Computers, Materials & Continua
DOI:10.32604/cmc.2021.019013
Article

Tech Science Press

Data and Machine Learning Fusion Architecture for Cardiovascular Disease Prediction

Munir Ahmad[1], Majed Alfayad[2], Shabib Aftab[1,3], Muhammad Adnan Khan[4,*], Areej Fatima[5], Bilal Shoaib[6], Mohammad Sh. Daoud[7] and Nouh Sabri Elmitwally[2,8]

[1]School of Computer Science, National College of Business Administration & Economics, Lahore, 54000, Pakistan
[2]College of Computer and Information Sciences, Jouf University, Sakaka, 72341, Saudi Arabia
[3]Department of Computer Science, Virtual University of Pakistan, Lahore, 54000, Pakistan
[4]Riphah School of Computing & Innovation, Riphah International University, Lahore Campus, Lahore, 54000, Pakistan
[5]Department of Computer Science, Lahore Garrison University, Lahore, 54000, Pakistan
[6]Department of Computer Science, Minhaj University Lahore, Lahore, 54000, Pakistan
[7]College of Engineering, Al Ain University, Abu Dhabi, 112612, UAE
[8]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, 12613, Egypt
[*]Corresponding Author: Muhammad Adnan Khan. Email: madnankhan@ncbae.edu.pk
Received: 28 March 2021; Accepted: 29 April 2021

Abstract: Heart disease, which is also known as cardiovascular disease, includes various conditions that affect the heart and has been considered a major cause of death over the past decades. Accurate and timely detection of heart disease is the single key factor for appropriate investigation, treatment, and prescription of medication. Emerging technologies such as fog, cloud, and mobile computing provide substantial support for the diagnosis and prediction of fatal diseases such as diabetes, cancer, and cardiovascular disease. Cloud computing provides a cost-efficient infrastructure for data processing, storage, and retrieval, with much of the extant research recommending machine learning (ML) algorithms for generating models for sample data. ML is considered best suited to explore hidden patterns, which is ultimately helpful for analysis and prediction. Accordingly, this study combines cloud computing with ML, collecting datasets from different geographical areas and applying fusion techniques to maintain data accuracy and consistency for the ML algorithms. Our recommended model considered three ML techniques: Artificial Neural Network, Decision Tree, and Naïve Bayes. Real-time patient data were extracted using the fuzzy-based model stored in the cloud.

Keywords: Machine learning fusion; cardiovascular disease; data fusion; fuzzy system; disease prediction

1 Introduction

The clinical investigation of heart disease, which is also known as cardiovascular disease, constitutes a major topic of interest for medical research, both historically and in contemporary times. According to the World Health Organization, around 23 million cardiovascular disease

patients die annually due to cardiac arrest and stroke [1], with a significant number of cases in developing countries. Heart diseases have a major influence not only on the life of an individual but also on the economies of countries. As such, heart health awareness programs significantly prevent disease by encouraging the adoption of a healthy lifestyle. Technology also provides remarkable support for the prevention of disease through medical applications of, for example, cloud computing and artificial intelligence. Cardiovascular diseases include all types of blood circulation problems and heart malfunctions.

Several underlying factors constitute the root causes of heart disease, including excessive intake of saturated fats, lack of exercise, and an imbalanced diet. In addition, genetic predisposition is increasingly recognized as a prominent cause [2]. Cloud computing provides applications and resources on an on-demand basis [3] and is compatible with modern tools and technologies. It can effectively support machine learning (ML) models and ultimately improve diagnostic analysis, as well as meet other needs of the healthcare industry [4]. Cloud-based applications are becoming the first choice for medical professionals and technicians, because they not only allow test reports to be updated instantly but also contribute to resolving the big data issues surrounding computerized tomography (CT) scans and radiology. However, this requires a tool to provide security, privacy and optimal accuracy along with enhancing the availability of information [5]. As a part of artificial intelligence, ML facilitates the accurate prediction of the likelihood of a particular event using the predefined dataset. In 2018, Khan et al. recommended a fuzzy inference system to predict the chances of heart disease [6], by examining examples from an array of research studies on heart health. For instance, in 2013, Kumar and Kaur conducted research on a heart disease diagnosis system using fuzzy logic and suggested that a fuzzy-based system could predict disease with 93.33% accuracy [7]. We proposed a cloud-based prediction model using ML techniques after considering the gravity of the problem and its fatal effects.

This paper organizes our approach into seven phases. Phase 1 concerns data collection. We collected datasets from geographically diffuse locations to ensure maximum coverage. Phase 2 consolidated all datasets into the fuzzy dataset. Phase 3 was a pre-processing layer involving the elimination of records with missing values; this included normalization and, ultimately, splitting training and testing data. Phase 4 concerned the training layer, in which we applied three algorithms: Artificial Neural Network (ANN), Decision Tree (DT), and Naïve Bayes (NB). Next, in Phase 5, we evaluated the data to obtain target accuracy. In the ML-fusion phase (Phase 6), the fuzzy-based system accepted data meeting our predefined criteria for two of three brains. Finally, in Phase 7, the fuzzy model was compared with the model stored in the cloud.

