Enhancing the Diagnosis of Cardiovascular Disease: A Comparative Examination of Support Vector Machine and Artificial Neural Network Models Utilizing Extensive Data Preprocessing Techniques

ANKUR KUMAR, ASIM ALI KHAN, JASPREET SINGH
Department of EIE,
Sant Longowal Institute of Engineering & Technology,
Sangrur, Punjab,
INDIA

Abstract: - This research delves into the classification of cardiovascular disease (CVD) utilizing state-of-the-art machine learning algorithms, namely Support Vector Machine (SVM) and Artificial Neural Network (ANN). Before model training, extensive data preprocessing techniques were implemented, including data cleaning, feature scaling, encoding, Feature selection, handling imbalanced data, normalization, and cross-validation. After data preparation, an extensive evaluation of performance was carried out against various parameters like accuracy, precision, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR-), and diagnostic odd ratio (DOR). The comparison of SVM and ANN techniques indicates that the SVM has a better sensitivity in detecting positive cases while ANNs have more accuracy in the classification. This paper not only documents the use of new methods but also highlights the advantages and disadvantages of SVM and ANN models, and therefore helps to improve the use of machine learning applications in making health care decisions on CVD diagnosis.

Key-Words: - CVD, Disease classification, SVM, ANN, Feature selection, Confusion Matrix.

Received: April 9, 2024. Revised: November 2, 2024. Accepted: December 2, 2024. Published: December 31, 2024.

1 Introduction

The blood circulation system consists of the heart, an important organ that acts as a muscle that moves blood, and a series of blood arteries such as arteries, veins, and capillaries. All of them constitute a closed system in which blood travels around the body. In every tissue and cell, the passage of blood through small capillaries is fundamentally important. Integration, regulation, and coordination are vital for the efficient functioning of the cardiovascular system, ensuring that blood is delivered to specific body areas according to demand, [1], [2].

Cardiovascular diseases (CVD) encompass various conditions affecting the functioning of the heart, such as abnormal heart rhythms (arrhythmias), aortic infections, Marfan syndrome, congenital heart defects, cardiomyopathy, and stroke. These diseases often share common risk factors, including age, unhealthy diet, gender, high blood pressure, diabetes mellitus, smoking, consumption of processed meats and alcohol, high sugar intake, family history, obesity, lack of physical activity, psychosocial factors, and air pollution, [3], [4].

Preventing CVD is a challenge, and the development of a robust data-driven system for predicting it will enhance our ability to detect it reliably, thus improving research and prevention efforts. This will ultimately enable more people to lead healthier lives. Numerous studies have demonstrated the advantages of ML techniques in predicting heart disease based on various prognostic and bio-clinical factors, including pulse rate, gender, age, and others, [5], [6], [7].

Another challenge in this domain is the abundance of features utilized in predicting CVD, posing considerable difficulty in the task. Moreover, the multitude of features complicates classification in machine learning, consequently impacting performance and diminishing the accuracy of ML systems, [8]. Hence, addressing this issue presents a substantial contribution to the advancement of heart disease diagnosis;

- The study focused on the classification of cardiovascular disease using the UCI heart failure dataset, addressing a critical area in healthcare research.
- Various data preprocessing techniques were employed, including data cleaning, feature

scaling, encoding, feature selection, handling imbalanced data, normalization, and cross-validation, to enhance the dataset's quality and suitability for machine learning model training.

- The research employed machine learning algorithms, SVM and ANN, for cardiovascular disease classification, conducting a comprehensive performance evaluation that considered metrics such as accuracy, precision, specificity, LR+, LR-, and DOR, and provided insights into the advantages and disadvantages of SVM and ANN models.

2 Literature Survey

Numerous researchers have explored methods for diagnosing cardiovascular disease (CVD). Here, are some recent works relevant to the proposed research. In [4] suggesting a cost-sensitive ensemble method comprising five diverse classifiers to enhance the efficacy of heart disease diagnosis and minimize misclassification costs. Through rigorous statistical tests, it was established that the ensemble outperformed individual classifiers significantly. Additionally, the application of the Relief algorithm further enhanced classification efficiency. In [6] proposed multi-tier ensemble (MTE) model, incorporating RF feature selection, demonstrated outstanding performance on the curated dataset, achieving an accuracy of 93.76%. This performance surpassed that of alternative classification models. The evaluation of experimental results encompassed various performance metrics, including accuracy, precision, recall, f-measure, and area under the curve (AUC). In [8] summary, the study explored various machine learning algorithms for classifying heart disease data. Random forest and SVM with grid search performed best on the Cleveland dataset, while logistic regression and naive Bayes were more effective on the Statlog dataset. Employing ANOVA F-test feature selection improved outcomes for both datasets except for naive Bayes. [9], introduced a hybrid approach, GAPSO-RF, which combines genetic algorithm (GA) and particle swarm optimization (PSO) techniques to optimize feature selection for heart disease prediction using random forest (RF) as the classifier. This method aims to enhance prediction accuracy by selecting optimal features. Compared to alternative methods, GAPSO-RF achieves superior accuracy, specificity, sensitivity, and AUC-ROC for heart disease prediction. In [10] Compared five machine learning algorithms (DT, RF, SVM, ANN, and Fuzzy Logic) for predicting heart disease using 15 medical parameters out of 76 collected parameters. The dataset used in the study consisted of 920 records from four different locations, and the data was split into an 80:20 ratio for training and testing the models. Overall, the paper demonstrated the effectiveness of various machine-learning algorithms in predicting heart disease using a subset of medical parameters. The RF algorithm showed the highest accuracy, and the choice of evaluation metrics and dataset ratio can impact the results.

