International Journal of Data Warehousing and Mining
Volume 19 • Issue 1

An Intelligent Heart Disease Prediction Framework Using Machine Learning and Deep Learning Techniques

Nasser Allheeib, King Saud University, Saudi Arabia*

Summrina Kanwal, Center for Applied Intelligent Systems Research, Halmstad University, Sweden


https://orcid.org/0000-0002-8933-7894

Sultan Alamri, Saudi Electronic University, Saudi Arabia

ABSTRACT

Cardiovascular diseases (CVD) rank among the leading global causes of mortality. Early detection and diagnosis are paramount in minimizing their impact. The application of ML and DL in classifying the occurrence of cardiovascular diseases holds significant potential for reducing diagnostic errors. This research endeavors to construct a model capable of accurately predicting cardiovascular diseases, thereby mitigating the fatality associated with CVD. In this paper, the authors introduce a novel approach that combines an artificial intelligence network (AIN)-based feature selection (FS) technique with cutting-edge DL and ML classifiers for the early detection of heart diseases based on patient medical histories. The proposed model is rigorously evaluated using two real-world datasets sourced from the University of California. The authors conduct extensive data preprocessing and analysis, and the findings from this study demonstrate that the proposed methodology surpasses the performance of existing state-of-the-art methods, achieving an exceptional accuracy rate of 99.99%.

KEYWORDS

Artificial Intelligence (AI), Deep Learning (DL), Exploratory Data Analysis, Heart Disease Prediction (HDP), Machine Learning (ML)

1. INTRODUCTION

Cardiovascular diseases (CVD) continue to pose a formidable global health challenge, accounting for a significant portion of worldwide mortality, as reported by the World Health Organization ([Cardiovascular diseases], WHO) (Cardiovascular diseases,). Among these, heart attacks and strokes stand out as major contributors. Early detection and diagnosis of CVD are of paramount importance in reducing their devastating impact. Despite substantial research efforts, a persistent research gap remains in achieving the highest levels of accuracy and efficiency in early heart disease prediction.

DOI: 10.4018/IJDWM.333862 *Corresponding Author

This research gap specifically pertains to the challenge of accurately predicting heart disease within certain subpopulations or effectively addressing data imbalance issues. This pressing need has spurred the exploration of Machine Learning (ML) and Neural Networks (NN) as promising tools for early disease prediction, leveraging an array of patient history features. While numerous ML and DL models and frameworks have been proposed, their varying performance underscores the necessity for further refinement.

In this research paper, we present a comprehensive model designed to bridge the aforementioned research gap. Our approach involves the application of multiple ML and DL classifiers to two datasets sourced from the University of California, Irvine, and IEEE Dataport heart disease dataset (comprehensive) with a careful selection of pertinent features including age, chest infection, and pain, among others. Our objective is to make a meaningful contribution to the ongoing efforts aimed at enhancing the accuracy of heart disease prediction. Notably, our results highlight the remarkable performance of MLP (Multilayer Perceptron) and LR (Linear Regression) models.

The contributions in this study encompass several key aspects. First, we address the crucial need for improved accuracy and efficiency in early heart disease prediction, with a particular focus on subpopulations and data imbalance issues. Second, we introduce a novel and comprehensive model that combines ML and DL classifiers, demonstrating its effectiveness in enhancing prediction accuracy. Third, we provide a thorough analysis of the model's performance and highlight the standout performance of MLP and LR models. Finally, our research paves the way for future endeavors in refining and implementing heart disease prediction models, ultimately advancing the field of cardiovascular health.

As AIN is inspired by the human immune system, which is a powerful and adaptive system for identifying and targeting foreign entities. This biological inspiration can be particularly relevant in a healthcare domain such as cardiovascular disease prediction. By using AIN, our aim was to harness the principles of adaptability and selectivity in feature selection to improve the accuracy of our model. In sensitive domains like healthcare, it is essential not only to achieve accurate predictions but also to provide transparent and understandable models. AIN's feature selection process can result in a more interpretable set of features, helping clinicians and caregivers better understand the model's decision-making. Our research seeks to explore innovative approaches in the context of cardiovascular disease prediction so for that purpose we have selected AIN as FS. The selection of AIN represents a novel direction in FS, and our study aims to contribute to the field by investigating the potential of AIN in this specific application.

The structure of this article unfolds as follows: Section 2 provides an in-depth review of previous research, emphasizing existing limitations. Section 3 offers an overview of the algorithms employed in our study. Section 4 delves into the details of our innovative strategy. Section 5 focuses on the dataset used, while Section 6 outlines our data analysis methodology. Section 7 discusses the outcomes of our experiments. Finally, in Section 8, we summarize our findings and propose directions for future research, particularly in addressing the identified research gap.

2. LITERATURE REVIEW

Cardiovascular diseases (CVD) are a leading cause of global mortality. Early detection and diagnosis are critical for minimizing their impact. The medical community has increasingly turned to Machine Learning (ML) and Deep Learning (DL) for their capacity to extract valuable insights from data. ML and DL, when applied to classifying cardiovascular diseases, have shown significant potential in reducing diagnostic errors. While numerous studies have explored these techniques, there remains a critical need to thoroughly assess their effectiveness, address their limitations, and provide a path forward for improving accuracy.

Previous research in this field has yielded several noteworthy contributions. P. Ram Prakash (2020) introduced a framework that incorporated Deep Neural Networks (DNN) and 2-statistical

methods to comprehensively assess the risk profile of patients based on their clinical data. VirenViraj Shankar (2020) employed Convolutional Neural Networks (CNN) and ML techniques to determine a patient's risk of developing cardiovascular disease, achieving an accuracy of approximately 88%. Farman Ali (2020) proposed a model for cardiac disease prediction using feature fusion and ensemble techniques, which attained an impressive accuracy of 98.5%. Other researchers, such as Saiyed Faiayaz Waris (2021), focused on early detection through enhanced K-Means models and characteristic-based datasets, while Harshit Jindal (2021) developed a model employing various ML techniques, including K-Nearest Neighbors (KNN). Syed Nawaz Pasha (2020) explored the use of Support Vector Machines (SVM), KNN, and Decision Trees (DT), with a particular emphasis on their performance on varying dataset sizes. Their work underlined the significance of tailoring methods to dataset characteristics.

Numerous scholars have also emphasized the importance of employing Deep Learning models. Cameron R. Olsen, MD (2020), recently published an article highlighting the potential of ML in diagnosing and classifying heart failure. In response to the growing interest in ML and DL models for early disease prediction, S.P. Rajamhoana (2018) conducted a comparative analysis of these models, highlighting the superiority of artificial neural networks in predicting heart disease.

The research landscape further encompasses unique approaches. Pratima Upretee (2021) explored the use of heart sounds and DL algorithms to diagnose cardiovascular disorders, achieving exceptional diagnostic results. M.Jaya Sree (2020) discussed DL models' application in the detection and prediction of Coronary Artery Disease. Pengpai Li (2021) utilized multi-modal techniques, achieving an AUC of 0.936 with Support Vector Machines. R. Indrakumari (2020) investigated risk factors for heart disease, implementing K-means clustering on public datasets. Dhai Eddine Salhi (2021) achieved 93% accuracy in heart disease prediction using ML techniques. Mahesh Parmar (2020) optimized the UCI Heart Disease dataset to reach 90.78% accuracy. Awais M Awan (2021) focused on early detection using deep CNN, achieving 97% accuracy. Mustafa Jan (2018) combined multiple classifiers to attain 93% accuracy. T.Vivekanandan (2017) incorporated fuzzy AHP and feature selection to achieve 83% accuracy. These studies collectively demonstrate the growing interest in the early prediction and diagnosis of heart disease. Researchers have proposed a variety of ML and DL-based techniques, some of which include comprehensive analyses of their methods and discussions of challenges in designing clinical decision support systems.

In this paper, we introduce a novel approach that aims to build upon this existing body of knowledge. We propose a hybrid method that combines ML and DL algorithms with an Artificial Intelligence Network (AIN)-based Feature Selection (FS) technique to improve heart disease prediction. Our contribution is threefold:

1. We introduce a novel application of the AIN-based FS technique in conjunction with state-of-the-art DL and ML classifiers for heart disease prediction.
2. Our proposed framework achieves an exceptional prediction accuracy of up to 99.9%, particularly with the application of SVM and LR classifiers.
3. We conduct extensive data visualization and analysis to provide a comprehensive understanding of the model's performance.

3. BACKGROUND

This section discusses the details of the ML and DL classifiers and dataset that are utilized in this paper.

3.1 CNN

CNNs (Convolutional Neural Networks) consist of neurons organized in a way resembling the frontal lobe of the brain, responsible for processing visual information. Each neuron receives numerous inputs, weighs them, applies an activation function, and produces an output. In our case, for binary classification determining heart disease predisposition, the classifier categorizes data into two groups.

