Journal of Computer Sciences

Original Research Paper

An Integrated Machine Learning Model for Heart Disease Classification and Categorization

[1]Srikanth Meda and [2]Raveendrababu Bhogapathi

[1]Department of Computer Science, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India
[2]Department of Computer Sciences, Sree Vidyanikethan Engineering College, Tirupati, Andhra Pradesh, India

Article history
Received: 29-07-2021
Revised: 11-09-2021
Accepted: 27-10-2021

Corresponding Author:
Srikanth Meda
Department of Computer Science, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India
Email: medasrikanth@gmail.com

Abstract: Progressions in the coordination among the machine learning algorithms, helped to achieve high accuracy and reliability in decision-making systems. Due to the impact and importance of heart diseases in real life, designing the efficient heart disease prediction model become a pivotal aspect today. Former research scholars applied the popular supervised machine learning models like Decision Trees, Naïve Bayes, ANN's, and FNN's to implement the heart disease prediction systems. As the heart disease prediction process is a multi-layered operation, each layer is expecting the optimal machine learning algorithm and the coordination among the algorithms of different layers to minimize the errors in prediction results. In this study, we proposed a new fuzzy neural-genetic algorithm to design an efficient and accurate heart disease prediction system. In our prediction system, we integrated the genetic algorithm, neural networks, and fuzzy logic technologies for training, classification, and categorization processes respectively. Experimental evaluations of Cleveland's heart disease dataset proved that the proposed fuzzy neural-genetic algorithm-based prediction system achieved high accuracy and low error rate when compared with the other machine learning models.

Keywords: Fuzzy Neural Genetic Algorithm, Heart Disease Prediction, Cleveland's Heart Disease Dataset, Machine Learning Classifiers

Introduction

Heart diseases are mainly developing in humans, due to the existence of various obstacles and abnormalities in the heart and blood vessels, which interrupts the blood pumping process negatively. Heart Attack, Coronary Heart Disease, Heart Arterial Disease, Congenital Heart Failure, and Rheumatoid Heart disease are some common and frequently occurred heart diseases. According to the statistics from the American Heart Association (AHA) (Salim and Virani, 2020), in 2017 total of 859,125 people died in the USA, due to various Cardio Vascular Diseases (CVD). A research program of the WHO revealed that 17.9 million global deaths happened in 2017 due to CVD and this value was 31% of the total global deaths in that year (WHO, 2021). The World Heart Federation (WHF) report (WHF, 2017) specified that in India among the noncommunicable disease mortalities, 60% of deaths happened in 2017, due to various CVDs only.

The above statistics on CVDs are alarming the need to focus on controlling the heart disease mortality rates. The early prediction of these heart diseases would help prevent the patients from hospitalization and mortalities. Heart disease prediction and severity values are evaluated based on the risk factors (Simons *et al*., 1998; Kannel, 2002) (i.e., bio-graphic data and medical data) associated with the humans like age, gender, chest pain, ECG, cholesterol levels, etc. By analyzing the relations, associations, and dependencies among these risk factors (attributes), heart diseases will be predicted and staged accurately. Machine learning algorithms (Alaa *et al*., 2019; Meda and Raveendra, 2017) are the most adaptable in this scenario for mining the relations and dependencies among the attributes and for predicting heart diseases.

Some former researchers (Shouman *et al*., 2011; Gupta *et al*., 2020; Javeed *et al*., 2019; Awan *et al*., 2018) designed the heart disease prediction models, which analyze the heart disease data sets. using the machine learning algorithms (Meda and Raveendra, 2017), to classify the heart diseases patient's data from the datasets. Shouman *et al*. (2011) conducted the experiments on heart disease datasets using various types of decision trees with Gini index, information gain, voting classifier, and error pruning techniques, to predict the heart diseases from

datasets. To enhance the disease prediction result accuracy, they integrated the two decision tree models voting classifier and discretization methods. Gupta *et al*. (2020) applied the six different machine learning classification models to the Cleveland heart disease dataset and observed that the Naïve Bayesian model recorded a high prediction accuracy than the other classification models. Backward elimination and Pearson correlation coefficient techniques were used with the Naïve Bayes to prioritize the high impact attributes and to downgrade the low impact attributes. This model reduced the burden of processing the low impact attributes in disease prediction, which improved the processing speed and result inaccuracy. Finally, the feature selection process is applied with all selected learning models to increase the prediction results accuracy by reducing the false negatives.

Javeed *et al*. (2019) integrated the random search algorithm and random forest tree methods to achieve high prediction accuracy. The grid search model they applied for feature selection and random forests generation is later used for the heart failure predictions. By integrating these two algorithms, the training data over-fit problem was addressed and the heart failure prediction accuracy was improved. Shahid *et al*. (2018) worked on different types of Artificial Neural Networks (ANNs) to predict the disease information from the Cleveland heart disease dataset. Simple Neural Networks, Multi-layer perceptron, Feed-forward neural networks, Recurrent Neural Networks, and the nonlinear autoregression are the types of neural networks they implemented in their research experiments. By applying the Principle Component Analysis (PCA) with ANNs, they achieved high prediction accuracy from the results.

