Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

CARDIOVASCULAR DISEASE DETECTION USING MACHINE LEARNING AND RISK CLASSIFICATION BASED ON FUZZY MODEL

GYAN PRASAD PAUDEL
Graduate School of Science and Technology, Mid-West University, Surkhet, Nepal.
Email: gyan.math725114@gmail.com

PARBATI KUMARI UPADHYAY *
Graduate School of Education, Mid-West University, Surkhet, Nepal.
*Corresponding Author Email: parbatipaudel725114@gmail.com

Abstract

The global prevalence of heart disease indicates a major public health issue. It causes shortness of breath, weakness, and swollen ankles. Early heart disease diagnosis is difficult with current approaches. Hence, a better heart disease detection tool is needed. Treatment requires more than just diagnosis. Risk classification is critical for accurate diagnosis and treatment. In this analysis, a novel cardiovascular disease (CVD) detection paradigm using machine learning (ML) and risk classification based on a weighted fuzzy system is proposed. The system is developed based on ML algorithms such as artificial neural network (ANN) and Long Short-Term Memory (LSTM) and uses standard feature selection techniques knowns as Principal Component Analysis (PCA). Furthermore, the cross-validation method has been used for learning the best practices of model assessment and for hyperparameter tuning. The accuracy-based performance measuring metrics are used for the assessment of the performances of the classifiers. Finally, the outcomes revealed that the proposed model achieved an accuracy of 94.01% which is higher than another conventional model developed in this domain. Additionally, the proposed system can easily be implemented in healthcare for the identification of heart disease.

AMS Subject Classification: *03E72; 92C50*

Keywords: *Cardiovascular Disease (CVD), Principal Component Analysis (PCA), Disease Classification, Fuzzy Model, Machine Learning (ML)*

1. INTRODUCTION

The incidence of heart disease has increased noticeably, and it has now surpassed all other causes of death as the leading cause of death in the majority of countries throughout the world [1]. Numerous distinct features of CVD can affect the structure or function of the heart [2]. Some disorders may be difficult for medical experts to diagnose quickly and accurately [3-4].

The treatment of cardiac disease is being performed by several systems, many of which depend on methods of soft computing that have been developed [5]. In particular, the combination of multiple different forms of soft computing to construct hybrid models has been examined as a means of producing results that are superior to those produced by a single kind of computational model [6]. In most cases, these models included two distinct states. In the first stage, approaches for selecting features are employed to pick a subset of those characteristics [7]. After that, the produced subset of characteristics is utilized as data for the categorization procedures that are employed in the second state [8]. Figure 1 shows various types of CVDs.

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

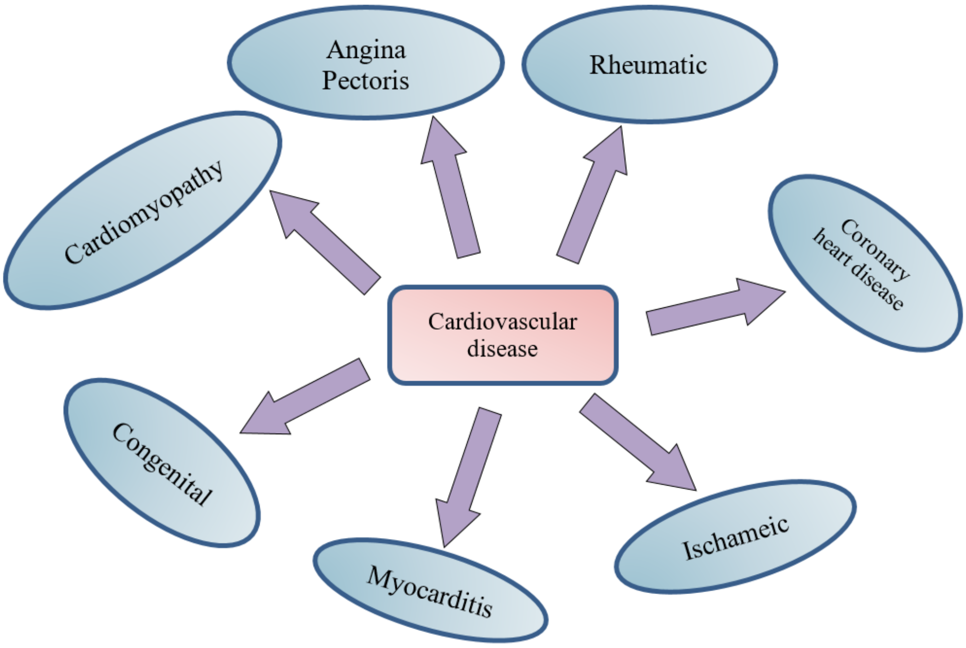


Figure 1: Types of CVD [9]

1.1 Fuzzy technique for risk classification

A set is said to be fuzzy if it permits its components to have varying levels of inclusion in the range [0, 1] [10]. The fuzzy categorization system provides an alternative crisp logic by analyzing data sets according to the individuals' membership in each category [11]. The concept of fuzzy membership is predicated on the idea that a person's participation in a particular group can vary from full membership (100%) to non-participation (0%), and it recognizes the possibility that a dataset might be divided into partial participation in two or more groups [12]. Figure 2 shows an illustration of fuzzy logic.

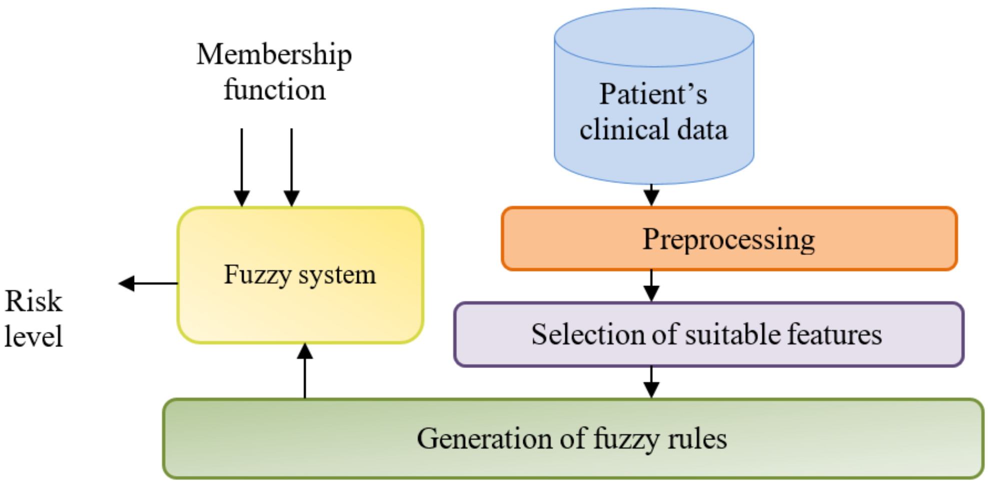


Figure 2: Block diagram of fuzzy logic [13]

ISSN : 1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

To represent the level of membership, fuzzy logic employs truth levels that range from 0.0 to 1.0 [14]. The values of the attributes are changed to fuzzy values. As an example, revenue is projected into the discrete classifications "low, medium, and high," and then fuzzy values are determined for each category. It is possible that more than one fuzzy number would be relevant to a certain fresh sample. Each criterion that is relevant casts a vote about membership in the respective groups. Typically, one would begin by adding up the truth values of each anticipated category [15]. This research has the following contributions.

- Initially the researcher try to address some problems related to early diagnosis of CVD diseases.
- Mostly the machine learning model, only detect or predict the CVD but not classify the level of risk. Therefore, for better treatment of the CVD disease a risk classification of CVD disease is analyzed using fuzzy system model.
- Thirdly, determine which weak characteristics in the dataset have an impact on the classifiers' performance.
- Fourthly, for optimize the performance of the classifier the ANN is ensembled with LSTM to provide more accurate detection of the disease.
- And at last, a complete model is designed which can perform both disease detection and risk classification as low, medium, and high as well.

In this research the remaining portions of the paper are organized as follows. The reviewed literature has been examined in section 2. Section 3, 4, and 5 discuss the background study, problem statement, and research objectives respectively. Section 6 focuses further on the mathematical and theoretical understanding of feature selection and detailed methodology. Section 7 analyses and discusses in depth the dataset as well as the outcomes of all trials. In the last part, 8 the conclusion and future scope of the study effort have been addressed in detail.