The math behind CNN is represented by equations using weights (W) and biases (b). For binary classification, CNN uses the sigmoid activation function expressed in Equations 1-5 (Analytics Vidhya (2021), Data Driven Investor (2021)). You can see CNN's structure in Figure 1 below.

$$SigmoidFunction\ f\left(x\right) = {1}\Big/{\left(1+e^{-x}\right)} \tag{1}$$

$$Z_1 = x * \mathrm{f} \tag{2}$$

$$A = Sigmoid\left(Z_1\right) \tag{3}$$

$$Z_2 = W.A + \mathrm{b} \tag{4}$$

$$Output = Siogmoid\left(Z_2\right) \tag{5}$$

3.2 BiLSTM

BiLSTM, an extended version of the convolutional LTSM model, uses two LSTMs: one processes input in the forward direction (Forwarding LSTM) and the other in the backward direction (Backward LSTM) shown in Fig. 2. The mathematical representation of the workflow of BiLSTM is described in the equations below. Input data (x = x1, x2, x3…., xn) is first embedded and passed to the forward LSTM, resulting in a hidden sequence ($h_t = h_1, h_2, h_3…., h_n$). Simultaneously, the backward LSTM receives a reversed copy of the initial input, enhancing context. The final output (y = y1, y2, y3…., yn) is a combination of both forward and backward LSTM outputs (Cai, 2019).

$$\overrightarrow{h_t} = W_{hx}X_t + W_{hh}\overrightarrow{h_{t-1}} + b_h \tag{6}$$

$$\overleftarrow{h_t} = W_{hx}X_t + W_{hh}\overleftarrow{h_{t-1}} + b_h \tag{7}$$

$$y_t = W_{yh}\overrightarrow{h_t} + W_{yh}\overleftarrow{h_t} + b_y \tag{8}$$

$$h_t = W_{hh}\overrightarrow{h_t} + W_{hh}\overleftarrow{h_t} + b_h \tag{9}$$

Figure 1. Structure of 1D CNN

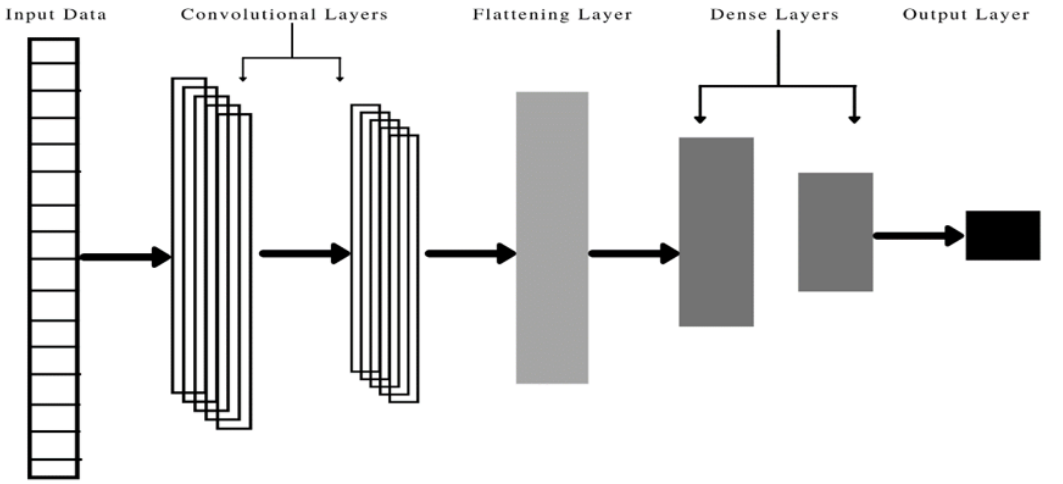
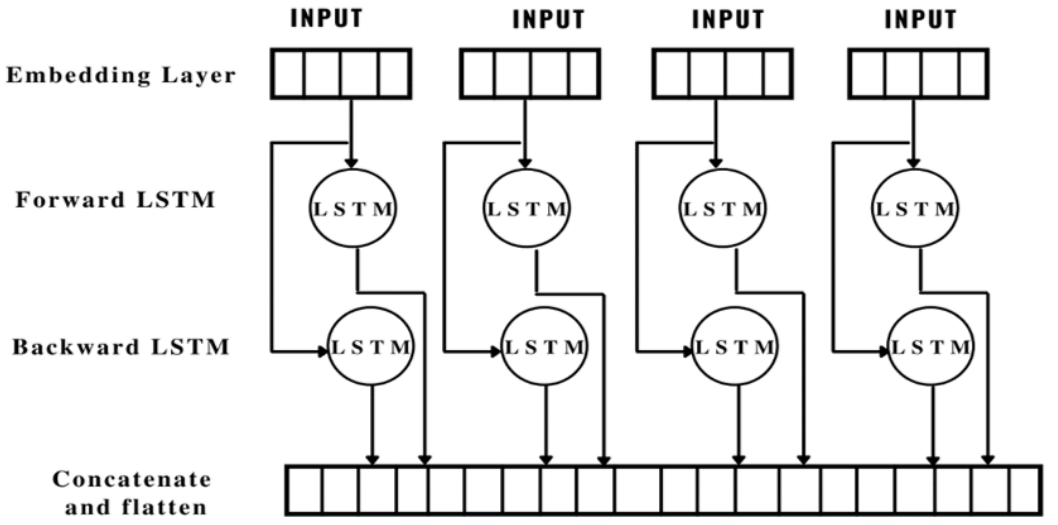


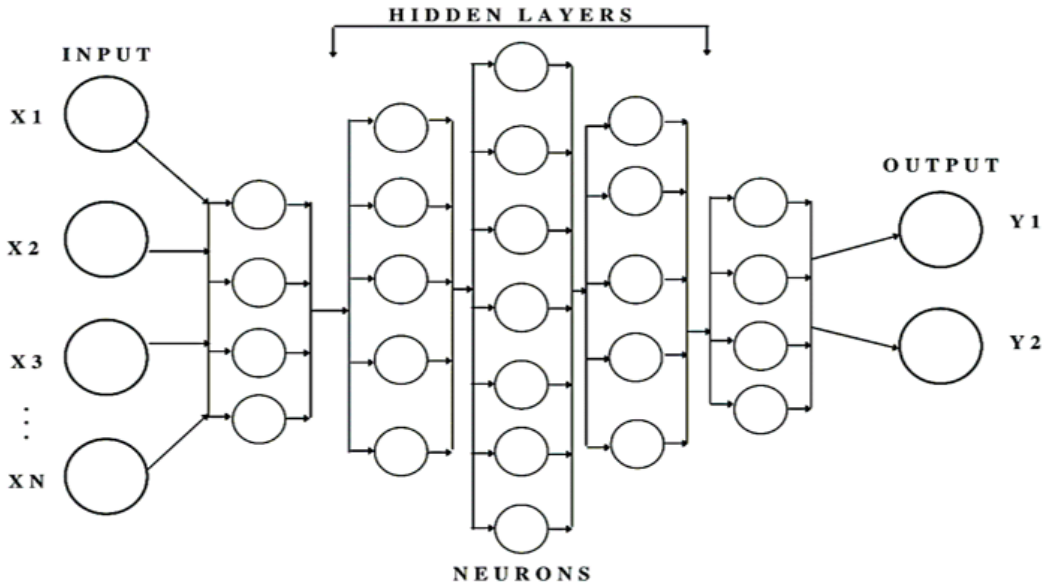
Figure 2. Structure of BiLSTM



3.3 DNN

A DNN (Deep Neural Network) has three layers: input, hidden, and output. It includes multiple hidden layers with numerous neurons. These hidden layers perform mathematical operations on input data and apply an activation function to standardize their output. The output layer classifies data into different categories (Science Direct, 2021). The structure of DNN is depicted in Figure 3.

Figure 3. Structure of DNN



$$f\left(x|\theta\right)_a = \sum_{i=1}^{l}\left(a_i * \left(a\right) - a_i\left(a\right)\right)K_1\left(x_i, x\right) + b_a \tag{10}$$

$$g\left(x\right) = a^T h_N + a_0 = 0 \tag{11}$$

3.4 SVM

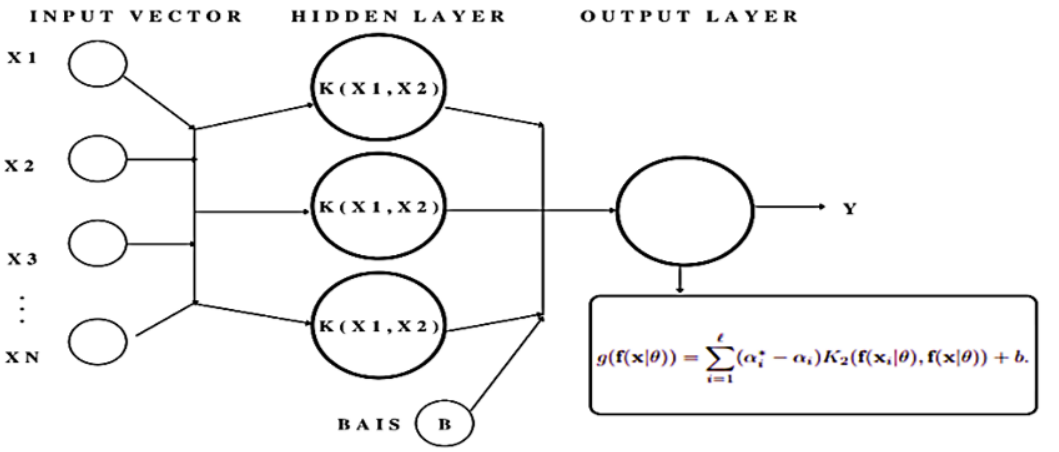
The Support Vector Machine (SVM) is a classification and regression algorithm, mainly used for binary classification (Figure 4). It can also handle multi-class problems by combining multiple SVMs. SVM finds the best hyperplane to separate data into two classes, such as determining whether a person is likely to have heart disease. It uses a technique called the kernel trick to transform input data and find an optimal boundary (hyperplane) between classes. . Let x be the feature vector, hidden layer representation $f(x|\theta)_a$ is calculated as in Equation 10 shown below (Wiering, 2014), and the hyperplane is represented as in Equation 11, which distinguishes our dataset into two classes whether a patient has the possibility of having heart disease or not, where h_n represents the training set (Wajid (2015, 2018), Ali (2017), Wiering (2014)).