Although many researchers (Shouman *et al*., 2011; Gupta *et al*., 2020; Javeed *et al*., 2019; Awan *et al*., 2018) were concentrated on designing the reliable heart disease prediction models, we noticed a few considerable limitations from them are: (1) Multi-layered disease prediction models are expecting the optimal algorithms at each layer, (2) None of the Individual algorithms recorded the high prediction accuracy (3) Disease categorization was not implemented and (4) false positive and false negative rate is high in results. Dey *et al*. (2016) conducted the two-phased experiments for heart disease predictions. In the first phase, they used the traditional machine learning algorithms for heart disease prediction and recorded the prediction result accuracy. In the second phase, they integrated the Principle Component Analysis (PCA) model with the traditional learning models, which returned a high accuracy than the traditional learning algorithms. Khan and Algarni (2020) integrated the Modified Slap Swarm Optimization (MSSO) based learning process with the advanced prediction model Adaptive Neuro-Fuzzy Inference System (ANFIS) to get the high accuracy in disease prediction. Khourdifi and Bahaj (2019), integrated the Particle Swarm Optimization (PSO) with the Ant Colony Optimization (ACO) algorithm to obtain high prediction accuracy using this integrated approach. From the above research works, we noticed that the integration of the optimal algorithms for disease prediction yielded better results than the individual algorithms.

To address the aforementioned limitations in the heart disease prediction process, in this study we proposed a new fuzzy neural-genetic algorithm, which helps in designing the accurate and reliable Heart Disease Prediction System (HDPS). For designing this new prediction model we integrated the three different machine learning technologies Genetic Algorithm, Artificial Neural Networks, and Fuzzy logic. At first, the Genetic Algorithm (GA) was appointed to train the ANNs with optimal input values in the training phase. Later the ANNs are appointed as a classification model, which classifies the input records as diseased and non-diseased, based on the supervised training knowledge obtained from the GA. Finally, fuzzy logic was used to categorize the disease range based on the ANN produced by each record-related disease prediction probability value. The major contributions and the advantages of the proposed fuzzy neural-genetic algorithm are:

1) Each processing layer of the prediction system is equipped with optimal algorithms
2) Various machine learning technologies were integrated and coordinated to achieve the high accuracy in predictions
3) Along with the disease prediction process, the disease range (severity) is also calculated
4) False-positive and false-negative rate from the predictions is reduced

To conduct the experiments using the proposed fuzzy neural-genetic algorithm, we selected the real-time dataset the Cleveland heart disease dataset. To implement the proposed fuzzy neural-genetic algorithm-based prediction system prototype, the python programming language libraries are used along with some external visualization tools. Experiments with the proposed prototype on the Cleveland dataset proved that the proposed fuzzy neural-genetic algorithm achieved a high accuracy than the other machine learning models.

Related Work

A. Heart Disease Dataset

In this study, the Cleveland heart disease dataset is used for conducting the experiments with the proposed fuzzy neural-genetic algorithm-based heart disease prediction model. Cleveland heart disease dataset contains a total of 303 health records with 75 attributes, which is a multivariate and popular heart disease dataset.

Among the 75 attributes, only the prominent 14 attributes (age, sex, cp, treetops, chol, FBS, resting, thalach, exang, old peak, slope, ca, thal and num) are chosen to participate in the heart disease prediction process. These 14 attributes are the considerable risk factors, whose values are helpful in prediction of the heart diseases. The values of these 14 attributes (presented in Table 1) are transformed as integers to make the prediction process smooth and feasible. Figure 1 is presenting some randomly collected records from the heart disease dataset before any preprocessing.

Dataset

B. Disease Prediction Algorithms

Machine learning modeled supervised classification algorithms are the adaptive algorithms for designing the heart disease prediction systems. Although various machine learning algorithms were used in designing the heart disease prediction systems, this section explores the popular classification models like Decision Trees, Naïve Bayes, Voting Classifiers, and ANNs in brief. As these algorithms were popular and most frequently used in disease prediction models, we selected them to perform the comparison with our prosed prediction model.

A decision tree (Shouman *et al*., 2011) is a supervised machine learning tree model, which is constructed with decision nodes and leaves at each level of the tree. Decision trees are mainly implemented as the classification trees and the regression trees. In the decision-making process, for the categorical (multivariate) data processing sake the binary classification trees are used, and for the continuous data values assessment sake, the regression trees are used. Decision trees will use the ID3, J48, CART, and C4.5 algorithms internally, which are useful for finding the multiple input attribute values and their relations to predict the outcomes.

