Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 32s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

A Data-Driven Ensemble Learning Model For Heart Disease Prediction Using Feature Representation And Classification

Nithya Shree A. P [1], Dr. R. Kannan [2]

[1]Research Scholar, Department of Computer Science, Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Bharathiar University, Coimbatore, Tamilnadu (St), India,
Email ID : navnaghul2013@gmail.com
[2]Associate Professor, Department of Computer science, Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Bharathiar University, Coimbatore, Tamilnadu (St), India,
Email ID : ramadosskannan@gmail.com

Cite this paper as: Nithya Shree A. P , Dr. R. Kannan , (2025) A Data-Driven Ensemble Learning Model For Heart Disease Prediction Using Feature Representation And Classification. *Journal of Neonatal Surgery*, 14 (32s), 5301-5311.

ABSTRACT
Heart disease is one of the most common causes of death in the general population. The prognosis of patients with heart conditions is greatly impacted by early detection. Several recognized factors can contribute to life-threatening cardiac problems, such as Age, sex, heart rate, cholesterol, and sugar. However, an expert may find it challenging to assess each patient while considering these factors due to the large number of variables. The work suggest assessing patients' risk of cardiovascular disease by combining Machine Learning (ML) and Deep Learning (DL) with feature augmentation techniques to form an ensemble model. The DenseNet, Gated Network Model (GNM) and Multi-Layer Perceptron's (MLP) are combined to form the ensemble model. The results of the proposed methods demonstrate a significant improvement, particularly for a condition that impacts a large population, surpassing previous methods by 4.4% and achieving a 95.89% accuracy rate

Keywords: *heart disease, prediction, machine learning, feature representation, accuracy*

1. INTRODUCTION

Worldwide, the leading cause of illness and mortality is cardiovascular disease (CVD). The term "cardiovascular disease" describes any condition that impacts the heart, blood vessels, or the body's capacity to pump and circulate blood [1]. Heart disease is a prevalent condition that has taken many lives. This is because it influences the function of the heart and can cause death. Heart disease, considered to be one of the primary causes of death, has become more prevalent in the population of the world during the past few decades [2]. Every year, heart disease kills over 17.7 million people. A specialist physician diagnoses heart illness, which is necessary for effective treatment. This drawback is that diagnoses are susceptible to human error and may not be objective. Because heart failure has been the focus of a great deal of research because of its intricate diagnosis procedure, a computer-aided decision support system, such as the one described in [3], is highly beneficial. The use of data mining tools has shortened the time required to make an accurate disease forecast. Particularly in underdeveloped nations, heart problems can result in a variety of consequences that can lower the quality of life or possibly cause death. Additionally, underdeveloped nations and those with less developed healthcare systems had greater rates of heart failure-related mortality [4]. This highlights the importance of creating a method to guarantee an accurate and prompt evaluation of a patient's risk for heart failure. Because of these factors, several writers have developed methods that, by considering various parameters, aid in identifying cardiac disease. Most of these methods employ machine learning methods to get around issues caused by statistical analytic methods that lose predictive information in sizable datasets that interact in multiple dimensions [5]. Large datasets that allow the identification of present ailments using historical data spanning a significant amount of time have often been beneficial to many of these investigations. However, until a few years ago, the findings concentrated more on identifying heart abnormalities without delving into whether they were potentially dangerous or harmless [6] . The researchers used a boosted decision tree technique to look for correlations between death and patient characteristics. Suppose a patient passed away shortly after being admitted to the hospital, and it was determined if they had a high or low probability of passing away. Taking into account the recommended labeling, the findings revealed a 0.88 area under the curve. Author et al. [7] introduced several techniques, like support vector machines (SVM), k-nearest neighbor (kNN), neural networks, decision trees, stochastic gradient descent (SGD), and the combined nomenclature (CN2) rule inducer. As a result, they could predict heart disease with an accuracy of up to 87.69%. Although validations have only been conducted on a few people, the article suggests a technique to diagnose heart disease satisfactorily [8]. In a recent study, the author looked for the most accurate machine learning classifiers for various diagnostic applications. A number of supervised machine learning algorithms were evaluated for their accuracy and efficacy in predicting cardiac disease. Making use of Decision Trees, Random Forest, and kNN, the results demonstrated 100% accuracy; however, they only displayed the best outcome obtained following a cross-validation procedure, which is not definitive [9]..

Since traditional machine-learning models demonstrated encouraging outcomes for this issue, several other writers have experimented with different machine-learning techniques. For example, a variety of machine learning techniques were employed to estimate the 30-day mortality risk for heart failure patients following their discharge. Multiple-variate adaptive regression splines, bootstrap regression tree aggregation, random forest classification, the arithmetic mean, main-terms Bayesian logistic regression utilizing a Cauchy prior, extended boosted regression, logistic regression of the primary terms with and without variable selection, and regression trees was utilized to determine each patient's marginal probability of death [10]. According to the results, ensemble models outperformed the benchmark models. To anticipate and identify cardiovascular disease recurrence, the researchers employed five classifier model techniques: support vector machines, artificial neural networks, Naive Bayes, regression analysis, and random forests. From the UCI collection, we used the Cleveland and Hungarian datasets. The random forest approach produced the best results (98.12% accuracy), demonstrating that ensemble algorithms outperformed individual methods. The use of ensemble algorithms has shown strong efficacy in diagnosing diseases such as diabetic retinopathy, hepatitis C and breast cancer in addition to heart failure [11]. Recently, neural network-based deep learning techniques have also been used to address medical conditions like the risk of heart failure. Early identification of heart failure risk using Convolutional Neural Networks (CNNs) and basic electrocardiograms (ECGs) [12]. The AUC for CNN was 0.78. The risk of cardiac failure was also predicted using adaptive multi-layer networks. Compared to regular neural networks, these networks fared better than hybrid and ensemble approaches in earlier years. The sample size was limited since only 297 patients were assessed because the Cleveland dataset was employed in this investigation [13]. Several well-known datasets, including Hungary (294 observations), Stalog (270 observations), Long Beach, Virginia (200 observations), Cleveland (303 observations), and Switzerland (123 observations), were combined to produce a new dataset last year [14]. This made it possible to train new methods that used comparatively few features to categorize this vast number of samples. Finding techniques for classifying cardiac issues that enable us to have a high success rate in early disease identification is the primary goal of our research [15]. Considering the assessed dataset, two secondary objectives have been accomplished:

Developing a novel ensemble approach to classification problems using a small number of features which combines DenseNet, Gated Network Model (GNM) and Multi-layer Perceptron's (MLP) for heart disease prediction;

Using ensemble neural network properties to enhance existing feature augmentation methods.