2. LITERATURE REVIEW

This strategy has been employed by a wide range of authors, who then presented their findings after doing a literature review:

Wojcik et al., (2023) [16] developed a health expert system for determining the severity of coronary artery tumors in individuals who suffer from coronary artery disease using fuzzy sets as the underlying data structure. The use of actual data in testing the intelligent system. In the end, it was found that the level of structural abnormality of the coronary artery in individuals with different kinds of coronary disease was 95%, according to the opinion of specialists.

Taylan et al., (2023) [17] discovered that early and correct identification of CVDs is of utmost significance to reduce the likelihood of suffering a myocardial attack. As a result, an approach that makes use of the adaptive neuro-fuzzy inference system (ANFIS) technology has been presented as a means of predicting, classifying, and enhancing the diagnostic efficiency of CVDs. According to the findings of the numerical study, the level of accuracy of prediction offered by ANFIS throughout the training phase is 96.56%.

Kharya et al., (2023) [18] initiated a novel idea known as the fuzzy-weighted Bayesian belief network (FWBBN), which was used to construct and create a healthcare diagnosis support tool based on the BBN. The fuzzy concept is being used for characteristics to cope with real-life circumstances to eliminate sharp boundary concerns that exist in the medical field. In conclusion, it was determined that FWBBN, in comparison to the traditional Bayesian model, is capable of being applied in a manner that is both highly effective and precisely precise in terms of high efficiency and low time intricacy.

ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

Seslier and Karakus (2023) [19] investigated that about 46% of the death of people in the world, excluding communicable diseases and accidents, are because of CVDs. In this analysis, various ML methods are used to determine heart disease. At last, it concluded that among various classifiers such as Logistics regression, Support Vector Machine (SVM), Naïve bayes, and Random Forest (RF), the SVM technique achieved the best accuracy outcomes at 87.91%.

Karthick et al., (2022) [20] indicated that ML is a viable tool for lowering and comprehending heart disease symptoms. A data visualization was made to show how the characteristics are connected. Experimental results show that for 303 data instances using 13 characteristics hand-picked from the Cleveland dataset, the RF method reaches 88.5% accuracy during validation.

Dong et al., (2022) [21] pointed out that diabetic kidney syndrome is the main reason for end-stage renal disease, CVD, and all-cause death and morbidity in individuals who have diabetes. 46 medical features that are readily accessible through the use of electronic medical records were utilized in the process of developing prediction models following seven different ML methods. At last, it concluded that the Light Gradient Boosting Machine (GBM) framework had the maximum Area under curve (AUC) of 0.815.

Ahmed et al., (2022) [22] emphasized the significance of disease prediction and early detection in the realm of medicine as a means of disease prevention. It is essential to have a thorough understanding of the disease's symptoms to make an accurate prognosis. At the moment, ML algorithms are useful for diagnosing diseases. Two distinct kinds of models make up the conceptual framework: SVM and ANN methods. In this paper, a model that employs a fused ML technique is shown. In the end, it concluded that the new fused ML framework has a predicted accuracy of 94.87, which is greater than the recently reported techniques.

Nadakinamani et al., (2022) [23] evaluated a wide variety of cutting-edge ML techniques to develop a CVD forecasting method that is very reliable. To determine which ML approach is the most appropriate, the efficiency of the suggested method was measured across several different criteria. The Decision Tree (DT) approach performed very well, with a maximum accuracy of 100%, when it came to forecasting individuals who would be diagnosed with CVD.

Cenitta et al., (2022) [24] recently introduced ischemic heart disease innovative missing value imputation techniques (IHDMIT) using fuzzy-rough sets and their expansions. The novel IHDMIT with RF classification against the state-of-the-art methods of expectation maximization, fuzzy C means, and fuzzy roughest is evaluated. According to the findings, the suggested IHDMIT RF classifier achieves a higher accuracy of 93%.

Yilmaz et al., (2022) [25] developed three distinct models for classifying coronary heart disease using RF, logistic regression LR, and SVM methods respectively. Accuracy served as the criterion for determining how well the models performed. At last, it concluded that the RF classifier had the greatest accuracy of 92.9% among all of the classifiers.

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

Table 1 shows the comparative analysis of the reviewed literature of different authors.

Table 1: Comparison of the reviewed literature

Author Name	Technique used	Outcomes
Wojcik et al., (2023) [16]	Fuzzy set	95% of patients with different kinds of coronary disease had anatomical coronary artery damage, according to specialists.
Taylan et al., (2023) [17]	ANFIS	According to a numerical study, ANFIS's training process prediction accuracy is 96.56%.
Kharya et al., (2023) [18]	FWBBN	Compared to the Bayesian model, FWBBN performs better and takes less time.
Seslier and Karakus (2023) [19]	SVM	The SVM approach produced the most accurate results, which came in at 87.91% overall.
Karthick et al., (2022) [20]	RF	By validating 303 data instances using 13 attributes from the Cleveland dataset, the RF technique obtains an accuracy of 88.5%.
Dong et al., (2022) [21]	Light (GBM) Model	The Light GBM framework had a maximum AUC of 0.815.
Ahmed et al., (2022) [22]	SVM, ANN	The suggested fused ML model outperforms the previously disclosed techniques with a forecasting accuracy of 94.87.
Nadakinamani et al., (2022) [23]	DT	The DT approach worked wonderfully in terms of patient prediction for cardiovascular illness, with the greatest accuracy of 100%.
Cenitta et al., (2022) [24]	IHDMIT	The result demonstrates that the suggested IHDMIT RF classifier provides greater accuracy, 93%.
Yilmaz et al., (2022) [25]	RF	In terms of accuracy, the RF classifier has outperformed all others with a score of 92.9%.

3. BACKGROUND STUDY

CVD is among the most difficult illnesses to treat, and it affects a considerable number of individuals throughout the world. A vital part of medical treatment, particularly in the specialism of cardiology, is the prompt and accurate identification of heart ailments. This research presents an ML-based algorithm that is efficient and accurate for the diagnosis of heart illness. The process would provide quicker and more precise outcomes. To improve the categorization accuracy and decrease the amount of time required for the execution of the classification system, feature-selecting algorithms are employed in the feature selection process. In addition, best practices in model evaluation and hyperparameter modification have been learned with the use of the cross-validation technique. Measures of performance are used as classifiers to determine efficiency. The characteristics chosen by the feature selection techniques have been used to evaluate the classifiers' performance. The experimental findings confirm the viability of the suggested feature selection approach using the classifier SVM for developing a sophisticated intelligent system for diagnosing CVD [26].

4. PROBLEM FORMULATION

Heart disease is a significant public health problem, as shown by the large number of individuals who have been affected by it all over the globe. It manifests itself with the typical symptoms of shortness of breath, general body weakness, and swollen feet. The present methods of diagnosing heart illness are not very successful in early-time classification. As a result, a method that is more effective in detecting heart disease needs to be developed. In addition, the identification of the condition alone

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

is insufficient for appropriate therapy. To provide an accurate diagnosis and provide appropriate therapy, risk categorization is essential. In light of this, the researcher of this study article offers an innovative approach for the diagnosis of diseases and the precise categorization of risks.

5. RESEARCH OBJECTIVES

- To develop and design an ML model to predict and detect heart disease.
- To design a model for further risk classification as low medium or high using the fuzzy technique.
- To prove the robustness of the proposed model by comparing it with another conventional model in terms of accuracy and other performance evaluation parameters.

6. RESEARCH METHODOLOGY

The concept of designed architecture is examined in the context of research methodology.