At each attribute level, the entropy is calculated to determine adaptive attributes and finally, the decision tree is constructed with the hierarchy of conditions to predict the results. Now the constructed decision trees are applied against the test dataset to predict the result values. To calculate the entropy value for the multiple attributes, where the T is the current state, E is entropy, P is a probability, and X as the selected attribute is represented as:

$$E(T,X)=\sum_{c\in X}P(c)*E(c)$$

Naïve Bayes (Gupta *et al*., 2020) is another reliable supervised machine learning model, which controls the classification errors under high dependency among the dataset attributes. It contains a set of data classification models internally that are homogeneous in nature. The input dataset is separated into the feature matrix and response vectors. Input data element X, which is the part of class C_i, and its posterior probability $P(C_i|x)$ is calculated as follows:

$$P(C_i\mid x)=\frac{P(x\mid C_i)*P(C_i)}{p(x)}$$

Voting Classifiers (Latha and Jeeva, 2019) are recently gained attention in classifying the multivariate data by incorporating several machine learning algorithms with the voting classifiers. Instead of the single classifier-based predictions, it simply allows multiple classifiers to participate in predictions. Finally, all classifier's related results are aggregated and the result generation class will be defined based on the votes gained by that class. The hard voting model of the voting classifier follows the voting majority, whereas the soft voting model follows the voting average in determining the result generation (prediction) class.

Artificial neural networks (Awan *et al*., 2018) are designed from the inspiration of the human brain mapping model, which is a layered architecture that converts the input vectors into the output. The input layer, hidden layers, and the output layer are the three different layers in this model. The input layer receives the vector inputs, the hidden layer applies the nonlinear functions (neurons) for processing the input data, and finally, the output layer returns the results as output as shown in Fig. 2.

The errors at each hidden layer of the processing will be forwarded to the next hidden layer to correct them using the processing functions. To determine the proximity between the predicted values and the actual values, entropy functions are appointed with ANNs. Feedforward and backpropagation are the two different ANN architectures. Feedforward networks will forward the errors to the next layer and don't allow the input cycles, whereas the backpropagation allows the input cycling based on the error rates by fine-tuning the weights.

Table 1: Cleveland heart disease dataset with 14 attributes

Attrib name	Description
Age	Patient's age (int)
Sex	Male/Female (0,1)
Cp	Chest pain type (0,1,2,3)
Trestbps	Resting blood pressure
Chol	Serum cholesterol level in mg/dl
FBS	Fasting blood sugar (0,1 at >120mg/dl)
Resting	Resting ECG results (0,1,2)
Thalach	Maximum heart rate achieved
Exang	Exercise-induced angina
Old peak	ST depression observed in ECG
Slope	Slope of the ST segment during peak exercise
Ca	Number of major vessels colored by fluoroscopy (0,1,2,3,4)
Thal	Normal, fixed defect, reversible defect (0,1,2,3)
Target	Abnormal/normal values of the heart disease (1/0)

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
74	0	1	120	269	0	0	121	1	0.2	2	1	2	1
61	1	0	138	166	0	0	125	1	3.6	1	1	2	0
45	1	1	128	308	0	0	170	0	0.0	2	0	2	1
62	0	0	150	244	0	1	154	1	1.4	1	0	2	0
65	1	0	120	177	0	1	140	0	0.4	2	0	3	1
41	1	2	112	250	0	1	179	0	0.0	2	0	2	1
59	1	0	110	239	0	0	142	1	1.2	1	1	3	0
63	0	0	150	407	0	0	154	0	4.0	1	3	3	0
55	0	1	135	250	0	0	161	0	1.4	1	0	2	1
60	0	2	120	178	1	1	96	0	0.0	2	0	2	1

Fig. 1: Sample records from Cleveland heart disease

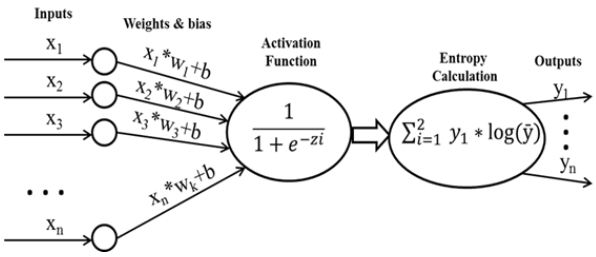

Fig. 2: Artificial neural networks processing model

Fuzzy Neural Networks with Genetic Algorithm

From the literature analysis on designing the Heart Disease Prediction Systems (HDPS), we noticed that the prediction systems are designed with multiple layers (modules) and each layer is dedicated to executing a part of the prediction process. Instead of using a single algorithm for the whole prediction process, the past research works (Javeed *et al*., 2019; Dey *et al*., 2016; Khan and Algarni, 2020; Khourdifi and Bahaj, 2019) are integrated multiple algorithms to attain more accuracy in prediction results. Inspired by the past integration models (Dey *et al*., 2016; Khan and Algarni, 2020), we proposed the heart disease prediction model (Fig. 3) with a fuzzy neural-genetic algorithm. In this, we integrated the Genetic Algorithm (GA), Artificial Neural Networks (ANNs) (Gad *et al*., 2018; Rivero *et al*., 2009), and Fuzzy Logic technologies to design a custom heart disease prediction model with different layers as shown in Fig. 3.

