Integrating Canonical Correlation Analysis with Random Forest for Heart Disease Prediction

[1]Vetrithangam D, [2]Sivaneasan Bala Krishnan, [3]Siva Shankar S and [4]Prasun Chakrabarti
[1,2]Singapore Institute of Technology, Singapore, Asia.
[3]Department of Computer Science of Engineering, KG Reddy College of Engineering and Technology, Hyderabad, Telangana, India.
[4]Sir Padampat Singhania University, Udaipur, Rajasthan, India.
[1]vetrigold@gmail.com, [2]sivaneasan@singaporetech.edu.sg, [3]drsivashankars@gmail.com, [4]drprasun.cse@gmail.com

Correspondence should be addressed to Vetrithangam D : vetrigold@gmail.com

ArticleInfo
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi : https://doi.org/10.53759/7669/jmc202404109
Received 22 March 2024; Revised from 01 July 2024; Accepted 29 August 2024.
Available online 05 October 2024.

Abstract – Heart disease, a leading global cause of death over the past several decades, encompasses a range of disorders affecting the heart. Researchers use various data mining and machine learning techniques to analyze complex medical data, aiding healthcare professionals in predicting cardiac conditions. Despite these advances, existing models often struggle with effectively modelling non-linear relationships, maximizing feature correlation, and addressing challenges related to dimensionality and overfitting. This research paper introduces the Hybrid CCRF model for heart disease prediction, which integrates Canonical Correlation Analysis (CCA) with Random Forest. The proposed model generates polynomial features to capture non-linear relationships and applies Canonical Correlation Analysis to identify canonical variables that maximize correlations between heart disease features and chronic condition features. By combining these canonical variables into a single feature set, the model enhances prediction accuracy. The objectives of the Hybrid CCRF model are threefold: 1) To capture complex non-linear relationships between heart disease and chronic condition features by integrating polynomial feature generation with Canonical Correlation Analysis, thereby improving the model's ability to represent intricate data patterns; 2) To use CCA to identify and integrate canonical variables that enhance feature correlation, creating a more informative feature set; and 3) To address high-dimensional data and overfitting issues by combining canonical variables with polynomial features in a Random Forest model, balancing complexity and performance for improved generalization and robustness across various datasets. The proposed model achieved an accuracy of 99.45%, with a sensitivity of 98.53%, specificity of 99.54%, precision of 95.73%, and an F1 Score of 0.9711, outperforming all existing models.

Keywords – Heart Disease, Disease Prediction, Canonical Correlation, Random Forest, Non-Linear Relationship.

I. INTRODUCTION

Heart disease has emerged as the primary cause of death worldwide and is a serious public health issue. Heart failure, coronary artery disease, vascular disease, irregular heartbeats, and many more conditions fall under the umbrella of heart disease. The signs of heart failure might currently manifest in the human body at any stage of life. The likelihood of experiencing this kind of symptom is higher in the elderly than in the younger population. Traditionally, a doctor would diagnose heart disease (HD) by reviewing the patient's medical history, the results of their physical examination, and an analysis of any concerning symptoms. However, the results of this diagnosis procedure do not accurately identify the HD patient. Additionally, the analysis is costly and computationally challenging. Finding hidden patterns and compiling pertinent data from a vast dataset is a good way to tackle real-world challenges. Machine learning is the foundational area of artificial intelligence, capable of deriving both linear and nonlinear patterns from vast amounts of data. However, medical data are also rather massive and complex. In light of all of this, machine learning (ML) has grown more useful and is being applied in the medical industry to identify or predict many diseases. The use of machine learning (ML) has greatly improved diagnostic processes in the field of medical application. In order to reduce the difference between expected and actual results, machine learning algorithms use data to identify complex and nonlinear patterns in the characteristics. Through the integration of multiple techniques with machine learning models, hidden patterns can be identified, and analytical structures can be established, including clustering, classification, regression, and correlation. SVM is trained to detect breast cancer and other kind of disease. The various machine learning models are analyzed [1]. Stacked ensemble model with Hawks Optimizer (HO) is used for classification process, and this approach used the 97 % of accuracy [2].

Selecting the proper algorithms influences the accuracy of the prediction model [3]. Machine learning tools use strong math to handle complex patterns in data, helping predict heart disease and guide prevention and treatment [4][5]. Social Determinants are included in prediction of cardiovascular disease prediction [6]. A multi-label active learning model is used for heart disease prediction [7]. In this context, using machine learning algorithms entails leveraging advanced computational methods to analyze vast amounts of medical data, identifying patterns and risk factors that can more accurately predict the likelihood of heart failure [8]. This can lead to better prevention strategies, personalized treatment plans, and ultimately improved patient outcomes. Principal component analysis can be used to analyze data related to heart disease [9]. Feature selection methods along with machine learning algorithms are **g**enerally used in detecting and predicting the heart disease [10][11]. Most existing models face technical issues related to:1) Modelling non-linear relationships effectively. 2)Maximizing feature correlation and integration. 3)Handling dimensionality and overfitting challenges.

To address these challenges and fill the gaps, we have developed a Hybrid CCRF model with the following objectives:

1. To develop a model that effectively captures complex non-linear relationships between heart disease features and chronic condition features by integrating polynomial feature generation with Canonical Correlation Analysis (CCA). This approach aims to enhance the model's ability to represent intricate patterns and interactions within the data.
2. To leverage Canonical Correlation Analysis (CCA) to identify and combine canonical variables that maximize the correlation between heart disease and chronic condition features. This objective focuses on improving feature representation by integrating these canonical variables with polynomial features, thereby creating a more comprehensive and informative feature set for prediction.
3. To mitigate issues related to high-dimensional data and overfitting by combining canonical variables derived from CCA with polynomial features in the Random Forest model. This approach aims to balance model complexity and performance, improving generalization and robustness across diverse datasets.