2.1 Motivation

Despite numerous advancements in diagnosing CVD using machine learning algorithms, there remain gaps that this research aims to address. The few studies that have been conducted, focusing on comparative analysis of different modeling techniques, have shown that multi-tier ensembles, random forests with feature selection, and hybrid models that use genetic algorithms with particle swarm optimization work. Also, other studies tend to analyze one specific dataset or apply a few evaluation metrics, which diminishes the scope of the research. Incorporating both issues, this paper intends to fill this gap by analyzing the performance of SVMs and ANNs on several evaluation metrics with the usage of comprehensive data preparation and validation methods. Doing this, it seeks to enhance the understanding of the advantages and disadvantages associated with these types of models which is a critical factor in developing effective CVD detection and health management processes.

3 Materials and Methods

This paper outlines a detailed methodology for machine learning analysis, which consists of several key stages illustrated in Figure 1. It starts with thorough data preprocessing, which includes careful management of missing values, detection of outliers, and standardization to ensure consistency across features. Feature selection is performed using Pearson correlation analysis to eliminate weakly correlated variables. Next, the pre-processed data is input into machine learning algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN). The performance of these models is evaluated using metrics like accuracy, sensitivity, precision, likelihood ratios (LR+ and LR-), and diagnostic odds ratio (DOR) to assess their predictive effectiveness. Finally, the results are compared with existing studies to confirm the methodology's effectiveness in improving model performance and the insights gained.

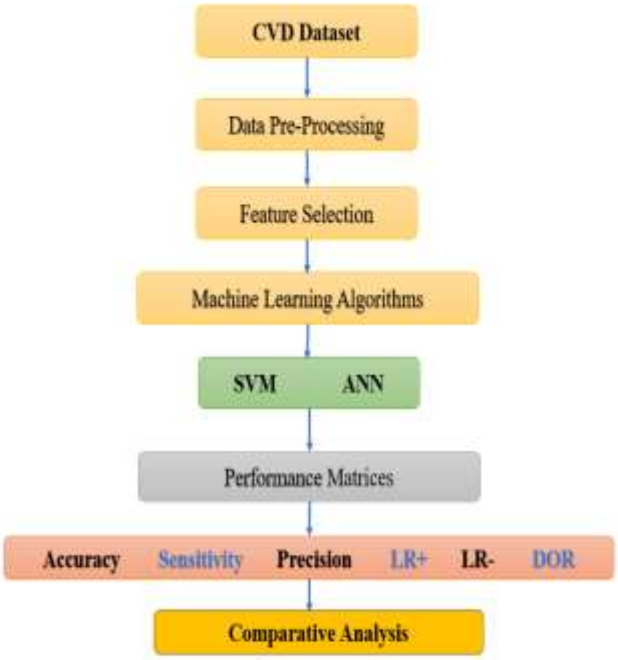


Fig. 1: Methodology Flowchart

3.1 Data Description

The dataset used in this study was sourced from the UCI ML repository [9] and includes 12 features that detail various aspects of patient health, as illustrated in Table 1. These features include age, reflecting the patient's age; anaemia, shows the presence of a specific hemoglobin level; creatinine phosphokinase, measuring the CPK enzyme level in the blood; diabetes, denoting the presence of diabetes; ejection fraction, representing the percentage of blood ejected from the heart during each contraction; high blood pressure, indicating the presence of hypertension; platelets, indicating the count of platelets in the blood; serum creatinine, measuring the serum creatinine level in the blood; serum sodium, gauging the serum sodium level in the blood; sex, denoting the patient's gender; smoking, shows whether the patient smokes; and time, representing the follow-up period in days. The dataset encompasses records of 299 patients, with boolean values represented as 0 for negative (No) and 1 for positive (Yes). The attribute "DEATH_EVENT" indicates whether the patient is deceased during the follow-up period. Furthermore, among these patients, 203 have experienced death events while 96 have remained safe throughout the follow-up period as shown in Figure 2.