The proposed HDPS is equipped with a total of four different layers are: The data preprocessing layer, dataset training Layer, ANN processing layer, and the results classification layer.

This section explores the execution process happening at each layer in detail.

A. Data Preprocessing Layer

Data preprocessing is the primary activity of any data mining task, which assures the data quality to generate reliable results. To preprocess the dataset, the cleansing, transformation, and reduction techniques were applied using the optimal preprocessing models (Tang *et al*., 2020). As part of the data cleansing process, the missing values are imputed using the probable values, and the noisy data is eliminated using the clustering process. In the data transformation process, the data normalization techniques were applied and the attributes were grouped into the personal data and the health data. Finally, the attribute with character values is transformed into the numeric (integral) values. In the data reduction process, attribute subsets were selected, records with outliers are eliminated and the duplication records were deleted from the dataset. After this dataset preprocessing, a total of 287 records were left as perfect ones and the other 5 records were deleted due to failure in satisfying the preprocessing conditions.

B. Dataset Training Layer

After preprocessing, the dataset records are partitioned into the training dataset and the test dataset. By following the common guidelines of data partitioning, we randomly selected 70% or 200 records for training purposes and the leftover 30% or 87 records are assigned for the testing purpose. Now the machine learning models can be used to obtain the knowledge from the training data, which further classifies the test data. In our proposed model, we selected the ANN (Gad *et al*., 2018) as the primary algorithm for executing the main classification process. But ANNs are by default created as the forward pass mechanisms for updating network weights among the hidden layers that existed with neurons. As the backward pass was not implemented by default with ANNs, it needs to create more hidden layers for updating the network weights, which impacts negatively on ANNs processing speed and accuracy. To overcome this issue with the ANN's training process, we adopted the genetic algorithm to optimize the ANN training network weights (Rivero *et al*., 2009).

A genetic algorithm (Katoch *et al*., 2021) is an unsupervised machine learning algorithm, which provides the optimal solutions based on the genetic and natural selection methodologies. Genetic algorithm is selected only for training the ANNs, because the GA allows the variance in results, multi-dimensional solutions to cover the problem and they won't force the population with a direct optimization algorithm in training. GA generates the optimal solutions after the basic five phases (Katoch *et al*., 2021) of processing are: Initial population, fitness function, selection, cross-over, and mutation. At the beginning of GA, the input data is generated as a set of possible population vectors, by

exchanging the binary chromosomes with the random gene (weight) values. Population vectors are unique in nature and are generated to cover the problem in all possible dimensions. Each population vector-related fitness value is calculated by the fitness function and the prediction function. In this process, the vector weights are adjusted based on the comparison against the sample outputs. To enrich the population vectors with more possibilities, GA applies the cross-over technique, which performs the mating among the parent vectors to generate the new population called the off-springs. To prevent premature convergence and to maintain the diversity among the off-springs, a part of the vector is selected and their binary values are flipped is called the mutation process of the GA. Finally, the GA returns the optimal population as input vectors to ANN processing layer, which is appointed to execute the heart disease main classification job.

C. ANN Processing Layer

This is the final layer, which contains the two modules' classification and categorization. Instead of self-training, ANN adopted the evolutionary genetic algorithm to train the neural networks called Evolutionary Artificial Neural Networks (EANN's). Regardless of the type of ANN model (i.e., either feed-forward or backpropagation), GA can train an ANN model with its fine-grained input population. The modeling and modularization process of the proposed fuzzy neural-genetic algorithm is shown in Fig. 4, which explains the optimal technologies integration model and their modules arrangement.

Genetic Algorithm

After the input optimization process, GA sends the selected population to ANN for the training and classification process. ANN decodes the GA input vectors as population networks to train the ANN model. With the help of this GA-based training knowledge, the ANN performs the classification process on test data to extract the desired result. Using the genetic population networks, ANN begins the optimization process iteratively to find the best solutions. Finally, the best fitness value is identified from the obtained solutions and the best fitness value associated with input data weights are considered as the ideal weights for the output prediction. The Fitness score is the input network fitness value, which plays a vital role in classification and the categorization of the networks. After calculating all networks related fitness scores using the trained knowledge, ANN applies the classification techniques using the arg-max function. For classification of the test data input networks, the forward propagation method is used to compare the obtained fitness (probability) score against each possible output class and the output class which is the nearest to the fitness value is considered the relevant output class. In this way based on the probability of input networks the input data classifies whether the person is having the heart disease (diseased) or not (non-diseased).