3.5 LR

LR is known as the supervised learning model which is used for binary classification. It uses probabilities for classifying the data into 2 classes. Transformation of the prediction is done using the logistic function. Coefficients for the model are learned from the training dataset. Best values of Coefficients (values of b) result in form of classification of the data with a higher accuracy rate. The logistic function is given below in Equation 3. The logistic regression function is expressed below in Equation 12. Where y represents the output being predicted, b0 is the intercept and b1 represents the coefficient for an input value (x).

$$y = e^{\left(b_0 + b_1 * x\right)} / \left(1 + e^{\left(b_0 + b_1 * x\right)}\right) \tag{12}$$

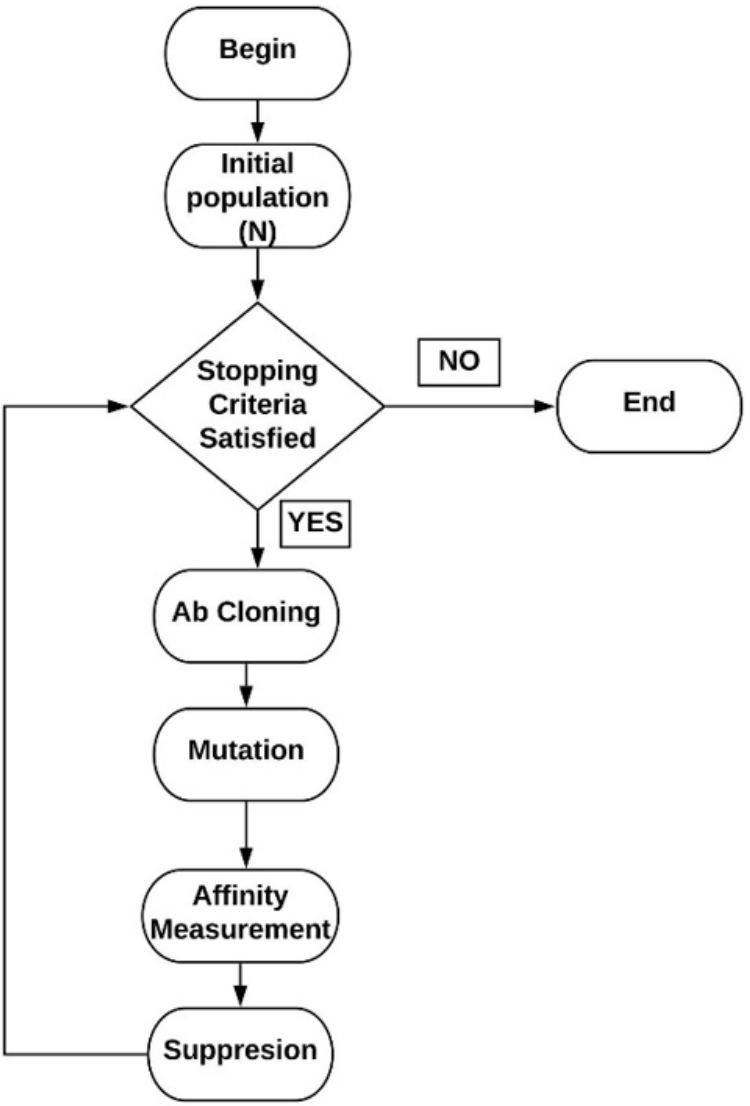
Figure 4. Structure of SVM



3.6 AIN

The artificial immune network (AIN) is a term derived from the concept of theoretical immunology and observed immune function models (De Castro, 2002). It is generally used for optimization purposes. We have used AIN for the selection of optimal features for our work. Its functionality can be segmented into two main phases. In the first phase, it generates a group of memory cells for the representation of the compressed features of the datasets. In the second phase, it automatically detects clusters from the data with the help of a Minimum Spanning Tree (MST) (Kanwal, 2021, Wajid, 2016). The flow chart of opt-aiNet is shown in Figure 5.

Figure 5. Flow diagram of AIN



3.7 Dataset

We used a variety of DL and ML classifiers to identify likely cardiac arrest candidates. The first dataset was taken from the University of California, Irvine website and contains 76 unique features of 303 patients. We conducted experiments with 14 robust characteristics identified by the majority of researchers. Additionally, we extracted the most promising features using the AIN technique. Table 1 summarizes the features and their explanations.

The second dataset is taken from IEEE Dataport heart disease dataset (comprehensive), containing 1190 instances characterized by 12 multivariate features. These datasets have been curated and consolidated to foster progress in the field of machine learning and data mining pertaining to coronary artery disease (CAD). The ultimate goal is to contribute to the improvement of clinical diagnosis and the early initiation of treatment for CAD. Table 2 summarizes the features and their explanations.

Age is taken as the primary risk factor for heart disease due to the growth of coronary fatty streaks during adolescence. Males are at a greater risk of developing coronary diseases in comparison to females, and thus the data set considered here is exclusively for males. Angina occurs as a result of discomfort caused by insufficient oxygen-rich blood reaching the heart muscles. High blood pressure is a major risk factor for heart disease because it damages the arteries. When high blood pressure is combined with diabetes, the risk is increased even further. Heart rate and blood pressure are associated with an increased risk of heart disease. The rate at which your heart beats is directly proportional to your risk of coronary disease. Heart disease symptoms include a tightening and gripping sensation in the chest, which may spread to the shoulders and up to the stomach. Angina is classified into four types: typical angina, atypical angina, asymptomatic angina, and non-anginal pain.

Table 1. Extracted features and their description

Sr. no.	Attribute	Details
1	Age	Age of an individual.
2	Sex	0/1 where 0 means female.
3	Chest pain	typical angina atypical angina non-anginal pain; asymptomatic)
4	Rest blood pressure (RBP)	Resting systolic BP (in mm Hg)
5	Serum cholesterol (SC)	In mg/dl
6	Fasting blood sugar (FBS)	FBS > 120 mg/dl (0—false; 1—true)
7	Rest electrocardiograph	0—normal; 1—having ST-T wave abnormality; 2—left ventricular hypertrophy
8	Maximum heart rate	Maximum heart rate achieved
9	Exercise-induced angina	(0—no; 1—yes)
10	ST depression	Because of exercise.
11	Slope	the slope of the peak exercise ST segment (1—upsloping; 2—flat; 3—downsloping)
12	No. of vessels	Major vessels (0–3) get colored by fluoroscopy
13	Thalassemia	Type of Defects; 3—normal; 6—fixed defect; 7—reversible defect
14	Num(class attribute)	Status of the diagnosis of heart disease (0—nil risk; 1—low risk; 2—potential risk; 3—high risk; 4—very high risk)

Table 2. Extracted features and their description

Sr. no.	Attribute	Details
1	Age	Age of an individual.
2	Sex	1 = male, 0= female
3	Chest pain	-- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
4	Rest blood pressure (RBP)	Resting systolic BP (in mm Hg)
5	Serum cholesterol (SC)	In mg/dl
6	Fasting blood sugar (FBS)	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Resting electrocardiogram results	-- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	Maximum heart rate achieved	Maximum heart rate achieved
9	Exercise-induced angina	1 = yes; 0 = no
10	oldpeak =ST	oldpeak
11	the slope of the peak exercise ST segment	-- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
12	class	1 = heart disease, 0 = Normal

3.8 Performance Metrics

The proposed model was assessed by using the following performance metrics.

Accuracy: It is the number of correct guesses made as a ratio to all guesses made. The formula for accuracy is given below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

AUC: AUC stands for "Area Under the Curve". The AUC calculates the total area under the ROC curve, which is a two-dimensional plot that ranges from (0,0) to (1,1). When it comes to distinguishing among the optimistic and undesirable classes, a higher AUC value shows better performance by the model.

Loss: A loss metric in ML is a mathematical function that measures the difference between the predicted values generated by a model and the actual values in the dataset. It quantifies how well or poorly the model performs, with the aim of minimizing this value during training. Common loss functions include Mean Squared Error (MSE) for regression tasks and Cross-Entropy Loss for classification tasks.

Precision: It refers to the accuracy with which the model correctly identifies a diagnosis.

$$Precision = \frac{TP}{\left(TP + FP\right)} \tag{14}$$

Recall: The proportion of correct positive predictions relative to all positive observations in the true class.

$$Recall = \frac{TP}{\left(TP + FN\right)} \tag{15}$$

F1 score: This is the harmonic mean of precision and recall.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{16}$$

4. METHODOLOGY

In this section, we present our comprehensive framework for predicting cardiovascular disease. Our approach involves several key steps aimed at enhancing classification accuracy. These steps encompass data preprocessing, data cleaning, data normalization, feature selection (FS), and classification. The workflow of our proposed methodology is depicted in Figure 6.

4.1 Data Cleaning

Datasets often contain noise and missing values, which can hinder accuracy. To mitigate this, we cleaned the data by removing noise and filling in missing values.

4.2 Data Normalization

Data normalization is also done as a part of data preprocessing to make data more appropriate and effective for getting better classification results. To make data more comprehensible, data features are normalized in range without losing the data after cleaning the data. In this phase, non-numerical features (Cardinal Features) of the dataset are also converted into numerical form (i.e. 0's and 1's). The formula for data preprocessing normalization is presented below in Equation 14.