Table 1. Data Description

S. No	Feature Descriptions
1.	**age:** Age of the patient
2.	**anaemia:** Haemoglobin level of patient (Boolean)
3.	**creatinine_phosphokinase:** Level of the CPK enzyme in the blood (mcg/L)
4.	**diabetes:** If the patient has diabetes (Boolean)
5.	**ejection_fraction:** Percentage of blood leaving the heart at each contraction
6.	**high_blood_pressure:** If the patient has hypertension (Boolean)
7.	**platelets:** Platelet count of blood (kilo platelets/mL)
8.	**serum_creatinine:** Level of serum creatinine in the blood (mg/dL)
9.	**serum_sodium:** Level of serum sodium in the blood (mEq/L)
10.	**sex:** Sex of the patient
11.	**smoking:** If the patient smokes or not (Boolean)
12.	**time:** Follow-up period (days)
13.	**DEATH_EVENT:** If the patient deceased during the follow-up period (Boolean)

[Attributes having Boolean values: 0 = Negative (No); 1 = Positive (Yes)]

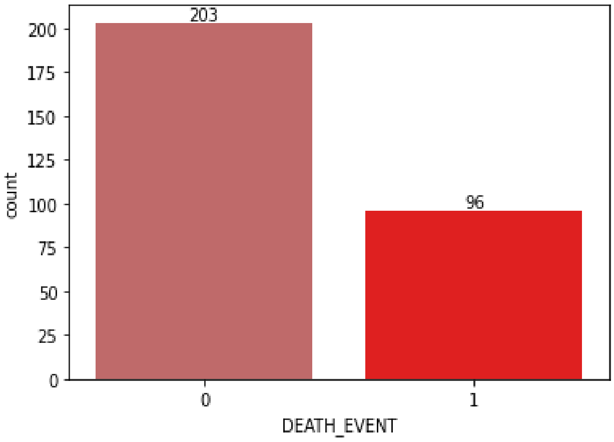


Fig. 2: Death Event Vs. Safe

3.2 Data Preprocessing

In the data preprocessing phase, the first step involved checking for null values in the dataset, which revealed the absence of any such values. After confirming the absence of null values, the dataset was examined for outliers. It is important to note that while some outliers were detected across multiple features, given the dataset's size and relevance, it was decided not to remove such outliers during data preprocessing unless they significantly affected statistical integrity.

Consequently, outliers were replaced with the mean value. Following outlier handling, the next step involved data standardization using the standard scale method, as illustrated in Figure 3.

Pearson correlation was used for feature selection as shown in Figure 4. Significantly, "time" emerged as the most important feature due to its inverse relationship with cardiovascular issues, emphasizing the importance of early diagnosis for timely treatment and reduced fatality risk. After that, "serum_creatinine" stood out as it directly impacts heart function due to its presence in blood.

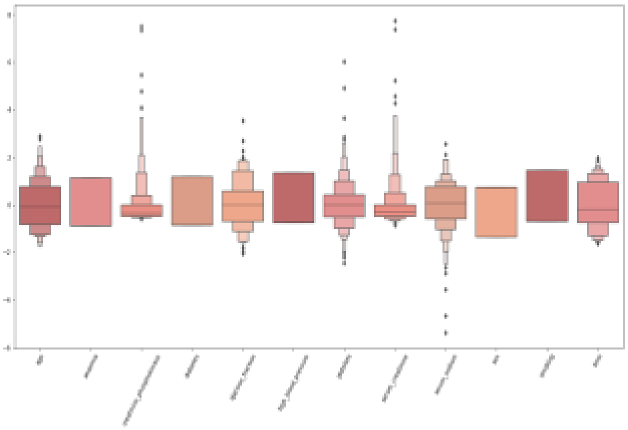


Fig. 3: Data Scaling

Additionally, "ejection_fraction" significantly influenced the target variable, reflecting the heart's efficiency. Furthermore, the inverse relationship observed with aging suggests a decline in heart function over time.