6.1 Technique Used

Two techniques are used in the proposed methodology. These techniques are given below:

- **Artificial Neural networks (ANN)**

An ANN is a kind of computational model that attempts to mimic the way a real brain operates. Applications that include prediction, categorization, and pattern recognition make up the majority of ANN deployments. Calculating the activation of a neuron using just the supplies to the network and the activation function, f(A), yields the following expression:

$$f(A) = f[\sum_{i=1}^{n} x_i w_i + b] \quad (1)$$

Where n is the total number of intakes, xi is the neurons that make up the input, w_i represents the weighted numbers, and b indicates the bias value. The initial weights that are given with each input are constantly being adjusted to reduce the amount of variance that exists between the goal and the projected values. The following is an updated version of the weights:

$$w_{ij}^{new} = w_{ij}^{new} + \Delta w_{ij} \quad (2)$$

To adjust the weights in an ANN, the fractional derivative or gradient, $\partial E/\partial w$, of a loss function 'E' is calculated for every weight 'w' in the network. This can be done for any weight in the system. It is recommended to use a gradient decline learning rule known as the delta rule:

$$\Delta w_{ij} = -\eta \frac{\partial E_n}{\partial w_{ij}} \quad (3)$$

In this equation, 'E' represents the ANN error function, and η denotes the learning rate of the network. The definition of 'E' can be written as:

$$E_n = \frac{1}{2} \sum_{s=1}^{n} \sum_{o=1}^{m} (T_{so} - Y_{so})^2 \quad (4)$$

Where s represents the total number of trials, o is the overall number of outputs, T is the desired output, and Y is the actual output respectively [27].

- **Fuzzy technique**

A fuzzifier, an inference engine, and a defuzzifier are the four components that make up a standard fuzzy system [28]. These components are shown in figure 3.

Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

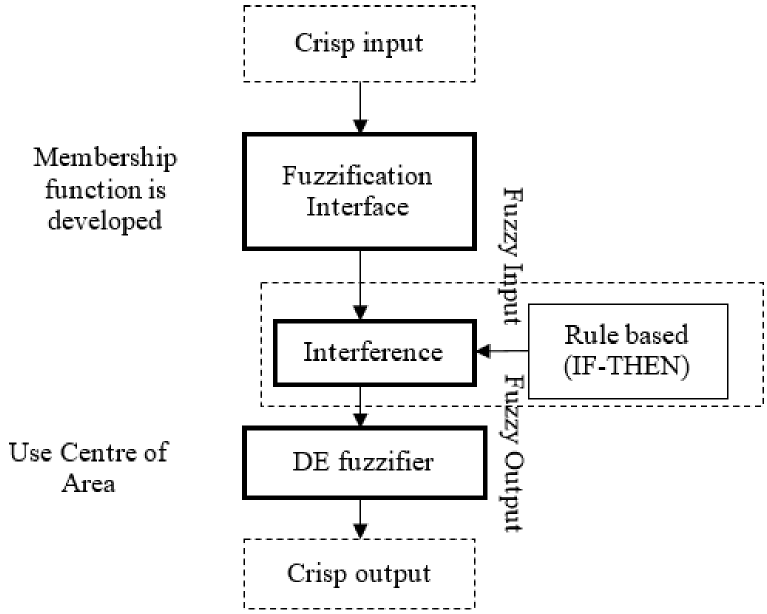


Figure 3: Fuzzy model [28]

The basic principle used in the fuzzy system.

- The establishment of fuzzier rules.
- Adjust the accuracy of the input data depending on their degree of membership.
- Combining fuzzy information with fuzzy rules to make more accurate judgments.
- The output is then de-fuzzified, resulting in a clear numerical value.

Below is an illustration of the principle used to assess the membership values:

$$f(a) = \begin{cases} 0 & if\ a \leq m \\ \frac{a-m}{n-1} & if\ m \leq a \leq n \\ \frac{x-a}{x-n} & if\ n \leq a \leq x \\ 0 & if a \geq x \end{cases} \tag{5}$$

Input variables are denoted by a, m n, and x. Fuzzy rule creation is a crucial task since it facilitates the process of input-to-output recording. High, medium, and low are examples of linguistic values that could be utilized to define a fuzzy rule, using attributes A1, A2..., A_n, and class labels C1, and C2. As a result, the fuzzy rule could be stated as follows:

- If A1 is superior to A2 and A3 is average, then C2 is the highest possible class.
- If A1 is low and A2 is medium and A3 is medium, then the class is C1.
- If A1 is high and A2 is medium and A3 is low, then the class is C2 [29].

- **Analytic Hierarchy Process (AHP)**

Analytic Hierarchy Processing, often known as AHP, is a strategy for generating decisions based on several criteria in an organized manner. This method considers a collection of assessment criteria or qualities as well as a set of alternatives to conclude. In the first step, the structure of the decision hierarchy is defined using a list of qualities and potential options. In the second step of the process,

ISSN：1674-2974 | CN 43-1061 / N
DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

the pair-wise comparative matrix is built, and the normalized matrix is calculated. In the third step, the Eigen numbers of the characteristics are computed so that the greatest Eigenvalues possible can be compared. At last, the characteristic weights are calculated based on an appropriate consistency ratio and consistency. Table 2 shows the weight of the attribute using the AHP method [27].

Table 2: attributes and their weights [27]

Attributes	Attribute weights
Age	0.0822
Sex	0.0287
Chest pain type	0.1333
blood pressure	0.0645
cholesterol	0.0560
Fasting blood pressure	0.0530
Resting electrocardiographic result	0.0452
Maximum heart rate	0.1235
Exercise-induced Angina	0.0696
Old Peak	0.0997
The slope of the peak exercise	0.0386
Number of major vessels	0.0849
Thallium scan	0.1208

- **Long Short-Term Memory**

The term LSTM was introduced to reconcile gradients that are disappearing or bursting in a recurrent neural network. The LSTM is equipped with an internal memory cell that is accessed by forgetting and input gate networks. In an LSTM layer, a forget gate controls how often memory should be transferred into the following time step. An input gate, on the other hand, scales fresh input to memory cells. LSTM can represent either long-term or short-term reliance on sequential data, depending on the states of both gates [30]. The following is the LSTM formulation:

$$i_t^l = \sigma\left(W_{xi}^l x_t^l + W_{hi}^l h_{t-1}^l + w_{ci}^l c_{t-1}^l + b_i^l\right) \tag{6}$$

$$f_t^l = \sigma\left(W_{xf}^l x_t^l + W_{hf}^l h_{t-1}^l + w_{cf}^l c_{t-1}^l + b_f^l\right) \tag{7}$$

$$c_t^l = f_t^l . c_{t-1}^l + i_t^l . \tan h\,(W_{xc}^l x_t^l + W_{hc}^l h_{t-1}^l + b_c^l) \tag{8}$$

$$O_t^l = \sigma(W_{xo}^l x_t^l + W_{ho}^l h_{t-1}^l + W_{co}^l c_t^l + b_o^l) \tag{9}$$

$$r_t^l = o_t^l . \tan h\,(c_t^l) \tag{10}$$

$$h_t^l = W_p^l . r_t^l \tag{11}$$

l denotes the layer index, while $i_t^l, f_t^l\ and\ o_t^l$ denote the input, forget, and output gates, respectively. These are successively multiplied by the input, memory cell, and hidden output to gradually open or close their interconnections. x_t^l represents input from the $(l-1)^{th}$ layer, h_{t-1}^l represents the l^{th} output layer at time $t-1$, and c_{t-1}^l represents the internal cell state at time $t-1$. W_p^l is a projection matrix that is used to lower the dimension of r_t^l.

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

6.2 Proposed algorithm

Start

Input: Read the electronic health record of the patient as input variable → E, Pearson correlation → PC.

Perform data preprocessing (Data cleaning, Normalization) on → E

Perform feature extraction (Statistical, Information entropy) on→ preprocessed (E)

Perform feature selection → using PC

Split → PC (E) as → (A, B) // Train → A, Test → B

Trained model using → E (A)

Test the model using → E (B)

If disease diagnosis

Print disease detected as positive (1)

Else

Print disease detected as negative (0)

End for

Assign a weight to → E using → AHP

The inputs and the respective member functions µ1 determine the fuzzy system.

Assign the applicable fuzzy set to → input (E)

Define inference system based on→ Knowledge (database, rule base)

Ascertain heart disease risk state as µ1 (Low) µ2 (medium) µ3 High.

``Match member function to input → E

For defuzzification assign crisp data to the resulting fuzzy set

If health risk state = µ1

Print Risk level = low

Else If health risk rate µ2 = medium

Else

Print risk level = high

End for

End

6.3 Proposed Methodology

The proposed technique is laid out in flowchart form in Figure 4. Further, the proposed methodology is explained step by step.

ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

Step 1: Data collection

Initially, the Electronic Health Record (EHR) of the patients is taken as the input variable. It contains various attributes that are used for training and testing the model. These health records were further submitted for pre-processing.