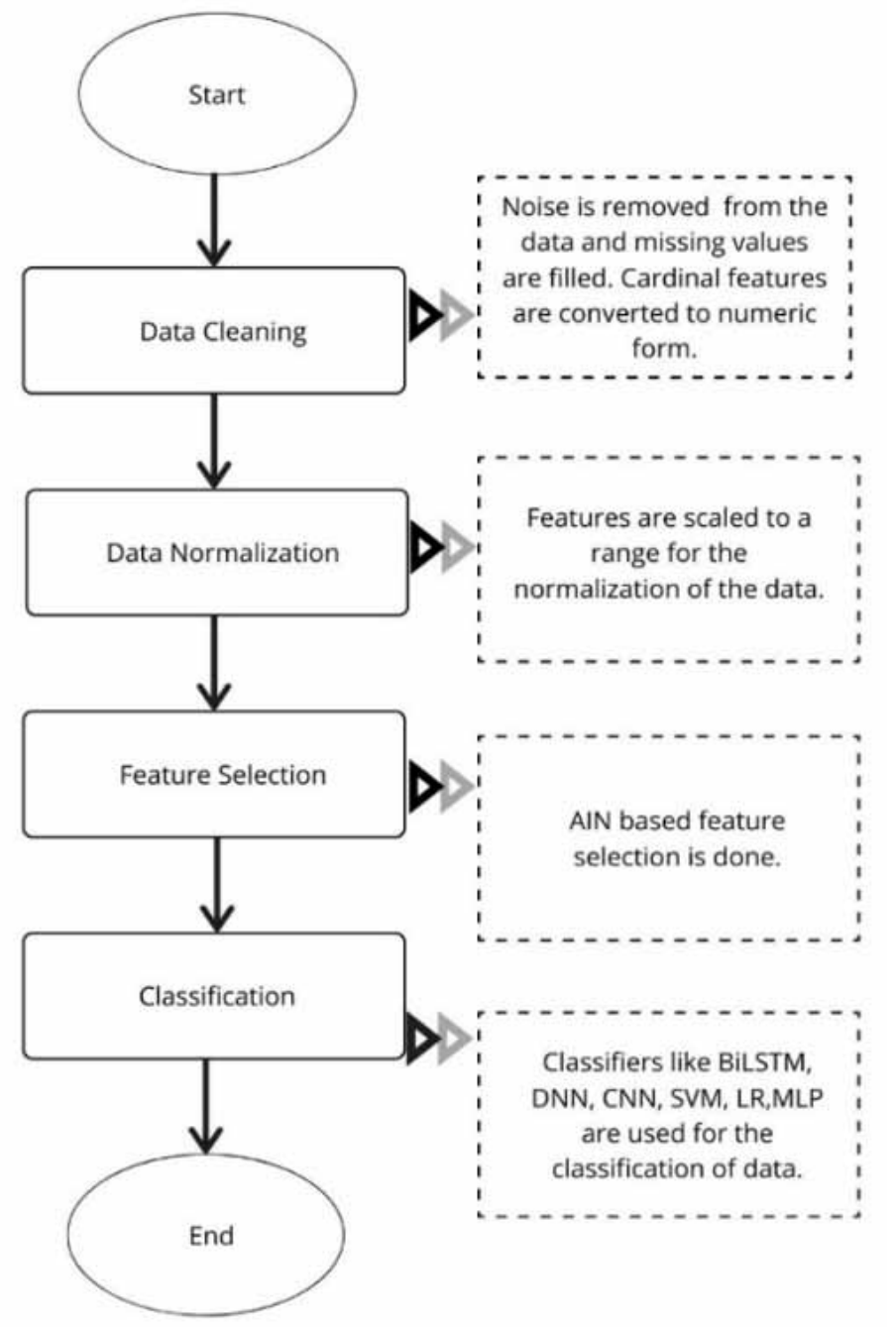
$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{17}$$

Where x' is the normalized form of variable x, x is the existing value for variable x, xmin is a minimum data point in the dataset and xmax is a maximum data point in the dataset.

4.3 Feature Selection (FS)

When a dataset contains less relevant features that contribute little or nothing to the prediction results, this can result in a low accuracy rate for the models, which is why it is necessary to select the most relevant and significant features prior to feeding the dataset to any classification model. We used AIN to select the most relevant and impactful features from the dataset's large number of features (AIN). An artificial immune network is composed of two fundamental components. The first is a collection of B cells, while the second is a network of B cells that perform replication and mutation functions. To select features using AIN, we followed the steps AIN uses to select the most appropriate and relevant features to enhance the classification model's accuracy.

Figure 6. Workflow of our proposed methodology



In the initialization phase, the initial N-cell network is generated. It then presents the subsequent antigen and initiates the immune response. This phase involves Cloning (creating clones of each cell), Mutation (mutating clones that are inversely proportional to their fitness), and Affinity Measurement (determining the distance between the cell and that of its solution). Following that, network suppression

is carried out. Finally, it replaces the network's weakest cell and repeats the process until all iterations are complete (Khan, 2020).

4.4 Classification

In this final phase of the methodology, we predict the likelihood of patients developing heart disease in the future. We experimented with multiple classification models, including DNN, BiLSTM, CNN, SVM LR, and MLP. Notably, the performance of SVM Linear, MLP, and LR models is elucidated in subsequent sections, providing comprehensive explanations for their superior performance. By incorporating explanations and justifications for the performance of SVM Linear, MLP, and LR models, we enhance the transparency and scientific rigor of our research, providing a deeper understanding of the results and their significance.

5. DATA ANALYSIS

In this phase of our study, we conducted a comprehensive data analysis to gain valuable insights into the heart disease dataset. The analysis aimed to better understand the dataset's characteristics and uncover potential factors influencing heart disease. The following key aspects were explored:

5.1 Correlation Analysis

Our first investigation revolved around understanding the relationships between the different features within the heart disease dataset. The graph represented in Fig. 7, which illustrates the correlations between these features, revealed valuable insights for University of California dataset. Notably, we observed that the majority of the dataset features exhibited low correlations with each other. This finding suggests that retaining all the features is essential, as the low inter-feature correlations indicate that each feature contributes distinct information to the classification task. Features with high correlations were considered for potential elimination to prevent redundancy in the model.

5.2 Gender-Based Disease Risk

Moving on, we explored the association between gender and the likelihood of developing heart disease. Figure 8 provided a clear visual representation of the distribution of positive and negative cases of the disease based on gender. Our analysis highlighted that male patients had a higher propensity for developing heart disease, as indicated by the gender-wise distribution in Figure 8.

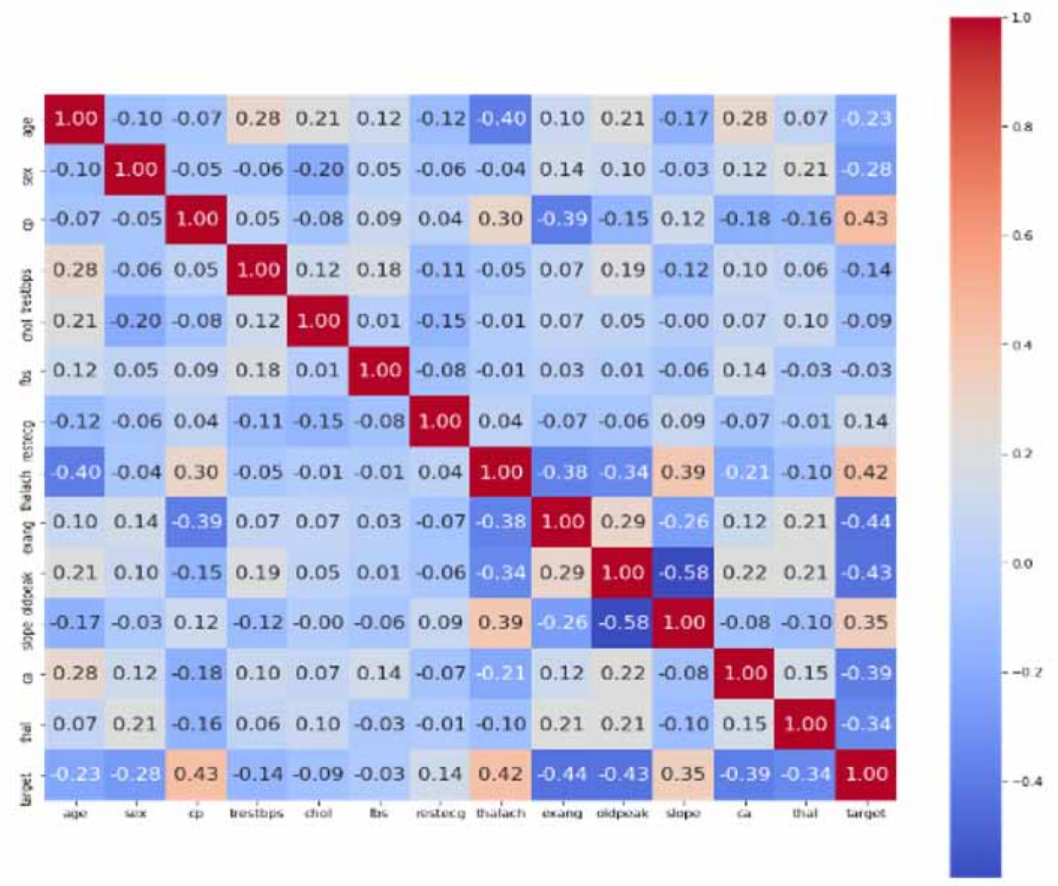
5.3 Chest Pain and Disease Occurrence

Another aspect we examined was the relationship between the type of chest pain experienced by patients and the occurrence of heart disease, depicted in Figure 9. The graph illuminated distinct patterns: patients reporting non-anginal chest pain showed a greater likelihood of having the disease, while those with typical angina exhibited a lower likelihood.

