Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Efficient Hybrid Classification Model for Heart Disease Prediction

Maaham Munsif
National University of Sciences and Technology
Mehvish Rashid
mehvish.rashid@seecs.edu.pk
National University of Sciences and Technology
Farzana Jabeen
National University of Sciences and Technology

Research Article

Keywords: Machine Learning, Deep Learning, Feature Selection, Classification, Heart Disease, Prediction, UCI, Support Vector Machine (SVM), Convolutional Neural Network (CNN), Genetic Algorithm (GA)

Posted Date: January 31st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3863899/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: Heart disease prediction is a critical healthcare task for identifying individuals at risk and enabling timely intervention. Accurate prediction models can help reduce morbidity and mortality rates associated with cardiovascular conditions. Various approaches, such as feature selection and hybrid models, have been proposed to improve the effectiveness and accuracy of the prediction of heart disease.

Method: The study employs the Genetic Algorithm-Support Vector

Machine-Convolutional Neural Network (GA-SVM-CNN) approach and evaluates it on three diverse datasets: UCI, Z-Alizadeh Sani, and Cardiovascular Disease Dataset. The genetic algorithm is utilized first to select the most relevant features from the datasets, effectively reducing dimensionality, eliminating irrelevant or redundant features, and choosing the most suitable ones. Subsequently, the hybrid SVM-CNN model is trained using the selected features, harnessing the complementary capabilities of both techniques to enhance prediction accuracy.

Results: The performance of the GA-SVM-CNN approach is assessed using the three benchmark datasets and models. On the UCI dataset, the approach achieves an impressive accuracy of 98%, indicating its effectiveness in accurately predicting heart disease. On the Z-Alizadeh Sani dataset, the approach achieves an accuracy of 97%. On the Cardiovascular Disease Dataset, the approach achieves an accuracy of 86%. These high accuracy rates across different datasets underscore the efficacy of the GA-SVM-CNN approach in heart disease prediction.

Conclusion: The combination of the genetic algorithm's feature selection and the hybrid SVM-CNN model's predictive power leads to superior performance in heart disease prediction. By accurately identifying individuals at risk of heart disease, this approach can enable timely interventions and contribute to more effective healthcare interventions.

Background

Cardiovascular disease is a critical medical condition that ranks as the third leading cause of death worldwide, resulting in approximately 18 million deaths annually [1, 2]. Early prediction of heart disease is crucial for improving patient outcomes, preventing complications, and reducing healthcare costs [3]. Traditional risk factors

provide valuable information, but the integration of multiple data sources and machine learning approaches has shown promise in enhancing prediction accuracy [4, 5, 6, 7]. Based on research, it has been identified that individuals experiencing heart disease commonly exhibit various symptoms [5]. These include chest pain or discomfort, known as angina, shortness of breath, cold sweat, nausea, and episodes of lightheadedness or sudden dizziness [8].

Machine learning has shown great contributions in the prediction of the risk of heart disease, enabling early intervention and personalized treatment plans. The current predictive models (e.g. SVM, Neural Networks, Hybrid, and ensemble models) used by healthcare professionals often rely on traditional approaches such as logistic regression (e.g. presence or absence of disease) or decision trees (e.g. high-risk or low-risk) [9, 10]. However, these approaches are limited in their accuracy, as they tend to rely on a single set of features and metrics that can be used to accurately predict the risk of heart disease. As such, there has been an increased interest in hybrid machine learning models that can combine features from various sources such as clinical history, lifestyle behaviors, and physiological data for better prediction accuracy. In recent years, there has been a notable emphasis on the development of hybrid classification models in the field of heart disease prediction [11]. These models aim to

enhance accuracy and performance by leveraging the strengths of multiple machine learning algorithms. Such studies aim to advance the accuracy and overall performance of heart disease prediction, leading to improved diagnostic and prognostic capabilities. [12].

In predicting heart disease, the hybridization of feature selection and classification techniques involves combining two essential steps [9]. The first step is the identification of relevant features from a pool of potential predictors, while the second step entails the utilization of a classification algorithm based on selected important features to predict the presence or absence of heart disease [9]. The integration of feature selection and classification techniques offers numerous advantages in heart disease prediction. Most notably, it increases the accuracy of prediction models by focusing solely on the most informative features (e.g. age, gender, cholesterol level, blood pressure, etc.). By doing so, these hybrid approaches can improve the efficiency and effectiveness of heart disease prediction systems [13]. The hybridization of feature selection and classification techniques holds significant promise in improving the accuracy and interpretability of heart disease prediction models. By identifying relevant features and utilizing effective classification algorithms, these hybrid approaches contribute to more accurate risk assessment, personalized treatment plans, and improved patient outcomes. Hybrid classification models are a promising approach for improving the accuracy and robustness of heart disease prediction. By combining the strengths of different machine learning algorithms and integrating other techniques, such as feature selection and data augmentation, we can develop more accurate and effective heart disease prediction models, enabling early intervention and personalized treatment plans. The integration of AI techniques in the form of hybrid ML models presents an exciting avenue for researchers. These models can aid in predicting and preventing heart disease, ultimately improving patient outcomes [14].

In our groundbreaking research study, we present an innovative and powerful hybrid technique aimed at revolutionizing heart disease prediction. The proposed

approach harnesses the collective strength of a genetic algorithm for feature selection and a cutting-edge hybrid SVM-CNN model for prediction. Our ultimate objective is to identify the most pertinent features through the genetic algorithm's intelligent search and subsequently, improve the overall performance of the prediction model. The genetic algorithm plays a pivotal role in this novel approach, as it employs a bio-inspired optimization process to search for the most relevant features within the vast pool of potential predictors. By iteratively selecting and evolving feature subsets based on their fitness and contribution to the prediction model, the genetic algorithm efficiently hones in on the key determinants of heart disease risk. This feature selection process is crucial for enhancing the prediction model's efficiency and interpretability by focusing on the most informative attributes while minimizing noise and irrelevant factors.

The core of our innovative approach lies in the integration of a hybrid SVM-CNN model for heart disease prediction. Support Vector Machines (SVM) have proven to be highly effective classifiers in various domains, including healthcare. By combining SVM with Convolutional Neural Networks (CNN), we capitalize on the CNN's exceptional ability to extract complex patterns from raw data, complementing the SVM's robustness in handling high-dimensional feature spaces. This synergistic fusion empowers our prediction model to capitalize on both global and local features within the data, resulting in superior accuracy and performance. The experimental results showcase remarkable outcomes, with our hybrid approach achieving significantly higher accuracy due to its ability to leverage critical and informative features extracted by the genetic algorithm. Our groundbreaking study opens up new horizons for heart disease prediction, providing healthcare professionals with an invaluable tool for early risk assessment and targeted interventions to combat this prevalent and life-threatening condition. The primary contributions of this research lie in the development of a novel hybrid technique that synergistically combines a

genetic algorithm for feature selection and a hybrid SVM-CNN model for heart disease prediction. This approach enhances the accuracy and interpretability of the predictive model, leading to more effective early risk assessment and targeted interventions. By advancing heart disease prediction research, the study opens new avenues for improved patient care and cardiovascular disease management.

The paper is structured into the background, literature review, methodology and proposed solution, results, and discussion, conclusion and future work sections.

Literature Review

The literature review section provides an overview of existing research on heart dis- ease prediction and related methodologies. It highlights previous studies, predictive models, and feature selection techniques used in this domain, demonstrating the research background and the need for the proposed hybrid approach. Heart disease is a complex and widespread condition that poses significant challenges in health- care, particularly in cardiology, detecting cardiac disease accurately and early is of utmost importance [15]. Researchers have developed various intelligent systems that can improve heart disease prediction [16]. Within the realm of heart disease pre- diction, a variety of machine learning and deep learning algorithms and approaches have been utilized. The classification process typically involves two key steps: the selection of relevant features and the training of the model.

The utilization of feature selection methods to enhance the accuracy of heart disease prediction models is addressed in a research paper [17]. The experimental results in [17] show that feature selection techniques can significantly gain accuracy up to 88% and reduce the dimensionality of heart disease prediction models. The paper [18] provides valuable insights into suitable machine-learning approaches for heart disease prediction and got up to 95% accuracy. In a study conducted by the researchers [19], the objective was to assess the performance of different feature selection techniques and classification algorithms for heart disease prediction. The author [19] found that the feature subset selected by the backward feature selection achieves the highest accuracy of 88.52% with the decision tree classifier. Through the use of a genetic algorithm for feature selection prior to classification, the scientists hope to increase the precision and effectiveness of heart disease prediction, the prediction accuracy increased from 72.9–91.9% by using the evolutionary algorithm for feature selection prior to classification [20].

Lakshmanarao et al. [21] used feature selection methods like information gain

and chi-square to select the most relevant features from the original dataset that contribute the most to the prediction accuracy. Ensemble learning techniques like Bagging, Ada-Boost, and Random Forest were applied to the selected features. This led to improved prediction accuracy compared to using a single classification model. This study [9] used 6 different machine learning classification algorithms on 2 different data sets to conduct an experimental investigation of the model's effectiveness. The Random-Forest classification algorithm performed best in both datasets, with a 93.25 percent accuracy in the second [9] dataset and a 77.25 percent accuracy in the first. Multiple heart disease datasets are utilized for experimentation analysis in this paper [22], which focuses on feature-selection strategies and algorithms and demonstrates the accuracy increase. In [23], Al-Tashi et al. suggest a hybrid feature selection strategy for diagnosing heart disease combining Support Vector Machines (SVM) and Grey Wolf Optimisation (GWO). It is employed to look for the best possible subset of features to use in order to improve classification accuracy [23].

Using the UCI heart illness dataset, one such work [12] suggested a hybrid classifi-

cation model for heart disease prediction. ANN is trained using swarm optimization, which optimizes its weight and biases to increase performance by up to 89.83%. An- other work [24] suggested a hybrid classification model for heart disease prediction. The SVM 85% UCI Cleveland heart disease dataset produced the highest accuracy. A study [25] used a filter algorithm for reducing low-value features and GA is used to select high-value features. Statlog dataset used from UCI Repository. Dissanayake et al. [19] found that the feature subset selected by the backward feature selection achieves the highest accuracy of 88.52% with the decision tree classifier. The main achievement of the paper [26] is its accuracy and effectiveness in predicting heart disease. The researchers tested the system on three datasets of heart disease patients and found that the random forest classifier achieved an average accuracy of around 90% when optimized using the evolutionary approach [26]. In [15] heart disease is predicted by using 10 different classifiers, six are single classifiers and one is an ensemble classifier with four different variations, the best accuracy scores are obtained as 91.78%, 85.47%, for the Z-Alizadeh Sani and Cleveland using MLP classifiers and 86.66% for the Statlog dataset using KNN classifier. In [27] proposed

a hybrid machine learning approach for heart disease prediction using a Genetic Algorithm with radial basis function (GA-RBF), using dataset Cleveland was taken from UCI Repository and this model achieved an accuracy of 94.20% while considering 9 features.