Step 2: Data pre-processing

Step 2 involves data preprocessing, such as data cleaning and tokenization, once the EHR has been obtained as the input variable. This allows for the removal of noisy data and other irregularities from the raw dataset.

Step 3: Feature extraction

After the data pre-processing of EHR in this step, feature extraction is performed. In feature extraction, the irrelevant data is reduced while maintaining the same level of accuracy. It is performed based on statistical analysis and information entropy.

Step 4: Feature selection

In step 4, after the feature extraction process, the feature selection process is performed using Pearson correlation. In feature selection, the important parameters are selected from extracted information.

Step 5: Training and testing

After the feature selection, in this step training and testing of the ML classifier are performed. The relevant data that is obtained after feature selection is split into 2 parts such as train and test data. The further machine is trained using train data and after training the testing of the machine is performed.

Step 6 Disease diagnosis

In step 6, after training the model the testing of the model is performed if the disease is detected it shows 1 and if the disease is not detected it shows 0 after the disease diagnosis the next step is to classify the level of the risk. The fuzzy technique is used to classify the risk level of level.

Step 7: Weight Determination

In step 7, the weight determination is performed using the Analytic Hierarchy Process (AHP). It is utilized to generate the attribute weights that are then put to use in the process of initializing the crisp input variable for fuzzification.

Step 8: Fuzzy inference

After the fuzzification process, the fuzzy inference system is designed using a knowledge base and a rule base. To make a conclusive decision on the optimal navigation approach, the inference step requires an in-depth analysis of both the present and the potential future.

Step 9: Defuzzification

In the last step, after designing the inference system in this step the output is defuzzied to classify the result. At last, the result is classified as the risk level of the disease is low, medium, or high.

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

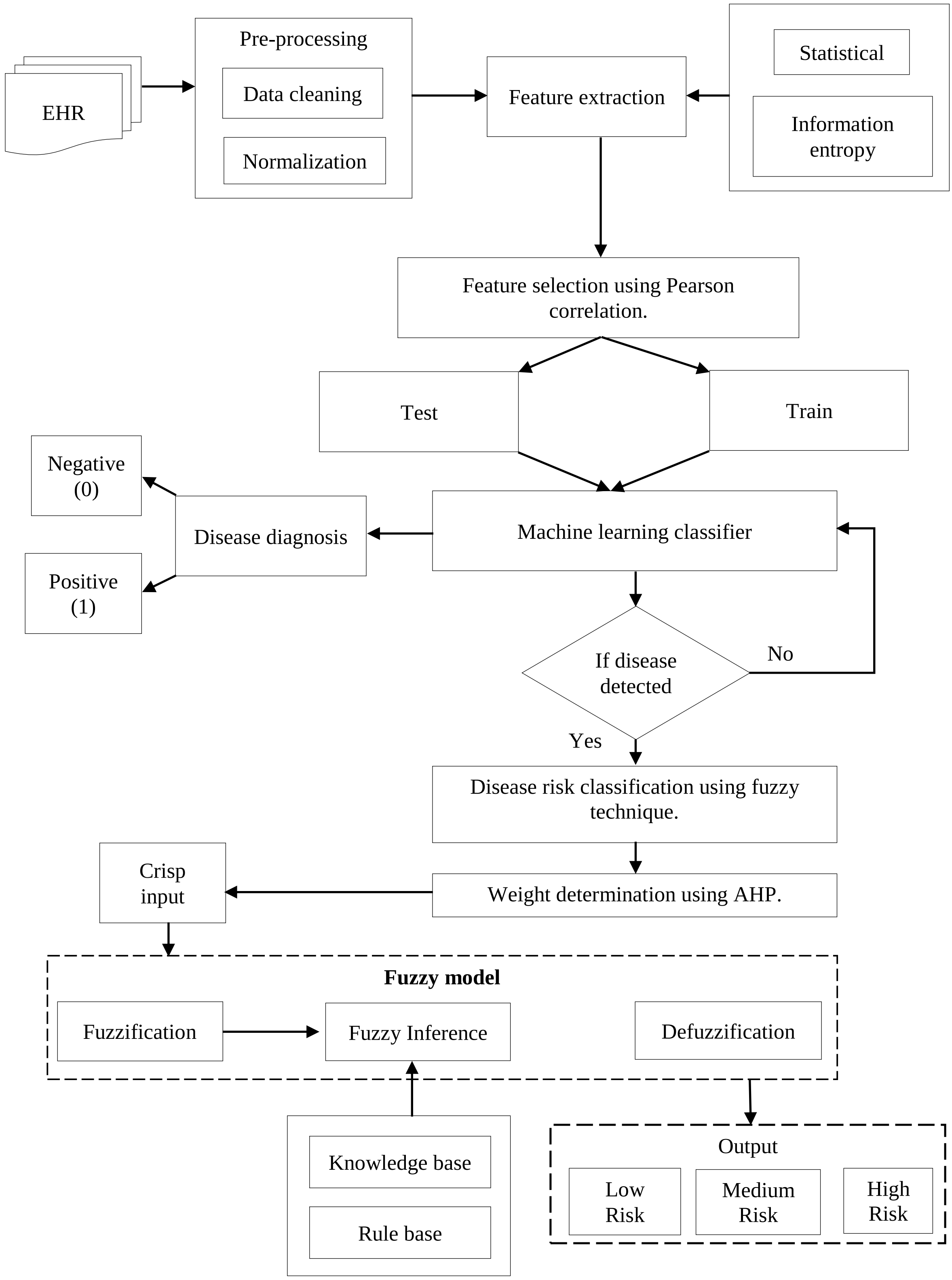


Figure 4: block diagram of the methodology

ISSN : 1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

7. RESULT AND DISCUSSION

In this section, a detailed description of the dataset and results are provided. In this analysis, the dataset that is used for testing and training is known as the Cleveland dataset and it is freely accessible on Kaggle. The performance measuring matrices of the proposed model is calculated at 400, 800, and 900 no. of epochs and then it is compared with other machine learning model.

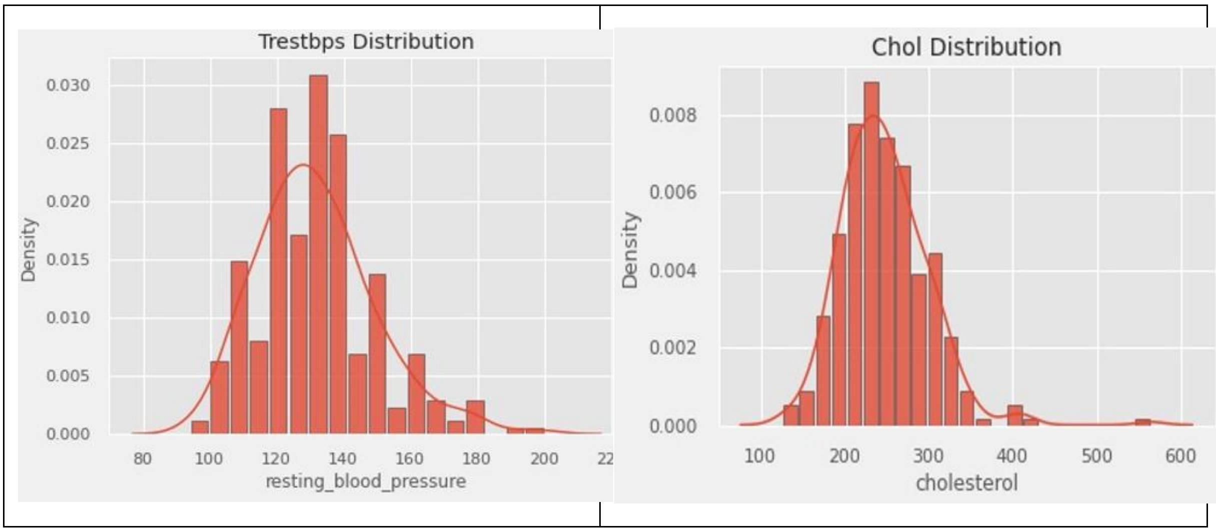
- **Dataset**

In this analysis, the dataset that is used is known as the Cleveland data set. This is an open-source dataset and freely accessible from the website of Kaggle. A total of 303 cases and 75 characteristics were considered during the data set's formation, but only 14 were used in any of the studies that were subsequently published. Six samples were omitted from the dataset as a result of the pre-processing efforts because of missing values. 297 samples are remaining, along with the 13 features and 1 output label from the original dataset. The dataset is described in great detail in Table 3.