II. LITERATURE REVIEW

This section explains existing machine learning models related to heart disease prediction, as well as various feature selection methods and other techniques. A Garg et al. [12] developed machine learning techniques for heart disease prediction. K-Nearest Neighbor (K-NN) and Random Forest algorithm produced the accuracy of 86.885% and 81.967%. The drawback of this model is that it produced less accuracy. Bhatt, C. M et al. [13] introduced k-modes clustering method and it can correctly predict cardiovascular diseases to reduce the fatality caused by cardiovascular diseases. Random forest model produced the accuracy of 87.05%. Because the model was validated based on a single dataset, it might not be applicable to different patient groups or populations. Subramani, S et al. [14] used stack of machine learning models to predict the cardio vascular disease and it produced the 96% of accuracy. This model's disadvantage is that a stacking technique requires more processing power for inference and training. Not every healthcare establishment will be able to accommodate this, especially in situations with limited resources. Taylan, O et al. [15] used various Machine learning models to predict and classify the cardio vascular disease. ANFIS's training procedure prediction accuracy is 96.56%. The system is complicated by the use of several models and methodologies, which could make it challenging to apply in a clinical environment. It determines the best threshold values. limiting its generalizability to real-world scenarios is drawback of this work. Elsedimy et al. [16] proposed a new method that uses support vector machines for cardiovascular disease prediction, and it produced an accuracy of 96.31%. It determines the best threshold values. The integration of QPSO with SVM and the use of an adaptive threshold method introduce complexity to the model. This complexity could make it more challenging to implement in clinical settings where simpler models are preferred for ease of interpretation and deployment. Khan, A et al. [17] used the machine learning (ML) algorithm, which is the RF algorithm, showing the best performance in terms of accuracy and sensitivity, but with relatively low specificity (43.48%) and notable misclassification errors (8.70%). Ambrish, G et al. [18] used logistic regression technique for prediction of cardiovascular disease. Features selection were performed by correlation with the target value for all the feature. The LR model obtained 87.10% accuracy. Although the study identified a 90:10 training-testing split as optimal for performance, this approach could lead to overfitting, where the model performs well on the training data but struggles with new data. The model lacks to identify a correlation between the features. Li, J. P et al. [19] proposed a new method that used conditional mutual information (FCMIM) feature selection algorithm for feature selection with support vector machines for heart disease prediction. The drawback of this work is that it takes more processing time for the diagnosis system and it achieved the accuracy of 92.37%. The model lacked generalization. Chang, V. et al. [20] used an artificial intelligence (AI) model along with machine learning to diagnose cardiac conditions and produced an accuracy rate of 83% over training data. It may not give better results in disease prediction. Ali F. et al. [21] introduced a smart healthcare system that used ensemble learning for heart disease prediction, and this system produced an accuracy of 98.5%. The study might be based on a specific dataset, and its results may not generalize well across different populations and regions. Ahmed, H et al. [22] proposed a new method to identify the heart disease, method used feature selection algorithms and machine learning algorithms and it produced the accuracy of accuracy at 94.9%. Sometimes selected features may not have the essential information to detect heart disease. Spencer, R t al. [23] used the BayesNet algorithm with feature selection for predicting heart disease. It achieved an accuracy of 85.00%. The drawback of this algorithm is that the performance and applicability of these models

may vary significantly with different datasets or in real-world settings, limiting their generalizability. Mienye, I. D et al. [24] proposed a new method that used the splitting method, classification, and regression tree to partition and predict heart disease. The proposed ensemble achieved classification accuracy of 93%. Overfitting may result in models. Following the literature review, several key issues are identified in existing models. These include limitations in accuracy, overfitting, and challenges related to interpretation and generalization. Many models struggle with effectively capturing non-linear relationships, maximizing feature correlation and integration, and addressing issues of dimensionality and overfitting. These challenges hinder the ability of models to provide robust and accurate predictions across diverse datasets and real-world scenarios. Addressing these issues is crucial for advancing predictive modelling techniques and improving overall model performance.

III. PROPOSED METHODOLOGY

This section provides a detailed overview of preprocessing, feature engineering, and Canonical Correlation Analysis (CCA). It outlines the methods used to prepare and transform the data, the techniques employed to generate and select relevant features, and the application of CCA to identify and maximize the relationships between different sets of features. These steps are crucial for ensuring the data is optimally structured and analyzed to enhance the accuracy and effectiveness of the predictive model**.**

Preprocessing

The data preprocessing steps involve separating the columns into numerical and categorical sets. Numerical columns (represented by Numerical_Cols) have missing values filled in with the mean value using the Imputer_Num object. Categorical columns (represented by Categorical_Cols) have missing values filled in with the most frequent value using the Imputer_Cat object for each individual column (iterated over by Col). Finally, one-hot encoding is applied to the categorical features using pd.get_dummies, creating a new dataset (DataOnehot) where each category is converted into a separate binary column. This step drops the first level of each category to avoid redundancy.

As shown in Equation 1, each column (denoted by j) is checked for missing values

$$\text{Missing\_values[j]} = \sum_{i=1}^{n} 1_{\{\text{data}_{ij}=\text{null}\}} \tag{1}$$

where n is the number of rows, and $1\{\cdot\}$ is the indicator function that equals 1 if the condition is true and 0 otherwise. Equation 2 demonstrates how to drop rows with missing values. D is the original dataset, and $D_{\text{cleaned is}}$ the dataset after dropping rows with missing values. where r_i is the i^{th} row in the dataset.

$$D_{\text{cleaned}} = \{ r_i \in D \mid \forall j, \text{data}_{ij} \neq \text{null} \} \tag{2}$$

$$C_{\text{num}} = \{ j \mid \text{data}_{ij} \in R \text{ for all } j \} \tag{3}$$

$$C_{\text{cat}} = \{ j \mid \text{data}_{ij} \notin R \text{ for all } j \} \tag{4}$$

Equations 3 and 4 introduce C_{num} and C_{cat}, representing the sets of numerical and categorical columns, respectively.

$$x_j = \frac{1}{n_j} \sum_{i=1}^{n} 1_{\{\text{data}_{ij}\neq\text{null}\}} \cdot \text{data}_{ij} \tag{5}$$

Equation 5 imputes missing values in numerical columns by replacing them with the mean of each respective column. where n_j is the number of non-null values in column j. One-hot encoding is a technique used to transform categorical variables into a numerical format suitable for machine learning algorithms. It works by creating a binary matrix, where each row represents a data point and each column represents a unique category within the original categorical variable. In this matrix, a value of 1 indicates the presence of that category for a particular data point, and a value of 0 indicates its absence. As described in Equation 6**,** a set of categorical columns can be represented for a dataset D with n rows and m columns.

$$C_{\text{cat}} \subseteq \{1,2,\ldots.,m\} \tag{6}$$

Each categorical column j in C_{cat} has a set of unique categories {c1, c2, ..., c_{kj}}. A new binary feature data_onehot_{ijc} is then created to represent these categories. This encoding scheme is shown in Equation (7).

$$\text{data\_onehot}_{ijc} = \{ 1 \text{ if data}_{ij} = c, 0 \text{ otherwise} \} \tag{7}$$

As shown in Equation (8), the one-hot encoding transformation for the entire dataset, D, can be represented as T(D).

$$\mathrm{T(D)} = \{\mathrm{data}_{\mathrm{onehot\ ijc}} \mid \mathrm{j} \in \mathrm{C_{cat}}\ \ \mathrm{i} \in \{1,2,..,\mathrm{n}\} \tag{8}$$

Feature Engineering

Let's denote the original features as follows

x1 = HighBP

$x2$ = HighChol

x3 = Stroke

x4 = Diabetes

The generated polynomial features will be all the unique products of these features taken two at a time, without the individual feature squares and without the bias term.