The proposed technique in [28] is tested on a small sample size and the perfor-mance may be compromised due to the high computational complexity of algo-rithms. In [29], a comparative analysis is performed among 3 proposed algorithms to measure the effectiveness and accuracy of the system. This study explores the use of an embedded feature selection method [30], combined with a deep neural network, to predict the presence of heart disease. The main contribution of the paper [30] is in demonstrating the effectiveness of an embedded feature selection method using linear-SVC and deep neural network for heart disease prediction, the authors concluded that the model was successful, with a prediction accuracy of 97%. The hybrid model (RF-DT) in [31] is applied to clinical data taken from the UCI repository named as Cleveland dataset to assess its accuracy and perfor- mance, using a hybrid approach 88.6% accuracy was obtained which is higher than DT and DF. Hybrid approaches for disease prediction and classification are very popular nowadays because of their performance and accuracy [32]. An advantage is taken in [33], by proposing a hybrid approach using an artificial bee algorithm and a genetic algorithm. Results show that employing and combing these algorithms can effectively improve the accuracy and performance of the feature selection pro- cess. A general disease prediction model proposed in [34], is based on the patient's symptoms. K-Nearest Neighbour (KNN) and Convolutional Neural Network (CNN) machine learning algorithms are used to accurately forecast disease. To handle the problem of data consistency and irregularity a new convolutional neural network (CNN) based on a multimodal disease risk prediction algorithm is proposed in [9]. The paper [35] discusses methods for identifying the most relevant features for prediction, as well as methods for eliminating redundant features, the Cleveland Heart dataset from the UCI repository was used in the study, and the different classification models are used in this study, the accuracy achieved 94.92% using Lasso-regression and ridge-regression. The paper [24] presents an overview of the performance of several supervised classification models, It concluded that the sup- port vector machine achieved the highest performance at 85%, and 87% in terms of accuracy and precision. In [36] examines the potential for a hybrid CNN and LSTM network to predict heart disease, improved prediction with 89% accuracy compared to traditional approaches.

The objective of this study [37], is to create an efficient method for predicting heart disease using an ensemble technique based on sequential feature selection. The authors use the gradient-boosting-based sequential feature selection technique to choose the most appropriate features from the training dataset. Table 1 shows the existing techniques, datasets used, and their limitations. An innovative way of predicting heart-disease risk using a hybrid

GA and PSO approach is presented in this paper [38]. For precise cardiac disease prediction, a Hybrid Trunicate Swarm Algorithm and Ensemble Deep Learning (TSA-EDL) technique are used [39]. In conclusion, the hybrid machine learning model shows promise for heart disease pre- diction, offering improved accuracy and robustness. By integrating diverse machine

Table 1
Existing Techniques

Study Reference & **Techniques**	**Datasets** & **Accuracy** & **Limitations**			
V. K. Sudha et al. 2023 [36]	**weight by SVM for featureselection CNN + LSTM for prediction**	**Clevelanddataset from UCI Repository**	**89%**	**It only evaluates the model on a single dataset**
Girish S. Bhavekar et al. 2022 [40]	HybridRNNand LSTM for prediction	Clevelanddataset from UCI Repository	95%.	small dataset,No pre-processing techniquesare defined.
Jafar Abdol- lahi et al. 2022 [25]	Filter algorithm and genetic algorithm for reduction and selec- tion of features.	Statlog dataset from UCI Repository	97%	Potential inaccu-racies due to un- equal feature im- portance
Senthil Murugan Nagarajan et al. 2022 [41]	genetic-based crow search algorithm for feature selection and DNN for classification	Statlog dataset from UCI Repository	88%	Accuracy is lim- ited by relying on a small sample set of data.
BP Doppala et al. 2021 [27]	GA-RBF	Clevelanddataset from UCI Repository	94.20%	single dataset.
Gárate- Escamila et al. 2020 [28]	Dimensionality reduc- tion method and prin- cipal component anal- ysis	Clevelanddataset from UCI Repository	98%.	Raw data com- puted lower re- sults and would require greater.
Karunakaran Velswamy et al. 2021 [42]	Modified bee algo- rithm for feature selection and SVM, Naïve Bayes, and KNN for classification	UCI Repository	85%	Withfewer features of the dataset
Dr.M. Kavitha et al. 2021 [31]	Hybrid DT-RF for pre- diction	Clevelanddataset from UCI Repository	88%	no pre-processing
Bindu M G et al. 2020 [33]	Artificial bee colony and Genetic algorithm for feature selection	Clevelanddataset from UCI Repository	91.8%	Limited in its data used to evaluate the effectiveness.

learning algorithms and techniques like feature selection, we can develop more effective models for early intervention and personalized treatment. However, further enhancements are needed to maximize performance.

Integrating multiple algorithms that capture different patterns can create a stronger predictive model for heart dis ease.

Methodology and Proposed Solution

The literature review process helps to understand and analyze the existing techniques in a comprehensive manner. The Review process assists in founding a re- search methodology process and basic pillars behind the whole study. After a de- tailed literature review, some RQs are drawn on the basis of previous techniques and information. The RQs help to understand the proposed work and methodology in a detailed manner. Following are the research questions for the proposed study.

RQ-1: What are current techniques for Heart Diseases prediction in Machine Learning? (answered in the Literature review)

RQ-2: Can the SVM-CNN be combined to outperform ML algorithms for heart disease prediction?

RQ-3: How can we evaluate and validate the criteria of the SVM-CNN model?

RQ-4: How can we compare the performance of the proposed SVM-CNN model with an existing model?

RQ2, RQ3, and RQ4 are discussed in the result and discussion section. Detailed information about the research process is explained in the methodology section, and the research methods used, and the data processing methods are explained in a detailed manner. The approach used in scientific research is composed of a few basic steps. Each study differs in some way due to the period of time, conditions, and location in which it is being conducted. The research process is a common and well-known phenomenon. A modification in one phase of the research process will also affect the other stages because each phase is interconnected. Researchers must evaluate the other stages after introducing a modification in one phase to ensure that changes are reflected in all phases.

The proposed model mainly focuses on identifying heart disease with better ac- curacy using selected features. The performance of the ML algorithm has been tested on three datasets UCI Dataset, Z-Alizadeh Sani Dataset, and Cardiovascular Dataset. The workflow of the model has been implemented in three stages shown in Fig. 1. Input phase, Feature Selection, and the last step is the classification and performance evaluation of heart disease. These are the following steps that we will discuss in this section.

Input Phase

In the input phase, various pre-processing techniques are applied to each dataset to enhance the proposed model's capabilities and performance. Following Data pre- processing techniques are applied.

1. Replace NaN values with the mean values of the column.
2. Removing Duplicate entries Handling Outliers
3. Label encoding is a common encoding technique for categorical information. This approach assigns a unique number to every label based on its alphabetical order. The label encoding gives each occurrence in the chosen datasets a special identification number.
4. Dataset 1 suffers from imbalanced class distribution, we use SMOTE technique to balance the classes and improve model performance.
5. Data is split into the ratio of 70:30. 70% is used for model training while the remaining 30% is used for testing purposes. The test set is used to evaluate the model performance on unseen data.

6. Feature Scaling is applied to the selected datasets. The Feature scaling is carried out by using standard scalers. In feature scaling the trained data is

Feature Selection Phase

After the pre-processing, pre-processed datasets are passed towards the genetic algorithm, one of the evolutionary algorithms. The genetic algorithm with CNN finds the optimal solution based on the F1 score criteria.

Genetic Algorithm GA is a useful, reliable strategy and search approach. The nat- ural selection of the test individuals used by nature to generate species served as the basis for the GA search algorithm. In that it determines the most suitable and effective solution to a given computer issue, a genetic algorithm is comparable to natural evolution. The biological ideas of evolution and the survival of the fittest serve as the foundation of the genetic algorithm. Because it offers a solid and re- liable answer that is graded against fitness criteria, this algorithm is significantly more effective and powerful than an exhaustive search algorithm. The fitness function measures how closely a particular solution comes to being optimum. Feature selection aims to reduce the dimensionality of the data by identifying the subset of features that are most predictive or influential for the problem at hand. This component helps improve the efficiency and accuracy of the solution by focusing on the most significant features and reducing noise or irrelevant information.

This method offers a decent and resilient solution that is graded against fitness criteria, making it far more powerful and efficient than an exhaustive search algorithm. A solution's closeness to optimality is measured using the fitness function. The possible solutions to the problem are represented by a set of chromosomes. A chromosome is a string of binary digits. Each digit is called a gene. The initial population can be created randomly. The Genetic Algorithm (GA) is a search and optimization technique inspired by the principles of natural evolution. It iteratively evolves a population of candidate solutions to a problem by mimicking the processes of selection, reproduction, and mutation.

1. Initialization: Start by creating an initial population of potential solutions to the problem. Each solution is represented as a set of parameters called chromosomes or genes. The population is typically generated randomly or using some heuristic. We used a population size and number of generations of 50,50 each.
2. Evaluation: Evaluate the fitness of each individual in the population. The fitness function measures how well each solution performs with respect to the problem at hand. The fitness function can be problem-specific and aims to quantify the quality or suitability of a solution
3. Selection: Select two individuals from the population at random and compare fitness values. The selection process is usually based on the fitness values of the individuals. Solutions with higher fitness have a higher probability of being selected, mimicking the concept of survival of the fittest.
4. Reproduction: Create offspring by combining the genetic material of the selected parents. This is typically done through genetic operators such as crossover and mutation. The probability of crossover is 0.8. Crossover involves exchanging genetic information between two parents to produce new individuals, while mutation introduces small random changes to the genetic material to promote exploration of the search space. The probability of mutation is

0.2. Choose crossover point and choose gene to mutate.

1. Replacement: Replace some individuals in the current population with newly created offspring. The replacement strategy can vary, but commonly the off- spring replace the least fit individuals, ensuring that the population maintains diversity and potential improvements.

2. Termination: Repeat steps 2–5 for a certain number of generations or until a termination criterion is met. Termination criteria can be based on the number of generations, reaching a satisfactory fitness level, or a predefined time limit. Or termination is also met when we analyze that the next step will overfit the data then we stop the process. Select the best individuals.

Now, selected features are used for classification and prediction.