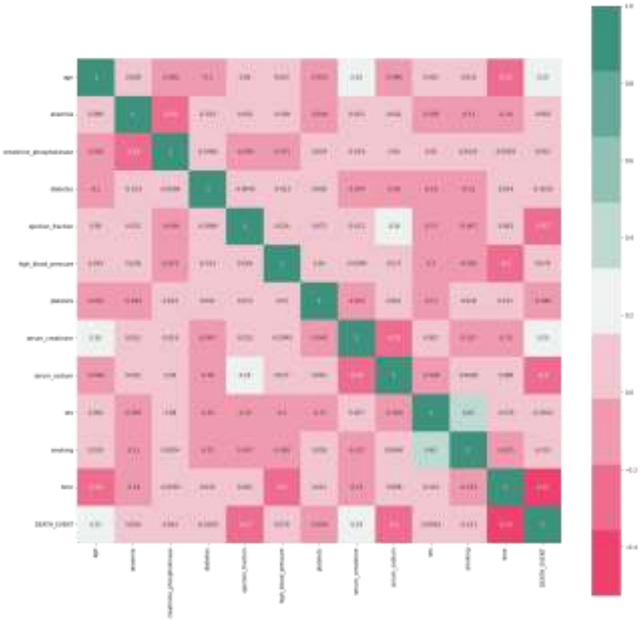


Fig. 4: Correlation heatmap of features vs. Target

3.3 Support Vector Machine (SVM) Classifier

SVM is a set of supervised learning methods widely used in medical diagnosis for both classification and regression tasks. SVM aims to simultaneously minimize empirical classification errors and maximize the geometric margin, earning the name "Maximum Margin Classifiers". It operates based on the statistical learning theory principle of structural risk minimization, providing guaranteed risk bounds. SVMs efficiently handle non-linear classification through the "kernel trick", which implicitly maps inputs into high-dimensional feature spaces, allowing the construction of the classifier without explicit knowledge of the feature space. In an SVM model, examples are represented as points in space, strategically mapped to ensure clear separation between categories with the widest possible gap. The optimal separating hyperplane (OSH) is identified, maximizing the distance between parallel hyperplanes, and minimizing misclassification risks for test examples. The pseudocode of SVM has been shown in Table 2, [10], [11].

Given labeled training data in the form (X_i, Y_i), where $Y_i = \frac{1}{-1}$ denotes the class to which the point belongs, and n represents the number of data samples, with each X_i being p-dimensional real vector, the SVM classifier maps input vectors to decision values and performs classification using an appropriate threshold value, [12], [13].

Table 2. Pseudocode for SVM

```
# Parameters:
# C - Regularization parameter
# kernel - Kernel function (e.g., linear, polynomial, RBF)
# max_iterations - Maximum number of iterations for training
# learning_rate - Learning rate for weight updates
Initialize weights and bias to small random values
for each iteration in max_iterations do
    for each (input, target) pair in training_dataset do
        # Compute the decision value based on the kernel function
        if kernel == "linear" then
            decision_value = dot_product(weights, input) + bias
        else if kernel == "polynomial" then
            decision_value = (dot_product(weights, input) + 1) ** degree + bias
        else if kernel == "RBF" then
            decision_value = exp(-gamma * (||weights - input|| ** 2)) + bias
        # Check if the sample is correctly classified with a margin
        if target * decision_value < 1 then
            # Misclassified, update weights and bias
            weights = weights + learning_rate * (target * input - 2 * C * weights)
            bias = bias + learning_rate * target
        else
            # Correctly classified, apply regularization only
            weights = weights - learning_rate * 2 * C * weights
        end if
    end for
end for
function dot_product(vector1, vector2)
    return sum(vector1[i] * vector2[i] for i in range(length(vector1)))
function euclidean_norm(vector)
    return sqrt(sum(vector[i] ** 2 for i in range(length(vector))))
```

3.4 Artificial Neural Network (ANN) Classifier

ANN classifier is a computational model inspired by how biological neural networks in the human brain process information. It consists of layers of interconnected nodes (neurons), where each node performs a weighted sum of its inputs, applies a non-linear activation function, and passes the output to the next layer. The network typically has an input layer, one or more hidden layers, and an output layer.

The ANN classifier processes input data through the network during training, generating predictions. The predictions are compared to the true class labels, and the error is computed with a loss function.

This error is then sent back through the network using backpropagation, which adjusts the weights to reduce the error. This process uses gradient descent optimization to gradually improve the weights. Once trained, the ANN classifier can effectively map new inputs to their corresponding class labels by leveraging the learned patterns and features in the data, making it a powerful tool for tasks such as image and speech recognition, and natural language processing.

In ANN for binary classification, the forward pass consists of computing the weighted sum of inputs at each layer, applying an activation function to add non-linearity, and producing the final output. Specifically, for a network with one hidden layer, the hidden layer's output is computed as $a^{(1)} = \sigma(W^{(1)} \cdot X + b^{(1)})$, σ is the sigmoid function. The output layer is then calculated as $a^{(2)} = \sigma(W^{(1)} \cdot a^{(1)} + b^{(2)})$.

The loss, measured by binary cross-entropy, quantifies the difference between the predicted output $a^{(2)}$ and the actual label y. During backpropagation, gradients of the loss concerning weights and biases are computed to update the parameters using gradient descent, with adjustments made to the weights $W^{(1)}$ and $W^{(2)}$and biases $b^{(1)}$ and $b^{(2)}$ to minimize the loss, the pseudocode of ANN is outlined in Table 3, [14], [15].

3.5 Performance Matrices

Evaluating ML models involves more than just accuracy. Metrics like precision measure the accuracy of positive predictions, while recall assesses the model's capability to capture all true positives. LR+ and LR- offer insights into test results, and the DOR combines them to evaluate a test's discriminatory power, [16], [17]. These metrics are derived from the confusion matrix as shown in Table 4. Understanding them aids in selecting the suitable evaluation method tailored to your specific ML objective, [18].