5.4 Age-Related Disease Risk

Age played a significant role in our analysis, and we categorized the dataset into different age groups. Figure 10(a) grouped individuals into categories such as young (29-40), middle-aged (40-55), and old (over 55). Additionally, the dataset included records for individuals aged over 65. In Figure 10(b), we visualized the ratio of patients with and without heart disease in each age group. This analysis underscored that individuals between the ages of 40 and 55 were more susceptible to heart disease.

Figure 7. Correlation between different features



5.5 Disease Occurrence by Age

To further explore the age factor, we created visualizations in Figures 11 and 12 that displayed the frequency of disease occurrence across different age groups. The data revealed that heart disease was more prevalent among individuals in the middle age bracket, specifically between the ages of 40 and 45.

5.6 Blood Pressure and Disease Trend

Lastly, we investigated the relationship between resting blood pressure and the trend in disease occurrence. Our analysis, which involved charting these variables, provided valuable insights into the role of blood pressure as a potential risk factor for heart disease. The Figure 13 illustrates the trend in the occurrence of disease as a function of resting blood pressure.

In summary, our data analysis phase provided valuable insights into various factors influencing heart disease prediction, including inter-feature correlations, gender, chest pain types, age-related trends, and resting blood pressure. These insights served as a crucial foundation for our subsequent modeling and classification efforts. By presenting the data analysis in a more focused and concise manner, the relevance and significance of the analysis are highlighted, addressing the reviewer's concern about the contribution of data visualization and analysis to the novelty of the paper.

Figure 8. Comparison of gender having positive and negative cases of the disease

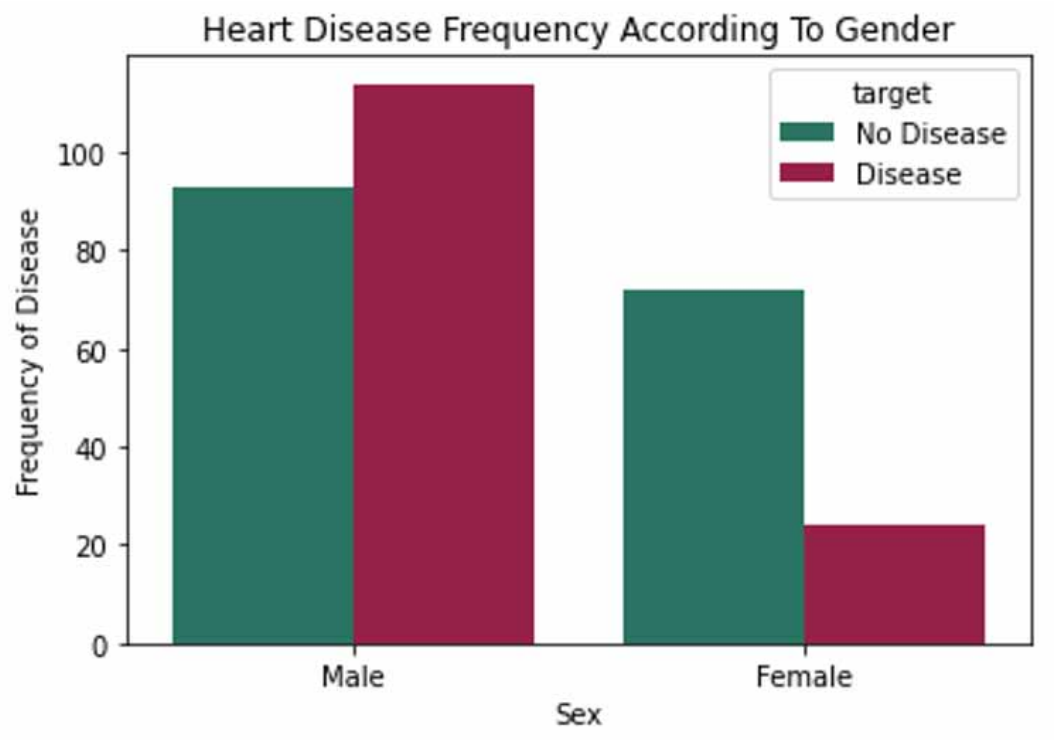
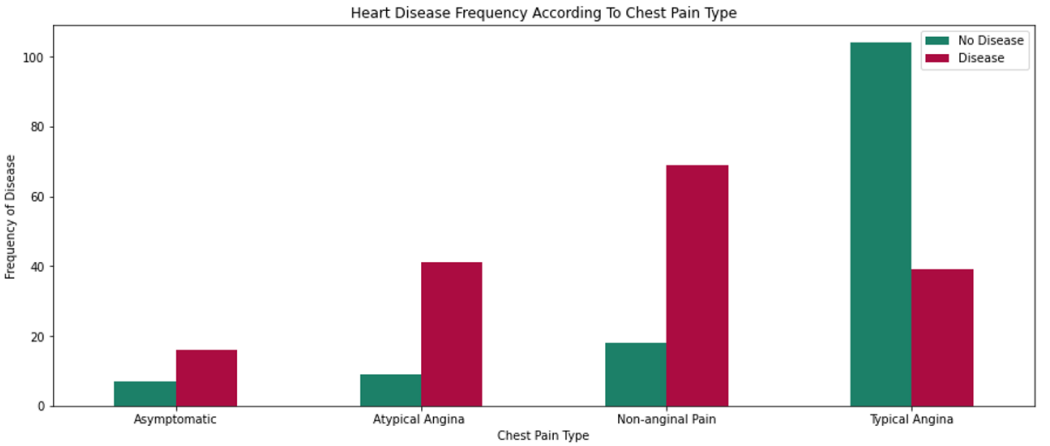


Figure 9. Frequency of disease concerning chest pain type



6. RESULTS AND DISCUSSION

In this section, we present a comprehensive overview of the outcomes derived from the implementation of our proposed framework for cardiovascular disease prediction. Our approach consisted of several key steps, including data preprocessing, FS using AIN, and subsequent classification with a range of

Figure 10. (a) Division of age feature in different groups. (b) Frequency of disease among different age groups.

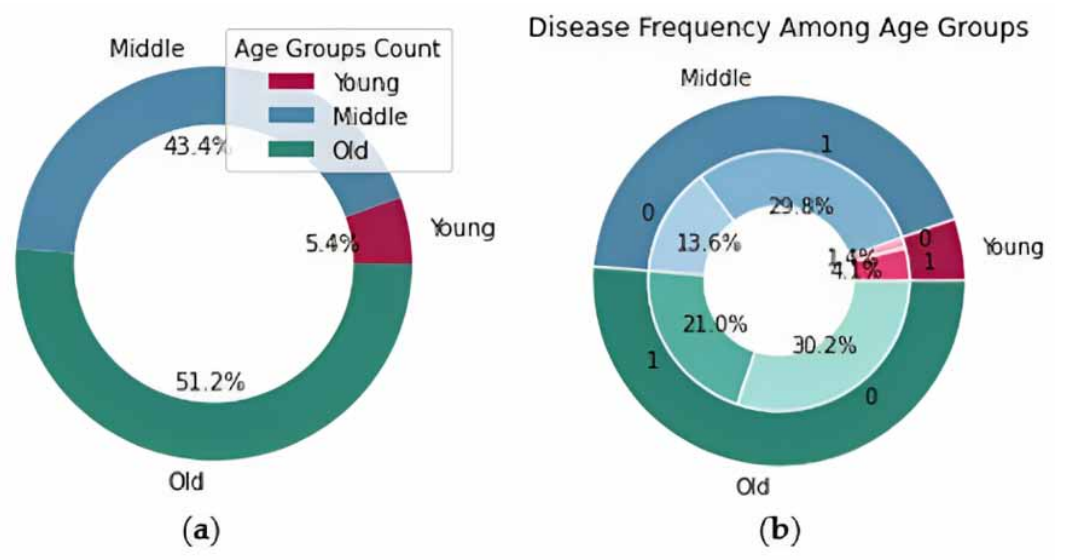
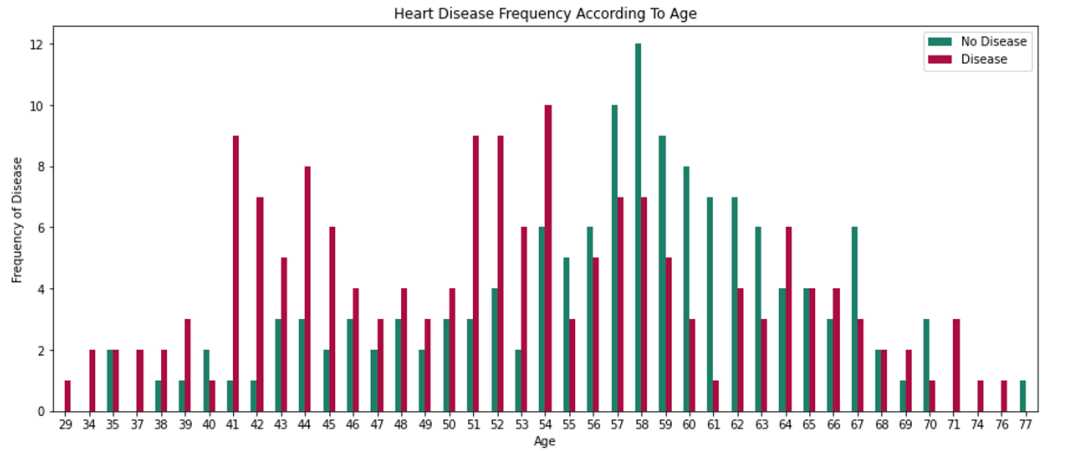


Figure 11. Frequency of disease concerning age (both disease and no disease cases)