Classification Phase Using Hybrid Model SVM-CNN

Set up the SVM model, including the choice of kernel function (e.g., linear, polynomial, or radial basis function) and hyperparameter selection. Train the SVM model using the training dataset. Define the architecture of the CNN model, including the arrangement and number of convolutional, pooling, and fully connected layers. Train the CNN model using the training dataset, optimizing the model's weights through forward and backward propagation. Determine how the output features or representations from the SVM and CNN models are combined or integrated to create a hybrid model. A hybrid approach that combines Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) for classification tasks. SVM is a machine learning algorithm used for supervised classification, while CNN is a deep learning technique commonly used for classification. By combining the strengths of both SVM and CNN, the hybrid approach aims to leverage the discriminative power of SVM and the ability of CNN to learn hierarchical features automatically from the input data. These components give an effective means of classifying or categorizing the input data based on the selected features, utilizing the power of both traditional machine learning and deep learning techniques.

Support Vector Machine

Support Vector Machine (SVM), is a classification and regression prediction model that makes use of machine learning theory for the best accuracy and to prevent model overfitting. Locating the hyperplane in N- dimensional space is the support vector machine's goal. For the separation of two classes, there are numerous potential hyperplanes from which to choose. Drawing the hyperplane with the greatest possible margin distance is the goal in order to

classify upcoming data points. Decision boundaries called hyperplanes aid in categorizing the data points. Support vectors are data points that relate to the sides of the hyperplane and which indicate the characteristics of the classes. The hyperplane will be a line that the attributes that belong to separate classes if the input features are two. A hyperplane is a two-dimensional plane if the input features are three. The hyperplanes become challenging when the input feature reaches three. Support vectors that are near the hyperplane have an impact on the hyperplane's position and orientation. The sigmoid function compresses the input in logistic regression into the range [0, 1]. Assign the input value a label of 1 if it exceeds the threshold value (0.5). The output of the linear function is taken into consideration while determining the class label in SVM.

A class label is assigned if the output is 1, and a different class label is assigned if the output is -1. There is a margin between threshold values, which range from 1 to -1. The margin value between the data points and the hyperplane should be maximized so that future data points can be classified more confidently. The kernel's goal is to make it possible for operations to be carried out in the input space. The data is transformed using the kernel functions, and based on these transformations, an ideal boundary between the potential outputs is discovered. We categorize the binary class data for the linearly separable data using the linear kernel. The non- linear SVM is utilized in the scenario of non-linear separable data though. The border determined by utilizing the non-linear kernels is not

a straight line in a non-linear SVM. Depending on the nature of the dataset, several kernels such as Gaussian, polynomial, sigmoid, or others, are employed in non-linear SVM.

Convolution Neural Network

A synthetic neural network with numerous concealed layers. CNN can extract greater order of connection by using more hidden layers. The incoming and outgoing layers of a CNN are split at different levels. Whatever their size or structure, neural networks must have neurons, connections, values, bias, and functionalities. Here's a detailed explanation of how CNNs work in our model and their components

Input Layer

The input layer of a CNN receives the output of SVM model.

Convolutional Layer

The convolutional layer is the core component of a CNN. It consists of a set of learnable filters (also called kernels) that convolve over the input data. Each filter detects specific features, by performing element-wise multiplication and aggregation. The outcome is a feature map that draws attention to the fact that certain features were present in the input.

Activation Function

After each convolutional operation, an activation function is applied element-wise to introduce non-linearity. Common activation functions used in CNNs include Rectified Linear Unit (ReLU), sigmoid, or hyperbolic tangent. ReLU is the most popular choice due to its simplicity and effectiveness in preventing the vanishing gradient problem. We use ReLU in our model.

Pooling Layer

The pooling layer reduces the spatial dimensions of the feature maps while retaining the most salient information. It achieves this by applying operations like max pooling or average pooling within a localized region. Pooling helps in reducing computational complexity and making the learned features more invariant to small spatial variations.

Fully Connected Layer

The fully connected layer, also known as the dense layer, takes the output from the previous layers and connects every neuron to every neuron in the subsequent layer. It learns high-level representations by combining the learned features from previous layers. These layers are similar to those in a traditional neural network and perform classification or regression tasks.

Dropout

Dropout is a regularization technique commonly used in CNNs to prevent overfitting. During training, a certain percentage of neurons in the fully connected layer are randomly ignored or "dropped out." This helps in reducing interdependent learning between neurons and forces the network to learn more robust features.

Output Layer

The output layer of the CNN provides the final predictions or out- puts. The activation function in the output layer depends on the nature of the problem being solved. For binary classification, a sigmoid function is typically used, while for multi-class classification, a softmax function is commonly used.

The proposed solution for heart disease classification involves utilizing a Genetic Algorithm (GA) for feature selection and a combination of Support Vector Machine (SVM) and Convolutional Neural Network (CNN) models (referred to as SVM- CNN) for prediction. This hybrid approach aims to enhance the accuracy and ro- bustness of heart disease classification by selecting relevant features and leveraging the strengths of both SVM and CNN algorithms.

Datasets

We use three different types of datasets and perform three experiments on each dataset. we use dataset 1 named as UCI Dataset with 920 instances and 76 features shown in Fig. 2 With 303 instances and 56 attributes shown in Fig. 3, dataset 2 also known as the Z-Alizadeh Sani Dataset is used in experiment 2. We use the Cardiovascular Disease Dataset shown in Fig. 4, a dataset containing 70,000 instances and 11 characteristics, in experiment 3. First of all pre-processing techniques are applied to clean the data that are already discussed in the methodology. Start by preparing the datasets one by one, ensuring it is labeled with the corresponding class labels for each instance. Here we did labeling on all heart disease Datasets that we are using UCI has 4 labels, Z-Alizadeh Sani has 2 and Cardiovascular has also 2 labels. We apply standardization from − 1 to 1 to transform data. Transform categorical variables into numerical data using a label encoder. Then we split the data into training and testing in the ratio of 70:30. UCI data was so imbalanced, for balancing the data we use SMOTE technique to balance the data. After balancing the data feature scaling is used to transform the data, we use a standard scaler to transform data from − 1 to 1. Preprocessing techniques are quite the same for the three datasets except for class imbalance, that are used only for the UCI dataset. The experiment consists of three phases. Input Phase, Feature Selection Phase, and Classification phase. Input phases consist of the data sets. The selected datasets are pre-processed to ensure the accuracy and efficiency of the pro- posed model. The input phase consists of three datasets having different numbers of instances and attributes. To check the performance of the proposed model it is trained and tested on different sizes of datasets. These datasets are sufficient and helpful in heart disease prediction for several reasons. They provide a diverse set of features, enabling the development of comprehensive prediction models. They offer large and diverse samples of patient data, ensuring that the models are trained on representative populations. They allow researchers to identify and investigate various risk factors associated with heart disease. The datasets can serve as baselines for evaluating new models and comparing their performance against established methods. These datasets also have some limitations. Datasets may have missing values or incomplete data for certain features, which can affect the model's performance and accuracy. We tackle this limitation by employing imputation techniques to fill in missing values based on statistical methods or machine learning algorithms. The data might contain errors, outliers, or noisy measurements due to various factors like data collection errors or inconsistent recording practices. Data cleaning and preprocessing steps are applied to remove outliers and correct errors in the data. We also use feature selection techniques to reduce the impact of noisy data on the model's performance.

Results and Discussions

The experiments are conducted in the setting and environment that were previously stated. The data sets are in numeric format. Data cleaning techniques are applied to the data before the experimentation process. Here the

performance of the model is evaluated by using multiple datasets. Comparison is performed through accuracy, precision, recall, and F1-score.

Here's a comparison of the three datasets based on their accuracy, precision, recall, and F1-score. As shown in the Fig. 5, The UCI dataset has the highest accuracy of 0.98, followed closely by the Z-Alizadeh Sani dataset with 0.97 accuracy. The Cardiovascular dataset has the lowest accuracy at 0.86. The UCI dataset achieves the highest precision of 0.95, indicating a relatively low false positive rate. The

Z-Alizadeh Sani dataset has a slightly lower precision of 0.93. The Cardiovascular dataset has the highest precision at 0.89. The Z-Alizadeh Sani dataset has the highest recall rate of 0.96, indicating a low false negative rate. The Cardiovascular dataset follows closely with a recall rate of 0.83, while the UCI dataset has the lowest recall at 0.81. The Z-Alizadeh Sani dataset achieves the highest F1-score of 0.95, indicating a good balance between precision and recall. The Cardiovascular dataset has the next highest F1-score at 0.86, and the UCI dataset has the lowest F1-score of 0.78. Overall, the UCI dataset performs well in terms of accuracy and precision but has a relatively lower recall and F1-score compared to the other datasets. The Z-Alizadeh Sani dataset shows high performance across all metrics, indicating a well- balanced classification performance. The Cardiovascular dataset, although having lower accuracy compared to the others, still achieves a good F1-score and has the highest precision among the three datasets. RQ3: We can evaluate and validate the criteria of the SVM-CNN model using accuracy, F1 score, precision, recall and statistical analysis (i.e. pairwise comparison and mean comparison using one way ANOVA). In each experiment, an optimized feature set is obtained using GA. The following feature sets are obtained respectively shown in Table 2.

Table 2
List of Selected Features

Dataset	Total No. of Features	No. of selected features	Selected Features
Heart-Disease-uci	76	06	'age', 'sex', 'fbs','restecg','thalch', 'slope'
Z-Alizadeh Sani	56	19	'Age', 'Length', 'Sex', 'BMI', 'HTN','EX-Smoker', 'BP','PR','TypicalChestPain','Atypical','Nonanginal','Q Wave', 'St Elevation', 'Poor R Progression', 'HDL', 'BUN','Neut','EF- TTE', 'Region RWMA'
Cardiovascular	11	09	'age' ,'gender', 'height' ,'weight', 'aphi','aplo'cholesterol', 'gluc', 'alco'

Results Comparison

The comparison is performed with state-of-the-art using UCI Heart Disease Dataset is plotted. Plotted accuracy of the dataset is the combination of training and testing. The proposed model and existing model are executed on the same dataset. The existing model is a combination of CNN and LSTM. At the start accuracy of the proposed model and the existing one is the same which is 0.2. After continuous training and testing proposed model outperforms

the existing model in accuracy. The model reaches an accuracy of 0.98 while the existing have only 0.89 shown in Fig. 6.
The accuracy comparison of Z-Alizadeh Sani Dataset is plotted. Plotted accuracy of the dataset is the reflection of training and testing. The standing model uses the technique of ensemble feature selection and classification on the same dataset, the suggested model and the current model are both run in the same environment. The proposed model's initial accuracy is 0.2, compared to the present model's initial accuracy of 0.15. The proposed solution outperforms the existing model in accuracy after repeated training and testing. The proposed model's accuracy is 0.97 as opposed to the old model's accuracy of 0.91 shown in Fig. 7.

The accuracy comparison of the cardiovascular Heart Disease Dataset is presented. The dataset's plotted accuracy is the result of both training and testing. On the same dataset, the suggested model and the current model both are executed on the same dataset. The current one uses the ensemble technique. The proposed model outperforms the existing model at the beginning of the process. The suggested model outperforms the existing model in accuracy after repeated training and testing. The proposed model's accuracy is 0.86 as opposed to the old model's accuracy of 0.75 shown in Fig. 8.