2 Related Work

Researchers have explored various alternative techniques for identifying cardiovascular disease. For example, some researchers have applied the neural method, obtaining results with 83% accuracy [8]. Meanwhile, in 2017, Kim and Kang applied ML techniques to predict coronary heart disease, with the recommended model viewed as a single layer. After performing 4146 tests, 3031 cases were deemed low-risk and 1115 were considered high-risk. The proposed model had 81.09% accuracy [9]. Elsewhere, researchers conducted a study predicting cerebral infarction disease, by developing convolutional neural network models to predict vulnerability relevant to structured and unstructured data from various sources. This was a unique experiment for the use of big data analysis in the medical sciences field. The proposed algorithm attained a 94.8% accuracy level [10].

Meanwhile, ANN techniques have been widely used to predict heart disease. Generalized regression neural networks and radial basis functions have been widely used to investigate heart

function problems, with experimental analysis proving that ANNs provide more accurate results than any other technique [11]. Recent medical research studies have also emphasized computational intelligence techniques for clinical investigation, developing models using deep extreme ML for diagnosing cardiovascular disease and concluding that more accurate and precise results can be achieved using these techniques [12]. Numerous techniques can probe the root causes of ailments, including fuzzy set, fuzzy deduction framework, and fuzzy connection. Research studies have highlighted the application of the latest approaches for therapeutic conclusions [13]. Many researchers have discussed ANN models and their relative importance for diagnosing heart disease at an early stage [14,15]. Meanwhile, multilayer perceptron and other data mining techniques have been successfully implemented for heart disease prediction, with one study using two distinct datasets featuring 303 and 270 cases. They identified 15 features for each patient that included smoking, body fat, hypertension, and gender. The accuracy of the DT was 99.62%, compared to 100% for multilayer perceptron [16].

3 Materials and Methods

Early-stage mild cardiovascular disease is curable through significant lifestyle changes, including adopting a more balanced diet [17]. However, this requires early identification of potential patients. Accordingly, this research considers cloud-based heart disease prediction using ML following the seven-phase methodology presented in Fig. 1.

Dataset selection [18] provided the foundation of the training layer. This study used a pre-labelled dataset of heart disease patients [19] for the implementation of the proposed framework. The selected dataset comprised 1190 cases and considered 12 features. Eleven of the features were independent and 1 was dependent, which represented the output class. The pre-processing layer involved data normalization, data cleaning, and data splitting, with the mean imputation method used to remove missing values before the data normalization process synchronized the values of the various features. These activities enabled the classification process to perform better and more accurately.

After the cleaning and normalization process, the dataset was divided into training data (70%) and test data (30%). Next, the classification process was started, which first involved training for the three classification techniques: ANN, NB, and DT. The classification process generated three predictions that were based on algorithms optimized to achieve maximum accuracy. A hidden layer was used with 12 neurons during the configuration of the ANN, with the weight back-propagation technique used to fine-tune the hidden layer. This involved multiple steps, including initialization of weight, feedforward, backpropagation of error and weight updating. In addition to the input and output layers, a multilayer perceptron was also used for at least one hidden layer. The sigmoid function for input and the hidden layer of the proposed back propagation neural network was expressed as follows:

$$\psi_n = b_1 + \sum_{m=1}^{t} (\omega_{mn} * r_m) \tag{1}$$

$$\varphi_m = \frac{1}{1 + e^{-\psi_m}}, \quad where\ m = 1, 2, 3 \ldots q. \tag{2}$$