Apart from this, our proposed model categorizes the disease severity among the 0 to 5 using the fuzzy membership functions (Al-Dmour *et al*., 2019). For this, we selected the fitness score and the ANN classifier predicted values as the inputs and selected the Mamdani fuzzy interference system (Çeven *et al*., 2020) for categorization. Using this model, initially, the fuzzy rules will be defined and the fuzzy membership function will evaluate the strength of the fuzzified inputs based on the fuzzy rules. Iteratively the consequent rules are identified and they are integrated finally to find the target output distribution. The proposed model will categorize the patient records based on the diseased information, fitness score, and attribute weights and values. The categorization falls into total 6 categories are: 0 (Not Diseased - No Risk), 1(Not Diseased - Early Warnings), 2(Not Diseased – Symptomatic), 3(Diseased – Mild Risk), 4(Diseased – Moderated Risk) and 5(Diseased – High Risk). In this way, the proposed fuzzy neural-genetic algorithm-based prediction model will classify and categorize the patient data, which is more useful in disease diagnosis and treatment.

D. Fuzzy Neural Genetic Algorithm

Along with the theoretical explanation using the layers and the activities of the proposed fuzzy neural-genetic algorithm, in this section, we presented the basic design prototype of the fuzzy neural-genetic algorithm using the sequential steps involved in it. After the routine data preprocessing and the separation (train and test data) process onwards, how the proposed algorithm is designed to predict the optimal solution is shown in algorithm-1.

Initially, the multivariate dataset Q contained x data records are transformed as the NN's input population vectors set $\{v_1, v_2, v_3 \ldots v_n\} \dot{\in}$ PV, which feeds the neural networks at an early stage. In the next step, the PV is passed as an input parameter to the procedure evalNN (), which annoys the neural networks model. The basic NN model with default configuration initiates the neural networks using the initiateNN () function and the weights $\{w_1, w_2, w_3 \ldots w_x\}$and bias $\{b_1, b_2, b_3 \ldots b_x\}$ values are randomly assigned to the population vectors. As part of the neural networks processing, each vector v_i related weighted sum *WS* (v_i) is calculated and passed through the activity functions (Nwankpa *et al*., 2018) like ReLu, Sigmoid, TanH, etc. These activation functions are part of hidden layers, which will help the neural networks to learn the complex non-linearity that existed among the input population. This learned knowledge plays a vital role in the training and classification process of NN.

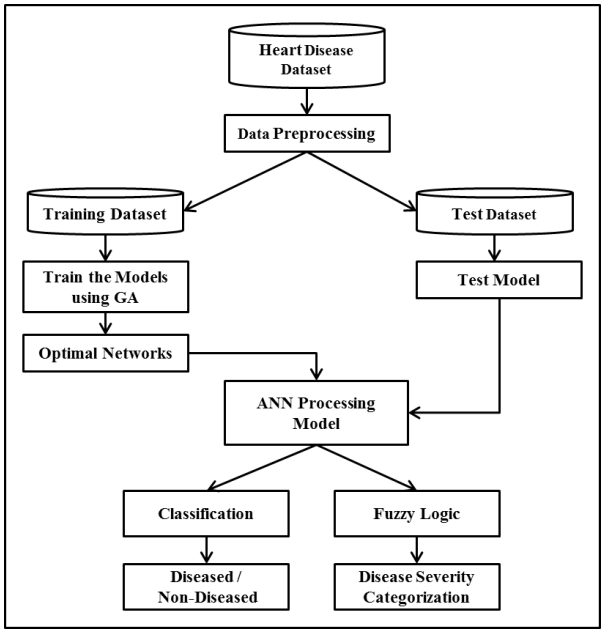

Fig. 3: Block diagram of the fuzzy neural-genetic algorithm in HDPS

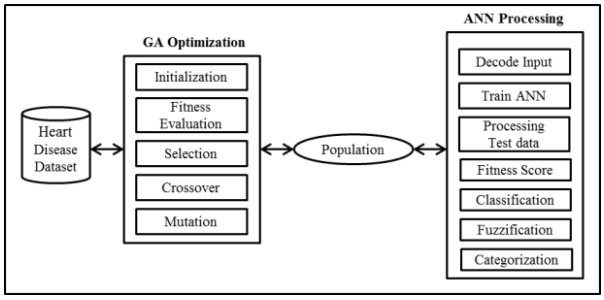

Fig. 4: Modeling and modularization of the fuzzy neural

The result (σ) of this active function compares against the threshold δ value and is passed to another hidden layer (activation function) as an updated population with the updated weights and bias values to obtain the improved results. The last hidden layer result will be forwarded to the output (predict) function to evaluate the fitness probability value. In our case, the prediction is a binary (disease or nondiseased) classification model, which can be implemented using the Sigmoid function (Nwankpa *et al*., 2018) and the result is $\sigma(z_i)$. To evaluate the prediction compliance with current weights and bias values the loss function is implemented which reveals the distance between the prediction value $\bar{y}$ and the actual value y. We selected the cross-entropy CE-based loss function to assess the performance of our binary classification model. At the end of NN processing, the fitness score of $\sigma(z_i)$ and the entropy value CE is compared against the expectations to determine the optimal solution ß among the generations.