The polynomial features generated by the transformation are:

1. x1×x2=HighBP×HighChol
2. $x1$×$x3$=HighBP×Stroke
3. $x1$×$x4$=HighBP×Diabetes
4. $x2$×$x3$=HighChol×Stroke
5. $x2$×$x4$=HighChol×Diabetes
6. $x3$×$x4$=Stroke×Diabetes

Let X be the input matrix with rows representing samples and columns representing features. The original matrix X is:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ .. & .. & .. & .. \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{bmatrix} = \begin{bmatrix} \mathrm{HighBP}_1 & \mathrm{HighChol}_1 & \mathrm{Stroke}_1 & \mathrm{Diabetes}_1 \\ \mathrm{HighBP}_2 & \mathrm{HighChol}_2 & \mathrm{Stroke}_2 & \mathrm{Diabetes}_2 \\ .. & .. & .. & .. \\ \mathrm{HighBP}_{n1} & \mathrm{HighChol}_{n2} & \mathrm{HighChol}_{n3} & \mathrm{HighChol}_{n4} \end{bmatrix} \tag{9}$$

The transformed matrix X_{poly} with interaction-only polynomial features is:

$$X_{poly} = \begin{bmatrix} x_{11}\ x_{12} & x_{11}\ x_{13} & x_{11}\ x_{14} & x_{12}\ x_{13} & x_{12}\ x_{14} & x_{13}\ x_{14} \\ x_{21}\ x_{22} & x_{21}\ x_{23} & x_{21}\ x_{24} & x_{22}\ x_{23} & x_{22}\ x_{24} & x_{23}\ x_{24} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1}\ x_{n2} & x_{n1}\ x_{n3} & x_{n1}\ x_{n4} & x_{n2}\ x_{n3} & x_{n2}\ x_{n4} & x_{n3}\ x_{n4} \end{bmatrix} \tag{10}$$

$$X_{poly} = \begin{bmatrix} \mathrm{HighBP}_1 * \mathrm{HighChol}_1 & \mathrm{HighBP}_1 * \mathrm{Stroke}_1 & \mathrm{HighBP}_1 * \mathrm{Diabetes}_1 & \mathrm{HighChol}_1 * \mathrm{Stroke}_1 & \mathrm{HighChol}_1 * \mathrm{Diabetes}_1 & Stroke_1 * \mathrm{Diabetes}_1 \\ \mathrm{HighBP}_2 * \mathrm{HighChol}_2 & \mathrm{HighBP}_2 * \mathrm{Stroke}_2 & \mathrm{HighBP}_2 * \mathrm{Diabetes}_2 & \mathrm{HighChol}_2 * \mathrm{Stroke}_2 & \mathrm{HighChol}_2 * \mathrm{Diabetes}_2 & Stroke_2 * \mathrm{Diabetes}_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathrm{HighBP}_n * \mathrm{HighChol}_n & \mathrm{HighBP}_n * \mathrm{Stroke}_n & \mathrm{HighBP}_n * \mathrm{Diabetes}_n & \mathrm{HighChol}_n * \mathrm{Stroke}_n & \mathrm{HighChol}_n * \mathrm{Diabetes}_n & Stroke_n * \mathrm{Diabetes}_n \end{bmatrix} \tag{11}$$

Thus, the polynomial feature transformation effectively constructs a new feature space where each new feature represents a pairwise interaction between two original features.

Canonical Correlation Analysis (CCA)

Step 1: Define the Datasets

Let X be the matrix of polynomial cardiovascular features and Y be the matrix of chronic condition features.

Step 2: Standardize the Data

Standardize X and Y to have zero mean and unit variance

$$\tilde{X} = \frac{X-\mu X}{\sigma X} \tag{12}$$

$$\tilde{Y} = \frac{Y-\mu Y}{\sigma Y} \tag{13}$$

where μX and σX are the mean and standard deviation of X, and μY and σY are the mean and standard deviation of Y.

Step 3: Calculate the Covariance Matrices

Calculate the covariance matrices of the standardized data:

$$S_{XX} = \frac{1}{n-1} \tilde{X}^T \tilde{X} \tag{14}$$

$$S_{YY} = \frac{1}{n-1} \tilde{Y}^T \tilde{Y} \tag{15}$$

$$S_{XY} = \frac{1}{n-1} \tilde{X}^T \tilde{Y} \tag{16}$$

$\tilde{X}^T$ denotes the transpose of the standardized X matrix $\tilde{Y}^T$ denotes the transpose of the standardized Y matrix.

Step 4: Solve the Generalized Eigenvalue Problem
Generalized eigenvalue problem to find the canonical weights.

$$\left(S_{XY} {S_{YY}}^{-1} S_{YX}\right) a = \lambda (S_{XX}) a \tag{17}$$

$$\left(S_{YX} {S_{XX}}^{-1} S_{XY}\right) b = \lambda (S_{YY}) b \tag{18}$$

Here, a and b are the canonical weight vectors for X and Y, respectively, and λ represents the canonical correlation.

Step 5: Compute Canonical Variates
Compute the canonical variates for X and Y:

$$U = \tilde{X} a \tag{19}$$

$$V = \tilde{Y} b \tag{20}$$

Step 6: Extract Canonical Variables
Combine the canonical variates U and V into a single feature set. For k canonical components:
Canonical Variables for X: $U_1, U_2, \ldots, U_k$
Canonical Variables for Y: $V_1, V_2, \ldots, V_k$
Combine them into a single DataFrame:

$$\text{Canonical Features} = [U_1, U_2, \ldots, U_k, V_1, V_2, \ldots, V_k] \tag{21}$$

Working Principle of The Proposed Model
The proposed Hybrid CCRF model for heart disease prediction follows a systematic workflow as shown in **Fig 1**. Initially, data preprocessing is performed on both the heart disease and chronic disease datasets to ensure cleanliness and consistency. The data preprocessing begins by separating the columns into numerical and categorical sets. For the numerical columns, missing values are filled with the mean of each column using the Imputer_Num object, as defined in Equation (5). The categorical columns have missing values filled with the most frequent value for each column using the Imputer_Cat object. Once imputed, categorical features undergo one-hot encoding using pd.get_dummies, transforming the categories into separate binary columns.

