p. 102233, 2023.

[16] B. Ghojogh and A. Ghodsi, "Recurrent neural networks and long short-term memory networks: Tutorial and survey," *arXiv preprint arXiv:2304.11461,* 2023.

[17] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, CNNs and vision transformers: A review," *IEEE Access,* vol. 11, pp. 35479-35516, 2023.

[18] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," *arXiv preprint arXiv:2305.17473,* 2023.

[19] S. F. Ahmed *et al.*, "Deep learning modelling techniques: current progress, applications, advantages, and challenges," *Artificial Intelligence Review,* vol. 56, no. 11, pp. 13521-13617, 2023.

[20] D. Jalal and A. M. Abdulazeez, "A Review on Heart Disease Detection Classification Based on Deep Learning Algorithm," *Indonesian Journal of Computer Science,* vol. 13, no. 2, 2024.

[21] B. F. Azevedo, A. M. A. Rocha, and A. I. Pereira, "Hybrid approaches to optimization and machine learning methods: a systematic literature review," *Machine Learning,* pp. 1-43, 2024.

[22] S. N. Pasha, D. Ramesh, S. Mohmmad, and A. Harshavardhan, "Cardiovascular disease prediction using deep learning techniques," in *IOP conference series: materials science and engineering*, 2020, vol. 981, no. 2: IOP Publishing, p. 022006.

[23] Y. Pan, M. Fu, B. Cheng, X. Tao, and J. Guo, "Enhanced deep learning assisted convolutional neural network for heart disease prediction on the internet of medical things platform," *Ieee Access,* vol. 8, pp. 189503-189512, 2020.

[24] J. Wankhede, P. Sambandam, and M. Kumar, "Effective prediction of heart disease using hybrid ensemble deep learning and tunicate swarm algorithm," *Journal of Biomolecular Structure and Dynamics,* vol. 40, no. 23, pp. 13334-13345, 2022.

[25] A. Kumar, S. S. Satyanarayana Reddy, G. B. Mahommad, B. Khan, and R. Sharma, "Smart healthcare: disease prediction using the cuckoo-enabled deep classifier in IoT framework," *Scientific Programming,* vol. 2022, no. 1, p. 2090681, 2022.

[26] R. Banoth, A. K. Godishala, R. Veena, and H. Yassin, "A healthcare monitoring system for predicting heart disease through recurrent neural network," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, 2022: IEEE, pp. 1-7.

[27] M. T. García-Ordás, M. Bayón-Gutiérrez, C. Benavides, J. Aveleira-Mata, and J. A. Benítez-Andrades, "Heart disease risk prediction using deep learning techniques with feature