In case, if the generated solution by NN is not up to mark, then the Genetic Algorithm (Katoch *et al*., 2021) prepares the next generation with new optimal population vectors $P\tilde{V}$ using the selection, crossover, and mutation operations. This input optimization and NN execution process will continue until the best solution ß is found. After the best solution is identified for the efficient classification of test data, the fuzzy fitness score $f(v_i)$ of each input vector $v_i \dot{\in}$ PV is evaluated and the range values $R(v_i)$ of each input vector is calculated based on the categorization rule set $\mathbb{R}$. After the defuzzification $\Delta f(v_i)$process, the obtained classification results with the best solution ß and the categorization results $R(v)$ will be presented.

Algorithm-1: Fuzzy neural genetic algorithm

Input: Training Data Set Q
Output: Optimal solution ß and rangeVector R(v)
Start
popSize :=x
population_vectors (PV) := initial_population(Q(x))
*procedure:*evalNN(population_vectorsPV)
nn := InitiateNN (PV)
nn := randomWeightsAndBias(nn)
forEach vector v_i in PV
$WS(v_i) = v_i * w_i + b_i$
σ = active_fun(WS(v_i))
if σ > δ than pass to predict()
else update weights and computeWS(v_i)
z_i = |WS(v_i) |

$$\sigma(z_i) = \frac{1}{1+e^{-zi}}$$

$$CE = -\sum_{i=1}^{2} y_1 * \log(\bar{y})$$

end
iffitness(σ(z_i) ± b) < y && loss (CE_i< avg (CE)) then
pass the σ(z_i) as optimal solution ß & break()
else
// GA Process
$\dot{S}$ = selection (PV)
$\acute{C}$ = crossover ($\dot{S}$)
$\acute{M}$ = mutation($\acute{C}$)
$P\tilde{V}$ = next generation($\acute{M}$)
evalNN($P\tilde{V}$)
// fuzzy process
foreach vector v_i in PV
f(v_i) =fuzzify (fitval (σ(z_i)))
R(v_i)=rangeCalc (f(k), $\mathbb{R}$)
Δf(v_i)= defuzzify (f(v_i))
print ß and R(v)
end
Stop

DOI: 10.3844/jcssp.2022.257.265

Experimental Analysis

In order to prove the accuracy and efficiency of the proposed fuzzy neural-genetic algorithm-based heart disease prediction systems, we implemented the proposed HDPS using the python libraries (van Rossum, 2018) and selected the Cleveland heart disease dataset (explained in section-2A) for experiments.

A windows7 based computer system with hardware 8GB RAM, 1TB hard disk, and i7 inter-processor is selected for the experiments. To implement, run and visualize the proposed HDPS, the Jupyter Notebook server (Randles *et al*., 2017) 6.1.3 with python 3.8.1 combination is chosen, as this collaboration is extremely compliant for the scientific application development process. To program the proposed system using python, many python machine learning libraries (Stančin and Jović, 2019) (i.e., pandas, numpy, scikit, pygad, etc.) are widely used along with our custom code.

The heart disease prediction experiments were planned to conduct with the proposed fuzzy neural-genetic algorithm and with the other popular machine learning models like Decision Trees (Shouman *et al*., 2011), Gupta *et al*. (2020), voting classifiers (Javeed *et al*., 2019) and ANN (Awan *et al*., 2018). Prediction accuracy, precision, recall, f score, and confusion matrix are selected as the metrics to measure the performance of each learning model implemented in experiments. Finally, the results of each metric regarding all learning models are compared to find the best learning model among the ones who participated in the experiments.

For this, the total data records are classified into the training dataset (70%) and the testing dataset (30%), where the training set is used to feed the learning models for training, and the test set is used to perform the predictions based on training knowledge. Finally, the predicted results of the learning model are compared against the actual values using various metrics to assess the performance of the learning model. In this way, we conducted the experiments on Cleveland dataset records using the selected learning models and the results were presented in Table 2.

The above Table 2 is presenting the results from various heart disease prediction models are proves that our proposed Fuzzy Neural Genetic Algorithm (FNN and GA) recorded the high-performance values in all dimensions like accuracy, precision, recall, and f_score. Apart from this, the four-fold confusion matrix is used to describe, how much the proposed FNN&GA model reduced the false positives and false negatives in disease predictions as shown in Table 3.