To avoid redundancy, the first level of each category is dropped. The resulting dataset, DataOnehot, represents the categorical features in a machine-readable binary format, as described in Equations (6)-(8). Additionally, missing values are handled as shown in Equations (1) and (2). Rows with missing values can be removed to form a cleaned dataset, D_cleaned, ensuring data integrity. Numerical and categorical columns are further classified using Equations (3) and (4), defining sets for numerical (C_num) and categorical (C_cat) columns, respectively. Following preprocessing, pairwise interaction terms are generated as part of a polynomial feature transformation to capture non-linear relationships between features. Canonical Correlation Analysis (CCA) is then applied, beginning with data standardization to ensure all features are on a comparable scale. Covariance matrices are computed to understand relationships within each dataset, and canonical weights are derived to maximize correlations between the datasets. These weights are used to compute canonical variates, which are transformed into canonical variables representing the most significant correlations. The canonical variables from both datasets are combined into a single feature set, which is then used to train a Random Forest classifier for heart disease prediction. The model outputs a prediction, where a value of 1 indicates the presence of heart disease and a value of 0 indicates its absence. This comprehensive approach integrates advanced data preprocessing, statistical techniques, and machine learning methods, improving the accuracy and predictive power of heart disease detection.

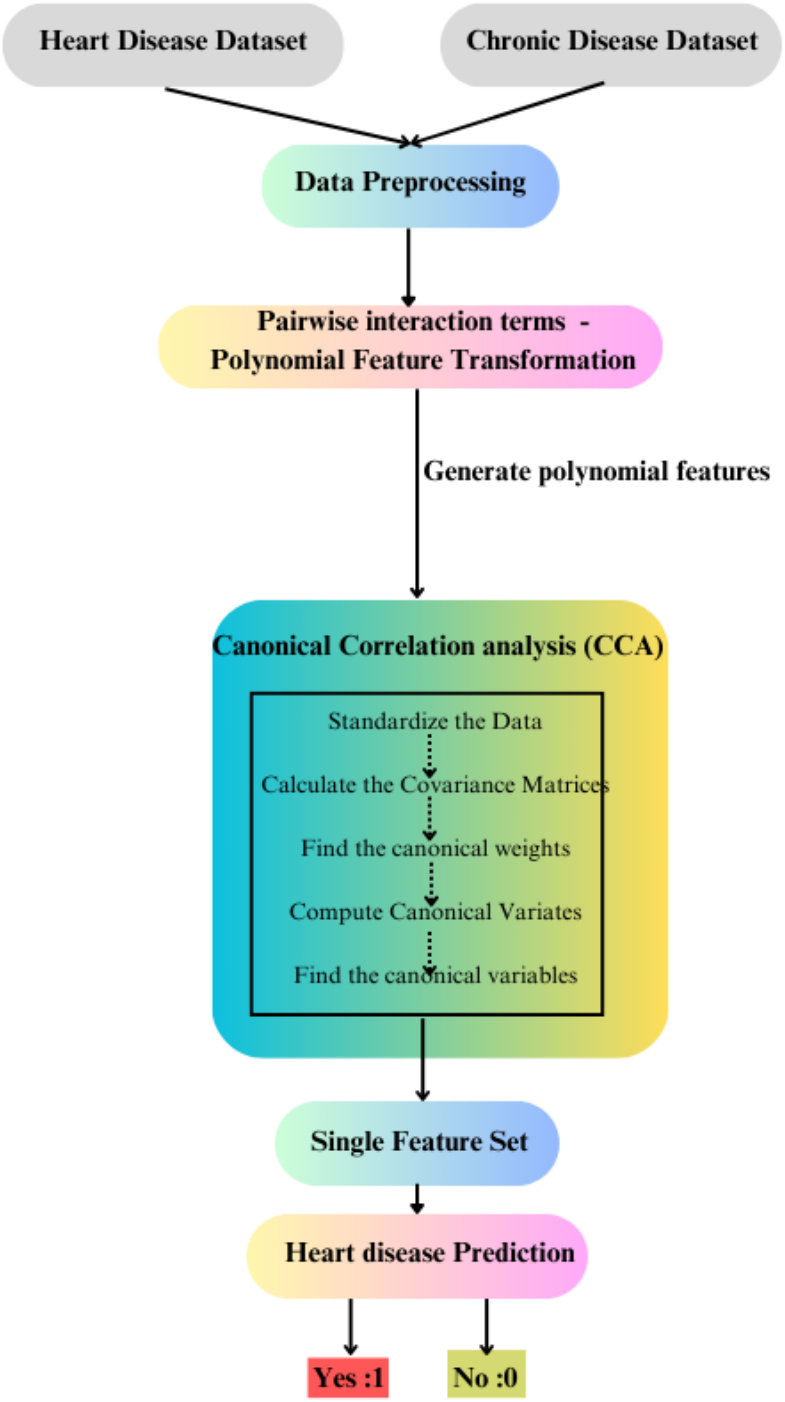

Fig 1. Workflow of the Proposed Hybrid CCRF Model.

IV. RESULTS AND ANALYSIS

Preprocessing

This section focuses on the data preparation process, encompassing both preprocessing and feature engineering. A summary of the DataFrame will be presented in **Table 1**, including the number of entries, column names, data types, and the number of non-null values in each column. This analysis allows for a comprehensive understanding of the dataset's structure and facilitates the identification of any columns containing missing data.