Table 3: Detailed description

S. No.	age	sex	cp	trestbps	Chol	FBS	ECG	thalach	exang	old peak	slope	ca	condition
1	69	1	0	160	234	1	2	131	0	0.1	1	1	0
2	69	0	0	140	239	0	0	151	0	1.8	0	2	0
3	66	0	0	150	226	0	0	114	0	2.6	2	0	0
4	65	1	0	138	282	1	2	174	0	1.4	1	1	1
5	…	…	…	…	…	…	…	…	…	…	…	…	…
6	40	1	3	152	223	0	0	181	0	0	0	0	1
7	39	1	3	118	219	0	0	140	0	1.2	1	0	1
8	35	1	3	120	198	0	0	130	1	1.6	1	0	1
9	35	0	3	138	183	0	0	182	0	1.4	0	0	0
10	35	1	3	126	282	0	2	156	1	0	0	0	1

This data set's histogram, depicted in Figure 5, shows the distribution of frequencies over time for events with values that fall within a narrow range and are consistently spaced apart.



Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

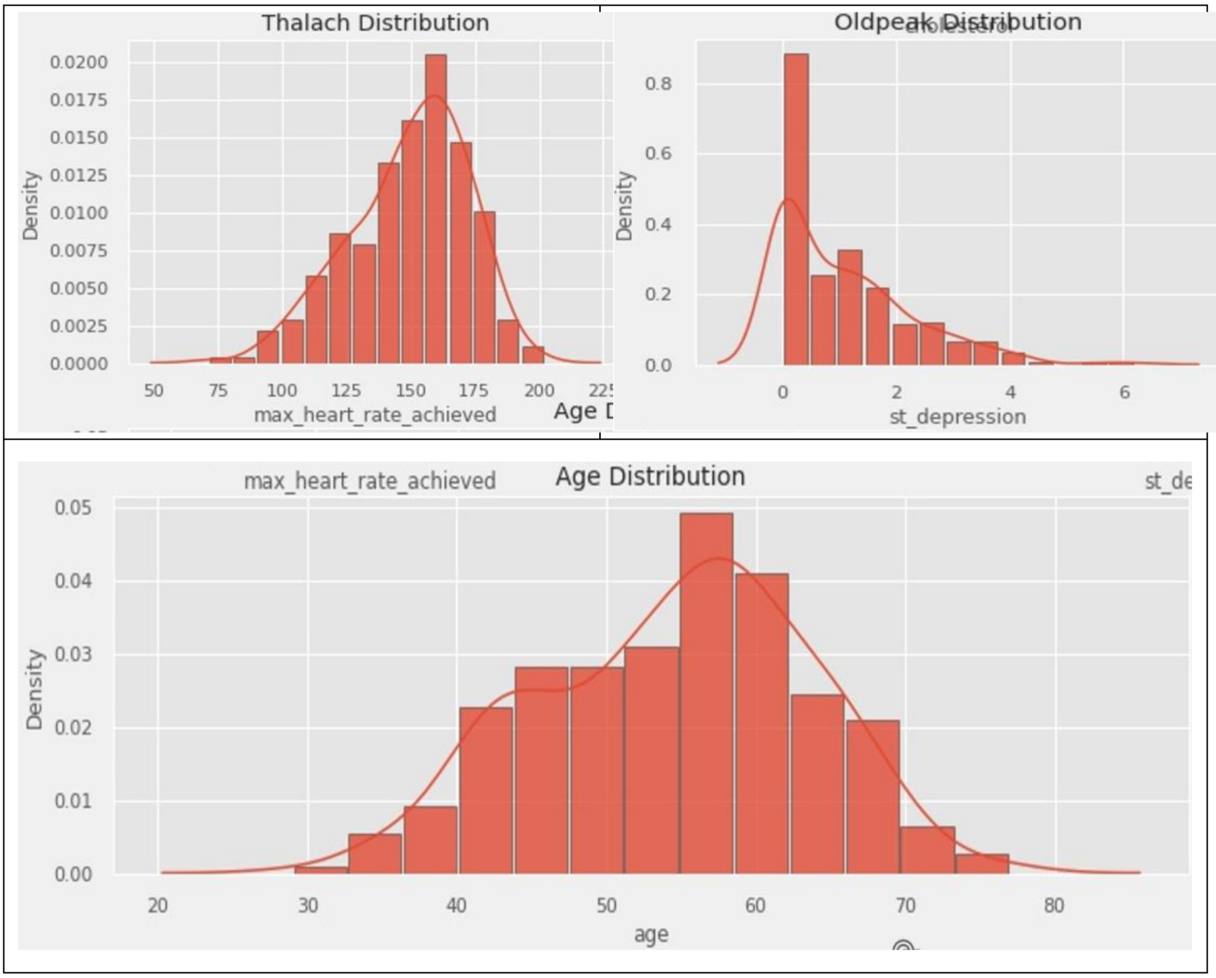


Figure 5: Histogram of heart disease dataset

Result 1:

A heat map depicting the relationships between the dataset characteristics is shown in Figure 6. The heat map is a graphical display of information in which different hues stand for different intensities. Every piece of data could be summarised in one easy-to-understand heat map. Better heat maps help the observer make sense of intricate data sets. It is also possible to see which combinations of categories contain more information by using a heatmap.

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

Figure 6: Correlation matrix Pearson method – categorical data

Result 2:

In this analysis, the Root Mean Square Error (RMSE) of the proposed model is calculated. The calculated value of LASSO cross-validation RMSE is 0.3398 as shown in figure 7. Lambda is a weight parameter, and the value of lambda lies between [0,1]. The RMSE of the proposed model can be calculated by the following formulae:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{k}(a_j - \hat{b}_j)^2}{n}} \quad (1)$$

Where, a_j is the actual value, and $\hat{b}_j$ is the predicted value and n is the total no. of values.

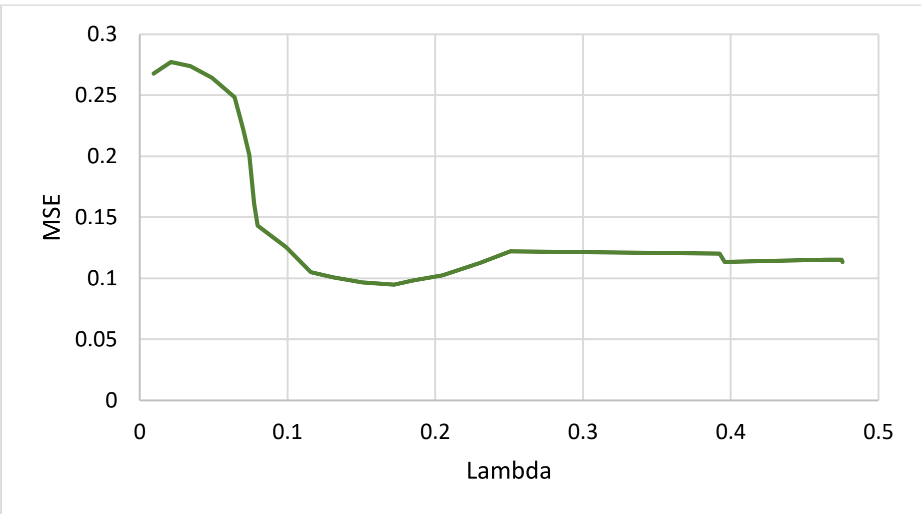


Figure 7: LAASO MSE graph

Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN : 1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

Result 3:

In this analysis, the sensitivity of the proposed model is calculated based on epochs such as 400, 800, and 900. It is seen in table 4 and figure 8 that as the no. of epochs of the dataset increases the sensitivity of the proposed model is also increased. The sensitivity of the proposed model can be calculated by the following formulae:

$$Sensitivity = \frac{true\ positive}{true\ positive + false\ negative} \tag{10}$$