HDPS

Each learning model-related confusion matrix values True Positives (TP), False Negatives (FP), False Positives (FP), and True Negatives (TN) will describe the prediction errors at the input records level. Table 3 presents confusion matrix comparison data specifying that the proposed FNN and GA model reduced the number of false negatives and the false positives dramatically when compared to its counterpart learning models.

Genetic Algorithm

Along with the classification process, our proposed model will do the disease severity categorization based on the calculated disease probability values, using the fuzzy logic technology (Al-Dmour *et al*., 2019) as shown in the piece of categorization results in Table 4. The prediction probability value, disease confirmation (0 or 1), disease range (0-5), and disease severity values are categorized using our model and this information will be very useful for doctors in treating the patients based on their disease severity. A total of 87 test records are categorized similarly and each patient record disease severity range (0-5) is presented in Fig. 5.

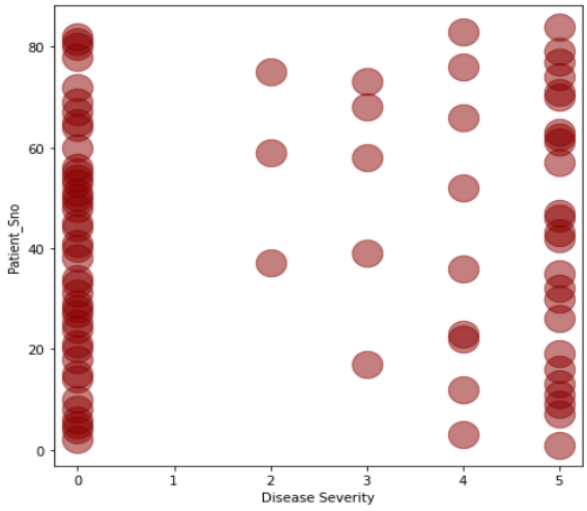

Fig. 5: Total test dataset severity categorization result

Table 2: Disease prediction results comparison of the HDPS

Prediction model	Accuracy	Precision	Recall	F_Score
Decision tree	80.45%	79.54%	81.39%	80.45%
Naïve bayes	83.90%	84.31%	87.75%	86%
Voting classifier	77.01%	71.69%	93.02%	79.16%
ANN	83.90%	83.33%	78.95%	81.08%
FNN and GA	96.55%	95.74%	97.83%	96.77%

Table 3: Comparison of the confusion matrix values of the

Prediction Model	TP	FN	FP	TN
Decision Tree	35	9	8	35
Naïve Bayes	30	8	6	43
Voting Classifier	29	15	5	38
ANN	30	8	6	43
FNN and GA	45	2	1	39

Table 4: Heart disease categorization using fuzzy neural

Prediction Value	Diseased (0)/ Non-Diseased (1)	Disease Stage	Disease-Severity
1.514	0	5	Diseased-High Risk
0.942	0	4	Diseased Moderate Risk
0.465	1	2	Non-Diseased-Symptomatic
-1.156	1	0	Non-Diseased-No Risk
0.623	0	3	Diseased
-1.617	1	1	Non-Diseased-Early Warnings
-1.779	1	0	Non-Diseased-No Risk
1.980	0	5	Diseased-High Risk
0.618	0	3	Diseased
-0.764	1	0	Non-Diseased-No Risk
1.494	0	5	Diseased-High Risk

Conclusion and Future Work

In this study, we proposed the fuzzy neural-genetic algorithm-based heart disease prediction model to improvise the prediction accuracy and efficiency. For this, the former disease prediction systems are thoroughly analyzed and considerable limitations are identified. At each layer of the disease prediction system, an efficient and compliance algorithm was appointed to get the best output from that layer. In this way, we integrated the neural networks, genetic algorithm, and fuzzy logic technologies to fulfill the objectives of this study. A genetic algorithm is used to train the neural networks with optimal input values and the neural networks are used for disease classification based on the training knowledge obtained from the genetic input population. Finally, fuzzy logic is used to categorize the disease severity based on prediction probability scores. Experimental results proved that the proposed fuzzy neural-genetic algorithm achieved high accuracy and low error rate (false positives and false negatives), when compared to the other machine learning models.

Although the proposed HDPS model is offering both classification and categorization operations, in the future we are planning to implement the advancements in the categorization process. In the future we are expecting to consider the input record level attribute values along with the prediction probability values, to make the categorization process more robust and reliable.

Author's Contributions

Srikanth Meda: Design and implementation of all the algorithms and result analysis and comparative analysis.

Raveendrababu Bhogapathi: Made significant contributions in the implementation of algorithms and experiment analysis

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues are involved.