Dataset Info

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):

Table 1. Summary of the DataFrame

S. No	Column	Non-Null Count	Dtype
0	HeartDiseaseorAttack	253680 non-null	float64
1	HighBP	253680 non-null	float64
2	HighChol	253680 non-null	float64
3	CholCheck	253680 non-null	float64
4	BMI	253680 non-null	float64
5	Smoker	253680 non-null	float64
6	Stroke	253680 non-null	float64
7	Diabetes	253680 non-null	float64
8	PhysActivity	253680 non-null	float64
9	Fruits	253680 non-null	float64
10	Veggies	253680 non-null	float64
11	HvyAlcoholConsump	253680 non-null	float64
12	AnyHealthcare	253680 non-null	float64
13	NoDocbcCost	253680 non-null	float64
14	GenHlth	253680 non-null	float64
15	MentHlth	253680 non-null	float64
16	PhysHlth	253680 non-null	float64
17	DiffWalk	253680 non-null	float64
18	Sex	253680 non-null	float64
19	Age	253680 non-null	float64
20	Education	253680 non-null	float64
21	Income	253680 non-null	float64

dtypes: float64(22)
memory usage: 42.6 MB
None

To provide an initial look at the data, the first five rows of the dataset are presented in **Table 2**. This initial view offer s a glimpse into the actual data by showcasing a sample of records.

Table 2. Example Rows of the Dataset

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0

	Diabetes	PhysActivity	Fruits	AnyHealthcare	NoDocbcCost	GenHlth
0	0.0	0.0	0.0	1.0	0.0	5.0
1	0.0	1.0	0.0	0.0	1.0	3.0
2	0.0	0.0	1.0	1.0	1.0	5.0
3	0.0	1.0	1.0	1.0	0.0	2.0
4	0.0	1.0	1.0	1.0	0.0	2.0

	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
1	0.0	0.0	0.0	0.0	7.0	6.0	1.0
2	30.0	30.0	1.0	0.0	9.0	4.0	8.0
3	0.0	0.0	0.0	0.0	11.0	3.0	6.0
4	3.0	0.0	0.0	0.0	11.0	5.0	4.0

[5 rows x 22 columns]

This step analyzes the data for missing values by printing a count of missing values in each column as presented in **T able 3**. This provides valuable insight into which columns have missing data and the extent of those missing values. It als o involves evaluating heart disease prediction using supervised machine learning algorithms, including performance anal ysis and comparison. Assessing the data's quality and determining how to handle missing values during the data preparati on process are crucial for improving model performance.

Table 3. A Summary of Missing Values

Missing Values:	
HeartDiseaseorAttack	0
HighBP	0
HighChol	0
CholCheck	0
BMI	0
Smoker	0
Stroke	0
Diabetes	0
PhysActivity	0
Fruits	0
Veggies	0
HvyAlcoholConsump	0
AnyHealthcare	0
NoDocbcCost	0
GenHlth	0
MentHlth	0
PhysHlth	0
DiffWalk	0
Sex	0
Age	0
Education	0
Income	0
dtype: int64	

This step tackles rows with missing values. Any rows containing missing entries are removed from the dataset. To vis ualize the impact of this removal, the resulting Data Frame's shape (number of rows and columns) is then displayed. Add itionally, the first few rows of this "cleaned" Data Frame are presented in **Table 4**. This allows us to examine the data wit hout the clutter of missing values and provides a clearer picture of the remaining information.

Table 4. Data Shape After Dropping Missing Values (253680, 22)

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0

	Diabetes	PhysActivity	Fruits	AnyHealthcare	NoDocbcCost	GenHlth
0	0.0	0.0	0.0	1.0	0.0	5.0
1	0.0	1.0	0.0	0.0	1.0	3.0
2	0.0	0.0	1.0	1.0	1.0	5.0
3	0.0	1.0	1.0	1.0	0.0	2.0
4	0.0	1.0	1.0	1.0	0.0	2.0

	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
1	0.0	0.0	0.0	0.0	7.0	6.0	1.0
2	30.0	30.0	1.0	0.0	9.0	4.0	8.0
3	0.0	0.0	0.0	0.0	11.0	3.0	6.0
4	3.0	0.0	0.0	0.0	11.0	5.0	4.0

[5 rows x 22 columns]

In this initial step, the data undergoes a categorization process. The columns are separated into two distinct groups: nu merical and categorical. This separation is based on the data type of each column. Following this separation, lists containi ng the names of the numerical and categorical columns are printed. This step serves a crucial purpose, as it lays the groun dwork for the imputation process. By identifying the data types, we can then employ appropriate imputation strategies tai lored to each data type, ultimately leading to a more effective imputation process.

Numerical Columns

```
Index(['HeartDiseaseorAttack', 'HighBP', 'HighChol', 'CholCheck', 'BMI',
       'Smoker', 'Stroke', 'Diabetes', 'PhysActivity', 'Fruits', 'Veggies',
       'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'GenHlth',
       'MentHlth', 'PhysHlth', 'DiffWalk', 'Sex', 'Age', 'Education',
       'Income'],
      dtype='object')
```

Categorical Columns

```
Index([], dtype='object')
```

In this step, we address missing values within the numerical columns of the dataset. The process utilizes the mean val ue of each column to fill in any missing data points. Following imputation, the first few rows of the resulting dataset are displayed in **Table 5**, allowing us to visualize the impact of this step on the data.

Table 5. Dataset After Imputing Numerical Columns

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0

	Diabetes	PhysActivity	Fruits	AnyHealthcare	NoDocbcCost	GenHlth
0	0.0	0.0	0.0	1.0	0.0	5.0
1	0.0	1.0	0.0	0.0	1.0	3.0
2	0.0	0.0	1.0	1.0	1.0	5.0
3	0.0	1.0	1.0	1.0	0.0	2.0
4	0.0	1.0	1.0	1.0	0.0	2.0

	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
1	0.0	0.0	0.0	0.0	7.0	6.0	1.0
2	30.0	30.0	1.0	0.0	9.0	4.0	8.0
3	0.0	0.0	0.0	0.0	11.0	3.0	6.0
4	3.0	0.0	0.0	0.0	11.0	5.0	4.0

[5 rows x 22 columns]

This following step employs a strategy of imputing these missing values with the most frequent value observed for ea ch individual category. The resulting dataset, with these imputed values in the categorical columns, are presented in **Tab le 6**. To provide a quick glimpse of the changes, the first few rows of the modified data is displayed.

As shown in **Table 7**, This step checks and presents the count of missing values in each column after imputation. The expected output should show zero missing values for all column.

As presented in **Table 8**, this converts categorical columns to numerical columns using one-hot encoding, creating bi nary columns for each category. The drop_first=True parameter helps avoid multicollinearity by dropping the first catego ry. The output shows the data types of all columns after conversion, indicating which columns have been converted to bi nary format.

This step serves as a final inspection of the transformed data. It showcases the first few rows of the dataset after one-h ot encoding has been applied to categorical features. This allows us to visualize how these categorical features have been converted into separate binary columns, providing a clear picture of the final pre-processed data as presented in **Table 9**.