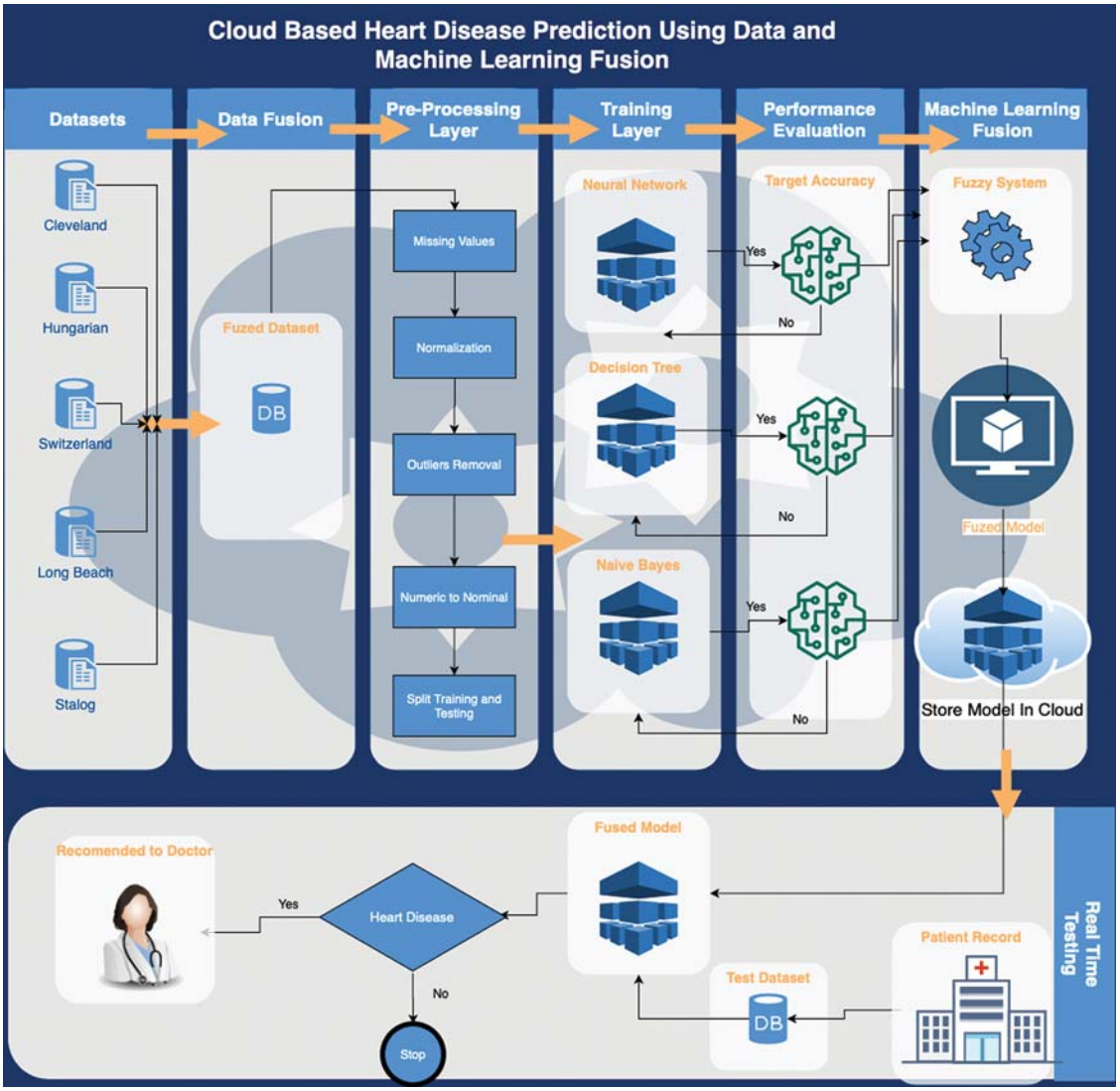

Figure 1: Proposed cloud-based heart disease prediction using a data and ML fusion model

The input derived from the output layer is given by:

$$\psi_t = b_2 + \sum_{i=1}^{n} (\upsilon_{it} * \varphi_i). \tag{3}$$

The output layer activation function is as follows:

$$\varphi_t = \frac{1}{1 + e^{-\psi_t}} \quad where \;\; t = 1, 2, 3 \ldots n \tag{4}$$

$$E = \frac{1}{2} \sum_{t} (\tau_t - \varphi_t)^2, \tag{5}$$

where τ_t and φ_t represent the desired output and estimated output, respectively. Eq. (6) describes the rate of weight change for the output:

$$\Delta W \propto -\frac{\partial E}{\partial W}$$

$$\Delta \upsilon_{m,n} = -\epsilon \frac{\partial E}{\partial \nu_{m,n}}. \tag{6}$$

After applying the chain rule method, this can be presented as:

$$\Delta \upsilon_{m,n} = -\epsilon \frac{\partial E}{\partial \varphi_n} \times \frac{\partial \varphi_n}{\partial \psi_n} \times \frac{\partial \psi_n}{\partial \nu_{m,n}}. \tag{7}$$

By substituting the values in Eq. (7), the value of weight change can be obtained using Eq. (8):

$$\Delta \upsilon_{m,n} = \epsilon(\tau_t - \varphi_t) \times \varphi_t(1 - \varphi_t) \times (\varphi_m)$$

$$\Delta \upsilon_{m,n} = \epsilon \zeta_n \varphi_m, \tag{8}$$

where

$$\zeta_t = (\tau_t - \varphi_t) \times \varphi_t(1 - \varphi_t).$$

Next, applying the chain rule for the updating of weights between input and hidden layers gives:

$$\Delta \omega_{m,n} \propto -\left[\sum_t \frac{\partial E}{\partial \varphi_t} \times \frac{\partial \varphi_t}{\partial \psi_t} \times \frac{\partial \psi_t}{\partial \varphi_n}\right] \times \frac{\partial \varphi_n}{\partial \psi_n} \times \frac{\partial \psi_n}{\partial \omega_{m,n}}$$

$$\Delta \omega_{m,n} = -\epsilon \left[\sum_t \frac{\partial E}{\partial \varphi_t} \times \frac{\partial \varphi_t}{\partial \psi_t} \times \frac{\partial \psi_t}{\partial \varphi_n}\right] \times \frac{\partial \varphi_n}{\partial \psi_n} \times \frac{\partial \psi_n}{\partial \omega_{m,n}},$$

where ϵ represents the constant and

$$\Delta \omega_{m,n} = \epsilon \left[\sum_t (\tau_t - \varphi_t) \times \varphi_t(1 - \varphi_t) \times (\nu_{n,t})\right] \times \varphi_t(1 - \varphi_t) \times \alpha_m$$

$$\Delta \omega_{m,n} = \epsilon \left[\sum_t (\tau_t - \varphi_t) \times \varphi_t(1 - \varphi_t) \times (\nu_{n,t})\right] \times \varphi_n(1 - \varphi_n) \times \alpha_m$$

$$\Delta \omega_{m,n} = \epsilon \left[\sum_t \zeta_t(\nu_{n,t})\right] \times \varphi_n(1 - \varphi_n) \times \alpha_m.$$