Table 4: Sensitivity at different learning percentages

Parameter	No. of epochs		
	400	800	900
Sensitivity	89.01	89.25	89.99

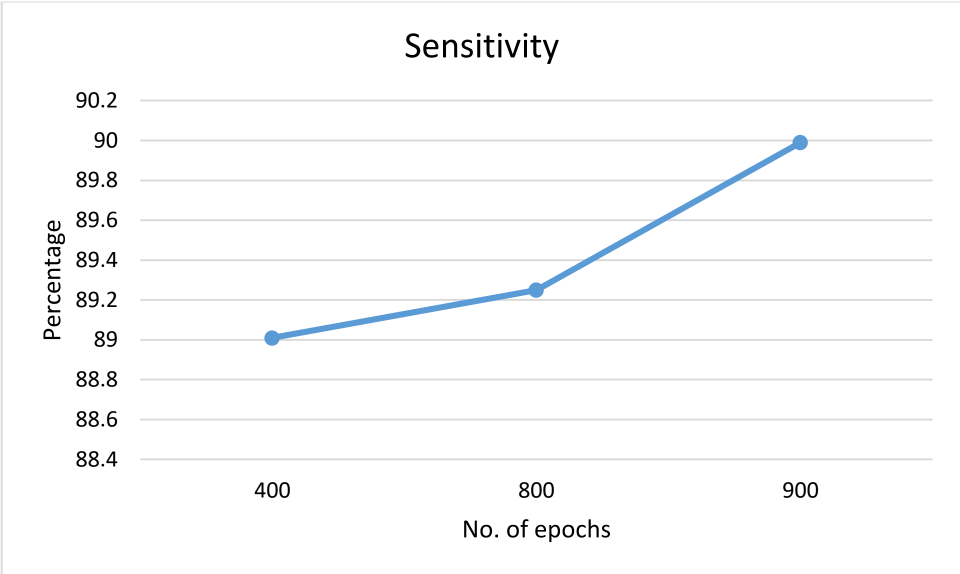


Figure 8: Graph showing the Sensitivity of the proposed model.

Result 4:

In this analysis, the specificity of the proposed model is calculated based on epochs such as 400, 800, and 900. It is seen in table 5 and figure 9 that as the no. of epochs of the dataset increases the specificity of the proposed model is also increased. The specificity of the proposed model can be calculated by the following formulae:

$$Specificity = \frac{Total\ negative}{total\ negative + false\ positive} \tag{11}$$

Table 5: Specificity at different learning percentages

Parameter	No. of Epochs		
	400	800	900
Specificity	90.05	90.31	92.05

ISSN : 1674-2974 | CN 43-1061 / N
DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

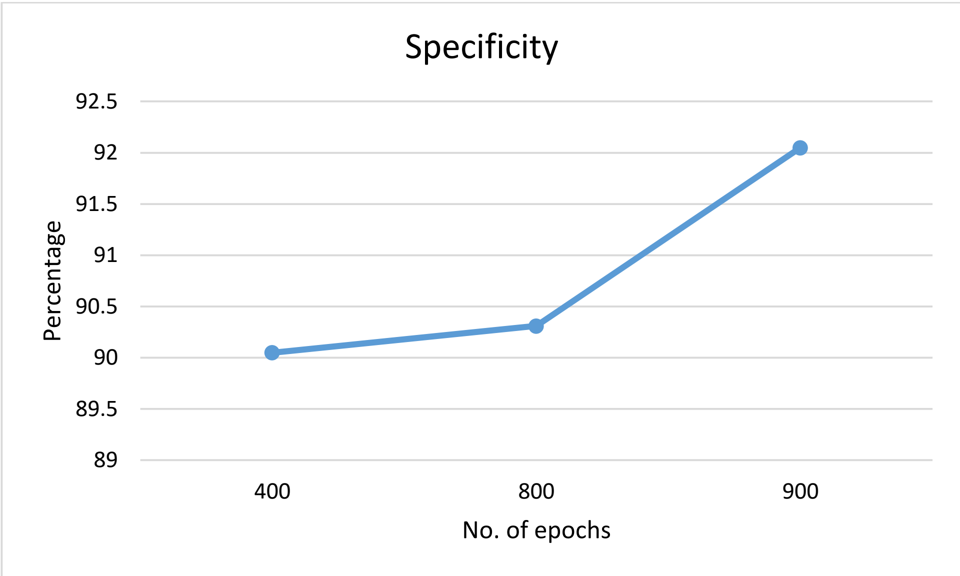


Figure 9: Graph showing the Sensitivity of the proposed model.

Result 5:

In this analysis, the accuracy of the proposed model is calculated based on epochs such as 400, 800, and 900. It is seen in table 6 and figure 10 that as the no. of epochs of the dataset increases the accuracy of the proposed model is also increased. The accuracy of the proposed model can be calculated by the following formulae:

$$Accuracy = \frac{true\ negative + true\ positive}{total\ no.of\ results} \quad (12)$$

Table 6: accuracy at different learning percentages

Parameter	No. of epochs		
	400	**800**	**900**
Accuracy	91.05	92.3	94.01

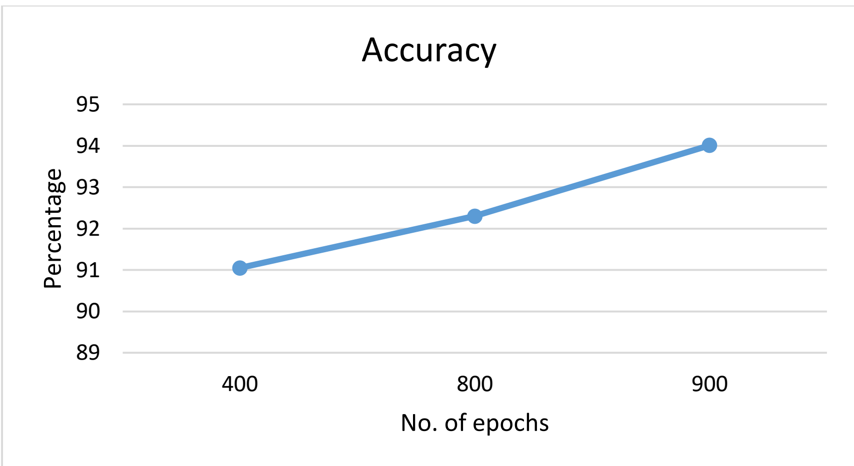


Figure 10: Graph showing the Accuracy of the proposed model.

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

Result 6:

In this analysis, the Mathews Correlation Coefficient (MCC) of the proposed model is calculated on various no. of epochs such as 400,800 and 900. It is seen in table 7 and figure 11 that as the no. of epochs of the dataset increases the MCC of the proposed model is also increased.

Table 7: MCC at different learning percentage

Parameter	No. of epochs		
	400	800	900
MCC	90.02	90.42	91.05

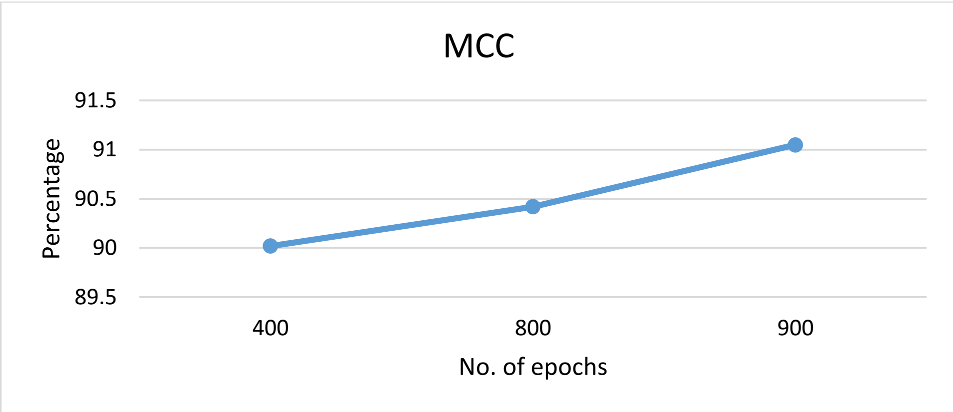


Figure 11: Graph showing MCC of the proposed model.