Table 3
Comparison of Proposed and Existing Technique

Proposed Hybrid Model				**Existing Model**		
Datasets	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Heart-Disease-uci	0.95	0.81	0.78	0.93	0.77	0.52
Z-Alizadeh Sani	0.93	0.96	0.95	0.91	0.92	0.91
Cardiovascular	0.89	0.83	0.86	0.86	0.80	0.81

Empirical Analysis

All research work empirical analysis performed which involves following steps: Ex- periments are performed on above describes three datasets

1. Data collection: All three datasets are collected from different resources.
2. Pre-processing: Previously described pre-processing steps are performed on these three datasets.
3. Feature Selection: Performed feature selection through Genetic Algorithm. Optimized feature set are obtained from these three previously described datasets.
4. Model Implementation: Proposed hybrid model is implemented, SVM is used as an initial classifier with hyper parameters tuning. And CNN is used as a resultant classifier with fully connected hidden layers.
5. Training: The preprocessed data is splitted into training and validation set with the ratio of 70% and 30% respectively. First of all the training is per- formed using SVM with previously described hyperparameters. Output of the SVM is passed to the CNN model. In the hidden layers of CNN 'relu' used as activation function while the output layer used 'sigmoid'.
   Furthermore, the SVM and CNN are also trained individually on all three datasets to produce the results for comparative analysis. As well as different ensemble methods are also used to compare the performance of the model.
6. Model Evaluation: The proposed model is evaluated using accuracy, precision, recall and F1-score measures.

7. Results and comparison: The comparison is performed with baseline models, state-of-the-art models, and other existing techniques. These comparisons are discussed in next slide.
8. Statistical Analysis: Performed statistical tests to evaluate the significance of the results.

Table 4 shows the empirical analysis of different existing classifiers and proposed classifiers and their performance on the basis of accuracy.

Table 4
Model Comparison

Classifiers	Dataset1 UCI	Dataset2 ZAl-izadeh	Dataset3 Cardio
Random Forest Classifier	0.77	0.78	0.68
LGBM Classifier	0.75	0.77	0.67
XGB Classifier	0.76	0.76	0.65
AdaBoost Classifier	0.74	0.78	0.64
Decision Tree Classifier	0.75	0.71	0.62
ExtraTrees Classifier	0.70	0.72	0.65
SVM	0.89	0.90	0.76
CNN	0.83	0.86	0.78
SVM + CNN	0.92	0.94	0.82

Statistical Analysis

Statistical Analysis shows the mean accuracies of different techniques used for heart disease prediction on "UCI, Zalizadehsani, and Cardio" and conducts a one-way ANOVA test to check if there are any statistically significant differences among the techniques. Additionally, it performs Tukey's Honestly Significant Difference (HSD) test as a post-hoc analysis to identify which pairs of techniques have significantly different mean accuracies. Following are the tests we performed in our model.

1. One-Way ANOVA: The one-way ANOVA test assesses whether there are any statistically significant differences in the mean accuracy scores across the ten techniques. The ANOVA provides an F-statistic and a p-value. We apply one-way ANOVA on three different datasets and used 10 different techniques shown in Figs. 9, 10 and 11 respectively. These figures show the mean accuracies on three different datasets among 10 different classifiers.
2. Tukey's HSD Test: If the one-way ANOVA indicates that there are statistically significant differences among the techniques, a post-hoc analysis is performed using Tukey's HSD test. This test conducts pairwise comparisons between all possible combinations of the ten techniques to identify which techniques have significantly different mean accuracy scores. It shows the pairwise comparison of three different datasets among 10 different classifiers shown in Fig. 12, 13, 14 respectively.
3. Tukey's HSD Test Plot: The plot shows horizontal lines connecting pairs of techniques that have significantly different mean accuracy scores. The y-axis represents the different techniques, while the x-axis displays the mean accu- racy scores.

RQ2: Based on the shown Table 4, it appears that the SVM-CNN (SVM + CNN) classifier has the highest accuracy on all three datasets: Dataset1 UCI, Dataset2 ZAlizadeh, and Dataset3 Cardio. It consistently outperforms the other classifiers, including SVM and CNN as standalone models, on all datasets. Therefore, the re- sults suggest that the combination of SVM and CNN in the SVM-CNN model is effective in improving heart disease prediction compared to traditional ML algo- rithms and standalone CNN.
RQ4: We can compare the performance of our model with the existing model using different performance metrics. We used accuracy, precision, recall, and F1-score to compare our model with the existing model. Table 3 shows the comparison of our model with the existing model using precision, recall, and F1-score. While Figs. 6, 7, 8 show the comparison on the basis of accuracy.

Limitations

The combination of a genetic algorithm for feature selection and a hybrid SVM- CNN model has proven to be a promising approach for heart disease prediction. Despite its promising results, the proposed approach has some limitations. Firstly, the genetic algorithm's effectiveness heavily relies on the quality and diversity of the initial feature pool, which could impact the performance of the model. Additionally, the feature selection step increases time complexity, leading to higher computational costs. Frequent data training and testing are required for the genetic algorithm to identify the optimal feature subset, which can be resource-intensive. Moreover, the performance of the SVM-CNN hybrid model may be influenced by dataset size, class imbalance, and the complexity of heart disease patterns present in the data. These factors could potentially affect the generalizability of the approach across different datasets and populations.

Conclusion and Future Work

The combination of a genetic algorithm for feature selection and a hybrid SVM- CNN model for heart disease prediction has shown promising results. This approach capitalizes on the strengths of both techniques, leveraging the genetic algorithm's ability to identify relevant features and the SVM-CNN hybrid model's capacity for accurate prediction. The genetic algorithm plays a crucial role in selecting the most informative features from a large pool of potential predictors. By iteratively evaluating the fitness of different feature subsets, it optimizes the selection process and reduces dimensionality. This feature selection step is critical for enhancing the model's performance, as it helps eliminate irrelevant or redundant features that may introduce noise or overfitting. The SVM-CNN hybrid model takes advantage of the power of both support vector machines (SVM) and convolutional neural networks (CNN).

SVMs excel at binary classification tasks by creating an optimal hyperplane to separate data points, while CNNs are well-suited for classification. By combining the two, we obtain a model that can effectively handle both structured and un- structured features for heart disease prediction. The hybrid model's performance is enhanced by the selected features, which are fed into the SVM for classification and then passed to the CNN for further analysis. This integration allows the model to

capture complex patterns and relationships present in the data, leading to improved accuracy and robustness. The results obtained from applying this approach to heart disease prediction are highly encouraging.

By leveraging the genetic algorithm's feature selection, we achieve a more efficient and accurate model, reducing computation time and improving predictive perfor- mance. The hybrid SVM-CNN model demonstrates its ability to handle diverse data types, capturing numerical features to make accurate predictions. Further research and validation are necessary to assess the generalizability and robustness of this approach across different datasets

and populations. Nonetheless, the combination of a genetic algorithm for feature selection and a hybrid SVM-CNN model holds great promise for enhancing heart disease prediction accuracy and contributing to the development of personalized healthcare solutions in the future.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Author contributions

MM wrote the main manuscript, while MM, MR and FJ collaborated in research. All authors reviewed the manuscript.

Funding

Not applicable

Acknowledgments

We heartily acknowledge Dr. Muhammad Ali Tahir, SEECS, NUST for his guidance and support.

Authordetails

[1,2,3] School of Electrical Engineering and Computer Science, National University of Science and Technology, 44000, Islamabad, Pakistan.

References

1. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed: 2023-005-22.
2. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BioMed Cent (2020).
3. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C;Catapano AL;Cooney MT;Corrà U;Cosyns B;Deaton C;Graham I;Hall MS;Hobbs FDR;Lщchen ML;Löllgen H;, Marques-Vidal P. ;Perk J;Prescott E;Redon J;Richter DJ;Sattar N;Smulders Y;Tiberi M;van der Worp HB;van Dis I;Versc, p.: 2016 European Guidelines on

Cardiovascular Disease Prevention in Clinical Practice: The Sixth joint task force of the European Society of Cardiology and other societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)developed with the special contribution of the European Association for Cardiovascular Prevention amp; Rehabilitation (EACPR). U.S. National Library of Mdicine. https://pubmed.ncbi.nlm.nih.gov/27222591/.

4. Chowdhury R, Khan H;Heydon E. ;Shroufi A;Fahimi S;Moore C;Stricker B;Mendis S;Hofman A;Mant J;Franco OH;, C.R.: Adherence to cardiovascular therapy: A meta-analysis of prevalence and clinical consequences. U.S. National Library of Medicine. https://pubmed.ncbi.nlm.nih.gov/23907142/.
5. P;M, M.D.E.A.D.M.M.S.F.S.J.H.V.M.C.M.S.B.J.L.S.R.D.D.E C. Heart disease and stroke statistics-2016 update: A report from the American Heart Association. U.S. National Library of Medicine. https://pubmed.ncbi.nlm.nih.gov/26673558/.
6. ;, R, C.V.B.M.R JP. T.: 21st-century hazards of smoking and benefits of cessation in the United States. U.S. National Library of Medicine. https://pubmed.ncbi.nlm.nih.gov/23343063/.
7. R.W.D.K.H.C.S.S.K.D.E.P.B.S.C.R.S.R.K JTWP.J.:, ACC/AHA/AAPA/ABC. /ACPM/AGS/apha/ash/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: Executive summary: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. U.S. National Library of Medicine. https://pubmed.ncbi.nlm.nih.gov/29133354/.
8. Yazdani A, Varathan KD, Chiam YK, Malik AW, Wan Ahmad WA. A novel approach for heart disease prediction using strength scores with significant predictors. BMC Med Inf Decis Mak. 2021;21(1):194.
9. Pemmaraju AG, Asish A, Das S. Heart disease prediction using feature selection and machine learning techniques. In: 2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS), pp. 28–33 (2022). IEEE.
10. Stevens D, Lane DA, Harrison SL, Lip GY, Kolamunnage-Dona R. Modelling of longitudinal data to predict cardiovascular disease risk: a methodological review. BMC Med Res Methodol. 2021;21(1):1–24.
11. Rehman S, Rehman E, Ikram M, Jianglin Z. Cardiovascular disease (cvd): assessment, prediction and policy implications. BMC Public Health. 2021;21(1):1–14.
12. Nandy S, Adhikari M, Balasubramanian V, Menon VG, Li X, Zakarya M. An intelligent heart disease prediction system based on swarm-artificial neural network. Neural Comput Appl, 1–15 (2021).
13. Zhenya Q, Zhang Z. A hybrid cost-sensitive ensemble for heart disease prediction. BMC Med Inf Decis Mak. 2021;21:1–18.
14. Almustafa KM. Prediction of heart disease and classifiers' sensitivity analysis. BMC Bioinformatics. 2020;21(1):1–18.
15. Kolukisa B, Bakir-Gungor B. Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis. Comput Stand Interfaces. 2023;84:103706.
16. Javeed A, Rizvi SS, Zhou S, Riaz R, Khan SU, Kwon SJ. Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. Mobile Information Systems 2020, 1–11 (2020).
17. Pathan MS, Nag A, Pathan MM, Dev S. Analyzing the impact of feature selection on the accuracy of heart disease prediction. Healthc Analytics. 2022;2:100060.
18. Reddy NSC, Nee SS, Min LZ, Ying CX. Classification and feature selection approaches by machine learning techniques: Heart disease prediction. Int J Innovative Comput 9(1) (2019).