This can be presented as Eq. (9) after simplification:

$$\Delta \omega_{m,n} = \epsilon \zeta_n \alpha_m, \tag{9}$$

where

$$\zeta_m = \left[\sum_k \zeta_k(v_{m,k})\right] \times \varphi_m(1-\varphi_m)$$

$$v_{m,n}^{+} = v_{m,n} + \lambda_F \Delta v_{m,n}. \tag{10}$$

Eq. (10) updates weights between hidden layers and outputs. Eq. (11) updates weights between the input and hidden layer:

$$\omega_{m,n}^{+} = \omega_{m,n} + \lambda_F \Delta \omega_{m,n} \tag{11}$$

In the DT, three optimizers were applied individually, including random search, Bayesian optimization, and grid search. The Bayesian optimization performed well and it was therefore selected for this framework:

$$E(S) = I_E(p_1, p_2, \ldots, p_n) = -\sum_{m=1}^{n} p_m \log_2 p_m. \tag{12}$$

The GINI index is provided by Eq. (13):

$$E(S) = I_G(p_1, p_2, \ldots, p_n) = 1 - \sum_{m=1}^{n} p_m^2. \tag{13}$$

Information gain is provided by Eq. (14):

$$\overbrace{IG(S,z)}^{\text{Information Gain}} = \overbrace{E(S)}^{\text{Entropy/Gini (parent)}} - \overbrace{E(S \mid z)}^{\text{Weighted Sum of Entropy/Gini (Children)}}$$

$$IG(S,z) = \text{Entrop } m(S) - \sum_z p(z)\text{Entrop } m(S \mid z) \tag{14}$$

$$z^{\star} = \arg\min_{z \in Z} \mathrm{f}(z). \tag{15}$$

Here, $f(z)$ demonstrates the aim of minimizing the error rate or the root mean square error, which is assessed as the validation set. z can take any value from domain Z and z^* is the set of hyper-parameters that represent the lowest value of the score. This approach sought the model hyper-parameters that could deliver the best score for the validation set metric. This model, which is known as the "surrogate" model, is represented as $p(z \mid n)$ for the objective function:

$$\mathrm{EI}_{z^*}(n) = \int_{-\infty}^{z^*} (z^* - z) p(z \mid n) \mathrm{dz}. \tag{16}$$

This is intended to optimize expected improvement with respect to the proposed set of hyperparameters n. Here, z^* is an edge value of the objective function, z depicts the actual value of the function using the set of hyperparameters n, and $p(z \mid n)$ is the surrogate probability model that states the probability of z given n. This enables finding the best set of hyperparameters under the function $p(z \mid n)$.

The hyperparameter does not expect to produce any improvement if $p(z \mid n)$ is zero in all cases that $z < z^*$. In contrast, the set of hyperparameters n is expected to produce a better result than the threshold value if the fundamental part is positive:

$$p(z \mid n) = \frac{p(n \mid z) * p(z)}{p(n)} \tag{17}$$

$p\ (n \mid z)$ function is expressed as:

$$p(n \mid z) = \begin{cases} l(n) & \text{if } \ n < n^* \\ g(n) & \text{if } \ n \geq n^*. \end{cases}$$

There are two different distributions for the hyperparameters in this equation, one where the value of the objective function is less than $l(n)$ and one where the objective function is greater than $g(n)$:

$$\mathrm{EI}_{z^*}(n) = \frac{\Upsilon y^* \ell(n) - \ell(n) \int_{-\infty}^{z^*} p(z)\mathrm{dz}}{\Upsilon \ell(n) + (1 - z) g(n)} \propto \left(\Upsilon + \frac{g(n)}{\ell(n)} (1 - \Upsilon) \right)^{-1} \tag{18}$$

For NB, the following three kernel types were used: box, Gaussian, and triangle:

$$\text{Probability of Outcome} \mid \text{Evidence(Posterior Probability)} = \frac{\text{Probability of Likelihood of Evidence} * \text{Prior}}{\text{Probability of Evidence}}.$$

The traditional NB classifier estimates probabilities by approximating the data through a function such as a Gaussian distribution:

$$P(S_r z) = \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left(-\frac{(s_r - \mu_r)^2}{2\sigma_z^2} \right), \tag{19}$$

where μ_t represents the mean of the values of an attribute S_t averaged over training points with class labels z and σ_z representing standard deviation. One-parameter Box-Cox transformations are defined as:

$$y_j^{(\lambda)} = \begin{cases} \dfrac{y_j^{\lambda} - 1}{\lambda} & \text{if } \ \lambda \neq 0 \\ \ln y_j & \text{if } \ \lambda = 0. \end{cases} \tag{20}$$

The two-parameter Box-Cox transformation is defined as:

$$y_j^{(\lambda)} = \begin{cases} \dfrac{(y_j + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \ \lambda_1 \neq 0 \\ \ln(y_j + \lambda_2) & \text{if } \ \lambda_1 = 0. \end{cases} \tag{21}$$

After each optimization, the optimized model was stored in the cloud before creating and implementing fuzzy logic on the results of the optimized classification algorithms as shown in Fig. 2. This involved using the results of the ANN, DT, and NB classifications to generate output using fuzzy rules as shown in Figs. 3 and 4; this output was again stored in the cloud.