- **Comparative analysis**

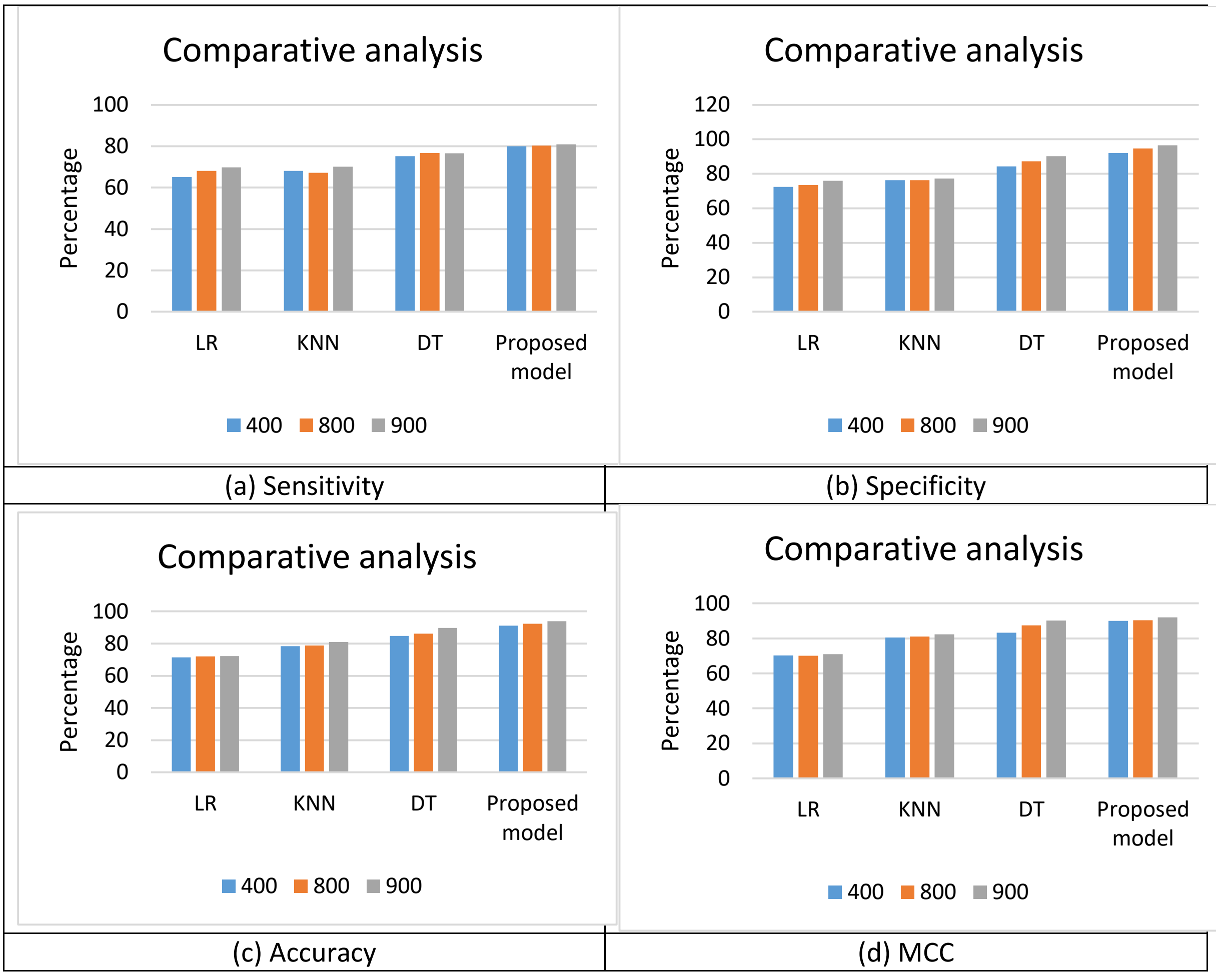
In this section, the proposed model is compared with other conventional methods such as LR, KNN, and DT by feature selected by Pearson correlation. It is compared based on positive metrics parameters such as sensitivity, specificity, accuracy, and MCC. Figure 12 (a) shows the comparison of the other machine learning technique with the proposed model based on sensitivity, and it is seen the sensitivity of the proposed model is higher among all the methods. Figure 12 (b) shows the comparison of the other machine learning technique with the proposed model based on specificity, and it is seen the specificity of the proposed model is higher among all the methods. Figure 12 (c) shows the comparison of the other machine learning technique with the proposed model based on accuracy, and it is seen that the accuracy of the proposed model is higher among all the methods. Figure 12 (d) shows the comparison of the other machine learning technique with the proposed model based on MCC, and it is seen the MCC of the proposed model is higher among all the methods. Table 8 shows the overall comparison of the proposed model and other machine-learning technique techniques in terms of sensitivity, specificity, accuracy, and MCC.

Table 8: Comparison table

Model	Sensitivity			Specificity			Accuracy			MCC		
	400	800	900	400	800	900	400	800	900	400	800	900
LR	65.12	68.07	69.83	72.51	73.59	75.93	71.5	72.05	72.25	70.25	69.99	71.02
KNN	68.15	67.17	70.05	76.25	76.36	77.26	78.41	78.87	81.01	80.55	81.01	82.25
DT	75.25	76.78	76.59	84.31	87.2	90.25	84.73	86.19	89.82	83.22	87.45	90.25
Proposed model	80.01	80.25	80.99	92.09	94.61	96.62	91.05	92.3	94.01	90.02	90.42	92.05

Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023



(a) Sensitivity | (b) Specificity

(c) Accuracy | (d) MCC

Figure 12: Comparison between models based on feature selection using PCA

To evaluate the efficacy of the suggested model (ANN+ LSTM+ fuzzy), it is compared to state-of-the-art methods currently used to diagnose heart disease. When compared to other existing methods, the suggested model outperformed them by an accuracy of 0.94. Table 9 and Figure 13 both visually depict the accuracy of the proposed technique and the present method, respectively.

Table 9: Comparison based on accuracy.

Method	Accuracy	References
NB based Diagnosis system	86.12	[31]
ANN diagnosis system	88.89	[32]
NNE	89.01	[33]
AFP	91.1	[34]
AWFSE	92.31	[35]
FCMIM-SVM	92.37	[26]
Proposed model (ANN+ LSTM+ Fuzzy) model	94.01	2023

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

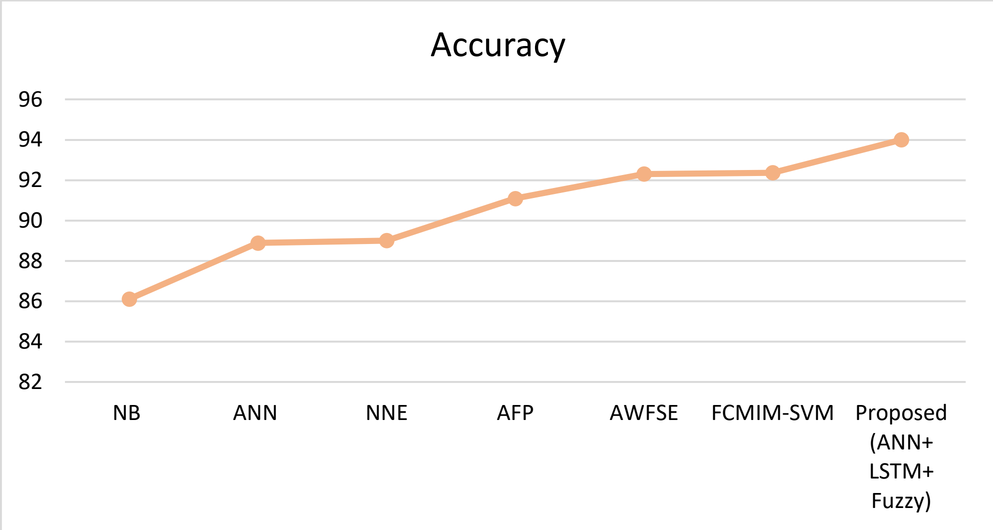


Figure 13: Comparison graph based on accuracy

8. CONCLUSION

CVD affects people worldwide, weakness, ankle swelling, and weariness are some types of CVD symptoms. It is difficult to detect CVD in an early stage. Hence, a better heart disease diagnosis system is needed. Risk assessment is essential for diagnosis and therapy. This novel paradigm for CVD identification uses ML and risk classification based on a weighted fuzzy approach. The system was built utilizing common feature selection techniques, including PCA and ML algorithms like ANN and LSTM. Furthermore, the cross-validation technique has been utilized to discover the most effective methods of model evaluation and hyperparameter tweaking. Classifiers' efficacies are evaluated with the use of accuracy-based performance indicators. In the end, the results showed that the suggested model had greater accuracy (94.01%) than a more traditional model designed for this area. Future research will likely make use of optimization strategies like Ant colony or Genetic algorithms to automatically produce an ideal set of hidden layer nodes and the related weighted linkages for the networks.