Table 3. Pseudocode of ANN

```
Initialize weights randomly
for each epoch in number_of_epochs do
for each (input, target) pair in training_dataset do
# Forward pass
input_layer_output = input
hidden_layer_output = activation_function(weighted_sum(input_layer_output, hidden_layer_weights))
output_layer_output = activation_function(weighted_sum(hidden_layer_output, output_layer_weights))
# Compute error (using Mean Squared Error as an example)
error = target - output_layer_output
# Backward pass (backpropagation)
output_layer_delta = error * activation_function_derivative(output_layer_output)
hidden_layer_error = dot_product(output_layer_delta, transpose(output_layer_weights))
hidden_layer_delta = hidden_layer_error * activation_function_derivative(hidden_layer_output)
# Update weights
output_layer_weights += learning_rate * outer_product(hidden_layer_output, output_layer_delta)
hidden_layer_weights += learning_rate * outer_product(input_layer_output, hidden_layer_delta)
end for
end for
function activation_function(x)
return 1 / (1 + exp(-x))  # Sigmoid function
function activation_function_derivative(x)
return x * (1 - x)  # Derivative of sigmoid function
function weighted_sum(inputs, weights)
return dot_product(inputs, weights)
function dot_product(vector1, vector2)
return sum(vector1[i] * vector2[i] for i in range(length(vector1)))
function outer_product(vector1, vector2)
return [[vector1[i] * vector2[j] for j in range(length(vector2))] for i in range(length(vector1))]
```

Table 4. Performance Matrice

S. No	Performance Metrics	Formula
1.	Accuracy	$\frac{(TN + TP)}{(FP + TP + FN + TN)}$
2.	Precision	$\frac{TP}{(FP + TP)}$
3.	Sensitivity	$\frac{TP}{(TP + FN)}$
4.	LR+	$\frac{TPR}{FPR}$
5.	LR-	$\frac{FNR}{TNR}$
6.	DOR	$\frac{LR +}{LR -}$

4 Result and Discussion

In this study, two ML algorithms, SVM and ANN were applied to a dataset having 299 patient records characterized by 12 features. The dataset underwent a partitioning into a 70/30 ratio for training and testing. Both models' performance was analyzed by performance metrics including accuracy, precision, specificity, LR-, LR+, and DOR. To validate the effectiveness of these models, ten-fold cross-validation was employed. Figure 5 and Figure 6 present the confusion matrix corresponding to the SVM and ANN models, respectively. These matrixes show an overview of the classification performance of each model, detailing the distribution of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions across different classes or categories.

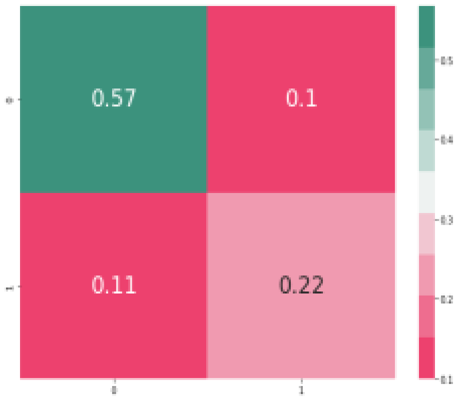


Fig. 5: Confusion matrix of SVM

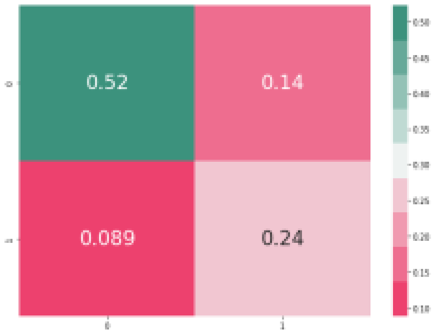


Fig. 6: Confusion matrix of ANN

Figure 7 and Figure 8 depict the training and validation loss, along with the training and validation accuracy, specifically for the ANN model being analyzed. These visualizations give insights into the model's performance and its ability to generalize during the training phase.

Results from Figure 9 indicate that both the SVM and ANN models achieve high overall accuracy, with SVM slightly surpassing ANN by a marginal 2%. However, in terms of the sensitivity (recall) metric, SVM shows superior performance in correctly identifying positive instances compared to ANN, implying a stronger capability in capturing instances belonging to a specific class. Conversely, ANN displays a higher precision, signifying its proficiency in accurately classifying positive instances among all instances it predicts as positive.

These findings collectively suggest that while SVM excels in correctly identifying instances of interest, ANN may offer better precision in its classifications, potentially providing a balanced performance across different aspects of model efficacy.