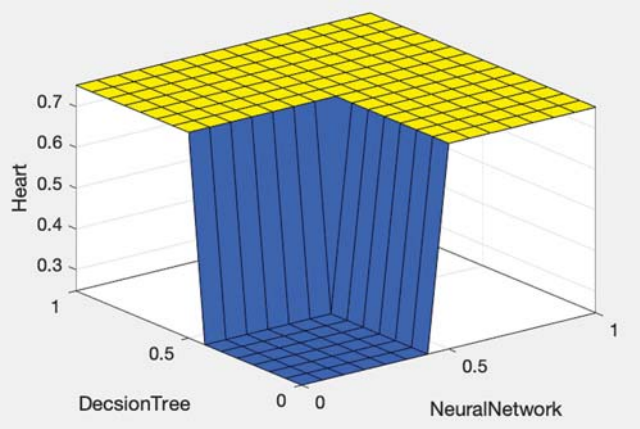

Figure 2: Proposed fuzzy output using the decision tree and artificial neural network classifications

Figure 3: Results showing the presence of heart disease

Conditional (if–then) statements are used to construct fuzzy logic. Fuzzy rules are then constructed based on this logic. In these statements, HD represents heart disease:

IF (ANN is yes, and NB is yes, and DT is also yes) THEN (HD is yes).

IF (ANN is yes, and NB is yes, and DT is no) THEN (HD is yes).

IF (ANN is yes, and NB is no, and DT is yes) THEN (HD is yes).

IF (ANN is no, and NB is yes, and DT is yes) THEN (HD is yes).

IF (ANN is no, and NB is no, and DT is also no) THEN (HD is no).

IF (ANN is yes, and NB is no, and DT is no) THEN (HD is no).

IF (ANN is no, and NB is no, and DT is yes) THEN (HD is no).

IF (ANN is no, and NB is yes, and DT is no) THEN (HD is no).

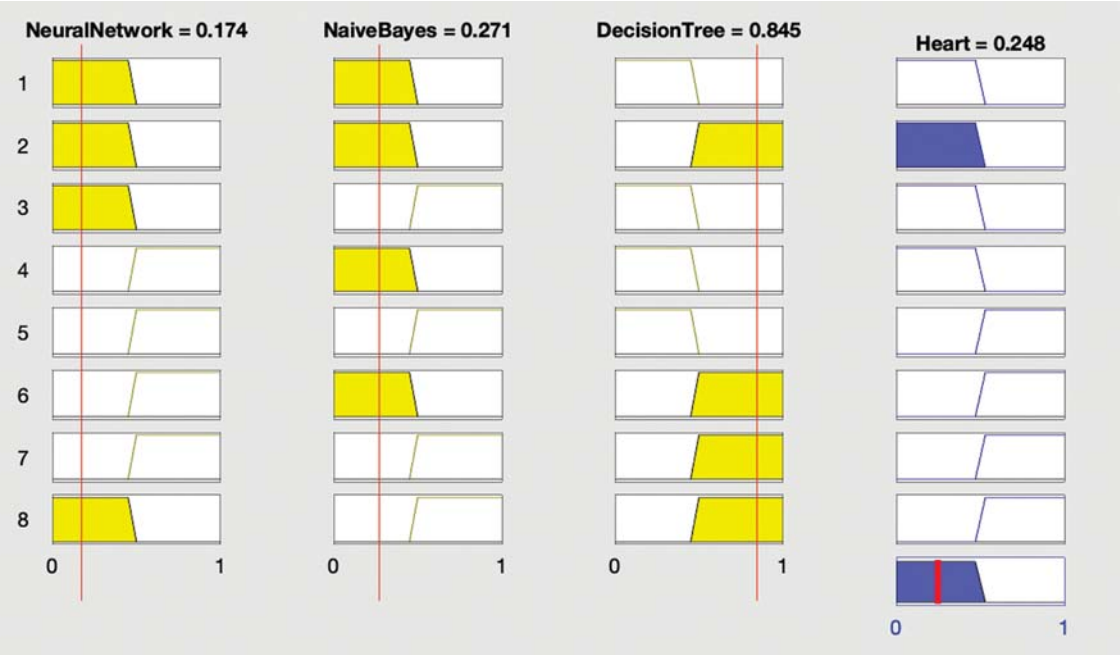

Figure 4: Results showing absence of heart disease

The rules indicate that if any two of the three supervised classification techniques are true then heart disease is considered present; if not, heart disease is not present.

The second layer of the recommended framework concerns the real-time classification of heart disease. Real-time patient data were inputted into the ML-fused model; hypothetically, the results can then be used to schedule appointments. Patients predicted to have cardiovascular disease could be given appointments on an emergency basis; patients predicted to have non-cardiovascular disease could be given a regularly scheduled appointment.