Table 6. Dataset After Imputing Categorical Columns

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0

	Diabetes	PhysActivity	Fruits	AnyHealthcare	NoDocbcCost	GenHlth
0	0.0	0.0	0.0	1.0	0.0	5.0
1	0.0	1.0	0.0	0.0	1.0	3.0
2	0.0	0.0	1.0	1.0	1.0	5.0
3	0.0	1.0	1.0	1.0	0.0	2.0
4	0.0	1.0	1.0	1.0	0.0	2.0

	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
1	0.0	0.0	0.0	0.0	7.0	6.0	1.0
2	30.0	30.0	1.0	0.0	9.0	4.0	8.0
3	0.0	0.0	0.0	0.0	11.0	3.0	6.0
4	3.0	0.0	0.0	0.0	11.0	5.0	4.0

[5 rows x 22 columns]

Table 7. A Summary of Missing Values after Imputation

Missing Values after Imputation	
HeartDiseaseorAttack	0
HighBP	0
HighChol	0
CholCheck	0
BMI	0
Smoker	0
Stroke	0
Diabetes	0
PhysActivity	0
Fruits	0
Veggies	0
HvyAlcoholConsump	0
AnyHealthcare	0
NoDocbcCost	0
GenHlth	0
MentHlth	0
PhysHlth	0
DiffWalk	0
Sex	0
Age	0
Education	0
Income	0
dtype: int64	

Table 8. Data Types after Conversion

Data Types after Conversion:
HeartDisease or Attack float64
HighBP float64
HighChol float64
CholCheck float64
BMI float64
Smoker float64
Stroke float64
Diabetes float64
PhysActivity float64
Fruits float64
Veggies float64
HvyAlcoholConsump float64
AnyHealthcare float64
NoDocbcCost float64
GenHlth float64
MentHlth float64
PhysHlth float64
DiffWalk float64
Sex float64
Age float64
Education float64
Income float64
dtype: object

Table 9. Sample One-Hot Encoded Dataset

First few rows of the one-hot encoded dataset:

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0

	Diabetes	PhysActivity	Fruits	AnyHealthcare	NoDocbcCost	GenHlth
0	0.0	0.0	0.0	1.0	0.0	5.0
1	0.0	1.0	0.0	0.0	1.0	3.0
2	0.0	0.0	1.0	1.0	1.0	5.0
3	0.0	1.0	1.0	1.0	0.0	2.0
4	0.0	1.0	1.0	1.0	0.0	2.0

	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
1	0.0	0.0	0.0	0.0	7.0	6.0	1.0
2	30.0	30.0	1.0	0.0	9.0	4.0	8.0
3	0.0	0.0	0.0	0.0	11.0	3.0	6.0
4	3.0	0.0	0.0	0.0	11.0	5.0	4.0

[5 rows x 22 columns]

Results of Polynomial Features and Canonical Correlation Analysis

The initial rows of the results generated by polynomial features are displayed in **Table 10**. These rows pro vide insight into how polynomial feature transformation expands the feature set by introducing new interaction terms and higher-order features, which help capture non-linear relationships in the data. This transformation pl ays a crucial role in improving the model's predictive capability by enabling it to better represent complex patt erns within the dataset.

Table 10. Example of Generated Polynomial Features

First Few Rows of Polynomial Features

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0

	Diabetes	PhysActivity	Fruits	Sex^2	Sex Age	Sex Education
0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0	0.0
2	0.0	0.0	1.0	0.0	0.0	0.0
3	0.0	1.0	1.0	0.0	0.0	0.0
4	0.0	1.0	1.0	0.0	0.0	0.0

	Sex Income	Age^2	Age Education	Age Income	Education^2
0	0.0	81.0	36.0	27.0	16.0
1	0.0	49.0	42.0	7.0	36.0
2	0.0	81.0	36.0	72.0	16.0
3	0.0	121.0	33.0	66.0	9.0
4	0.0	121.0	55.0	44.0	25.0

	Education Income	Income^2
0	12.0	9.0
1	6.0	1.0
2	32.0	64.0
3	18.0	36.0
4	20.0	16.0

[5 rows x 275 columns]

The first few rows of the canonical variables are displayed in **Table 11**. These canonical variables are derived through Canonical Correlation Analysis (CCA), which identifies relationships between heart disease features and chronic conditio n features. By transforming the original feature sets into a new space, the canonical variables maximize the correlation be tween the two datasets, allowing the model to capture meaningful patterns and dependencies. The canonical correlation c oefficient, calculated at 0.99, indicates a nearly perfect correlation between the two sets of variables. This step is essential in improving the overall prediction accuracy by integrating the most relevant information from both feature sets. The co mmon features contributing to the high correlation include Smoker, BMI, Physical Activity, Heavy Alcohol Consumptio n, General Health, Mental Health, Physical Health, Difficulty Walking, Age, and Sex. These features are critical in both heart disease and chronic condition datasets. Their inclusion in the Canonical Correlation Analysis (CCA) process enhan ces the model's ability to capture meaningful relationships between the two datasets, ultimately improving the accuracy o f heart disease predictions.

Table 11. First Few Rows of Canonical Variables

	C1_X	C2_X	C1_Y	C2_Y
0	2.375713	1.165770	2.432980	1.463676
1	-1.251166	0.943882	-0.114703	1.035340
2	1.397902	-0.211091	2.429492	1.408970
3	-0.277770	-0.768525	-0.024889	-0.855688
4	0.272771	-1.708305	-0.214919	-1.301108

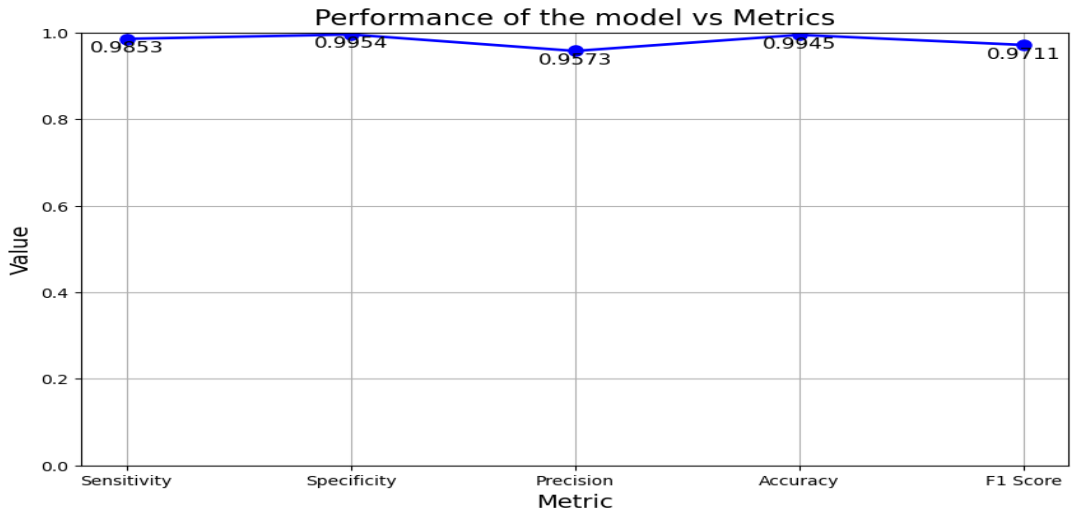

Fig 2 . Performance of the Proposed Hybrid CCRF Model.