To accomplish these, an ensemble-based architecture are presented in this study for the data processing and analysis that follows. The architecture attains up to 98% precision, which is a noteworthy development and a massive aid in assessing the risk of cardiac conditions.

The rest of the study is displayed as follows: A comprehensive review of earlier studies to predict heart disease is provided in Section 2. The suggested ensemble model for feature analysis and prediction is highlighted in the methodology section. The experimental analysis and numerical findings are displayed in Section 4. Under Section 5 the job summary is presented.

2. Related works

The prediction of heart disease (HD) and its associated problems is a complex topic, and using machine learning approaches to address it is becoming more and more common these days. The need for an effective HD prediction system has grown due to a shortage of doctors and incorrect or delayed diagnoses. Hospital administration uses automated methods to handle the massive volume of data created daily [16]. The data is represented in charts, photographs, and sequence numbers to track and assess the patient's health information. As a result, the current era requires that the HD datasets be categorized for effective predictions and that the main pattern be identified through feature selection. Feature selection and classification techniques can be used to do this. Furthermore, redundant features have been removed from the data using fast correlation-based feature selection (FCBF). Artificial neural networks (ANN), support vector machines (SVM), KNN, and naïve-Bayes Random Forest (NBRF) are the following classification techniques used. Particle swarm optimization (PSO) and ant colony optimization (ACO) are two examples of specialized optimization techniques that are used to improve classification algorithms [17] – [18]. This procedure would yield precise prediction results and increase the accuracy of HD categorization. As a result, this method is crucial for diagnosing HD. The dataset's features have been chosen for prediction using the genetic algorithm (GA). A range of feature selection algorithms, such as the one-attribute filtered-attribute methods, correlation-based feature selection (CFS), consistency-subset, filtered subset, info-gain, and relief technique, have also been used in this work to compare the outcomes of the suggested framework. SVM-based optimization function was used [19]. According to the jack-knife cross-validation (CV) process, feature selection algorithms and classification approaches employ several internal processes, including feature extraction methods, data discretization and feature reduction with the application of the heuristic rough-set reduction technique, to obtain prediction results from the data. The accuracy rate of the whole result of these processes is 92% [20]. These machine-learning algorithms produce accurate results for cardiac disease prediction and have shown themselves to be an effective method for medical optimization. Nevertheless, some of the less effective strategies should be improved by merging them with other classifiers [21]. Apart from enhancing the precision of inadequate classification systems, the method suggested can also predict heart disease early on. When assessing HD risks, the ensemble techniques, which include bagging and boosting, perform better and increase the prediction accuracy of weak classification techniques. Combining the optimization algorithms is another way of hybridizing machine learning techniques [22]. This situation was discussed where a hybrid strategy combining the PSO optimization algorithm and the slap swarm algorithm (SSA) was suggested [23]. The prediction method uses the hybrid approach known as SSAPSO, which was created by fusing

the two algorithms mentioned. It is evident from the methods presented for HD prediction that hybrid algorithms provide a high level of exploration efficacy throughout the prediction phase [24] – [25].

2. METHODOLOGY

These 11 clinical characteristics include the type of chest discomfort (atypical, non-anginal, atypical, or asymptomatic), age, sex, and Age, sex, resting blood pressure (mmHg), serum cholesterol (mm/dl), whether or not exercise causes angina, and whether the peak workout ST segment slopes upward, downward, or flat, and the old peak (number value observed in depression) are all factors to consider. The output class, 0 (normal) or 1 (heart disease), is shown in column 12. Previously available separately, the Stalog, Long Beach, Switzerland, Cleveland, and Hungarian datasets had never been merged into a single dataset as of September 2021. The final dataset for heart disease consists of 918 samples, each class having the same number of cases, which includes 508 cases in the heart disease class and 410 instances in the healthy class. Therefore, there has been no requirement for ways to deal with classes that are not balanced [26].

3.1. Prediction

The suggested ensemble models, comprise various DL models. A detailed discussion of each deep model utilized in the suggested ensemble models is provided in this section.

a) DenseNet

DenseeNet is a CNN architectures. Seven layers make up the architecture of DenseNet: two pooling, two fully connected, and three convolutional layers. Five of the seven layers in the model are learnable, as the name implies. Fig 2 depicts the architecture of DenseNet. Its primary function is the recognition of handwritten digits. It is a popular architecture due to its ease of use and effective pattern recognition capabilities. DenseNet is utilized in this study to detect cardiac disease. Convolutional layers are crucial to DenseNet's ability to uncover complex linkages and hidden trends within the information.