4 Results and Discussion

Each stage systematically interacts with the next stage. We generated a dataset comprising five databases to initiate the model. For greater accuracy, we optimized geodemographic diffusion.

Our experiment comprised 1190 cases and considered 12 attributes shown in Tab. 1. We further refined the data by identifying distorted data, including conflicting records or missing values, after the consolidation of the dataset into a single fuzzy database. At this stage, we eliminated these data to achieve more accurate predictions. Refined data were then classified into two broad categories: testing and training. The training layer was initiated using the selected data, with the three most appropriate ML techniques implemented: ANN, DT, and NB.

Table 1: Cardiovascular data set attributes

No. attributes	Attributes	No. of attributes	Attributes
1	Patient age	7	Resting electrocardiogram
2	Sex	8	Max heart rate
3	Chest pain	9	Exercise angina
4	Blood pressure	10	Old peak
5	Cholesterol	11	ST slope
6	Fasting blood sugar	12	Target

The following mathematical equations were applied to obtain results:

$$\text{Miss rate} = \frac{(R_1/E_0 + R_0/E_1)}{E_0 + E_1} \tag{22}$$

$$\text{Accuracy} = \frac{(R_0/E_0 + R_1/E_1)}{E_0 + E_1} \tag{23}$$

$$\text{Positive Prediction Value} = \frac{R_1/E_1}{(R_1/E_1 + R_0/E_1)} \tag{24}$$

$$\text{Negative Prediction Value} = \frac{R_0/E_0}{(R_0/E_0 + R_1/E_0)} \tag{25}$$

$$\text{Specificity} = \frac{R_0/E_0}{(R_0/E_0 + R_0/E_1)} \tag{26}$$

$$\text{Sensitivity} = \frac{R_1/E_1}{(R_1/E_0 + R_1/E_1)} \tag{27}$$

$$\text{False Positive Ratio} = 1 - \text{Specificity} \tag{28}$$

$$\text{False Negitive Ratio} = 1 - \text{Sensitvity} \tag{29}$$

$$\text{Likelihood Ratio Positive} = \frac{\text{Sensitivity}}{(1 - \text{Specificity})} \tag{30}$$

$$\text{Likelihood Ratio Negative} = \frac{(1 - \text{Sensitivity})}{\text{Specificity}}. \tag{31}$$

First, we used a neural network to classify the data, which involved establishing an ANN structure using 70% of the cases for training data (833 of 1190) and the remaining 30% of cases (357) for testing data. As shown in Tab. 2, 393 of the records used for training were negative and 440 were positive; the training process classified 351 as negative and 400 as positive, which indicates an accuracy of 90.20% and a miss rate of 9.80%. For the testing data, 144 records were negative and 28 were positive, with the testing process producing an accuracy of 85.40% and a miss rate of 14.60%.

The NB classification shown in Tab. 3 classified 337 training records as negative and 366 as positive, which indicates an accuracy of 84.40% and a miss rate of 15.60%. For testing data, NB classified 142 records as negative and 158 as positive, which indicates 84.00% accuracy and a miss rate of 16.00%.

Table 2: Artificial neural network

	Training data			Testing data		
	$N = 833$ (No. of samples)	Result (output) (R_0, R_1)		$N = 357$ (No. of samples)	Result (output) (R_0, R_1)	
INPUT	Expected output (E_0, E_1)	R_0 (Negative)	R_1 (Positive)	Expected output (E_0, E_1)	R_0 (Negative)	R_1 (Positive)
	$E_0 = 393$ (Negative)	351	42	$E_0 = 168$ (Negative)	144	24
	$E_1 = 440$ (Positive)	40	400	$E_1 = 189$ (Positive)	28	161

Table 3: Naïve Bayes

	Training data			Testing data		
	$N = 833$ (No. of samples)	Result (output) (R_0, R_1)		$N = 357$ (No. of samples)	Result (output) (R_0, R_1)	
INPUT	Expected output (E_0, E_1)	R_0 (Negative)	R_1 (Positive)	Expected output (E_0, E_1)	R_0 (Negative)	R_1 (Positive)
	$E_0 = 393$ (Negative)	337	56	$E_0 = 168$ (Negative)	142	26
	$E_1 = 440$ (Positive)	74	366	$E_1 = 189$ (Positive)	31	158

The DT classification shown in Tab. 4 classified 358 training records as negative and 399 as positive, which indicates 90.90% accuracy and a miss rate of 9.10%. For testing data, DT classified 141 records as negative and 174 as positive, which indicates 88.20% accuracy and a miss rate of 11.80%.

Table 4: Decision tree

	Training data			Testing data		
	$N = 833$ (No. of samples)	Result (output) (R_0, R_1)		$N = 357$ (No. of samples)	Result (output) (R_0, R_1)	
INPUT	Expected output (E_0, E_1)	R_0 (Negative)	R_1 (Positive)	Expected output (E_0, E_1)	R_0 (Negative)	R_1 (Positive)
	$E_0 = 393$ (Negative)	358	35	$E_0 = 168$ (Negative)	141	27
	$E_1 = 440$ (Positive)	41	399	$E_1 = 189$ (Positive)	15	174

Subsequent test data records were used for the fuzzy-based system along with the output class to arrive at the final classification. The fuzzy-based system classified 150 records as negative and 176 records as positive (Tab. 5). A comparison of the output of the fuzzy-based system with the expected output revealed an accuracy of 89.30% and a miss rate of 10.70%.