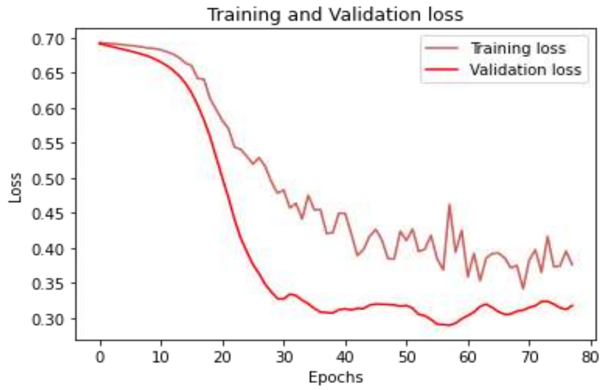
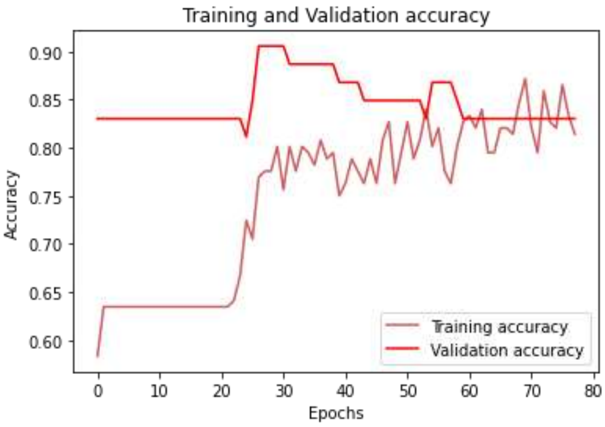
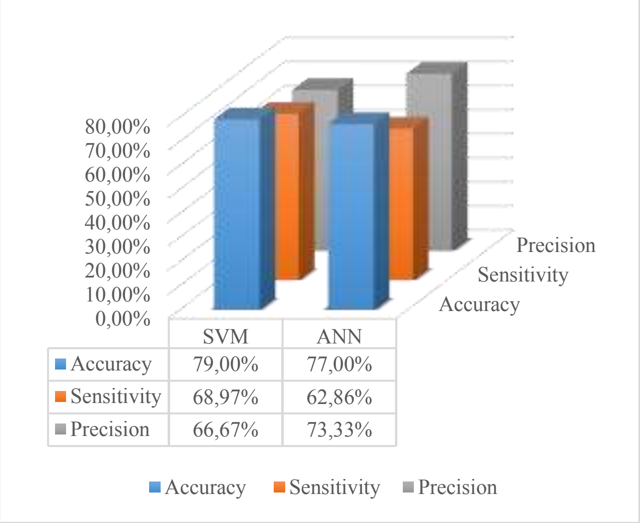


Fig. 7: Training and validation loss of the ANN model





	SVM	ANN
■ Accuracy	79,00%	77,00%
■ Sensitivity	68,97%	62,86%
■ Precision	66,67%	73,33%

Fig. 8: Training and validation accuracy of the ANN model

Fig. 9: Performance matrices of SVM and ANN

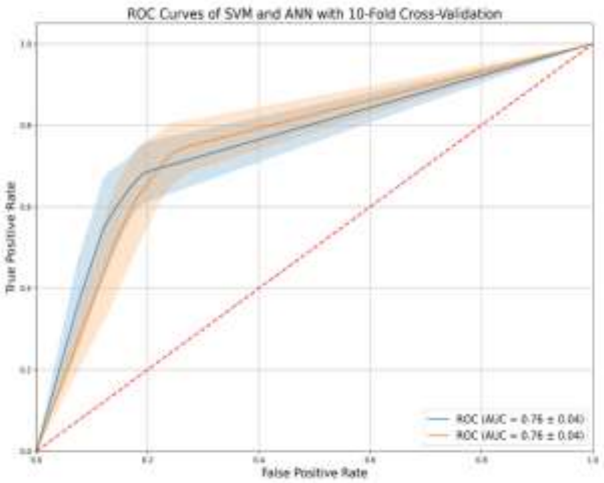


Fig. 10: ROC curve of SVM and ANN with 10-fold cross-validation

Figure 10 presents the receiver operating curves (ROC) for SVM and ANN models using tenfold cross-validation, depicting the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). These curves demonstrate how well the models can distinguish between different classes at various threshold levels, with the Area Under the Curve (AUC) acting as a crucial performance metric higher AUC values indicate superior model performance. The tenfold cross-validation process enhances reliability by dividing the dataset into ten segments, training on nine of them, and validating the remaining one in a repeated manner, which gives a broader evaluation of the model's effectiveness. These ROC curves provide additional insights into the models' classification abilities, complementing the performance metrics illustrated in Figure 9.

Other than the above-mentioned performance measures, LR+, LR-, and DOR give an overview of how well the models would work for healthcare improvement have been shown in Figure 11. Examining these metrics, SVM also promises good performance, though ANN gives a slightly higher LR+ of 4.32 compared to SVM's LR+ of 4.21, which means the ability to correctly classify a positive case is slightly more prominent with ANN. However, ANN has a lower LR- of 0.4347 than SVM's LR- of 0.3712, meaning it has a better ability to identify negative cases. Moreover, even though SVM has a higher DOR of 11.33, which means better overall discriminatory power, ANN still has a DOR of 9.94, which is a great indicator of its effectiveness in diagnostic decision-making.

These findings highlight the ability of both SVM and ANN models to enhance healthcare by supporting accurate diagnoses and informed decision-making processes.