References

1) Bhatnagar, Prachi, Kremlin Wickramasinghe, Julianne Williams, Mike Rayner, and Nick Townsend. "The epidemiology of the cardiovascular disease in the UK 2014." *Heart* 101, no. 15 (2015): 1182-1189.

2) Maeder, Micha T., Otto D. Schoch, and Hans Rickli. "A clinical approach to obstructive sleep apnea as a risk factor for cardiovascular disease." *Vascular health and risk management* 12 (2016): 85.

3) Alugubelli, Raghunandan. "Exploratory Study of Artificial Intelligence in Healthcare." *International Journal of Innovations in Engineering Research and Technology* 3, no. 1 (2016): 1-10.

4) Long, Nguyen Cong, Phayung Meesad, and Herwig Unger. "A highly accurate firefly-based algorithm for heart disease prediction." *Expert Systems with Applications* 42, no. 21 (2015): 8221-8231.

5) Chitra, R., and V. Seenivasagam. "Heart disease prediction system using supervised learning classifier." *Bonfring International Journal of Software Engineering and Soft Computing* 3, no. 1 (2013): 01-07.

6) Jurado, Sergio, Àngela Nebot, Fransisco Mugica, and Narcís Avellana. "Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques." *Energy* 86 (2015): 276-291.

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

7) Rostami, Mehrdad, Kamal Berahmand, Elahe Nasiri, and Saman Forouzandeh. "Review of swarm intelligence-based feature selection methods." *Engineering Applications of Artificial Intelligence* 100 (2021): 104210.

8) Bouktif, Salah, Ali Fiaz, Ali Ouni, and Mohamed Adel Serhani. "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches." *Energies* 11, no. 7 (2018): 1636.

9) Thiyagaraj, M., and G. Suseendran. "Survey on heart disease prediction system based on data mining techniques." *Indian Journal of Innovations and Developments* 6, no. 1 (2017): 1-9.

10) Sola, Humberto Bustince, Javier Fernandez, Hani Hagras, Francisco Herrera, Miguel Pagola, and Edurne Barrenechea. "Interval type-2 fuzzy sets are a generalization of interval-valued fuzzy sets: Toward a wider view on their relationship." *IEEE Transactions on Fuzzy Systems* 23, no. 5 (2014): 1876-1882.

11) Mokeddem, Sid Ahmed. "A fuzzy classification model for myocardial infarction risk assessment." *Applied Intelligence* 48, no. 5 (2018): 1233-1250.

12) Burrough, Peter A. *Natural objects with indeterminate boundaries*. CRC Press, 2020.

13) Kora, Padmavathi, K. Meenakshi, K. Swaraja, A. Rajani, and Md Kafiul Islam. "Detection of cardiac arrhythmia using fuzzy logic." Informatics in Medicine Unlocked 17 (2019): 100257.

14) Dhanalakshmi, R., K. Anitha, D. Rukmani Devi, and T. Sethukarasi. "Association rule generation and classification with fuzzy influence rule based on information mass value." *Journal of Ambient Intelligence and Humanized Computing* 12, no. 6 (2021): 6613-6620.

15) Krishnaiah, V., G. Narsimha, and N. Subhash Chandra. "Heart disease prediction system using data mining techniques and intelligent fuzzy approach: a review." *International Journal of Computer Applications* 136, no. 2 (2016): 43-51.

16) Wójcik, Waldemar, Iryna Mezhiievska, Sergii V. Pavlov, Tomasz Lewandowski, Oleh V. Vlasenko, Valentyn Maslovskyi, Oleksandr Volosovych et al. "Medical Fuzzy-Expert System for Assessment of the Degree of Anatomical Lesion of Coronary Arteries." *International Journal of Environmental Research and Public Health* 20, no. 2 (2023): 979.

17) Taylan, Osman, Abdulaziz S. Alkabaa, Hanan S. Alqabbaa, Esra Pamukçu, and Víctor Leiva. "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy, and Statistical Methods." *Biology* 12, no. 1 (2023): 117.

18) Kharya, Shweta, Sunita Soni, and Tripti Swarnkar. "Fuzzy weighted Bayesian belief network: a medical knowledge-driven Bayesian model using fuzzy weighted rules." *International Journal of Information Technology* (2023): 1-9.

19) SESLİER, Tansu, and Mücella ÖZBAY KARAKUŞ. "In Healthcare Applications of Machine Learning Algorithms for Prediction of Heart Attacks." *Journal of Scientific Reports-A* 051: 358-370.

20) Karthick, K., S. K. Aruna, Ravi Samikannu, Ramya Kuppusamy, Yuvaraja Teekaraman, and Amruth Ramesh Thelkar. "Implementation of a heart disease risk prediction model using machine learning." Computational and Mathematical Methods in Medicine 2022 (2022).

21) Dong, Zheyi, Qian Wang, Yujing Ke, Weiguang Zhang, Quan Hong, Chao Liu, Xiaomin Liu, et al. "Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records." Journal of Translational Medicine 20, no. 1 (2022): 1-10.

22) Ahmed, Usama, Ghassan F. Issa, Muhammad Adnan Khan, Shabib Aftab, Muhammad Farhan Khan, Raed AT Said, Taher M. Ghazal, and Munir Ahmad. "Prediction of diabetes empowered with fused machine learning." IEEE Access 10 (2022): 8529-8538.

DOI: 10.5281/zenodo.10433960
Vol: 60 | Issue: 12 | 2023

23) Nadakinamani, Rajkumar Gangappa, A. Reyana, Sandeep Kautish, A. S. Vibith, Yogita Gupta, Sayed F. Abdelwahab, and Ali Wagdy Mohamed. "Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques." *Computational Intelligence and Neuroscience* 2022 (2022).

24) Cenitta, D., R. Vijaya Arjunan, and K. V. Prema. "Ischemic heart disease multiple imputation techniques using a machine learning algorithm." *Engineered Science* (2022).

25) YILMAZ, Rüstem, and Fatma Hilal YAĞIN. "Early detection of coronary heart disease based on machine learning methods." *Medical Records* 4, no. 1 (2022): 1-6.

26) Li, Jian Ping, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, and Abdus Saboor. "Heart disease identification method using machine learning classification in e-healthcare." *IEEE Access* 8 (2020): 107562-107582.

27) Kishore, A. H., and V. E. Jayanthi. "Neuro-fuzzy based medical decision support system for coronary artery disease diagnosis and risk level prediction." *Journal of Computational and Theoretical Nanoscience* 15, no. 3 (2018): 1027-1037.

28) Che Osmi, Siti Fatimah, Marlinda Abdul Malek, M. Yusoff, N. H. Azman, and W. M. Faizal. "Development of river water quality management using fuzzy techniques: a review." *International journal of river basin management* 14, no. 2 (2016): 243-254.

29) Reddy, G. Thippa, M. Praveen Kumar Reddy, Kuruva Lakshmanna, Dharmendra Singh Rajput, Rajesh Kaluri, and Gautam Srivastava. "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis." *Evolutionary Intelligence* 13 (2020): 185-196.

30) Sharma, Rohit, Sachin S. Kamble, Angappa Gunasekaran, Vikas Kumar, and Anil Kumar. "A systematic literature review on machine learning applications for sustainable agriculture supply chain performance." Computers & Operations Research 119 (2020): 104926.

31) Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." In 2008 IEEE/ACS international conference on computer systems and applications, pp. 108-115. IEEE, 2008.

32) Olaniyi, Ebenezer Obaloluwa, Oyebade Kayode Oyedotun, and Khashman Adnan. "Heart diseases diagnosis using neural networks arbitration." International Journal of Intelligent Systems and Applications 7, no. 12 (2015): 72.

33) Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36, no. 4 (2009): 7675-7680.

34) Samuel, Oluwarotimi Williams, Grace Mojisola Asogbon, Arun Kumar Sangaiah, Peng Fang, and Guanglin Li. "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction." Expert Systems with Applications 68 (2017): 163-172.

35) A. V. S. Kumar, ‘‘Diagnosis of heart disease using fuzzy resolution mechanism,’’ J. Artif. Intell., vol. 5, no. 1, pp. 47–55, Jan. 2012.