Table 5: Proposed fuzzy model (testing)

$N = 357$ (No. of samples)	Result (Output) (R_0, R_1)	
Expected output (E_0, E_1)	R_0 (Negative)	R_1 (Positive)
$E_0 = 168$ (Negative)	150	18
$E_1 = 189$ (Positive)	13	176

The consolidated results of all classification techniques and the proposed model are presented in Tab. 6. The fuzzy model performed better based on accuracy measurements.

Table 6: Consolidated results

ML algorithm	Type	Specificity (SPEC) %	Sensitivity (SEN) %	False positive value (FPV) %	False negative value (FNV) %	Likelihood ratio positive (LRP)	Likelihood ratio negative (LRN)	Positive prediction value (PPV) %	Negative prediction value (NPV) %
Naïve Bayes	Training	(0.8199) 81.9	(0.8673) 86.7	(0.1800) 18.0	(0.1327) 13.3	4.82	0.16	(0.8318) 83.2	(0.8575) 85.8
	Testing	(0.8208) 82.1	(0.8587) 85.9	(0.1792) 17.9	(0.1413) 14.1	4.79	0.17	(0.8360) 83.6	(0.8453) 94.5
Decision tree	Training	(0.8972) 89.7	(0.9194) 91.9	(0.1028) 10.3	(0.0806) 8.1	8.94	0.09	(0.9068) 90.7	(0.9109) 91.1
	Testing	(0.9038) 90.4	(0.8657) 86.6	(0.0962) 9.6	(0.1343) 13.4	9.00	0.15	(0.9206) 92.1	(0.8393) 83.9
Artificial neural network	Training	(0.8977) 89.8	(0.9049) 90.5	(0.1023) 10.2	(0.0950) 9.5	8.85	0.12	(0.9090) 90.9	(0.8931) 89.3
	Testing	(0.8372) 83.7	(0.8702) 87.0	(0.1628) 16.3	(0.1297) 12.9	5.35	0.15	(0.8519) 85.2	(0.8571) 85.7
Proposed fuzzy model	Testing	(0.9202) 92.0	(0.9072) 90.7	(0.0798) 7.9	(0.0928) 9.3	11.38	0.1	(0.9312) 93.1	(0.8929) 89.3

Further analysis of the model in relation to input parameters was provided by the decision support system. Accordingly, the specific predictions of the three classifiers along with the results derived from the fuzzy-based system are presented in Tab. 7.

Table 7: Prediction comparison of human *vs.* proposed ML approach

INPUT											Human *vs.* ML				
Patient ID	Sex	Chest pain type	Resting PBs	Cholesterol	Fasting blood sugar	Resting electrocardiogram	Maximum heart rate	Exercise angina	Old peak	ST slope	Class	NN	NB	DT	Fuzzy-based system
48	1	2	130	245	0	2	180	0	0.2	2	0	0	0	0	0
44	1	2	120	263	0	0	173	0	0	1	1	1	0	1	1
41	1	2	110	235	0	0	153	0	0	1	0	0	0	0	0
55	0	2	135	250	0	2	161	0	1.4	2	0	0	0	0	0
41	0	2	105	198	0	0	168	0	0	1	0	0	1	0	0
56	1	4	120	85	0	0	140	0	0	1	1	0	0	0	0
52	1	3	172	199	1	0	162	0	0.5	1	0	0	0	0	0
54	1	4	140	239	0	0	160	0	1.2	1	0	0	0	0	0
47	1	3	130	253	0	0	179	0	0	1	0	0	0	0	0
59	1	3	130	318	0	0	120	1	1	2	0	1	1	1	1
54	0	3	110	214	0	0	158	0	1.6	2	0	1	0	1	1
44	0	4	120	218	0	1	115	0	0	1	0	0	0	0	0
54	1	3	133	203	0	1	137	0	0.2	1	1	0	0	0	0
62	1	2	128	208	1	2	140	0	0	1	0	0	0	0	0
52	1	2	128	205	1	0	184	0	0	1	0	0	0	0	0

Finally, the proposed framework is compared with frameworks described in previous research (Tab. 8). The results obtained from the proposed framework in this study is compared with Hybrid random forest linear model (HRFLM) [20], NB [20], DT [20], Support vector machine with the Radial basis function (SVM RBF) [21], Logistic Regression [9], and Framingham Risk Score [9]. The accuracy results of the proposed fuzzy framework are significantly higher than those obtained from previous research.

Table 8: Comparison with state-of-the-art methods

Algorithms	Accuracy (%)	Miss rate (%)
HRFLM [20]	88.40	11.60
Naïve Bayes [20]	75.80	24.20
Decision Tree [20]	85.00	15.00
SVM (RBF) [21]	88.00	12.00
Logistic regression [21]	89.00	11.00
Logistic regression [9]	86.11	13.89
Framingham risk score (FRS) [9]	87.04	12.96
Proposed fuzzy-based ML	91.30	08.70

5 Conclusion

Accurately predicting heart disease using ML techniques is a challenge. This research paper proposed a cloud-based prediction model that used ML techniques. The approach features seven phases: dataset collection, data fusion, pre-processing, training, performance evolution, ML fusion, and real-time testing. Three widely used ML techniques were used: ANN, DT, and NB. The combined results of the ANN, NB, and DT classifications were tested using a fuzzy-based system.