As shown in **Fig 2**, the model's performance is evaluated using several key metrics. The sensitivity, calculated at 0.98 53, indicates that the model correctly identifies 98.53% of actual heart disease cases, demonstrating its effectiveness in m inimizing false negatives. The specificity of 0.9954 shows that the model accurately classifies 99.54% of non-disease cas es, reducing false positives and ensuring that healthy individuals are correctly identified. With a precision of 0.9573, 95.7 3% of the cases predicted as heart disease are true positives, indicating the reliability of the model in predicting positive o utcomes. The overall accuracy of the model is 0.9945, meaning that it correctly predicts the presence or absence of heart disease in 99.45% of cases, highlighting its robust performance. Additionally, the F1 Score, which balances both precisio n and recall, is 0.9711, further underscoring the model's reliability in handling both positive and negative cases. These pe rformance metrics collectively demonstrate that the Hybrid CCRF model is highly accurate and effective in heart disease prediction, making it a strong candidate for medical diagnostic applications.

Analysis

This section explains the analysis of the proposed model in comparison with existing models.

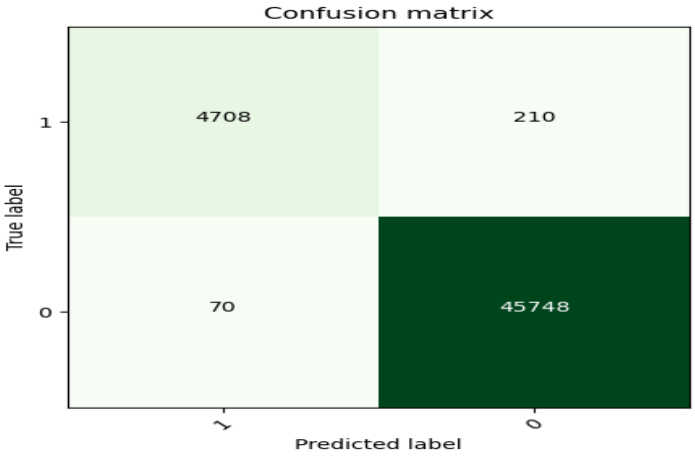

Fig 3. Confusion Matrix of the Proposed Hybrid CCRF Model.

In this research paper, heart disease prediction is conducted using a hybrid model that integrates Canonical Correlatio n Analysis (CCA) and Random Forest, referred to as the Hybrid CCRF model. The prediction outcomes were assessed us ing a confusion matrix, which provides a summary of the prediction performance by comparing actual and predicted labe ls. In this binary classification task, a value of 1 indicates the presence of heart disease, while 0 represents the absence of the disease. As shown in **Fig 3**, The confusion matrix results are as follows: the model correctly predicted heart disease in 4,708 cases, known as true positives (TP). It incorrectly predicted heart disease in 210 cases where no disease was presen t, categorized as false positives (FP). The model missed heart disease in 70 cases where the condition was actually presen t, referred to as false negatives (FN). Lastly, it correctly predicted no heart disease in 45,748 cases, classified as true nega tives (TN). These results highlight the model’s capability in distinguishing between heart disease and non-disease cases e ffectively.

Table 12. Comparison of Existing Models with the Proposed Hybrid CCRF Model

Model/ References	**Accuracy**
Machine learning techniques(KNN & Random forest) [12]	86.89% & 81.97%
k-modes clustering method [13]	87.05%
Machine Learning [14]	96%
An adaptive neuro-fuzzy inference system [15]	97%
QPSO with SVM [16]	96.31
Logistic Regression [18]	87.10%
Conditional mutual information (FCMIM) Feature selection algorithm [19]	92.31%
AI model using machine learning [20]	83%
Ensemble Learning[21]	98.50%
Feature selection algorithms and machine learning algorithms[22]	94.90%
Ensemble Learning classification[23]	93%
Deep neural network [25]	98.15%
Hybrid Random forest with a linear model (HRFLM) [26]	88.70%
Heart disease prediction model (HDPM) [27]	96%
Hybrid decision support system [28]	86.60%
Machine Learning model with Feature Selection [29]	73%
Proposed Hybrid CCRF model	99.45%

As shown in **Table 12**, the proposed Hybrid CCRF model is compared with several existing heart disease prediction models based on their accuracy. The **Table 12** highlights a range of machine learning techniques and hybrid models used in prior research, showing varying levels of effectiveness in predicting heart disease. For example, traditional machine le arning techniques like K-Nearest Neighbours (KNN) and Random Forest achieved accuracies of 86.89% and 81.97%, res pectively [1], while models employing k-modes clustering [2] and logistic regression [12] yielded accuracies around 87% . More advanced methods, such as an adaptive neuro-fuzzy inference system [5], achieved an accuracy of 97%, and an en semble learning approach reached 98.50% [17]. Deep neural networks also performed well, with an accuracy of 98.15% [23]. However, the proposed Hybrid CCRF model outperformed all of these methods with an accuracy of 99.45%, demon strating superior performance in heart disease prediction. This significant improvement can be attributed to the integratio n of Canonical Correlation Analysis (CCA) with Random Forest, which effectively captures both linear and non-linear re lationships between features, enhancing predictive capabilities. The comparison clearly illustrates that the Hybrid CCRF model offers a robust and reliable solution for heart disease prediction, surpassing many state-of-the-art models in terms of accuracy.