References

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., & Van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PloS one, 14(5), e0213653. doi.org/10.1371/journal.pone.0213653

Al-Dmour, J. A., Sagahyroon, A., Al-Ali, A. R., & Abusnana, S. (2019). A fuzzy logic–based warning system for patient's classification. Health informatics journal, 25(3), 1004-1024. doi.org 10.1177/1460 458217735674

Awan, S., Riaz, M., & Khan, A. (2018). "Prediction Of Heart Disease Using Artificial Neural Network", VFAST Transactions on Software Engineering, Volume 13, pp 102-112, Sep-Dec, 2018.

Çeven, S., Albayrak, A., & Bayır, R. (2020). Real-time range estimation in electric vehicles using fuzzy logic classifier. Computers & Electrical Engineering, 83, 106577. doi.org/10.1016/j.compeleceng.2020.106577

Dey, A., Singh, J., & Singh, N. (2016). Analysis of supervised machine learning algorithms for heart disease prediction with reduced number of attributes using principal component analysis. International Journal of Computer Applications, 140(2), 27-31.

Gad, A. F., Gad, A. F., & John, S. (2018). Practical computer vision applications using deep learning with CNNs. New York, NY, USA: Apress. doi.org link.springer.com/book/10.1007/978-1-4842-4167-7?noAccess=true

Gupta, A., Kumar, L., Jain, R., & Nagrath, P. (2020). Heart disease prediction using classification (naive bayes). In Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019) (pp. 561-573). Springer, Singapore. doi.org/10.1007/978-981-15-3369-3_42

Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., & Nour, R. (2019). An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access, 7, 180235-180243. 180235-180243, 2019. doi.org/10.1109/ACCESS.2019.2952107

Kannel, W. B. (2002). Coronary heart disease risk factors in the elderly. The American journal of geriatric cardiology, 11(2), 101-107. doi.org/10.1111/j.1076-7460.2002.00995.x

Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: Past, present, and future. Multimedia Tools and Applications, 80(5), 8091-8126. doi.org/10.1007/s11042-020-10139-6

Khan, M. A., & Algarni, F. (2020). A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. IEEE Access, 8, 122259-122269. 122259-122269, 2020, doi.org/10.1109/ACCESS.2020.3006424

Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. International Journal of Intelligent Engineering and Systems, 12(1), 242-252. doi.org10.22266/ijies2019.0228.24

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203. doi.org/10.1016/j.imu.2019.100203

Meda, S., & Raveendra, B. B., "Identification and Predicting Heart Disease with Data Mining methods – A Survey", International Journal of Institutional & Industrial Research ISSN: 2456-1274, Vol. 2, Issue 2, May-August 2017, pp.11-16 .

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378. doi.org/10.48550/arXiv.1811.03378

Randles, B. M., Pasquetto, I. V., Golshan, M. S., & Borgman, C. L. (2017, June). Using the Jupyter notebook as a tool for open science: An empirical study. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 1-2). IEEE. doi.org/10.1109/JCDL.2017.7991618

Rivero, D., Dorado, J., Fernández-Blanco, E., & Pazos, A. (2009, June). A genetic algorithm for ANN design, training and simplification. In International Work-Conference on Artificial Neural Networks (pp. 391-398). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-02478-8_49

Salim, S., & Virani, A. A. (2020). Heart Disease and Stroke Statistical Update Fact Sheet At-a-Glance", American Heart Association Research. doi.org/10.1161/CIR.000000 0000000757

Shouman, M., Turner, T., & Stocker, R. (2011). "Using Decision Tree for Diagnosing Heart Disease Patients", Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, Volume 121, December 2011, Pages 23-30.

Simons, L. A., McCallum, J., Friedlander, Y., & Simons, J. (1998). Risk factors for ischemic stroke: Dubbo Study of the elderly. Stroke, 29(7), 1341-1346. doi.org/10.1161/01.STR.29.7.1341

Stančin, I., & Jović, A. (2019, May). An overview and comparison of free Python libraries for data mining and big data analysis. In 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 977-982). IEEE. doi.org/abstract/document/8757088

Tang, S., Yuan, S., & Zhu, Y. (2020). Data preprocessing techniques in convolutional neural network based on fault diagnosis towards rotating machinery. IEEE Access, 8, 149487-149496. doi.orgieeexplore.ieee.org/abstract/document/9149875

van Rossum, G. (2018). The Python development team, "Python Tutorial-Version 3.7.0", by Python Software Foundation, September 02, 2018. https://bugs.python.org/file47781/Tutorial_EDIT.pdf

WHF, (2017). "Fact sheet: Cardiovascular diseases in India". World Heart Federation. https://www.world-heart-federation.org/wpcontent/uploads/2017/05/Cardiovascular_diseases_in_Indi a .pdf

WHO, (2021). "The Key Facts of Cardiovascular diseases (CVDs)". World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)