19. Dissanayake K, Md Johar MG. Comparative study on heart disease prediction using feature selection techniques on classification algorithms. Applied Computational Intelligence and Soft Computing 2021, 1–17 (2021).
20. Kanwal S, Rashid J, Nisar MW, Kim J, Hussain A. An effective classification algorithm for heart disease prediction with genetic algorithm for feature selection. In: 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), pp. 1–6 (2021). IEEE.
21. Lakshmanarao A, Srisaila A, Kiran TSR. Heart disease prediction using feature selection and ensemble learning techniques. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 994–998 (2021). IEEE.
22. Bashir S, Khan ZS, Khan FH, Anjum A, Bashir K. Improving heart disease prediction using feature selection approaches. In: 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 619–623 (2019). IEEE.
23. Al-Tashi Q, Rais H, Jadid S. Feature selection method based on grey wolf optimization for coronary artery disease classification. In: Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018), pp. 257–266 (2019). Springer.
24. Ogundepo EA, Yahya WB. Performance analysis of supervised classification models on heart disease prediction. Innov Syst Softw Eng, 1–16 (2023).
25. Abdollahi J, Nouri-Moghaddam B. A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation. Iran J Comput Sci. 2022;5(3):229–46.
26. Jabbar M, Deekshatulu B, Chandra P. Intelligent heart disease prediction system using random forest and evolutionary approach. Journal of network and innovative computing 4(2016), 175–184 (2016).
27. Doppala BP, Bhattacharyya D, Chakkravarthy M, Kim T-h. A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset. Distrib Parallel Databases, 1–20 (2021).
28. Gárate-Escamila AK, Hajjam El Hassani A, Andrès E. Classification models for heart disease prediction using feature selection and pca. Inf Med Unlocked. 2020;19:100330. 10.1016/j.imu.2020.100330.
29. Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience 2021 (2021).
30. Zhang D, Chen Y, Chen Y, Ye S, Cai W, Jiang J, Xu Y, Zheng G, Chen M. Heart disease prediction based on the embedded feature selection method and deep neural network. Journal of healthcare engineering 2021, 1–9 (2021).
31. Kavitha M, Gnaneswar G, Dinesh R, Sai YR, Suraj RS. Heart disease prediction using hybrid machine learning model. In: 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 1329–1333 (2021). IEEE.
32. Geluvaraj B, Santhosh K, Prabhu N, Reddy A, Sandhya T, Bhaskar S. A hybrid approach for predicting diseases using clustering and classification techniques. In: 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), pp. 1–6 (2022). IEEE.
33. Bindu M, Sabu M. A hybrid feature selection approach using artificial bee colony and genetic algorithm. In: 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), pp. 211–216 (2020). IEEE.

34. Dahiwade D, Patle G, Meshram E. Designing disease prediction model using machine learning approach. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211–1215 (2019). IEEE.
35. Panda D, Ray R, Abdullah AA, Dash SR. Predictive systems: Role of feature selection in prediction of heart disease. In: Journal of Physics: Conference Series, vol. 1372, p. 012074 (2019). IOP Publishing.
36. Sudha V, Kumar D. Hybrid cnn and lstm network for heart disease prediction. SN Comput Sci. 2023;4(2):172.
37. Chaurasia V, Chaurasia A. Novel method of characterization of heart disease prediction using sequential feature selection-based ensemble technique. Biomedical Mater Devices, 1–10 (2023).
38. El-Shafiey MG, Hagag A, El-Dahshan E-SA, Ismail MA. A hybrid ga and pso optimized approach for heart-disease prediction based on random forest. Multimedia Tools and Applications. 2022;81(13):18155–79.
39. Wankhede J, Sambandam P, Kumar M. Effective prediction of heart disease using hybrid ensemble deep learning and tunicate swarm algorithm. J Biomol Struct Dynamics. 2022;40(23):13334–45.
40. Bhavekar GS, Goswami AD. A hybrid model for heart disease prediction using recurrent neural network and long short term memory. Int J Inform Technol. 2022;14(4):1781–9.
41. Nagarajan SM, Muthukumaran V, Murugesan R, Joseph RB, Meram M, Prathik A. Innovative feature selection and classification model for heart disease prediction. J Reliable Intell Environ. 2022;8(4):333–43.
42. Velswamy K, Velswamy R, Swamidason ITJ, Chinnaiyan S. Classification model for heart disease prediction with feature selection through modified bee algorithm. Soft Comput, 1–9 (2021).

Figures

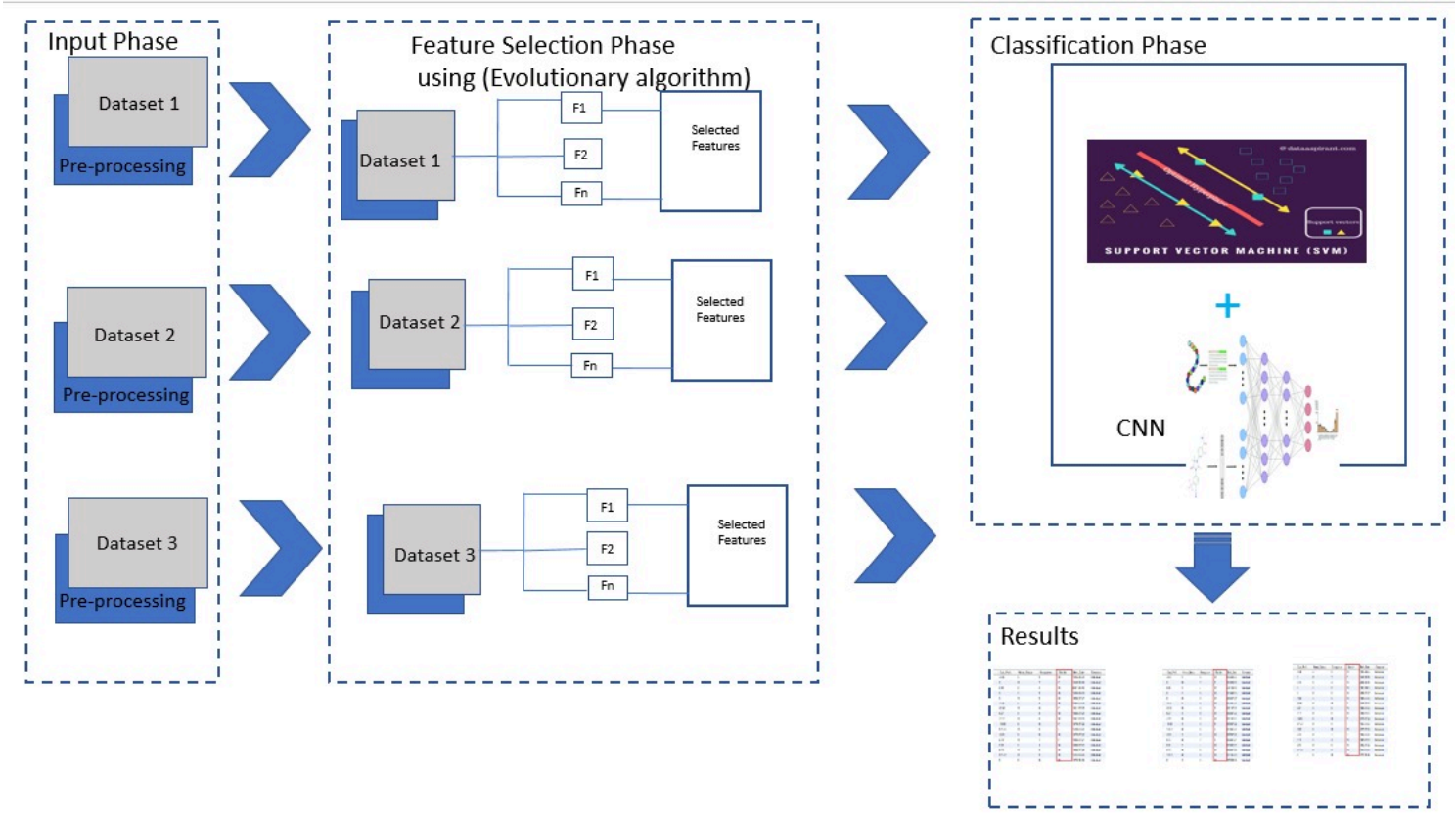


Figure 1

Proposed Model

	id	age	sex	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	1	3	145.0	233.0	2	0	150.0	0	2.3	0	0.0	0	0
1	2	67	1	0	160.0	286.0	0	0	108.0	2	1.5	1	3.0	1	2
2	3	67	1	0	120.0	229.0	0	0	129.0	2	2.6	1	2.0	2	1
3	4	37	1	2	130.0	250.0	0	1	187.0	0	3.5	0	0.0	1	0
4	5	41	0	1	130.0	204.0	0	0	172.0	0	1.4	2	0.0	1	0

Figure 2

Snippet of UCI dataset

```
<bound method NDFrame.sample of      Age  Weight  Length  Sex        BMI  DM  HTN  Current Smoker  EX-Smoker
0     53      90     175    1  29.387755   0    1               1          0
1     67      70     157    0  28.398718   0    1               0          0
2     54      54     164    1  20.077335   0    0               1          0
3     66      67     158    0  26.838648   0    1               0          0
4     50      87     153    0  37.165193   0    1               0          0
..   ...     ...     ...  ...        ...  ..  ...             ...        ...
298   58      84     168    1  29.761905   0    0               0          0
299   55      64     152    0  27.700831   0    0               0          0
300   48      77     160    0  30.078125   0    1               0          0
301   57      90     159    0  35.599858   1    0               0          0
302   56      85     170    0  29.411765   0    1               1          0

     FH  ...    K   Na    WBC  Lymph  Neut  PLT  EF-TTE  Region RWMA  VHD  \
0     0  ...  4.7  141   5700     39    52  261      50            0    1
1     0  ...  4.7  156   7700     38    55  165      40            4    1
2     0  ...  4.7  139   7400     38    60  230      40            2    3
3     0  ...  4.4  142  13000     18    72  742      55            0    2
4     0  ...  4.0  140   9200     55    39  274      50            0    2
..   ..  ...  ...  ...    ...    ...   ...  ...     ...          ...  ...
298   0  ...  4.8  146   8500     34    58  251      45            0    1
299   0  ...  4.0  139  11400     16    80  377      40            0    3
300   1  ...  4.0  140   9000     35    55  279      55            0    1
301   0  ...  3.8  141   3800     48    40  208      55            0    1
302   0  ...  4.4  147   6000     32    55  302      55            0    1
```