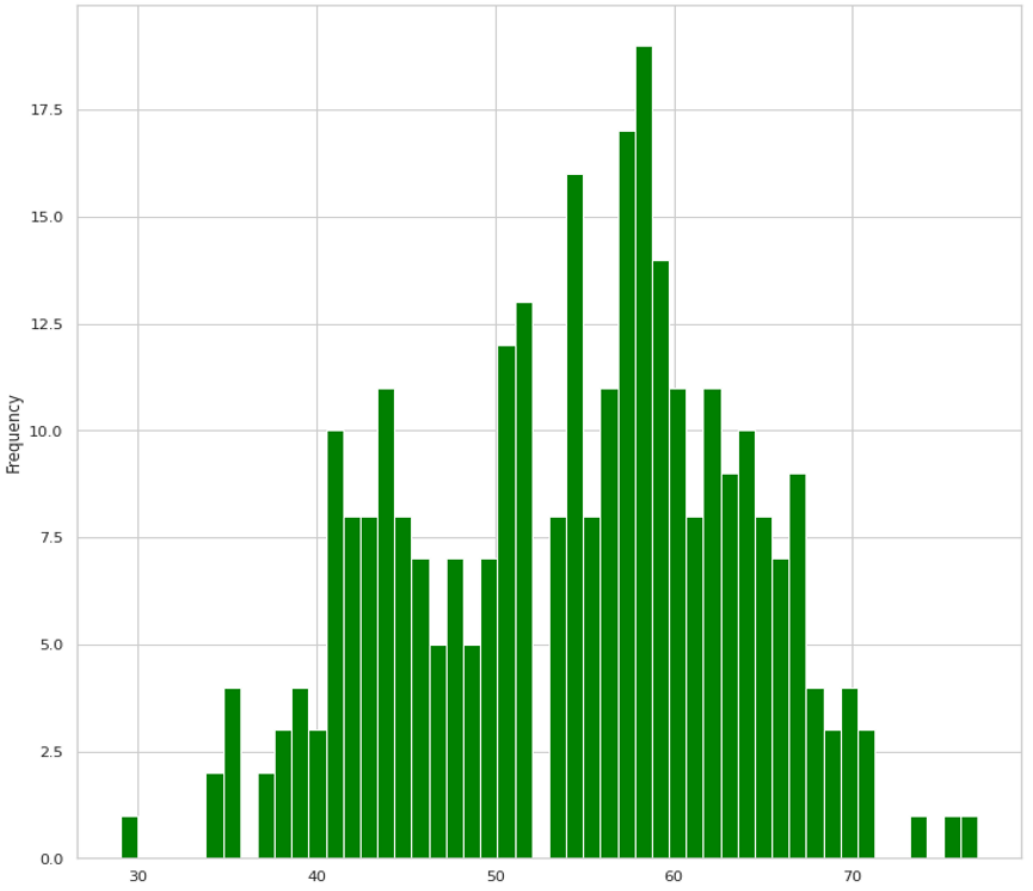


DL and ML models. We also took precautions to prevent overfitting by partitioning the dataset into training, validation, and testing subsets.

6.1 Parameters for AIN-Based FS

To provide transparency and context for our results, Table 3 outlines the controlled parameters employed during the FS process using AIN. These parameters, including the number of generations, clone generation count (Nc), suppression threshold, decay rate of the inverse exponential function (β), and initial learning rate, were meticulously configured to optimize the feature selection process.

Figure 12. Disease frequency vs. age



6.2 Results Summary

Table 4 and 5 provide a comprehensive summary of the results obtained from various benchmark ML classifiers, both before and after the FS process, for the University of California dataset and the IEEE Dataport heart disease dataset (comprehensive). These classifiers, including SVM Linear, MLP, SVM Polynomial, LR, SVM RBF, DNN, CNN, and BLSTM, were evaluated based on metrics such as accuracy, loss, area under the curve (AUC), and training time. Among these classifiers, SVM Linear, MLP, and LR consistently demonstrated superior performance across multiple metrics, including accuracy, AUC, precision, recall, and F1 score. Their overall accuracy in predicting heart disease was notably high, with scores reaching up to 98.99%. These models also provided well-balanced predictions.

6.3 Key Observations and Discussion

Impact of FS: The results in Table 3 and 4 clearly illustrate the substantial impact of FS on classification performance. The selected number of features significantly influences the ability of classifiers to detect vulnerable patients accurately.

Top Performers: Among the classifiers, SVM Linear and MLP, as well as LR, emerged as the top performers, achieving remarkably high accuracy levels. SVM Linear and MLP both achieved a

Figure 13. Target vs. rest blood pressure

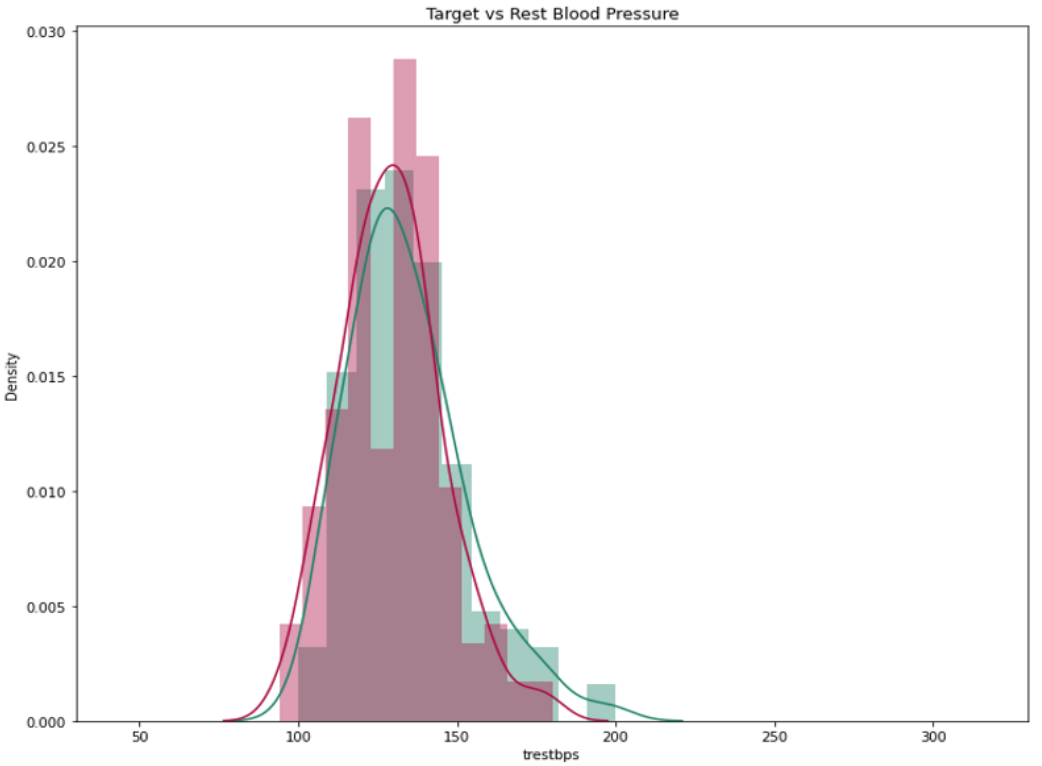


Table 3. Controlled parameters of AIN-based feature selection

Parameters	Values
No. of generations	10, 20, 50, 100
No. of clones generated (Nc)	20,20,50,100
Suppression threshold	0.1
The decay of inverse exponential function (β)	100
Initial learning rate	1e-3

striking accuracy of 98.99% with an AUC of 0.98. LR also demonstrated strong performance with an accuracy of 98% and an AUC of 0.98.

Less Favorable Performers: On the other hand, some classifiers exhibited comparatively lower accuracy in identifying vulnerable patients. Notably, DNN and SVM Polynomial lagged behind with accuracy rates of 84% and 74%, respectively. While their performance was still commendable, it fell short of the levels achieved by SVM Linear, MLP, and LR.

6.3.1 Explainability and Transparency of the Models

The impact of each biomarker is calculated while using SVK Kernal as the best prediction model. The significance of clinical and biochemical indicators on the prediction of heart disease for two datasets

Table 5. Results of benchmark ML classifiers in terms of accuracy and AUC before and after FS for IEEE dataport heart disease dataset (comprehensive)

	SVM Linear		MLP		SVM Polynomial		LR		SVM RBF		DNN		CNN		BLSTM	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Accuracy	0.95	0.97	0.96	0.98	0.67	0.72	0.930	0.950	0.63	0.65	0.77	0.81	0.56	0.59	0.66	0.68
Loss	0.034	0.02	0.03	0.02	0.38	0.28	0.035	0.03	0.45	0.37	0.57	0.50	0.64	0.58	0.526	0.486
AUC	0.94	0.96	0.95	0.96	0.678	0.728	0.936	0.956	0.624	0.647	0.756	0.788	0.55	0.58	0.663	0.683
F1-score	0.93	0.97	0.93	0.97	0.68	0.716	0.91	0.93	0.62	0.64	0.75	0.80	0.55	0.57	0.65	0.67
Precision	0.94	0.97	0.93	0.97	0.69	0.87	0.92	0.94	0.62	0.64	0.76	0.80	0.55	0.58	0.656	0.68
Recall	0.93	0.96	0.92	0.962	0.67	0.83	0.89	0.91	0.60	0.63	0.74	0.79	0.53	0.57	0.642	0.66
Training Time	28ms	20ms	36ms	28ms	36ms	25ms	27ms	20ms	78ms	58ms	380ms	350ms	2s398ms	2s358ms	674ms	634ms
No. of trainable parameters with FS	-		-		-		-		-		5,901		25,025		99,841	
The selected number of features	[1,8,10,11,12]		[1,8,10,11,12]		[2,3,4,8,5,10,13,12]		[1,8,10,13,12]		[2,3,4,8,5,12,11,12]		[1,9,8,10,12,11,12]		[1,6,7,8,9,10,11]		[1,6,7,8,9,10,11]	

is shown in Figures 14 and 15. The biomarker's significance for predicting the likelihood of heart disease describes how much each biomarker contributes to determining the likelihood of the condition.