augmentation," *Multimedia Tools and Applications,* vol. 82, no. 20, pp. 31759-31773, 2023.
[28] M. I. A. Hossain, A. Tabassum, and Z. U. Shamszaman, "Deep edge intelligence-based solution for heart failure prediction in ambient assisted living," *Discover Internet of Things,* vol. 3, no. 1, p. 11, 2023.
[29] R. Jayasudha, C. Suragali, J. Thirukrishna, and B. Santhosh Kumar, "Hybrid optimization enabled deep learning-based ensemble classification for heart disease detection," *Signal, Image and Video Processing,* vol. 17, no. 8, pp. 4235-4244, 2023.
[30] M. Venkatachala Appa Swamy *et al.*, "Design and Development of IoT and Deep Ensemble Learning Based Model for Disease Monitoring and Prediction," *Diagnostics,* vol. 13, no. 11, p. 1942, 2023.
[31] S. M. Ganie, P. K. D. Pramanik, M. B. Malik, A. Nayyar, and K. S. Kwak, "An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms," *Comput. Syst. Sci. Eng.,* vol. 46, no. 3, pp. 3993-4006, 2023.
[32] A. U. Rahman, Y. Alsenani, A. Zafar, K. Ullah, K. Rabie, and T. Shongwe, "Enhancing heart disease prediction using a self-attention-based transformer model," *Scientific Reports,* vol. 14, no. 1, p. 514, 2024.
[33] S. Balasubramaniam, C. V. Joe, C. Manthiramoorthy, and K. S. Kumar, "ReliefF based feature selection and Gradient Squirrel search Algorithm enabled Deep Maxout Network for detection of heart disease," *Biomedical Signal Processing and Control,* vol. 87, p. 105446, 2024.
[34] N. A. Karandikar, " Advanced Heart Disease Prediction: Deep Learning-Enhanced Convolutional Neural Network in the Internet of Medical Things Environment," *The Journal of Electrical Systems (JES),* no. 20(1s), pp. 1-10, 2024.
[35] S. Julkaew, T. Wongsirichot, K. Damkliang, and P. Sangthawan, "DeepVAQ: an adaptive deep learning for prediction of vascular access quality in hemodialysis patients," *BMC Medical Informatics and Decision Making,* vol. 24, no. 1, p. 45, 2024.
[36] N. Conrad *et al.*, "Trends in cardiovascular disease incidence among 22 million people in the UK over 20 years: population based study," *bmj,* vol. 385, 2024.
[37] W. Ahmed, T. Muhammad, C. Maurya, and S. N. Akhtar, "Prevalence and factors associated with undiagnosed and uncontrolled heart disease: A study based on self-reported chronic heart disease and symptom-based angina pectoris among middle-aged and older Indian adults," *Plos one,* vol. 18, no. 6, p. e0287455, 2023.
[38] Y. Lyu, H. Li, M. Sayagh, Z. M. Jiang, and A. E. Hassan, "An empirical study of the impact of data splitting decisions on the performance of AIOps solutions," *ACM Transactions on Software Engineering and Methodology (TOSEM),* vol. 30, no. 4, pp. 1-38, 2021.
[39] I. Muraina, "Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts," in *7th international Mardin Artuklu scientific research conference*, 2022, pp. 496-504.
[40] N. Goel, "Optimized Prognostic Models for Oral Cancer Survival using Feature Selection Methods," *Procedia Computer Science,* vol. 235, pp. 1832-1840, 2024.
[41] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access,* vol. 9, pp. 19304-19326, 2021.
[42] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics,* vol. 2, p. 927312, 2022.
[43] 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," *Applied Sciences,* vol. 13, no. 6, p. 4006, 2023.
[44] M. Dubey, J. Tembhurne, and R. Makhijani, "Improving coronary heart disease prediction with real-life dataset: a stacked generalization framework with maximum clinical attributes and SMOTE balancing for imbalanced data," *Multimedia Tools and Applications,* pp. 1-30, 2024.
[45] A. M. Carrington *et al.*, "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 1, pp. 329-341, 2022.
[46] M. A. Al-Hadi, G. H. Al-Gaphari, I. A. Al-Baltah, and F. B. Julian, "A Promising Smart Healthcare Monitoring Model based on Internet of Things and Deep Learning Techniques," *Sana'a University Journal of Applied Sciences and Technology,* vol. 2, no. 2, pp. 147-153, 2024.
[47] M. A. Al-Hadi, G. H. Al-Gaphari, I. A. Al-Baltah, F. B. Julian, and A. A. Al-Hadi, "IoT-Based Healthcare Monitoring System," in *2024 1st International Conference on Emerging Technologies for Dependable Internet of Things (ICETI)*, 2024: IEEE, pp. 1-7.

Bio of authors

MARWA AL-HADI has been a lecturer in the field of Computer and Information Technology at Sana'a University since 2013. She received the B.S. degree in Information Systems and the M.S. degree in Computer Science from Sana'a University, faculty of Computer and Information Technology in 2019. She is currently pursuing a Ph.D. with the Department of Computer Science, Faculty of Computing and IT. She concentrates on research papers. Her research interests include enterprise resource planning systems, artificial intelligence, data mining, deep learning, machine learning, cloud computing, edge computing, blockchain, fog computing, and IoT.
marwa.alhadi@su.edu.ye

GHALEB AL-GAPHARI is a professor of Artificial Intelligence in the Department of Computer Science at Sanaa University. Professor of Artificial Intelligence in the Computer Science Department and chairman of the Software Development and Construction Unit at the Computer and Information Technology Faculty at the University of Sana'a, 2016–until now. Professor of Artificial Intelligence in the Computer Science Department and vice dean for academic affairs at the Computer and Information Technology Faculty at the University of Sana'a, 2014–2016. Associate Professor of Artificial Intelligence in the Computer Science Department and vice dean for academic affairs at the Computer and Information Technology Faculty at the University of Sana'a, 2011–2014. He had published a lot of papers in the computer science field. His research interests include artificial intelligence, deep learning algorithms, NLP, optimization, and cloud computing.
drghalebh@gmail.com