The ratio of training data to testing data was set to 70:30, which enabled accurate prediction. The classification process for all of the techniques was combined with results obtained by the fuzzy-based system, and the processes were conducted until accuracy levels could be observed. The results demonstrated that the proposed fuzzy-based model is 91.30% accurate.

Acknowledgement: We are grateful to our families and colleagues for their emotional support.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] I. Javid, A. K. Z. Alsaedi and R. Ghazali, "Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, pp. 540–551, 2020.

[2] M. E. Pierpont, M. Brueckner, W. K. Chung, V. Garg, R. V. Lacro *et al.*, "Genetic basis for congenital heart disease: Revisited: A scientific statement from the American heart association," *Circulation*, vol. 138, no. 21, pp. 1–12, 2018.

[3] A. Prasanth, M. Bajpei, V. Shrivastava and R. G. Mishra, "Cloud computing: A survey of associated services," *Cloud Computing: Reviews, Surveys, Tools, Techniques and Applications*, vol. 13, pp. 1–15, 2015.

[4] N. C. Reddy, S. S. Nee, L. Zhi Min and C. Xin Ying, "Classification and feature selection approaches by machine learning techniques: Heart disease prediction," *International Journal of Innovative Computing*, vol. 9, no. 1, pp. 10–17, 2019.

[5] H. Xia, I. Asif and X. Zhao, "Cloud-ecg for real time ecg monitoring and analysis," *Computer Methods and Programs in Biomedicine*, vol. 110, no. 3, pp. 253–259, 2013.

[6] A. Hassan, H. M. Bilal, M. A. Khan, M. F. Khan, R. Hassan *et al.*, "Enhanced fuzzy resolution appliance for identification of heart disease in teenagers," in *Int. Conf. on Information Technology and Applications*, Bahawalpur, Pakistan, pp. 28–37, 2019.

[7] S. Kumar and G. Kaur, "Detection of heart diseases using fuzzy logic," *International Journal of Engineering Trends and Technologies*, vol. 4, no. 6, pp. 2694–2699, 2013.

[8] I. P. Atamanyuk and Y. Kondratenko, "Calculation method for a computer's diagnostics of cardiovascular diseases based on canonical decompositions of random sequences," in *Information and Communication Technologies in Education Conf.*, Lviv, Ukraine, pp. 108–120, 2015.

[9] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," *Journal of Healthcare Engineering*, vol. 2017, pp. 1–11, 2017.

[10] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, no. 1, pp. 8869–8879, 2017.

[11] R. R. Kouser, T. Manikandan and V. V. Kumar, "Heart disease prediction system using artificial neural network, radial basis function and case based reasoning," *Journal of Computational and Theoretical Nanoscience*, vol. 15, no. 9, pp. 2810–2817, 2018.

[12] S. Y. Siddiqui, A. Athar, M. A. Khan, S. Abbas, Y. Saeed *et al.*, "Modelling, simulation and optimization of diagnosis cardiovascular disease using computational intelligence approaches," *EAI Endorsed Transactions on Internet of Things*, vol. 5, no. 18, pp. 1–15, 2019.

[13] A. Manimaran, V. M. Chandrasekaran and B. Praba, "A review of fuzzy environmental study in medical diagnosis system," *Research Journal of Pharmacy and Technology*, vol. 9, no. 2, pp. 177–184, 2016.

[14] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang and G. Li, "An integrated decision support system based on ann and fuzzy_ahp for heart failure risk prediction," *Expert Systems with Applications*, vol. 68, no. 10, pp. 163–172, 2017.

[15] B. Zebardast, R. Rashidi, T. Hasanpour and F. S. Gharehchopogh, "Artificial neural network models for diagnosing heart disease: A brief review," *International Journal of Academic Research*, vol. 6, no. 3, pp. 73–78, 2014.

[16] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44–48, 2012.

[17] J. H. O. Keefe, N. Bergman, P. Carrera-Bastos, M. F. Villalba, J. J. D. Nicolantonio *et al.*, "Nutritional strategies for skeletal and cardiovascular health: Hard bones, soft arteries, rather than vice versa," *Open Heart*, vol. 3, no. 1, pp. 1–8, 2016.

[18] M. Siddhartha, "Heart disease dataset (Comprehensive) statlog+Cleveland+Hungary dataset," Kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/sid321axn/heart-statlog-cleveland-hungary-final.

[19] F. Fernandez, "Diabetes from DAT263x lab01, predict who has and who doesn't have diabetes from physical data," Kaggle.com, 2018. [Online]. Available: https://www.kaggle.com/fmendes/diabetes-from-dat263x-lab01.

[20] S. Mohan, C. Thirumalai and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[21] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun *et al.*, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information System*, vol. 2018, pp. 1–12, 2018.