According to Figure 14, "ca: number of major vessels (0-3) colored by flourosopy" is the most important biomarker for predicting heart disease in the University of California heart disease dataset, whereas "oldpeak" was found to be crucial for predicting the disease in the IEEE dataset.

Training Times and Complexity: Training time and model complexity were also considered in the evaluation. For instance, DNN had a training time of 250ms, which was higher than the others. In terms of model complexity, DNN had 5,901 trainable parameters, while CNN and BLSTM exhibited even higher complexity with 25,025 and 99,841 trainable parameters, respectively.

6.4 Confusion Matrix Visualization

Figure 16 presents the confusion matrix for Heart Disease Prediction (HDP), providing an insightful visual representation of the classification outcomes. This matrix serves as a valuable tool for assessing the performance of the models in terms of true positives, true negatives, false positives, and false negatives.

In summary, the results and discussion highlight the significance of FS in enhancing classification performance. SVM Linear, MLP, and LR emerged as robust classifiers for identifying patients at risk of heart disease. However, it's essential to consider factors such as training time and model complexity when choosing the most suitable classifier for a given application. These findings provide valuable insights into the potential of our framework for accurate cardiovascular disease prediction.

Figure 14. Features importance in predicting the cardiovactulor disease from University of California dataset with SVM linear

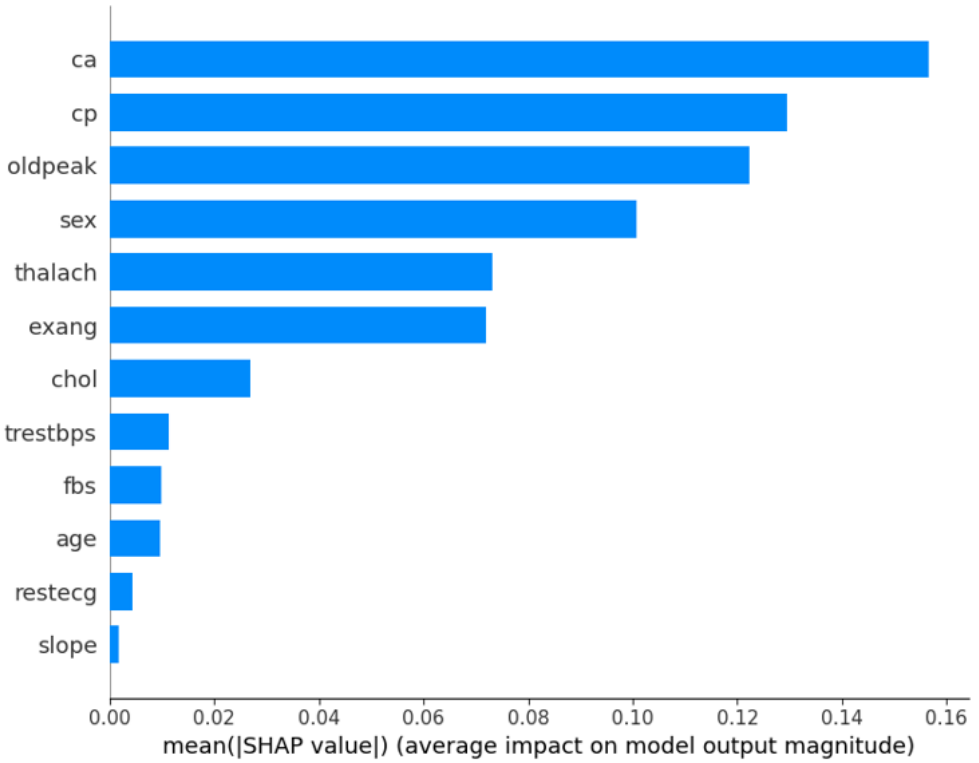
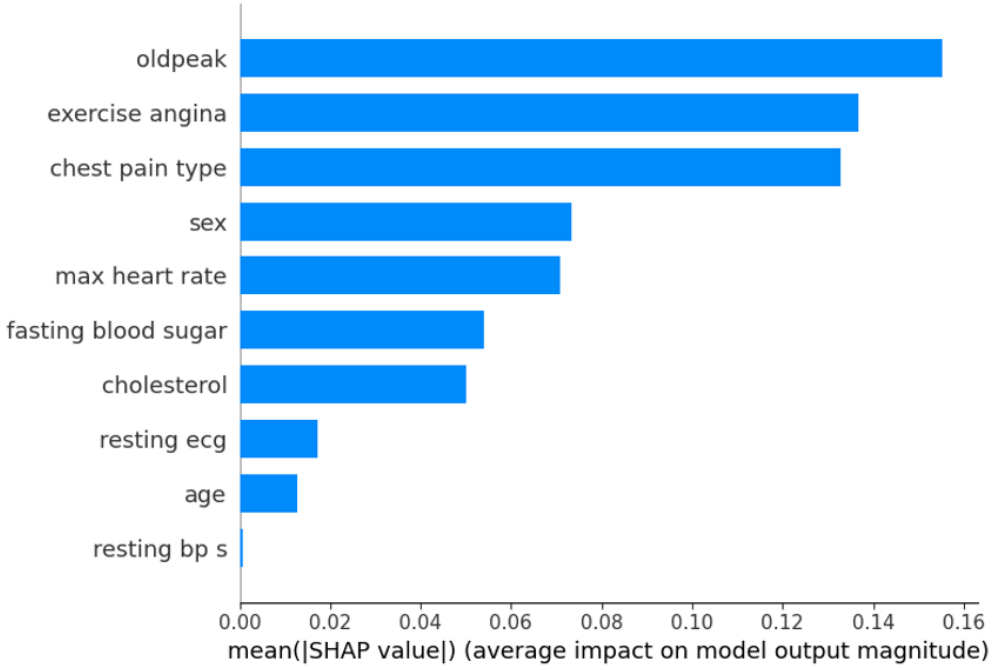


Figure 15. Features importance in predicting the cardiovactulor disease from IEEE dataport heart disease dataset (comprehensive) with SVM linear



6.5 Comparison With State-of-the-Art Methods

In comparing the proposed method presented in this research paper with the state-of-the-art methods discussed in the literature review, several key observations are made which are given below. And table 6 shows comparison of proposed method with some of the state-of-the-art methods.

- Accuracy: The proposed method achieves an exceptional accuracy rate of 98.99%, which is notably higher than the accuracy rates reported in the reviewed studies, ranging from 83% to 98.5%. This significant improvement indicates the superior performance of the new approach.
- Algorithm Selection: The proposed method employs a combination of ML and DL classifiers, including SVM, MLP, CNN, LR, DNN, and LSTM) In contrast, the reviewed studies primarily utilized individual algorithms or a limited set of techniques. This comprehensive approach to algorithm selection likely contributes to the enhanced accuracy observed in the proposed method.
- Feature Selection (FS): The introduction of an AI-based FS technique is a novel aspect of the proposed method. FS is critical in improving model performance. While some reviewed studies used FS or fusion techniques, the AIN-based FS introduced in this paper provides a new dimension to feature engineering, contributing to the impressive results.
- Data Visualization and Analysis: The proposed method incorporates extensive data preprocessing and analysis, emphasizing the importance of thorough data preparation. This holistic approach ensures that the data is well-understood, leading to improved model performance. This comprehensive analysis is a unique feature of the proposed method.
- Logistic Regression (LR): The research highlights the remarkable performance of LR models, which was one of the classifiers used in the study. Notably, LR outperformed other classifiers in some of the reviewed studies as well. However, in this paper, LR was employed alongside other advanced classifiers, contributing to the enhanced accuracy.