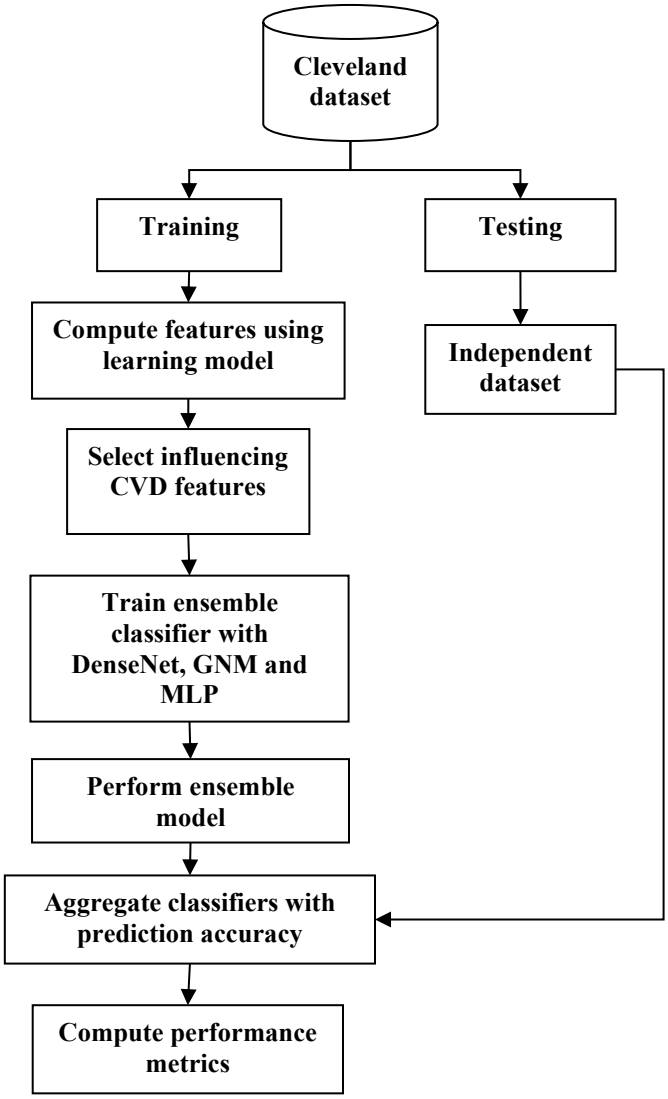

Fig 1 Flow diagram of Ensemble classifier model

After the convolutional layers, the network is made less complex using the pooling layers, which reduce the feature maps' dimensionality without compromising the most critical data. The first convolutional layer applies a 1D convolution to 1D heart disease data using five kernels and thirty-two neurons. A pooling layer then follows. With eight neurons and the same kernel size, the second convolutional layer improves the feature maps even more. Sixteen neurons are used in a 1×1 convolution in the third convolutional layer. Three completely connected layers with 32, 16, and 4 neurons each follow. In the final dense layer, a sigmoid activation function is employed, whereas a tanh activation function is used in the first five learnable layers. The probability of heart disease is output by this previous layer. This combination of activation functions is used to maintain a balance between network non-linearity and computational efficiency [27]. DenseNet is an easy-to-implement and successful model. DenseNet is, therefore, a well-liked option in the DL space. The DenseNet architecture's alternating convolutional and pooling layers, which are succeeded by fully connected layers, offer a strong foundation for problems involving predicting cardiac disease.

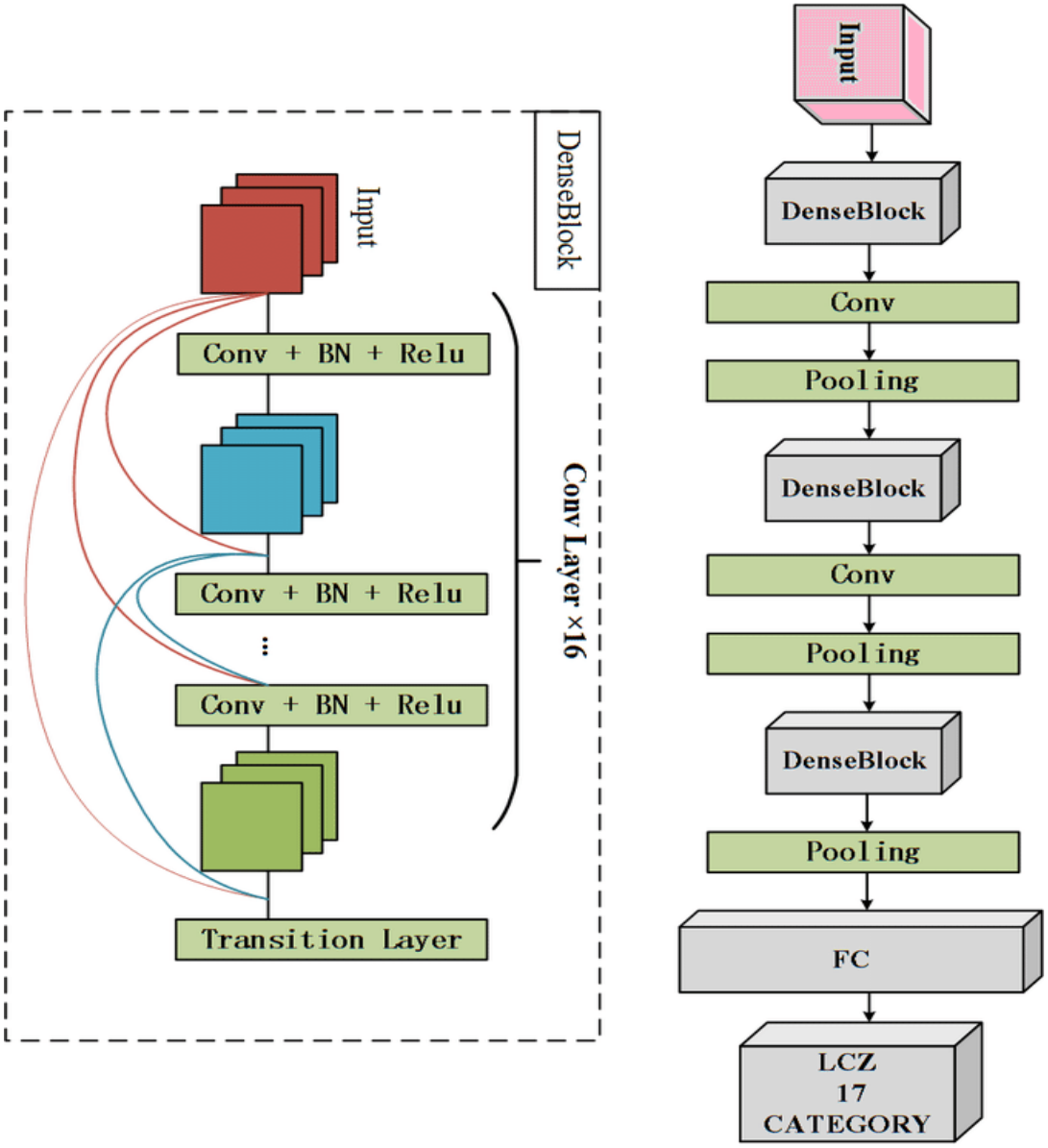

Fig 2. DenseNet architecture

b) Gated Network Model (GNM)

GNM is a type of RNN that aims to lower the quantity of parameters and streamline the gating mechanism, enhancing computing performance and making training easier. The gated procedures also control and manage the information flow in the NNs. It was created to overcome some of RNN's drawbacks, such as disappearing gradients, while successfully capturing sequential dependencies. GNM makes it possible to extract long-term dependencies from enormous sequential data without deleting information from the preceding segment of the data sequence. As illustrated in Fig 3, GNM transfers data across cells via a reset gate, an update gate, and a hidden state. The reset gate establishes the amount of earlier data that should be

lost, whereas the update gate establishes the amount that should be saved. The hidden state is a memory, storing the data from previous time steps. As GNM handles sequential input, it changes with time. A reset gate receives the current input (x_t) and the previous hidden state (h_{t-1}).

$$r_t = \sigma[W(r)x_t + U(r)h_{t-1}] \quad (1)$$

In the meantime, the upgrade gate adds fresh data to the hidden state. The GNM determines a potential hidden state by fusing the previously identified hidden state with new information from the current input. When sequential data processing is needed in the DL field, GNM works well. In contrast to LSTM and other RNN variations with gating mechanisms, the GNM has a simpler design and fewer parameters. This increases the GNM's computational efficiency while working with constrained resources [28].

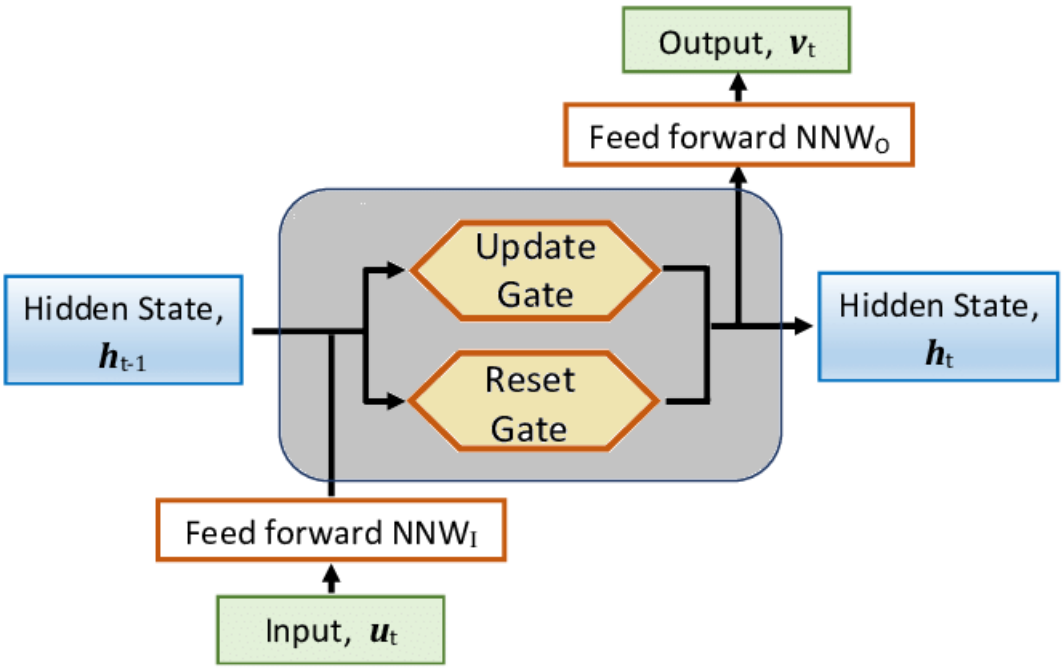

Fig 3. GNM architecture

c) Multi-Layer Perceptron's

Multiple layers of interconnected neurons make up the MLP type of ANN. A feed-forward neural network's data flow is unidirectional, which means it starts at the input layer, progresses via hidden levels, and ends at the output layer. To produce precise predictions, the network learns to modify the weights assigned to each node-to-node connection during training. An input layer, an output layer, and a buried layer or levels are the three primary layer types that make up an MLP. The essential architecture of MLP is shown in Fig 4. Nodes representing the input data's characteristics make up the input layer. The input layer's output, represented by a^0, is the same as the input values for the input features $x_1, x_2, x_3, \ldots, x_n$.

$$a^0 = [x_1, x_2, \ldots, x_n] \quad (2)$$

The MLP's input and output layers are divided by one or more hidden layers. A hidden layer's nodes are linked to all nodes in its adjacent layers, both before and after. Each node's output in the hidden layer is determined through the application of an activation function to the weighted sum of the inputs. Tanh, ReLU, and the sigmoid function are examples of standard activation functions. The network's output layer generates the final output. The weighted sum and an activation function are used in a calculation comparable to the hidden layers.

$$z_k^L = \sum_{j=1}^{N^{L-1}} w_{kj}^L a_j^{L-1} + b_k^L \quad (3)$$

$$a_k^L = \sigma(z_k^L) \quad (4)$$

The preceding layer's node count is N^{L-1}, the connection weight is w_{kj}^L, the output of the j^{th} node in the previous layer is an a_j^{L-1}, the bias term is b_k^L, the activation function is σ, and the output nodes are indexed by k. A supervised learning technique trains the MLP architecture, altering the biases and weights to reduce a loss function. To reduce error, gradient descent and back-propagation are typically employed [29]. Algorithm 1 uses x as the input, w for weights, h for hidden layers, and l for

the number of hidden layers.

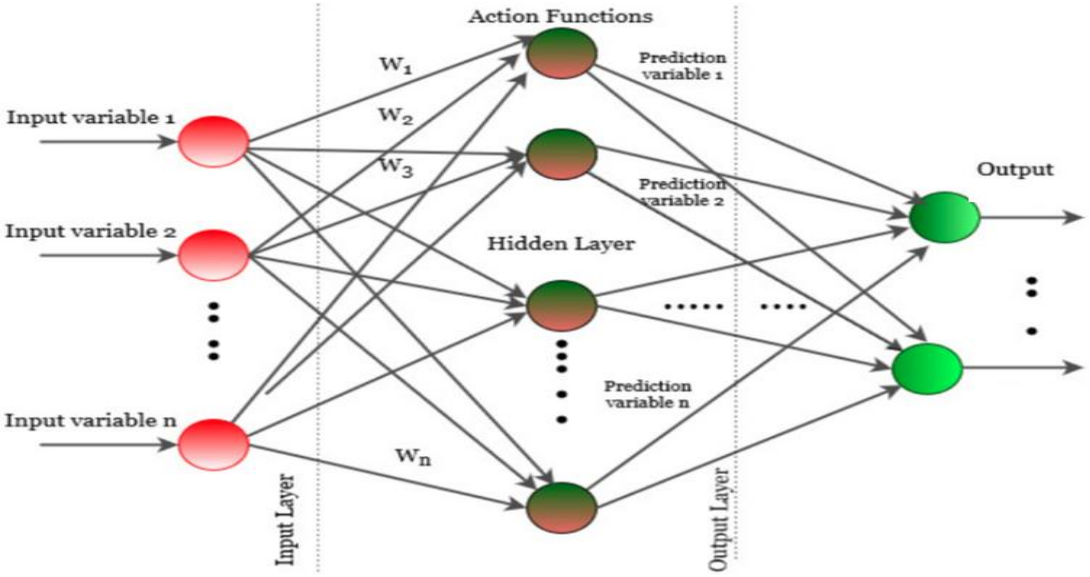

Fig 4. MLP architecture

Algorithm 1:
Initialize: Cleveland dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$
1. Perform data partitioning based on features and labels X_{train} and Y_{train}
2. Input X;
3. Initialize $DenseNet$ and GNM with network architecture and parameters;
4. Provide dataset input X via Densenet;
5. $Y_{DenseNet} = DenseNet\ (X)$;
6. Provide input X via GNM;
7. Concatenate GNM and DenseNet;
8. Perform concatenation $Y_{concat} = concatenation\ (Y_{DenseNet,} Y_{GRU})$;
9. Perform concatenation output via fully connected MLP layer;
10. $Y_{concat} = DenseNet\ (Y_{concatenation})$;
11. Perform sigmoid function to acquire output;
12. $\hat{Y} = sigmoid\ (Y_{output})$
13. Perform final prediction outcomes;

4. Experimental setup

The confusion matrix evaluates these proposed models' performance and the individual models. As shown in Table 1, True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) comprise the confusion matrix. It helps assess performance indicators like F1-score, recall, accuracy, and precision. To begin with, a pre-processing step was done to clean up the dataset and extract more valuable information. Three new columns reflecting the age range of young, adult, and senior were added when the Age was removed. Three new low, medium, and high blood pressure columns were created using the resting blood pressure feature. Three columns were created from the cholesterol feature: low, medium, and high, representing risk. A single hot encoding technique was used for the ST Slope, RestingECG, and Chest Pain Type features. Lastly, a label encoder was used to handle Exercise Angina and Sex. After this approach, we are left with a dataset consisting of 24 features. Ten folds of k-fold CV have been performed on each trial to eliminate randomness. Analysis of each model has been done using a thorough grid search with hyper-parameters. The results provide the score for the hyper-parameter setup with the optimal ten-cycle mean value. One classifier are both trained simultaneously, make up neural network topologies for feature augmentation. We've implemented two distinct settings. The first creates a bi-dimensional matrix using the latent space, whereas the second uses an MLP classifier to build a 2D convolutional neural network [30]. For training in both cases, the ADAM optimizer was used due to its ability to handle sparse gradients, invariance to gradient rescaling, and short optimization time greatly enhancing neural network performance. The loss functions selected for the classifier subnet were binary cross-entropy and mean squared error for the decoder. The following is the definition of binary cross-entropy, and is utilized since it is the same as applying the greatest likelihood to fit the model. You can define mean squared error as

follows:

$$-(ylog\ (p) + (1-y)\log(1-p)) \quad (5)$$

$$\sum_{i=1}^{D}(x_i - y_i)^2 \quad (6)$$

Additionally, it is used because it heavily penalizes significant errors, which, when applied to autoencoders, is interesting for feature reconstruction. By contrasting different latent space sizes, the significance of this parameter in the final classification was investigated.

4.1. Result analysis

The authors have analyzed the traditional machine learning techniques covered to compare the suggested and other strategies. As previously mentioned, the optimal hyper-parameters for each method have been determined via a grid search. Fig. 5, which shows the mean of accuracy using the ideal arrangement, illustrates this result. With the highest accuracy of 86%, these data show how well the neural network performed. Comparable results from the RF or the SVM ensemble approach followed this.

Table 1. Comparison of proposed vs. existing approaches

Dataset	ML algorithm	Accuracy (%)		Sensitivity	Specificity	F1-score	No. of features		Time (min)	AUROC
		Without feature	With feature				Without feature	With feature		
Cleveland	LR	86	92	0.75	0.71	0.15	14	10	0.11	92.85
	RF	80	87	0.78	0.69	0.14	14	9	0.11	92.35
	L-GBM	84	88	0.72	0.70	0.15	14	11	0.79	93.85
	SVM	80	86	0.75	0.73	0.15	14	7	6.81	91.45
	BRT	79	87	0.86	0.90	0.16	14	12	1.1	92.35
	Ensemble Model	87	93	0.91	0.93	0.17	14	8	0.05	95.90

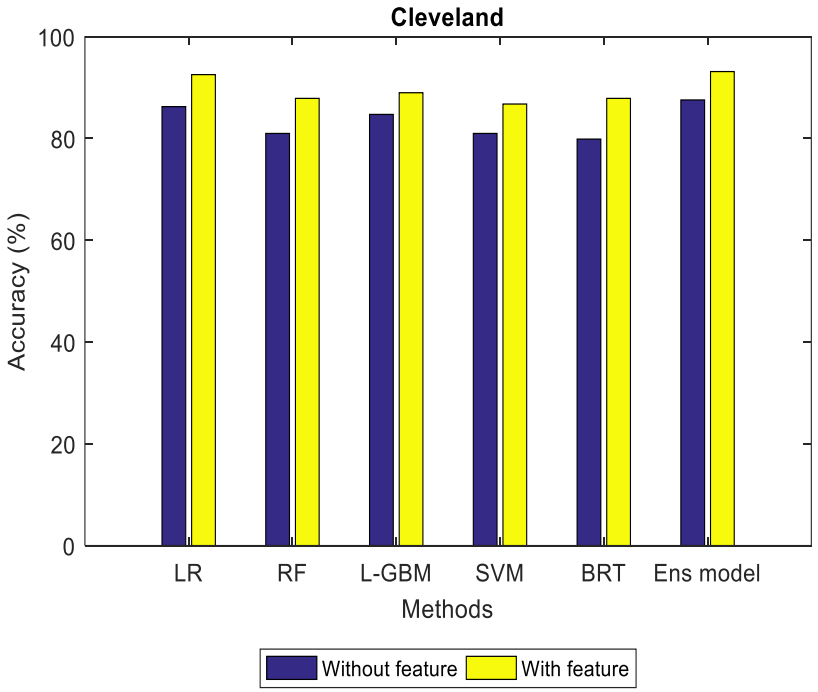

Fig 5. Accuracy comparison

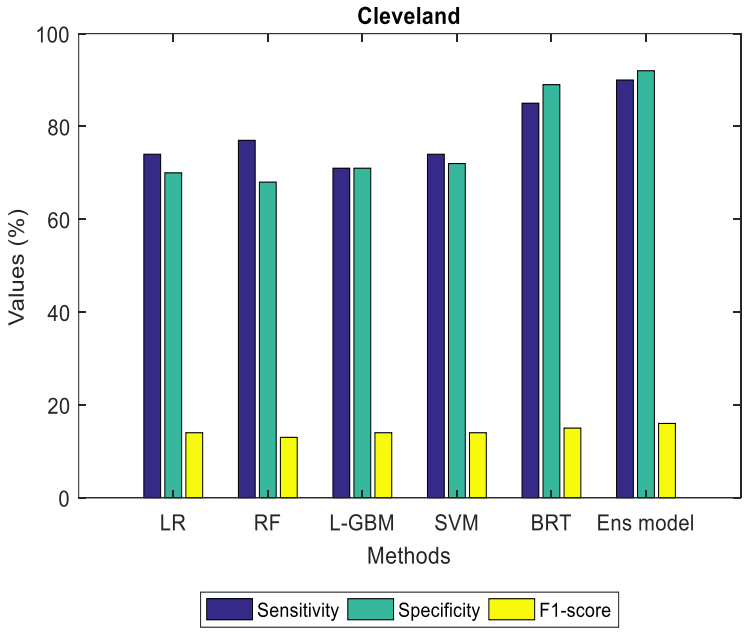

Fig 6. Sensitivity, specificity and F1-score comparison

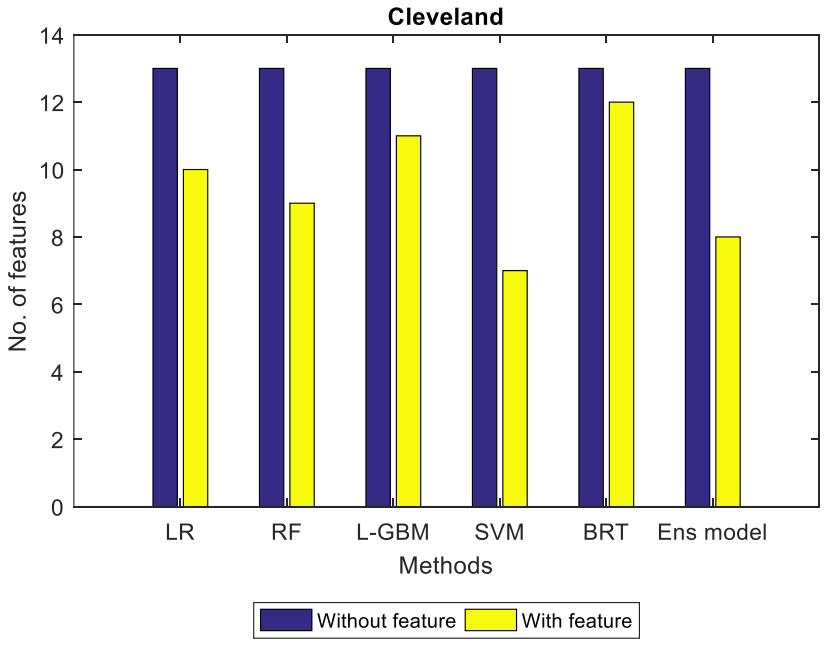

Fig 7. Feature representation

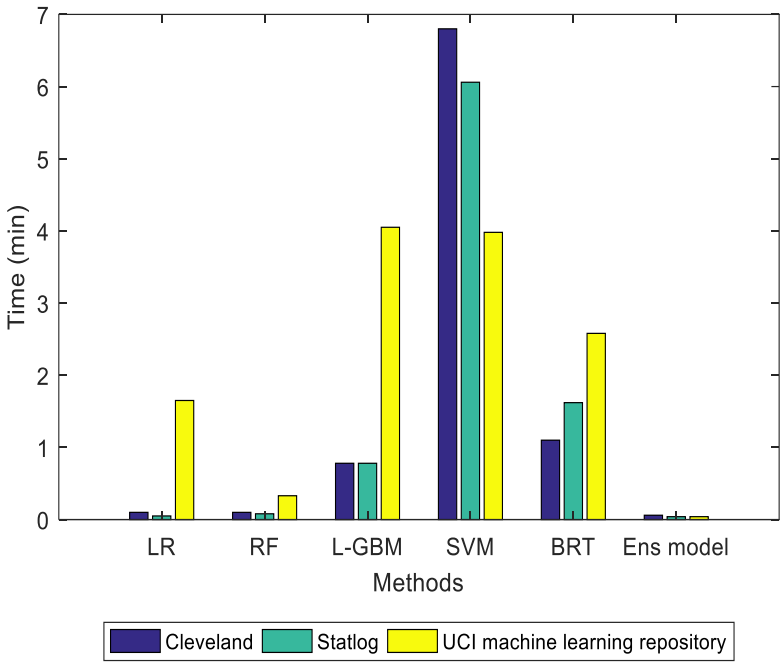

Fig 8. Execution time comparison

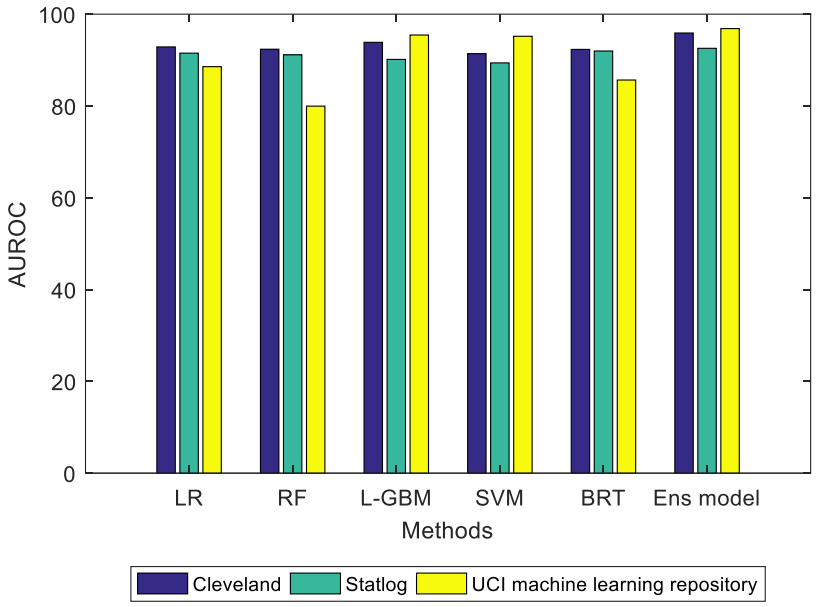

Fig 9. AUROC comparison

On the other hand, decision trees performed the poorest achieving an accuracy of 78%, which was 8.46% less than the proposed model. Because neural networks perform so well, we have used an ensemble model for training and paired an MLP with feature augmentation. The outcomes obtained with different latent space sizes are shown in Fig 6. The average accuracy achieved with a 10-fold CV is displayed in each graphic. When the latent space had 100 features rather than the original 11 characteristics, the classification accuracy rose from 3.78% to 89.543%. Since preliminary results showed that training the ensemble model simultaneously with a multi-tasking NN improved classification, a fresh round of trials with a network in

place of the MLP classifier has been carried out. In addition to being the most effective method for identifying cardiac issues, this strategy makes it possible to extract additional features from the dataset. This approach handles structured data by reorganizing the data using the ensemble model, merging features to create new ones, and adding the optimal spatial representation. Fig 7 shows findings with different latent space sizes for comparison. The best result in this new series of trials was achieved with an accuracy of 98% and a latent space of 200 additional features. When it comes to a condition that can lead to patients experiencing severe problems or even dying, this is a pretty intriguing boost in performance of over 4.5% over the standard MLP. Our suggestions perform better than the vanilla MLP and the random forest model, as demonstrated in Fig 8 and Fig 9. The results obtained in the original dataset are improved by feature augmentation techniques as demonstrated.

Furthermore, accuracy is slightly increased by reorganizing the additional features integrating DenseNet with MLP. The Kolmogorov-Smirnov test and the Independent Samples t-test, two statistical tests, verified that the suggested strategy produces better results than more conventional approaches. These tests were conducted by grouping the accuracy results into two groups: (group I) existing techniques, which comprised traditional MLP and machine learning techniques, and (group II) strategies suggested in this research, which included DenseNet with GNM in conjunction with MLP. Following the application of these tests, the results showed that, in contrast to the accuracy achieved with conventional techniques, the accuracy advantage attained by the recommended strategy is statistically significant with a value of $p < 0.001$. Furthermore, the dataset used in this work has been used in a few recent publications. A study employing various machine learning approaches was carried out. A stacking strategy that combines logistic regression and K-NN yielded the best results with an accuracy of 87.24% when the vote classifier output from the previous methods was used to perform a K-NN classification. RF achieved an accuracy of 86%. For example, by integrating the output of RF, MLP, Boost, Decision Trees, and Logistic, the final result is obtained by training a meta classifier using regression. This stacking technique was recently introduced. Comparing this method to our idea, it obtained an accuracy of 89% under the same experimental settings. A comparison between our concept and the findings of the state of the art is shown in Table 1. Our ensemble classifier with DenseNet, GNM and MLP results is superior to all other reported approaches.

3. CONCLUSION

This work presents deep learning-based methods for predicting heart problems by combining classification and feature augmentation tasks using a dataset of patient records from five hospitals. Only 11 of the 918 samples in this collection exhibit clinical traits. The DenseNet, GNM and MLP are combined to form an ensemble approaches in a novel architectural method. Since only eleven features are in the dataset, the DenseNet has been used to extract further features through feature augmentation. GNM can be trained using the many features we have gathered by rearranging them into a 2D array. These two procedures use the classifier data acquired as feedback in the back-propagation technique to produce a complicated net that combines the classifier (MLP or GNM) which has been used to increase feature extraction capabilities. The suggested ensemble model outperforms MLP by 0.6% while training the ensemble model concurrently. By forcing it to employ spatial position information in order to extract more pertinent features, DenseNet impedes the feature extraction process. The convolutional network performs significantly better even after altering the feature extraction process. A thorough analysis showed that 200 neurons are the ideal number of neurons in the DenseNet, which represents the new properties. This study indicates that having more neurons does not always result in better outcomes because the results decline at a specific size. Compared to typical classifiers trained on the same dataset and under the same conditions, our performance of 98% was 4.4% better. The state-of-the-art methods which employed stacking and a combination of approaches was also exceeded. Furthermore, it is a computationally very costly technique because ensemble requires studying multiple models sometimes to achieve the desired outcome. Since identifying a cardiac condition in a patient can result in survival, the advancements described in this publication and the suggested approach are highly relevant to the field's experts.

REFERENCES

[1] Jensen, L. J. Jensen, and S. Brunak, ‘‘Mining electronic health records: Towards better research applications and clinical care,’’ Nature Rev. Genet., vol. 13, no. 6, pp. 395–405, Jun. 2012.

[2] Li, C. Bai, and C. K. Reddy, ‘‘A distributed ensemble approach for mining healthcare data under privacy constraints,’’ Inf. Sci., vol. 330, pp. 245–259, Feb. 2016.

[3]] Khedr, Z. A. L. Aghbari, and I. Kamel, "Privacy-preserving decomposable mining association rules on distributed data," Int. J. Eng. Technol., vol. 7, nos. 3–13, pp. 157–162, 2018.

[4] Khedr and R. Bhatnagar, ‘‘New algorithm for clustering distributed data using K-means,’’ Comput. Information., vol. 33, pp. 1001–1022, Oct. 2014.

[5] Subhashini and M. K. Jeyakumar, ‘‘OF-KNN technique: An approach for chronic kidney disease prediction,’’ Int. J. Pure Appl. Math., vol. 116, no. 24, pp. 331–348, 2017.

[6] Nathan, P. M. Kumar, P. Panchatcharam, G. Manogaran, and R. Varadharajan, "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease," Des. Automat. Embedded Syst., vol. 22, no. 3, p. 225, 2018

[7] Hsu, G. Manogaran, P. Panchatcharam, and S. Vivekanandan, "A new approach for prediction of lung

carcinoma using backpropagation neural network with decision tree classifiers," in Proc. IEEE 8th Int. Symp. Cloud Service Comput. (SC), Nov. 2018, pp. 111–115

[8] Ramasamy and K. Nirmala, "Disease prediction in data mining using association rule mining and keyword-based clustering algorithms," Int. J. Comput. Appl., vol. 42, no. 1, pp. 1–8, Jan. 2020

[9] Bakar, Z. Kefli, S. Abdullah, and M. Sahani, ''Predictive models for dengue outbreak using multiple rulebase classifiers,'' in Proc. Int. Conf. Electr. Eng. Informat., Jul. 2011, pp. 1–6

[10] Hariharan, R. Umadevi, T. Stephen, and S. Pradeep, ''Burden of diabetes and hypertension among people attending health camps in an urban area of Kancheepuram district,'' Int. J. Community Med. Public Health, vol. 5, no. 1, p. 140, Dec. 2017.

[11] [Tun, G. Arunagirinathan, S. K. Munshi, and J. M. Pappachan, ''Diabetes mellitus and stroke: A clinical update,'' World J. Diabetes, vol. 8, no. 6, pp. 235–248, Jun. 2017

[12] Ley, O. Hamdy, V. Mohan, and F. B. Hu, ''Prevention and management of type 2 diabetes: Dietary components and nutritional strategies,'' Lancet, vol. 383, no. 9933, pp. 1999–2007, Jun. 2014

[13] Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, ''Comparison of three data mining models for predicting diabetes or prediabetes by risk factors,'' Kaohsiung J. Med. Sci., vol. 29, no. 2, pp. 93–99, Feb. 2013

[14] Bashir, U. Qamar, and F. H. Khan, ''IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework,'' J. Biomed. Inform., vol. 59, pp. 185–200, Feb. 201

[15] Nai-Arun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," Procedia Comput. Sci., vol. 69, pp. 132–142, Nov. 2015

[16] Calheiros, K. Ramamohanarao, R. Buyya, C. Leckie, and S. Versteeg, ''On the effectiveness of isolation-based anomaly detection in cloud data centers,'' Concurrency Comput., Pract. Exper., vol. 29, no. 18, p. e4169, Sep. 2017.

[17] Chen, X. Shi, and Y. D. Wong, ''Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data,'' Accident Anal. Prevention, vol. 129, pp. 156–169, Aug. 2019.

[18] Shin, A. Abraham, and S. Y. Han, ''Improving kNN text categorization by removing outliers from training set,'' in Computational Linguistics and Intelligent Text Processing, vol. 3878, A. Gelbukh, Ed. Berlin, Germany: Springer, 2006, pp. 563–566.

[19] [Alfian, M. Syafrudin, B. Yoon, and J. Rhee, "False-positive RFID detection using classification models," Appl. Sci., vol. 9, no. 6, p. 1154, Mar. 2019

[20]] Farquad and I. Bose, ''Preprocessing unbalanced data using support vector machine,'' Decis. Support Syst., vol. 53, no. 1, pp. 226–233, Apr. 2012.

[21] Hardiman and K. Uchida, "Data- and algorithm-hybrid approach for imbalanced data problems in deep neural network," Int. J. Mach. Learn. Comput., vol. 8, no. 3, pp. 208–213, Jun. 2018

[22] Lemaitre, F. Nogueira, and C. K. Aridas, ''Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning,'' J. Mach. Learn. Res., vol. 18, no. 1, pp. 559–563, 2017

[23] Gupta, R. Kumar, H. Singh Arora, and B. Raman, ''MIFH: A machine intelligence framework for heart disease diagnosis,'' IEEE Access, vol. 8, pp. 14659–14674, 2020.

[24]] Pasha and E. S. Mohamed, "Bio-inspired ensemble feature selection (BEFS) model with machine learning and data mining algorithms for disease risk prediction," in Proc. 5th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA), Pune, India, Sep. 2019, pp. 1–

[25] Sabahi, ''Bimodal fuzzy analytic hierarchy process (BFAHP) for coronary heart disease risk assessment,'' J. Biomed. Informat., vol. 83, pp. 204–216, Jul. 2018.

[26] Maji and S. Arora, ''Decision tree algorithms for prediction of heart disease,'' in Information and Communication Technology for Competitive Strategies, vol. 40. Singapore: Springer, 2019, pp. 447–454.

[27] Reddy, M. P. K. Reddy, K. Lakshmana, D. S. Rajput, R. Kaluri, and G. Srivastava, "Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis," Evol. Intell., vol. 13, no. 2, pp. 185–196, Jun. 2020.

[28] [Zhu, Z. Ni, L. Ni, F. Jin, M. Cheng, and J. Li, ''Improved discrete artificial fish swarm algorithm combined with margin distance minimization for ensemble pruning,'' Comput. Ind. Eng., vol. 128, pp. 32–46, Feb. 2019.

[29] Tanveer, C. Gautam, and P. N. Suganthan, ''Comprehensive evaluation of twin SVM based classifiers on UCI datasets,'' Appl. Soft Comput., vol. 83, Oct. 2019, Art. no. 105617.

[30] [Paul, P. C. Shill, M. R. I. Rabin, and K. Murase, ''Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease,'' Int. J. Speech Technol., vol. 48, no. 7, pp. 1739–1756, Jul. 2018.

[31] UCI Machine Learning Repository. (2015). Chronic_Kidney_Disease Data Set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ chronic_kidney_disease.