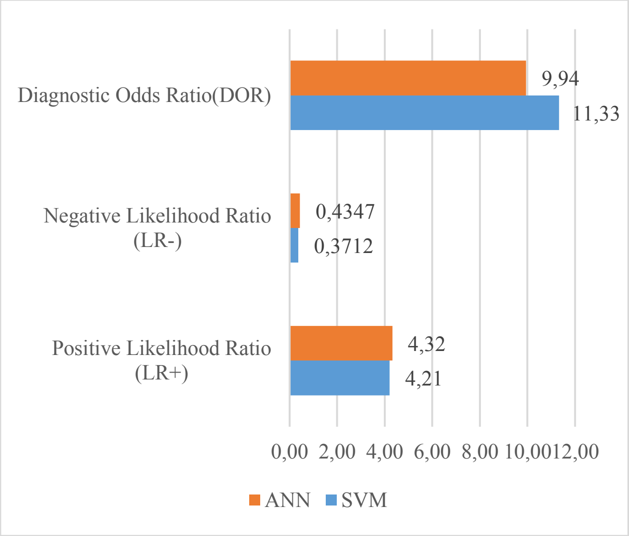


Fig. 11: Diagnostic matrices for SVM and ANN

5 Comparative Analysis

After the result analysis, a comprehensive comparative study was done to compare the different methodologies with the available literature as shown in Table 5. [14], used an outlier elimination pre-processing technique along with an SVM classifier and achieved an accuracy of 76%. Similarly, [15], used data normalization followed by an SVM classifier and achieved a slightly higher accuracy of 78%. Another study by [19] also normalized the data using an SVM classifier but their accuracy was marginally low at 70%.

Contrary to the proposed method, this had multiple features integrated pre-processing steps: data scaling, outlier handling, and correlation analysis before the application of the SVM classifier. The highest accuracy of 79% in the classification of CVD was thus achieved. Data scaling ensures that all features are equal contributors to the model; outlier handling enhances the robustness of the model by filtering out noise; and correlation analysis performs feature selection on the most relevant feature correlation sets.

This comparative analysis introduces a method's effectiveness: emphasizing the comprehensive preprocessing of data enhances the overall performance of the SVM classifier in predictive outcomes for CVD. The accuracy improved to 79%, and that was with a proper multi-aspect pre-processing strategy. It is essential that for better model performance than methods relying on fewer or one pre-processing technique, appropriate and broad preprocessing should be applied.

Table 5. Comparative Analysis

Authors	Pre-processing	Classifiers	Accuracy (%)
[14]	Elimination of outliers	SVM	76
[15]	Data normalization	SVM	78
[19]	Data normalization	SVM	70
Proposed	Data Scaling, outlier handling, correlation	SVM	79

6 Conclusion and Future Scope

The authors have classified CVD by using the UCI heart failure dataset with two machine learning algorithms, namely SVM and ANN. The performances of these models were tested with several performance metrics including accuracy, precision, specificity, LR+, LR-, and DOR. Overall, both SVM and ANN display very high accuracy, although SVM performs better than ANN with a margin of 2%. However, when considering sensitivity (recall), SVM demonstrated superior performance in correctly identifying positive instances compared to ANN. This suggests that SVM has a stronger capability in capturing instances belonging to a specific class. Conversely, ANN displayed higher precision, indicating its proficiency in accurately classifying positive instances among all instances it predicts as positive. These findings collectively suggest that while SVM excels in correctly identifying instances of interest, ANN may offer better precision in its classifications. Following the result analysis, a comparative examination between the proposed and existing studies was conducted. It was observed that the proposed approach outperforms the existing study in terms of accuracy.

In the future, there is potential for integrating the employed algorithms with IoT (Internet of Things) devices for healthcare decision-making. This integration could enable real-time monitoring of health parameters and facilitate timely interventions. By incorporating ML algorithms such as SVM and ANN into IoT devices, healthcare professionals can access predictive analytics and decision support tools, aiding in the early detection of cardiovascular disease and personalized patient care. Additionally, IoT-enabled healthcare devices could offer continuous data collection, allowing for long-term trend analysis and proactive management of cardiovascular health.

Acknowledgement:
I would like to express my gratitude to my senior, Dr. Sanjay Dhanka, and head of the department Dr. Surita Maini for their invaluable assistance throughout my research.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process:
During the preparation of this work, the authors used Quillbot to address grammatical errors and enhance readability. After the use of this tool, the authors thoroughly reviewed and edited the content as necessary and take full responsibility for the accuracy and integrity of the publication.

References:

[1] A. Tiwari, A. Chugh, and A. Sharma, “Ensemble framework for cardiovascular disease prediction,” *Comput Biol Med,* vol. 146, Jul. 2022, doi: 10.1016/j.compbiomed.2022.105624.

[2] S. Gupta Dogiparthi, “A Comprehensive survey on Heart Disease Prediction using Machine Intelligence,” *International Journal of Medical Research and Health Sciences*, 10 (2021): 60-68.