Figure 3

Snippet of Z Alizadeh Sani Dataset

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	\
0	0	18393	2	168	62.0	110	80	1	1	0	
1	1	20228	1	156	85.0	140	90	3	1	0	
2	2	18857	1	165	64.0	130	70	3	1	0	
3	3	17623	2	169	82.0	150	100	1	1	0	
4	4	17474	1	156	56.0	100	60	1	1	0	

	alco	active	cardio
0	0	1	0
1	0	1	1
2	0	0	1
3	0	1	1
4	0	0	0

Figure 4

Snippet of Cardio Dataset

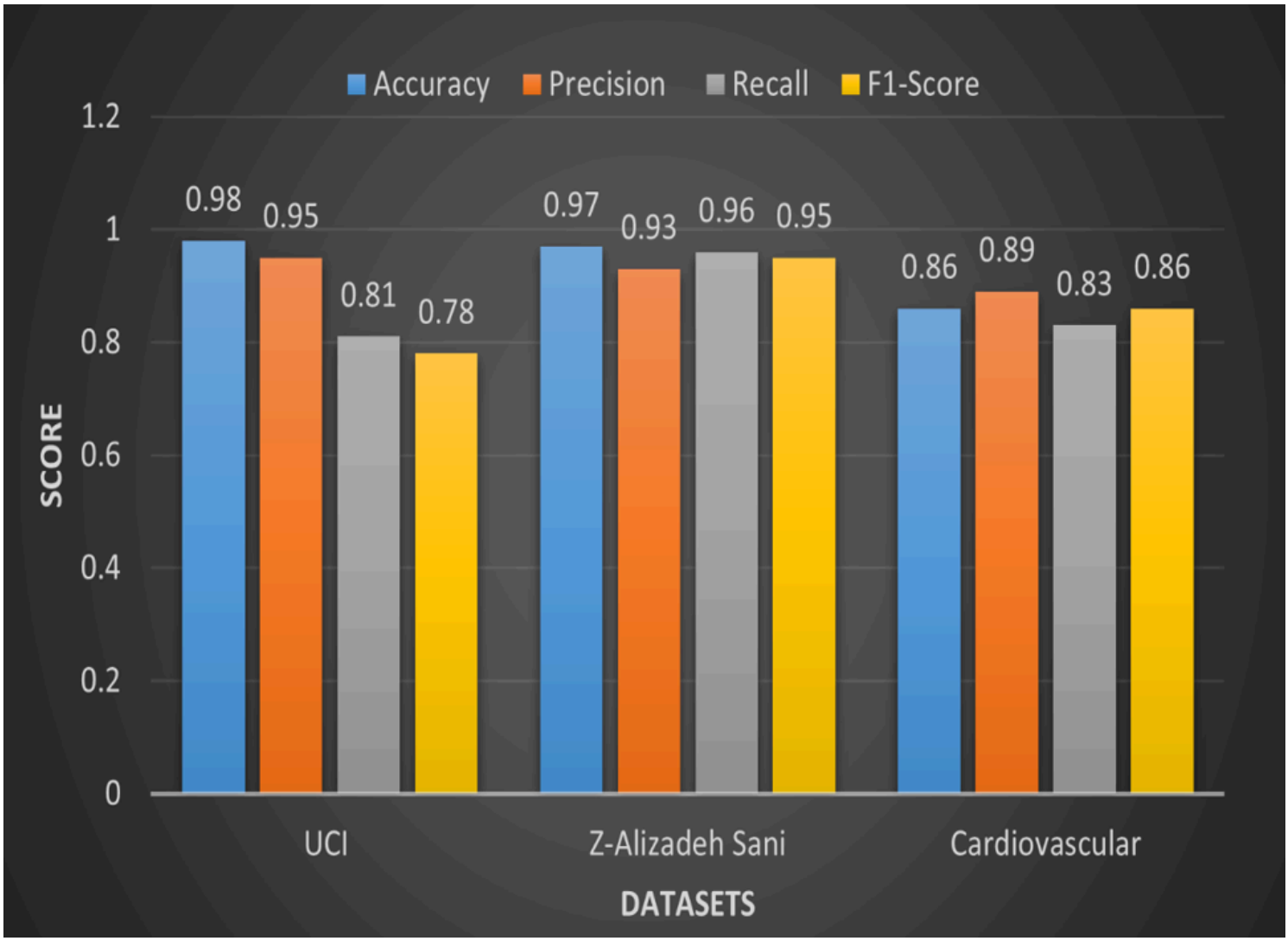


Figure 5

Performance of proposed model on three Datasets

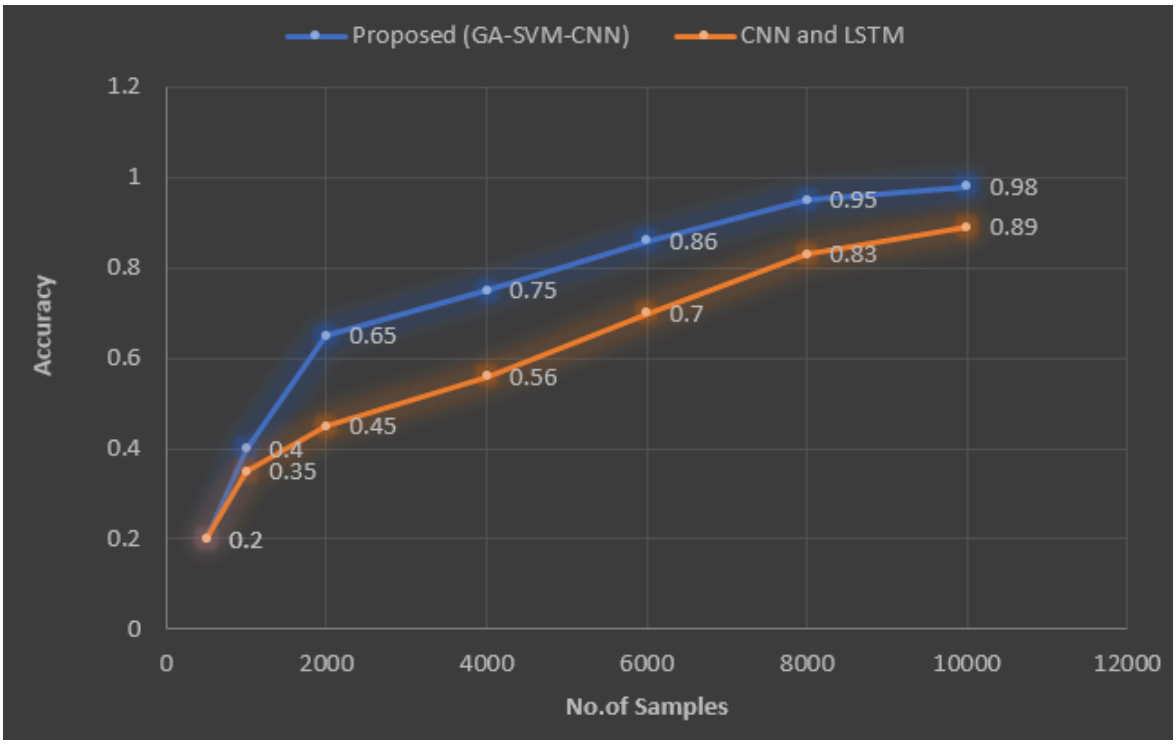


Figure 6

Accuracy comparison using UCI

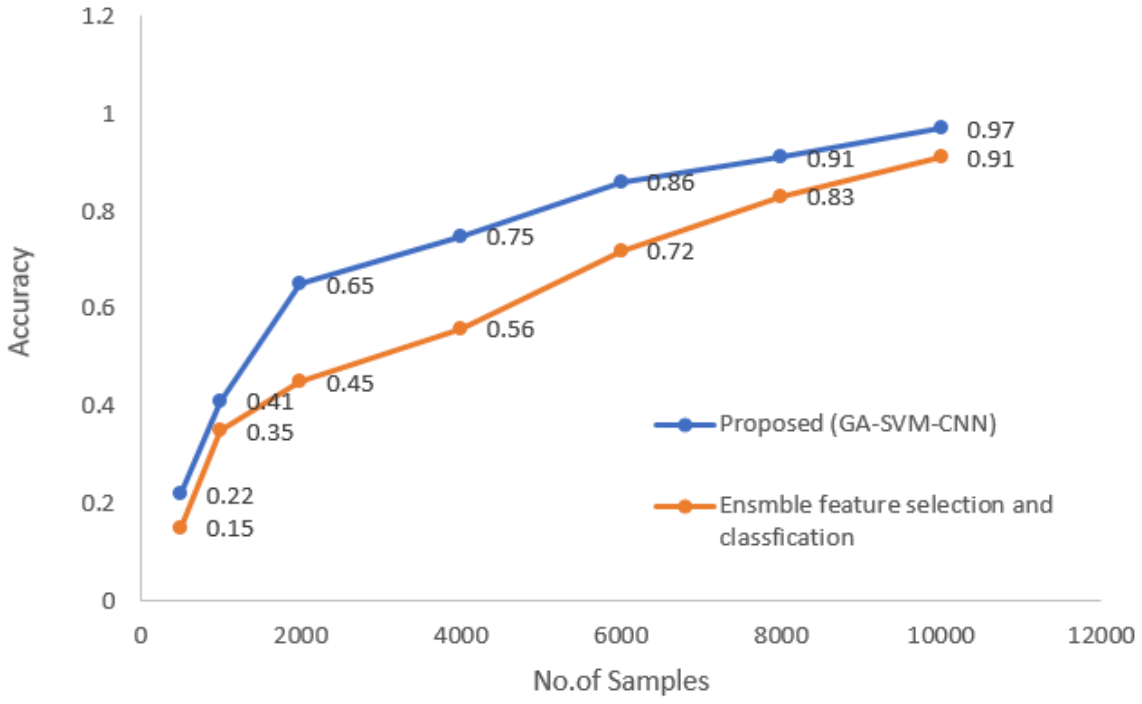


Figure 7

Accuracy comparison using Z-Alizadeh Sani

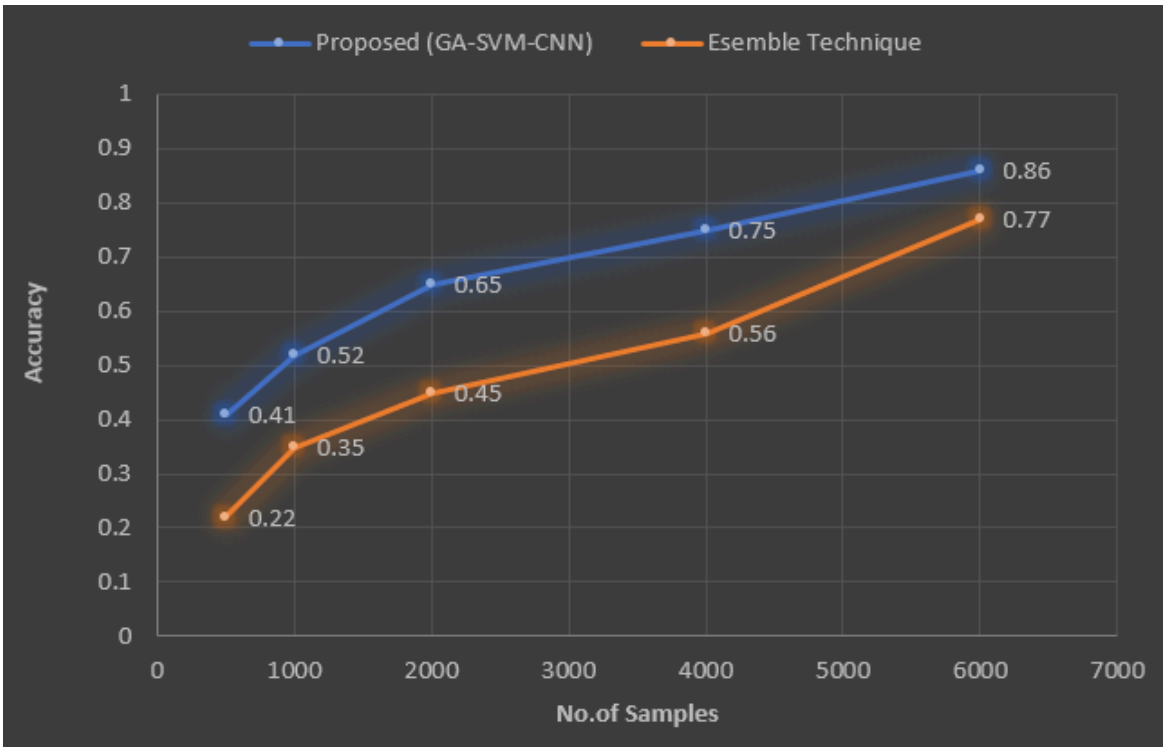


Figure 8

Accuracy comparison using Cardio

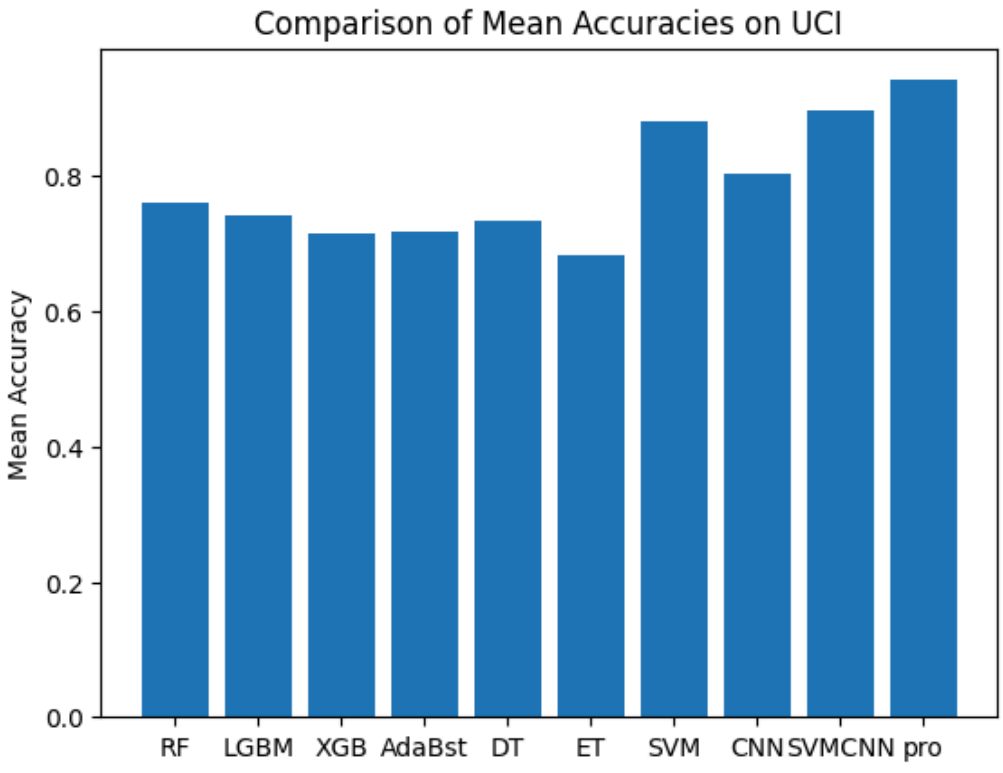


Figure 9

comparison of mean accuracies on UCI

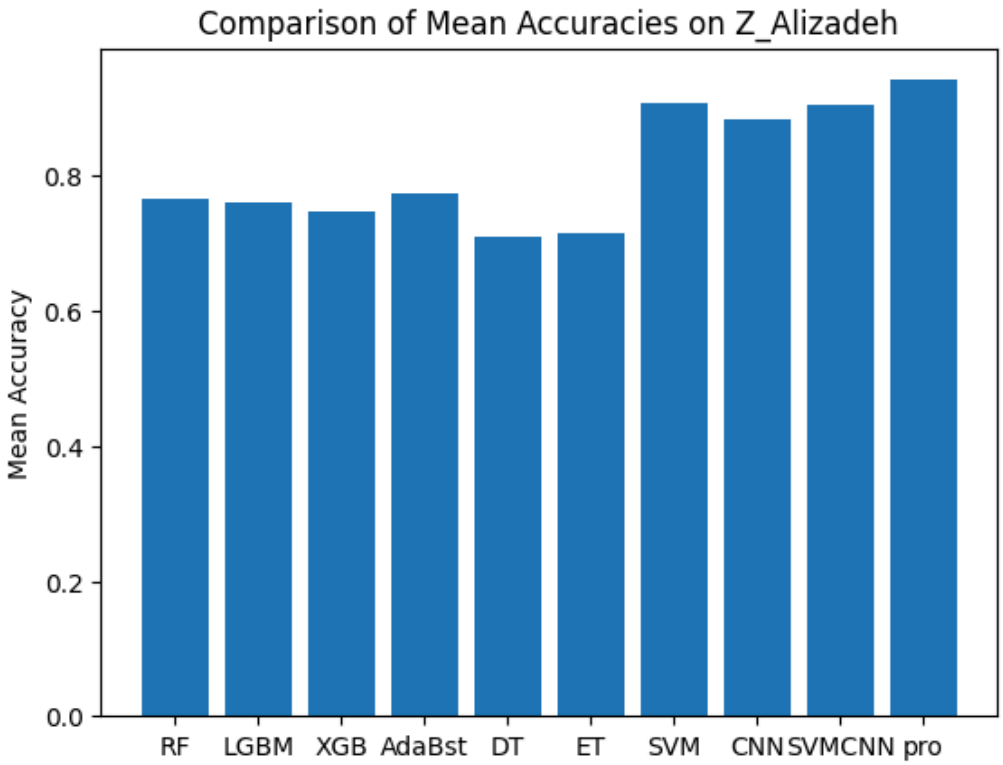


Figure 10

comparison of mean accuracies on ZAlizadehSani

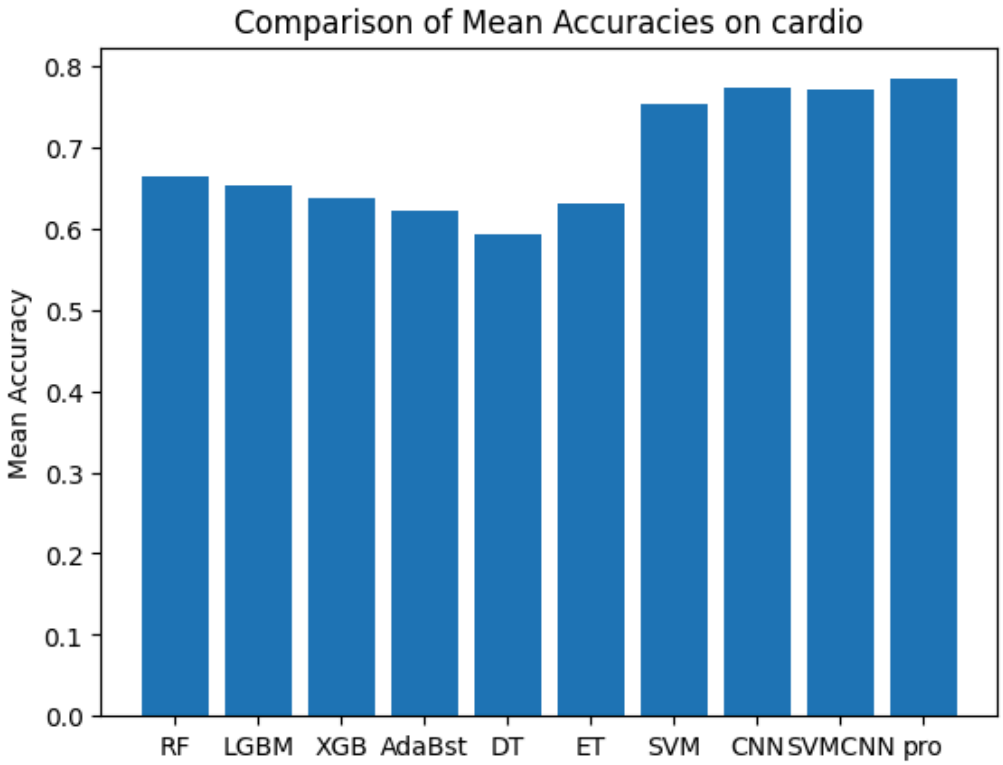


Figure 11

comparison of mean accuracies on Cardio

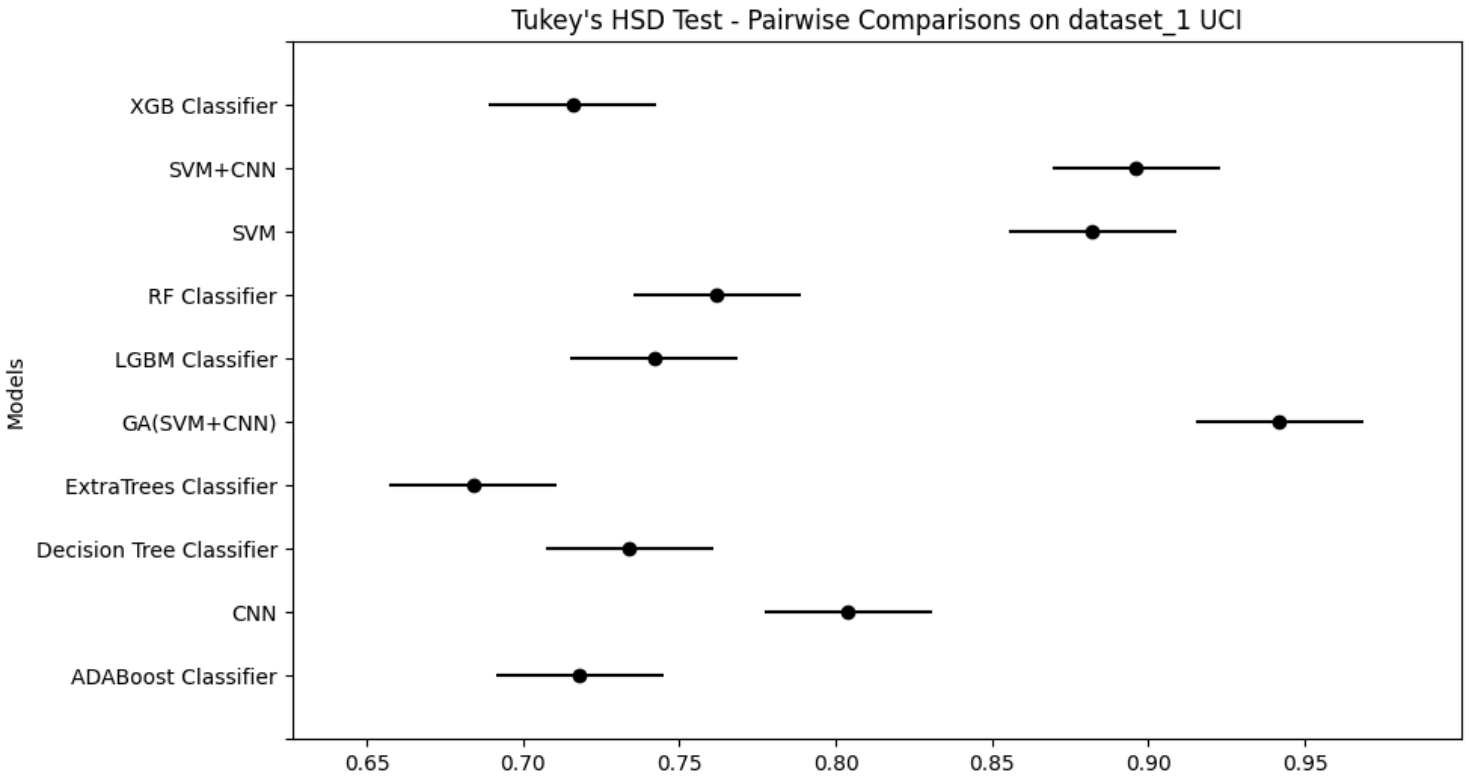


Figure 12

Pairwise Comparison on UCI

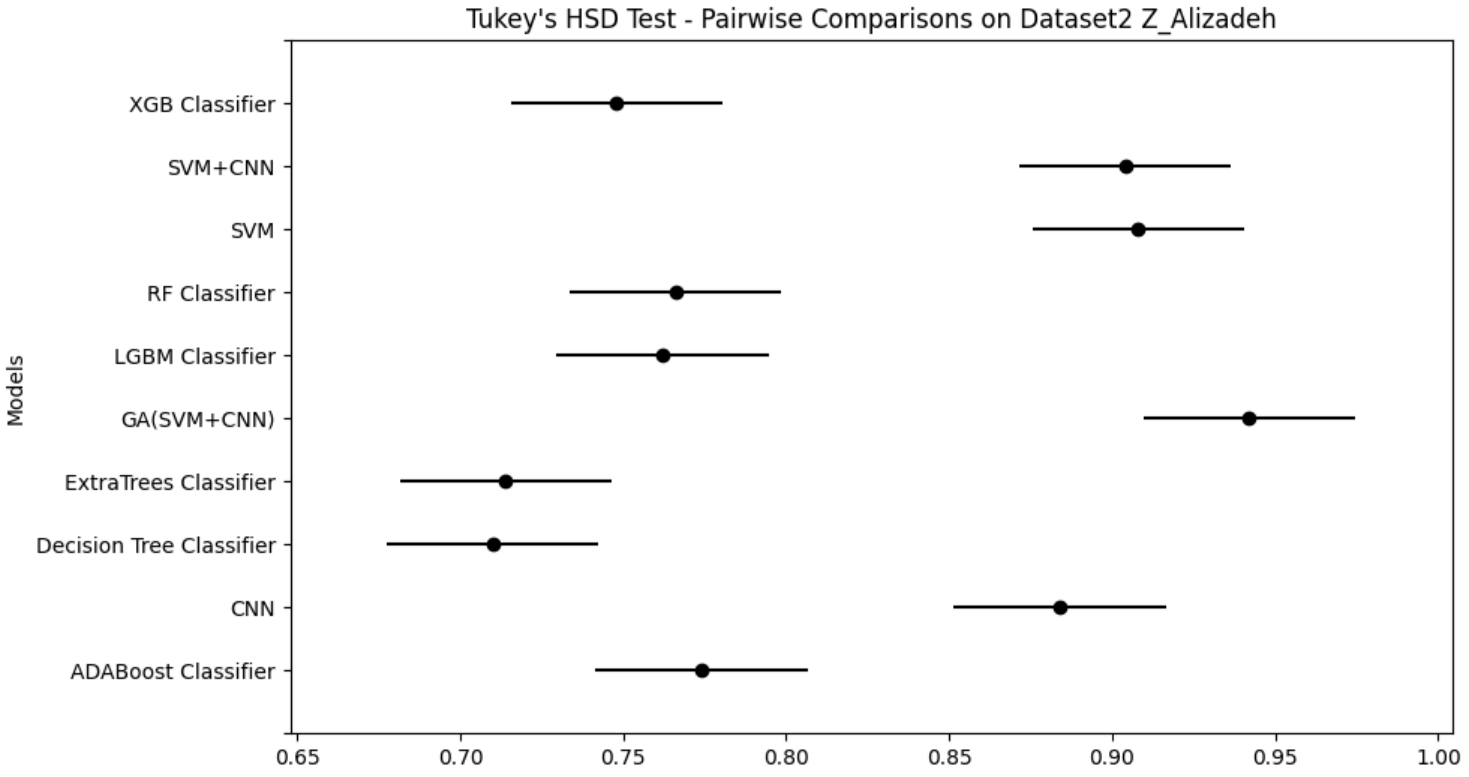


Figure 13

pairwise comparison on Z-AlizadehSani

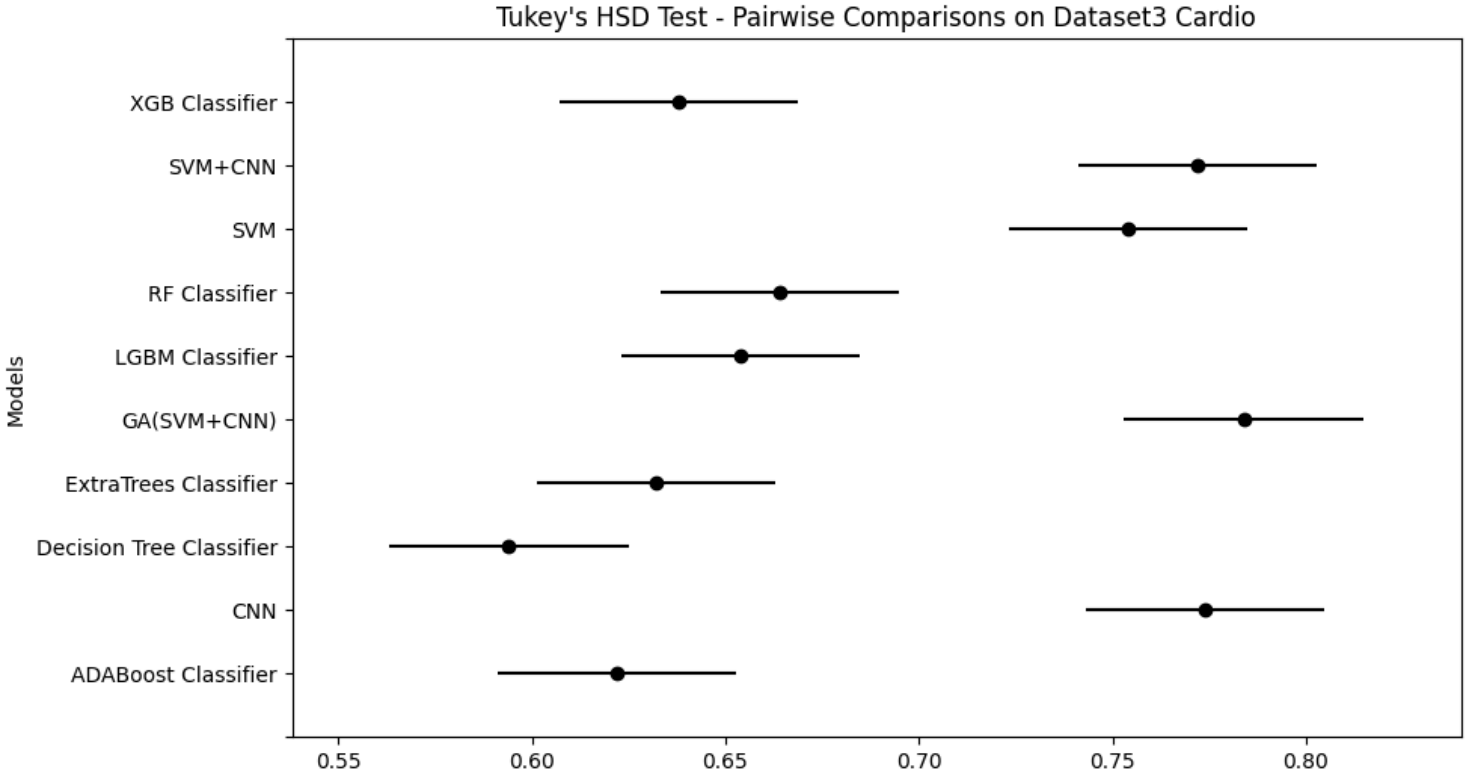


Figure 14

pairwise comparison on cardio