V. CONCLUSION AND FUTURE WORK

The Hybrid CCRF model, which combines Canonical Correlation Analysis (CCA) with Random Forest, exhibits excepti onal performance in heart disease prediction. By integrating polynomial feature generation with CCA, the model effectiv ely captures complex non-linear relationships and maximizes feature correlation between heart disease and chronic condi tion features. This approach results in a single, enhanced feature set that significantly boosts prediction accuracy. The mo del achieved a remarkable accuracy of 99.45%, with a sensitivity of 98.53%, specificity of 99.54%, precision of 95.73%, and an F1 Score of 0.9711, surpassing the performance of existing models. These results highlight the model's effectiven ess in overcoming challenges related to non-linearity, dimensionality, and overfitting.

Future research could explore several directions to further enhance the Hybrid CCRF model. Incorporating additional data modalities, such as genetic or lifestyle factors, may provide a more comprehensive view of heart disease and improv e model performance. Advanced techniques in feature selection and dimensionality reduction could further refine the mo del, addressing any remaining issues related to high-dimensional data. Additionally, assessing the model's performance a cross diverse populations and real-world clinical settings would validate its generalizability and robustness. Integrating th e Hybrid CCRF model with real-time monitoring systems could also enable early detection and timely intervention, pote ntially advancing predictive healthcare in cardiology.

Data Availability

No data was used to support this study.

Conflicts of Interests
The author(s) declare(s) that they have no conflicts of interest.

Funding
No funding agency is associated with this research.

Competing Interests
There are no competing interests.

References

[1]. Rubini P. E., Dr. C. A. Subasini, Dr. A. Vanitha Katharine, V. Kumaresan, S. Gowdham Kumar, T. M. Nithya, "A Cardiovascular Disease Prediction using Machine Learning Algorithms", Annals of RSCB, vol. 25, no. 2, pp. 904–912, Mar. 2021.
[2]. A. S. Kumar and R. Rekha, "An improved hawks optimizer based learning algorithms for cardiovascular disease prediction," Biomedical Signal Processing and Control, vol. 81, p. 104442, Mar. 2023, doi: 10.1016/j.bspc.2022.104442.
[3]. C. Krittanawong et al., "Machine learning prediction in cardiovascular diseases: a meta-analysis," Scientific Reports, vol. 10, no. 1, Sep. 2020, doi: 10.1038/s41598-020-72685-1.
[4]. W. Sun, P. Zhang, Z. Wang, and D. Li, "Prediction of Cardiovascular Diseases based on Machine Learning," ASP Transactions on Internet of Things, vol. 1, no. 1, pp. 30–35, May 2021, doi: 10.52810/tiot.2021.100035.
[5]. M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Computers in Biology and Medicine, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.
[6]. Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, "Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review," American Journal of Preventive Medicine, vol. 61, no. 4, pp. 596–605, Oct. 2021, doi: 10.1016/j.amepre.2021.04.016.
[7]. I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction," Sensors, vol. 22, no. 3, p. 1184, Feb. 2022, doi: 10.3390/s22031184.
[8]. E. D. Adler et al., "Improving risk prediction in heart failure using machine learning," European Journal of Heart Failure, vol. 22, no. 1, pp. 139–147, Nov. 2019, doi: 10.1002/ejhf.1628.
[9]. A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Informatics in Medicine Unlocked, vol. 19, p. 100330, 2020, doi: 10.1016/j.imu.2020.100330.
[10]. Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," Scientific Reports, vol. 10, no. 1, Nov. 2020, doi: 10.1038/s41598-020-76635-9.
[11]. Vetrithangam, D., Senthilkumar, V., Kumar, A. R., Naresh, P., & Sharma, M, "Coronary Artery Disease Prediction Based on Optimal Feature Selection Using Improved Artificial Neural Network with Meta-Heuristic Algorithm." Journal of Theoretical and Applied Information Technology, vol.100. no.24, p.4771-4782, (2022).
[12]. A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, p. 012046, Jan. 2021, doi: 10.1088/1757-899x/1022/1/012046.
[13]. C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, vol. 16, no. 2, p. 88, Feb. 2023, doi: 10.3390/a16020088.
[14]. S. Subramani et al., "cardiovascular diseases prediction by machine learning incorporation with deep learning," Frontiers in Medicine, vol. 10, Apr. 2023, doi: 10.3389/fmed.2023.1150933.
[15]. O. Taylan, A. Alkabaa, H. Alqabbaa, E. Pamukçu, and V. Leiva, "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods," Biology, vol. 12, no. 1, p. 117, Jan. 2023, doi: 10.3390/biology12010117.
[16]. E. I. Elsedimy, S. M. M. AboHashish, and F. Algarni, "New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization," Multimedia Tools and Applications, vol. 83, no. 8, pp. 23901–23928, Aug. 2023, doi: 10.1007/s11042-023-16194-z.
[17]. A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction," Health & Social Care in the Community, vol. 2023, pp. 1–10, Feb. 2023, doi: 10.1155/2023/1406060.
[18]. A. G, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," Global Transitions Proceedings, vol. 3, no. 1, pp. 127–130, Jun. 2022, doi: 10.1016/j.gltp.2022.04.008.
[19]. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/access.2020.3001149.
[20]. V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," Healthcare Analytics, vol. 2, p. 100016, Nov. 2022, doi: 10.1016/j.health.2022.100016.
[21]. F. Ali et al., "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," Information Fusion, vol. 63, pp. 208–222, Nov. 2020, doi: 10.1016/j.inffus.2020.06.008.
[22]. H. Ahmed, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark," Future Generation Computer Systems, vol. 111, pp. 714–722, Oct. 2020, doi: 10.1016/j.future.2019.09.056.
[23]. R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," Digital Health, vol. 6, p. 205520762091477, Jan. 2020, doi: 10.1177/2055207620914777.
[24]. I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," Informatics in Medicine Unlocked, vol. 20, p. 100402, 2020, doi: 10.1016/j.imu.2020.100402.
[25]. S. I. Ayon, Md. M. Islam, and Md. R. Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques," IETE Journal of Research, vol. 68, no. 4, pp. 2488–2507, Jan. 2020, doi: 10.1080/03772063.2020.1713916.
[26]. S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/access.2019.2923707.
[27]. N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," IEEE Access, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/access.2020.3010511.
[28]. P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," Journal of Reliable Intelligent Environments, vol. 7, no. 3, pp. 263–275, Jan. 2021, doi: 10.1007/s40860-021-00133-6.
[29]. M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," Healthcare Analytics, vol. 2, p. 100060, Nov. 2022, doi: 10.1016/j.health.2022.100060.