[3] Karadayı Ataş, P. “Exploring the molecular interaction of PCOS and endometrial carcinoma through novel hyperparameter-optimized ensemble clustering approaches” *Mathematics*, *12*(2), 295,2024, https://doi.org/10.3390/math12020295.

[4] Q. Zhenya and Z. Zhang, “A hybrid cost-sensitive ensemble for heart disease prediction,” *BMC Med Inform Decis Mak,* vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01436-7.

[5] Abha Marathe, Virendra Shete and Dhananjay Upasani, “A Knowledge Based Framework for Cardiovascular Disease Prediction” *International Journal of Advanced Computer Science and Applications(IJACSA)*, 14(5), 2023. http://dx.doi.org/10.14569/IJACSA.2023.0140556.

[6] Dhanka, S., Maini, S. “HyOPTXGBoost and HyOPTRF: Hybridized Intelligent Systems using Optuna Optimization Framework for Heart Disease Prediction with Clinical Interpretations”. *Multimed Tools Appl.* 83, 72889–72937, (2024). https://doi.org/10.1007/s11042-024-18312-x.

[7] Arora, S., Vedpal & Chauhan, N. “Polycystic Ovary Syndrome (PCOS) diagnostic methods in machine learning: a systematic literature review”. *Multimed Tools Appl* (2024). https://doi.org/10.1007/s11042-024-19707-6.

[8] Kumar, A., Dhanka, S., Singh, J., Ali Khan, A., & Maini, S. (2024). Hybrid machine learning techniques based on genetic algorithm for heart disease detection. *Innovation and Emerging Technologies*, *11*, 2450008,(2024), https://doi.org/10.1142/S2737599424500087

[9] S. Y. Hera, M. Amjad, and M. K. Saba, “Improving heart disease prediction using multi-tier ensemble model,” *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 11, no. 1, Dec. 2022, doi: 10.1007/s13721-022-00381-3.

[10] V. Avasthi, A. Kumar, A. Bhardwaj and T. Jain, "Empowering Women's Health: Machine Learning for PCOS Detection and Prediction," *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, Bengaluru, India, 2024, pp. 1-6, doi: 10.1109/ICDCOT61034.2024.10516171.

[11] U. A. Musa and S. A. Muhammad, “Enhancing the Performance of Heart Disease Prediction from Collecting Cleveland Heart Dataset using Bayesian Network,” *Journal of Applied Sciences and Environmental Management,* vol. 26, no. 6, pp. 1093–1098, Jun. 2022, doi: 10.4314/jasem.v26i6.15.

[12] Chicco, G. (2020). *Heart failure clinical records*. University of California, Irvine Machine Learning Repository, doi: https://doi.org/10.24432/C5XW24.

[13] Faris, N. N., & Miften, F. S. (2022). Detection of PCOS Based on Genetic Algorithm Coupled with SVM. *Journal of Education for Pure Science-University of Thi-Qar*, *12*(2), 73-84, doi: 10.32792/utq.jceps.12.02.08.

[14] Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR)*, 6(2), 1-12, [Online]. https://www.researchmanuscripts.com/March2015/2.pdf (Accessed Date: December 1, 2024).

[15] C. Jegan, V. A. Kumari, and R. Chitra, “Classification Of Diabetes Disease Using

Support Vector Machine," vol. 3, pp. 1797–1801, [Online]. https://www.researchgate.net/publication/320395340 (Accessed Date: December 1, 2024).

[16] R. Azziz et al., "Position statement: Criteria for defining polycystic ovary syndrome as a predominantly hyperandrogenic syndrome: An androgen excess society guideline," *Journal of Clinical Endocrinology and Metabolism*, vol. 91, no. 11, pp. 4237–4245, 2006, doi: 10.1210/jc.2006-0178.

[17] S. Dhanka and S. Maini, "Random Forest for Heart Disease Detection: A Classification Approach," in *2021 IEEE 2nd International Conference on Electrical Power and Energy Systems, ICEPES 2021, Institute of Electrical and Electronics Engineers Inc.*, 2021. doi: 10.1109/ICEPES52894.2021.9699506.

[18] Kaur, S., Taneja, S., Khetarpal, V., Garg, K., Sadana, S., Aggarwal, K. (2024). Diagnosis of Polycystic Ovary Syndrome Using Feature Selection-Based Machine Learning Algorithms. In: Hassanien, A.E., Anand, S., Jaiswal, A., Kumar, P. (eds) Innovative Computing and Communications. *ICICC 2024. Lecture Notes in Networks and Systems*, vol 1043. Springer, Singapore. https://doi.org/10.1007/978-981-97-4228-8_26.

[19] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," in *Procedia Computer Science*, Elsevier B.V., 2016, pp. 962–969. doi: 10.1016/j.procs.2016.05.288.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Ankur Kumar: Data curation, Resources, Investigation, Formal analysis, Conceptualization, Methodology, Project administration, Prepared original draft, Review and editing, Validation.
- Jaspreet Singh and Asim Ali Khan: Methodology, Supervision, Project administration, Review and editing.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors declare that they have no conflict of interest.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US