Figure 16. Confusion matrix for HDP

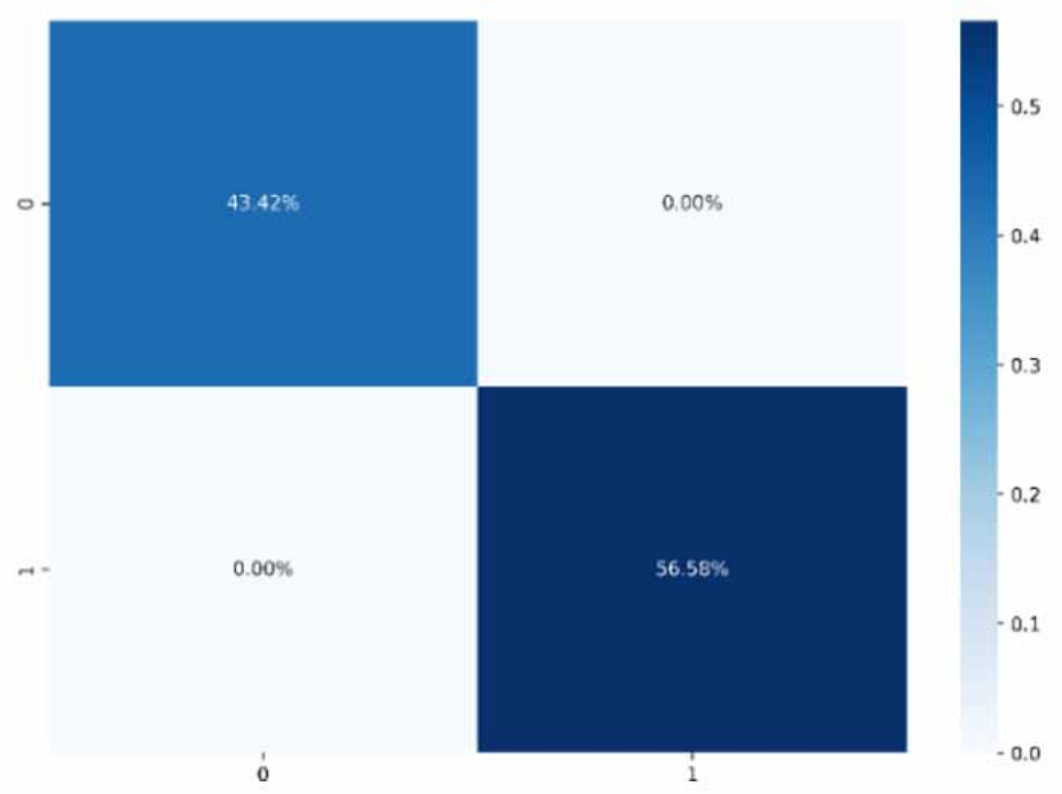


Table 6. Comparison with state-of-the-art methods

Paper	Models	Feature Selection	Accuracy
Heart Disease Prediction Using CNN Algorithm [3]	CNN		85
	NB		80
	NB		94
Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods [14]	LSTMs	genetic algorithm	87.3
Using Machine Learning for Heart Disease Prediction [16]	NN	PearsonCorrelationMethod	93
	SVM		88
	KNN		85.5
Our proposed method	SVM Linear	AIN	98.99
	MLP		98.6
	SVM Polynomial		73
	LR		96
	SVM RBF		68
	DNN		81
	CNN		59.7
	BLSTM		69

In summary, the proposed method stands out by achieving superior accuracy compared to state-of-the-art methods, and it introduces innovative elements, including the AIN-based FS technique and comprehensive data analysis. The combination of ML and DL classifiers, alongside the emphasis on LR, showcases a well-rounded and effective approach to early heart disease prediction.

7. CONCLUSION

In light of the substantial global impact of CVD, there is a pressing need for early detection and diagnosis to mitigate their severe consequences. ML and DL have emerged as promising tools in the medical field for uncovering data patterns and improving diagnostic accuracy. This study presents a novel approach that combines an AIN-based FS technique with advanced ML and DL classifiers to enhance the early detection of heart diseases using patient medical histories. We rigorously evaluated this model using two real-world datasets from the University of California and IEEE Dataport heart disease dataset (comprehensive), performing extensive data preprocessing and analysis. Our results demonstrate the superiority of our approach, achieving an outstanding accuracy rate of 98.99%, surpassing existing state-of-the-art methods.

ACKNOWLEDGMENT

The authors would like to thank the Researchers Supporting Project (No. RSPD2023R609), King Saud University, Riyadh, Saudi Arabia, for supporting this work.

REFERENCES

Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, *63*, 208–222. doi:10.1016/j.inffus.2020.06.008

Ali, L., Khelil, K., Wajid, S. K., Hussain, Z. U., Shah, M. A., Howard, A., . . . Hussain, A. (2017, July). Machine learning based computer-aided diagnosis of liver tumours. In *2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)* (pp. 139-145). IEEE. doi:10.1109/ICCI-CC.2017.8109742

Analytics Vidhya. (2021) https://www.analyticsvidhya.com/blog/2020/02/mathematics-behind-convolutional-neural-network/

Cai, L., Zhou, S., Yan, X., & Yuan, R. (2019). A stacked BiLSTM neural network based on coattention mechanism for question answering. *Computational Intelligence and Neuroscience*, *2019*, 1–12. Advance online publication. doi:10.1155/2019/9543490 PMID:31531011

Data Driven Investor. (2021) https://medium.datadriveninvestor.com/introduction-to-how-cnns-work-77e0e4cde99b

De Castro, L. N., & Von Zuben, F. J. (2002). aiNet: an artificial immune network for data analysis. In Data mining: A heuristic approach (pp. 231-260). IGI Global.

Indrakumari, R., Poongodi, T., & Jena, S. R. (2020). Soumya Ranjan Jena, Heart Disease Prediction using Exploratory Data Analysis. *Procedia Computer Science*, *173*, 130–139. doi:10.1016/j.procs.2020.06.017

Jan, M., Awan, A. A., Khalid, M. S., & Nisar, S. (2018). Ensemble approach for developing a smart heart disease prediction system using classification algorithms. *Research Reports in Clinical Cardiology*, 33-45.

JayaSree, M., & Rao, L. K. (2020). *WITHDRAWN: Survey on-Identification of Coronary Artery Disease using Deep Learning*. Academic Press.

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series. Materials Science and Engineering*, *1022*(1), 012072. doi:10.1088/1757-899X/1022/1/012072

Kanwal, S., Hussain, A., & Huang, K. (2021). Novel Artificial Immune Networks-based optimization of shallow machine learning (ML) classifiers. *Expert Systems with Applications*, *165*, 113834. doi:10.1016/j.eswa.2020.113834

Khan, F., Kanwal, S., Alamri, S., & Mumtaz, B. (2020). Hyper-parameter optimization of classifiers, using an artificial immune network and its application to software bug prediction. *IEEE Access : Practical Innovations, Open Solutions*, *8*, 20954–20964. doi:10.1109/ACCESS.2020.2968362

Li, P., Hu, Y., & Liu, Z. P. (2021). Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods. *Biomedical Signal Processing and Control*, *66*, 102474. doi:10.1016/j.bspc.2021.102474

Mehmood, A., Iqbal, M., Mehmood, Z., Irtaza, A., Nawaz, M., Nazir, T., & Masood, M. (2021). Prediction of heart disease using deep convolutional neural networks. *Arabian Journal for Science and Engineering*, *46*(4), 3409–3422. doi:10.1007/s13369-020-05105-1

Olsen, C. R., Mentz, R. J., Anstrom, K. J., Page, D., & Patel, P. A. (2020). Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. *American Heart Journal*, *229*, 1–17. doi:10.1016/j.ahj.2020.07.009 PMID:32905873

Pasha, S. N., Ramesh, D., Mohmmad, S., Harshavardhan, A., & Shabana, . (2020, December). Cardiovascular disease prediction using deep learning techniques. *IOP Conference Series. Materials Science and Engineering*, *981*(2), 022006. doi:10.1088/1757-899X/981/2/022006

Rajamhoana, S. P., Devi, C. A., Umamaheswari, K., Kiruba, R., Karunya, K., & Deepika, R. (2018, July). Analysis of neural networks based heart disease prediction system. In *2018 11th international conference on human system interaction (HSI)* (pp. 233-239). IEEE.

Ramprakash, P., Sarumathi, R., Mowriya, R., & Nithyavishnupriya, S. (2020, February). Heart disease prediction using deep neural network. In *2020 International Conference on Inventive Computation Technologies (ICICT)* (pp. 666-670). IEEE. doi:10.1109/ICICT48043.2020.9112443

Salhi, D. E., Tari, A., & Kechadi, M. T. (2021). Using machine learning for heart disease prediction. In *Advances in Computing Systems and Applications: Proceedings of the 4th Conference on Computing Systems and Applications* (pp. 70-81). Springer International Publishing. doi:10.1007/978-3-030-69418-0_7

Science Direct. (2021). https://www.sciencedirect.com/topics/computer-science/deep-neural-network

Shankar, V., Kumar, V., Devagade, U., Karanth, V., & Rohitaksha, K. (2020). Heart disease prediction using CNN algorithm. *SN Computer Science*, *1*(3), 170. doi:10.1007/s42979-020-0097-6

Sharma, S., & Parmar, M. (2020). Heart diseases prediction using deep learning neural network model. *International Journal of Innovative Technology and Exploring Engineering*, *9*(3), 2244–2248. doi:10.35940/ijitee.C9009.019320

Upretee, P., & Yüksel, M. E. (2021). Accurate classification of heart sounds for disease diagnosis by using spectral analysis and deep learning methods. In *Data Analytics in Biomedical Engineering and Healthcare* (pp. 215–232). Academic Press. doi:10.1016/B978-0-12-819314-3.00014-8

Vivekanandan, T., & Ch Sriman Narayana Iyengar, N. (2017). Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. *Computers in Biology and Medicine*, *90*, 125–136. doi:10.1016/j.compbiomed.2017.09.011 PMID:28987988

Wajid, S. K., & Hussain, A. (2015). Local energy-based shape histogram feature extraction technique for breast cancer diagnosis. *Expert Systems with Applications*, *42*(20), 6990–6999. doi:10.1016/j.eswa.2015.04.057

Wajid, S. K., Hussain, A., & Huang, K. (2018). Three-dimensional local energy-based shape histogram (3D-LESH): A novel feature extraction technique. *Expert Systems with Applications*, *112*, 388–400. doi:10.1016/j.eswa.2017.11.057

Wajid, S. K., Hussain, A., Luo, B., & Huang, K. (2016). An investigation of machine learning and neural computation paradigms in the design of clinical decision support systems (CDSSs). In *Advances in Brain Inspired Cognitive Systems: 8th International Conference, BICS 2016,* Beijing, China, November 28-30, 2016*, Proceedings 8* (pp. 58-67). Springer International Publishing. doi:10.1007/978-3-319-49685-6_6

Waris, S. F., & Koteeswaran, S. (2021). *WITHDRAWN: Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python*. Academic Press.

Wiering, M. A., & Schomaker, L. R. (2014). Multi-layer support vector machines. *Regularization, optimization, kernels, and support vector machines*, 457-475.

World Health Organization. (n.d.). *Cardiovascular diseases*. World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases