Empirical Study of Intelligence Techniques for Cardio Vascular Disease

Nisha Mary, Bilal Khan, Qaiser Ishfaq, Muhammad Zakir Khan
City University of Science and Information Technology

Abstract: Human life is in troubles due to different types of disease they are facing on daily basis. Mostly, the cardio vascular diseases that is very common in aged people as well in the youngsters too. It is very important to predict such kind of disease in the early stages. For this purpose, a variety of test are conducted to diagnose these diseases. Data mining plays a pivotal role in the prediction of different kind of disease with lesser error rate. This study presents the comparative analysis of different data mining classifiers on heart disease dataset taken from UCI machine learning repository. Going through comparative analysis, nine classifiers, NB, LR, RF, Bagging, RT, ANN, J48, KNN, and SVM are evaluated using MAE, RMSE, RAE, and RRSE evaluation measures. Overall results show that NB performs better in compared with the other aforementioned techniques by reducing error rate considerably.

Keywords: *Heart Disease, Classification Techniques, UCI Machine Learning Repository, Evaluation Measures*

Corresponding authors:*nishaphilomina21@gmail.com, bilalsoft63@gmail.com, qaiser.ishfaq@hotmail.com*

1. Introduction

Heart Failure accounted for bulk of deaths which occurs in old patients and often causes lower quality of life. It is necessary to predict such diseases [1]. The number one cause of death is heart disease entirely [2]. Also estimation is made that deaths would reach up to 15.36% in Pakistan linked to Coronary Heart Disease [3]. There are an estimated 23.6 million People will get affected from cardiovascular disease by the year 2030 [2],[3]. Healthcare data mining solves such problem by putting sensitive real-world data into meaningful data for diagnose and treatment [2]. Beneficial information from large amount of databases can be extracted to contribute data mining, as it is becoming an essential element of putting meaningful information together [4],[5]. In particular, healthcare industry is accountable for bulky records of data [6]. Medical physicians generate bulky data which is not properly being used with secluded information present, and it's not properly being used effectively for prognosis. Due to this purpose, the data gets distilled into an exhaustive dataset using different data mining techniques [7]. These techniques are useful for prognosis of diseases in order to match with particular symptoms in the medical field [5],[6]. Such oppression prediction performs an important role in data mining.

In prognosis of heart disease, it requires cost on medical tests. With the aid of data mining techniques, it can lower the number of tests and thus reduce time consumption requires during treatment. An increasing number of tests might slack performance and time. It is, therefore, a crucial attempt because that permits doctors to negotiate with valuable attributes like age, sex, family history etc. This boosts diagnosing the disease adroitly [8]. Cardio vascular diseases can be predicted via data mining in its initial stage. Researchers are trying to apply the data mining approaches, since few decades they have proposed many techniques for early diagnosis of such diseases. However, the intent of this research is the comparative analysis of contrasting data mining techniques and finding the best technique among the existing techniques for the early prediction of cardio vascular disease with lesser amount of error rate. For benchmarking of existing techniques, this research focuses on MAE, RMSE, RSE, and RAE evaluation metrics.

The remaining of the paper is organized as follows: Exhaustive study of related works is discussed in Sect.2. The dataset and evaluation measures are discussed in Sect. 3. In Sect. 4, overview of employed techniques has been presented. Results and discussion have been figured out inSect.5. The conclusion is presented in Sect.6.

1.1. Problem Background

Since last decade, many data mining and machine learning techniques are tested for cardio vascular diseases prediction. Some of these achieved in term of reducing the error rate in the evaluation metrics and improving the performance in term of accuracy. But still there is uncertainty in the evaluated data and

the results need to be improved. Researchers have mentioned that we can use some latest data mining or machine learning algorithms to improve the results according to our desire. They have mentioned that it can be done through improving the existing algorithm, hybridization, or proposing new solution.

2. Related Work

According to information gain from literature review, there are various data mining techniques for diagnosing and predicting cardiovascular disease with less error rates and higher accuracy. Different types of studies have been conducted to target the prediction of heart disease, which includes the following related work:

In 2015, Shafique, Umair Campus, Lyallpur et al. [3] perform comparison among data mining algorithm like J48, Artificial Neural Network and Naïve Bayes. The dataset obtained from UCI which contains 13 features and concluded Naive Bayes and J-48 has accuracies 82.9 %, 77.2 %. Evaluation metrics is based on Accuracy, True Negative and True Positive rate, ROC curve and the time.

In 2015, Kim, Jaekwon Lee, Jongsik Lee, Youngho et al. [9] project risk prediction for prognosis of Coronary Heart disease in Korea. The projected model is based on a rule base and fuzzy membership Functions to contribute a fuzzy-logic-based inference. There were nine attributes used as dataset obtained from UCI. Confusion Matrix was used for evaluation criteria and concluded that accuracy is 69.51% .The proposed solution is excelling than ANN and SVM. Also comparison performed among ANN, SVM, logistics regression, and decision tree C5.0.Future work is to enhance the specificity and accuracy.

In 2015, Dewan, Ankita Sharma, Meghna et al. [10] propose hybrid technique (Back Propagation Technique, Genetic Algorithm) with ability to solve complex skepticism which is indispensable for prognosis of cardiac disease which may help doctors diagnose the condition. The proposed hybrid technique is based on dataset which is capture from UCI repository having 13 Attributes. Evaluation matrix used are Accuracy and sensitivity. Comparisons perform among decision tress, Naive Bayes, SVM& ANN. Results show ANN is best among all the classification for non-linear data. No future work discussed.

In 2016, Verma, Luxmi Srivastava, Sangeet Negi, P. C.et al. [11] propose an amalgam methodology to anticipate Congestive Artery Disease. The Hybrid method is feature selection with further techniques like particle swam optimization (PSO) search method, K-means clustering algorithms, Multi-layer perceptron (MLP), multinomial logistic regression (MLR), fuzzy unordered rule induction algorithm (FURIA) and C4.5. The amalgam solution is based on dataset taken from UCI repository comprises of 22 attributes. Evaluation criteria include accuracy and concluded Multinomial logistic regression (MLR) has highest prognosis accuracy of 88.4 %. Future work include enhancement of accuracy with more data instances.

In 2016, R, Theresa Princy Thomas et al. [12] provide the survey around contrasting classification techniques used for prognosis the peril level individually. Classifier module is trained via KNN. In prognosis module data is tested and predicted via ID3. Input attributes, key attribute and Prediction Attributes were used obtained from UCI.Evalaution criteria is accuracy and concluded that using KNN and ID3, the peril rate of heart disease was detected successfully. Future work include enhancement in accuracy with reduced number of attributes.

In 2019, Muppalaneni, Naresh BabuMa, Maode Gurumoorthy, Sasikumar et al. [13] compare Random Forest, SVM, Logistic Regression, gradient boosting with receiver operating characteristic (ROC) curve for prognosis and identification of congestive problem. The 14 attributes obtained from UCI Cardiac Datasets. Results show logistic regression is superior to others with accuracy 87%.Future work target prediction using tensor flow of deep learning algorithms with further dataset.

In 2019, Diller,A. Kempny,S.Babu Naryan et al. [14] assess utility of ML Algorithms on measuring prognosis and maneuver therapy in adult congenital heart disease (ACHD). ECG parameters, cardiopulmonary exercise testing, and selected laboratory markers where composed and incorporate in deep learning (DL) algorithms. For assessment of paper, data was accessed from multi-institutional datasets. Evaluation criteria is accuracy. Results show Deep Learning algorithms can mount to multi-institutional datasets to enhance accuracy ultimately.

In 2019, Diego R. Mazzotti, Brendan T. Keenan, Ryan Urbanowicz, Allan I. Pack et al. [15] evaluate the performance of different supervised machine learning methods like Naive Bayes (NB), Logistic Regression (LR), ANN, Decision Tree, Random Forest, Extreme Gradient Boosting, K-Nearest Neighbors, and Support Vector Machine (SVM) to predict Congestive disease. For evaluation of classifiers, data obtained from UCI Cardiac Datasets. Comparison performed among Naives Bayes, Logistic Regression, elastic-net regularized general linear model (enGLM), ANN, decision tree, extreme gradient boosting, KNN, and SVM. All methods, except SVM, had an irresistible progress in disease prediction.

In 2019, Shamsollahi, M Badiee, A Ghazanfari et al. [16] project a model developed using the blend of descriptive and predictive techniques of data mining. First, the number of clusters is determined using clustering indices. Next, some types of decision tree methods and artificial neural network are applied to each cluster. The data for given model used in this work is real, collected from a heart clinic database. The results achieved demonstrate that the C&RT decision tree method performs best on all the data used in this work with 0.074 error.

In 2019, Kodati, Sarangam Vivekanandam, R Ravi et al. [8] perform resemblance of unsupervised clustering algorithms like farthest first, filtered cluster hierarchical cluster, OPTICS, simple k-means

approach which are evaluated in Weka tool and discover which algorithms may be most promising for the heart disease dataset. Evaluation Criteria include Time , true fine and real negative values .Outcomes demonstrate that off first clustering algorithm demand minimum time performed to form the cluster and also it is the simplest k-means algorithm as compared to the other algorithms. Future work focuses that it can incorporate other medical attributes.

In 2019, Makumba, Dominic Obwogi Cheruiyot, Wilson Ogada, Kennedy et al. [17] propose model has been developed to support decision making in heart disease prognosis based on data mining techniques. The Decision Tree, Naive Bayes, KNN (K-Nearest Neighbors) and WEKA API (Waikato Environment for Knowledge Analysis application programming interface) were the various data mining methods that were used. Data for proposed model has been accessed from UCI with 13 attributes. Evaluation criteria are based on Confusion matrix. Results show that projected technique gives the exact conclusion of coronary illness than the current strategies.

3. Dataset and Evaluation Measures

The comparative analysis is heavily based on dataset procure from UCI Machine Learning Repository for scrutiny of machine learning algorithms. The database contains 303 instances with 14 features that are listed in Figure 2.

Table 1 List of Attributes

S No	Variables	Description
1	Age	Age of patient in Year
2	Sex	Sex of patient (value 1: Male; value 0: Female)
3	Cp	chest pain type (value 1: typical type 1 angina, value 2: typical type angina, value3: non-angina pain; value 4: asymptomatic)
4	Trestbps	resting blood pressure
5	Chol	cholesterol level
6	Fbs	Fasting blood sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
7	Trestecg	maximum heart rate achieved
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina (value 1: yes; value 0: no)
10	Oldpeak	ST depression induced by exercise relative to rest(1-3)
11	Slope	the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping)
12	ca	number of major vessels colored by fluoroscopy (value 0-3)
13	Thal	Normal, fixed defect, reversible defect (value 3: normal; value 6: fixed defect; value 7: reversible defect)
14	Target	Absence, Presence

3.1. Evaluation Measures

Model evaluation is the core task of any research work. To achieve satisfactory results of any model it is important to evaluate it with some standard evaluation measures/models. In this study the following evaluation metrics are used:

3.1.1. Mean Absolute Error (MAE)

Mean absolute error can be determine by taking disparity of continuous variables, for instance predicted and observed values, final time versus initial time [18].

3.1.2. Relative Absolute Error (RAE)

Absolute error and exploratory or experimental values are the two variables on which relative absolute error relies on. To measure the relative error, these two criterions must be recognized. Relative error is obtained by the ratio of absolute error and the experimental value. Percentage or fraction are used to indicate relative absolute error because it has no units [18].

3.1.3. Root Mean Squared Error (RMSE)

The difference between values predicted by a model or an estimator and the values observed can be used to calculate root-mean-square deviation (RMSD) or root-mean-square error (RMSE). It is a generally used measure. The RMSD expresses the square root of the second instance moment of the differences between prognosis values and recognized values or taking the quadratic mean of these differences, known as residuals. When computed out-of-sample, the calculations execute over the data instances for estimation, known as errors. The RMSD serve to aggregate the weightage of the errors in predictions for various times into a single measure [19].

3.1.4. Root Relative Squared Error (RRSE)

The root relative squared error is related to elementary predictor. Elementary predictor can be determined by taking the average or mean of the actual values. By taking the overall squared error and dividing by the total squared error of the simple or elementary predictor, the error can be reduced to the some dimensions as the quantity being predicted by taking the square root of the relative squared error [18].

4. Overview of Employed Techniques

Going through several published literatures, it was found that there are numerous numbers of classifiers used nowadays. For this, we compare classifiers, namely Naive Bayes, J48 of Decision Tree, Bagging algorithm, REPTree, K-Nearest Neighbor, Artificial Neural Network, Random Forest, Logistic Regression and Support vector machine. Depending on comparisons made among them, some classifiers outperformed than others. The overview of some algorithms for comparative analysis is discussed below.

4.1. Naive Bayes

Bayesian Theorem provides the basis of Naive Bayes. In this, individual parameters contribute independently to the probability. For instance fruit is an apple contributes independently the probability of apple, despite of any possible correlations between the color, roundness, and diameter features for classification. For classifying spatial datasets, the naive bayes algorithm is preferable. This approach performs conditional independency. An attribute value is independent of other attributes to estimate conditional independency, So that it proves fruitful for investigating and getting information [17].

4.2. J48 Decision Tree

Exploring attributes-vector behavior or interact for a number of objects is Decision Tree Algorithm. Producing rules and constraints for the prediction of the target variable is performed by this algorithm. J48 is a continuation of ID3. Exploring missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. are advanced features of J48. By comparing, classification of other algorithms is performed recursively until every single leaf is pure and accountable for perfect data. The intent is to generalize decision tree more and more until it gains balance between flexibility and accuracy [20].

4.3. Multilayer Perceptron

Multi-layer Perceptron (MLP) is a feed-forward artificial neural network technique. The artificial neurons are called perceptron's. There are multiple tiers including hidden layers in artificial neurons are employed in the multilayer perceptron algorithm. A perceptron uses an activation function for each neuron. The activation function sketches the weighted inputs of each neuron causing the number of layers reduced to two layers. Activation function is applied on non-linear or broad except the input nodes. Back propagation for training is used. These algorithms are mainly used for binary classification problems [21].

4.4. K-nearest Neighbors

K-Nearest Neighbors (KNN)'s deals with majority voting among neighbors which is a kind of occurrence-based learning. Its arrangement involves class cooperation and arranged by ordering based on leading part vote of its neighbors. Due to their large number, they tend to be familiar among the k-nearest neighbors. For statistical measurements and pattern recognition, K-Nearest Neighbors have been used [17].

4.5. Support Vector Machine

In the Support Vector Machine (SVM) model, the examples are represented as points in space and designed or mapped. Each level is divided by a clear difference or gap. New examples are mapped into the same space and predict to which category they belong based on side of the gap they contain [22].

4.6. REPTree

REPTree (RT) creates multiple trees in different repetitions. After that it chooses best one from all generated trees. It uses the regression tree logic. Using information gain, REPTree builds a decision tree. The measure used is the mean square error on the predictions made by the tree is used to trim the trees which is also called pruning the tree. It is an agile decision tree learner. The tree is pruned using reduced error pruning with back fitting [23].

4.7. Bagging

To enhance the equilibrium and accuracy of machine learning algorithms used in Statistical classification and regression, Bagging is designed. To avoid over fitting by reducing variance. Employing bagging on classifiers especially on decision trees, neural networks enhance accuracy of classification. Bagging plays indispensable role in the domain of heart disease diagnosis [24].

4.8. Logistic regression

To evaluate the features of a logistic model is logistic regression to probe the probability of a certain class or event. Logistic regression algorithm is a regression and classification method for examining the dataset in which it contains one or more independent variables that conclude an outcome. Divided variable is used to measure the outcome [13].

4.9. Random Forests

The way of averaging or obtaining mean from several deep decision trees with the intent of reducing the variance, trained on different parts of the same training set, is known as Random Forest. It comprises a forest by random tree. It contains a direct relationship and dependency between the numbers of trees in the forest, in addition to the outcomes it can get. It can be used for both classification and regression tasks. Over fitting is excluded in this classifier. In case there are enough trees in the forest, the classifier will not over fit the model. Exploring missing values, categorical values can be modeled on Random Forest classifiers are some of its features [13].

5. Results and Discussions

This section presents results and analysis performed on machine learning algorithms. In this work, nine classifiers including NB, J48, RT, RF, LR, Bagging, SVM, KNN and NN are conducted. Data divided into trainset that is 80% and testset that is 20% respectively. The training set is used to build the classifier and test set used to validate it. Different error rates are measured which is merely much more valuable for judgment criteria.

The experimental results show that NB classifier perform well as compare to other classifiers in reducing the error rate that are 63.25% for RRSE, 34.29% for RAE, 0.31 for RMSE and 0.17 for MAE as shown in Table 2.

Table 2 Experimental Results of Classifiers Employed

S No	Technique	RRSE %	RAE %	RMSE	MAE
1	NB	63.25	34.29	0.31	0.17
2	LR	68.26	46.4	0.344	0.23
3	RF	69.82	49.74	0.35	0.24
4	Bagging	73.05	50.54	0.36	0.25
5	RP	73.88	51.98	0.37	0.26
6	ANN	84.33	41.8	0.425	0.209
7	J48	89.67	53.17	0.45	0.26
8	KNN	107.29	59.24	0.54	0.29
9	SVM	141.38	101.44	0.712	0.508

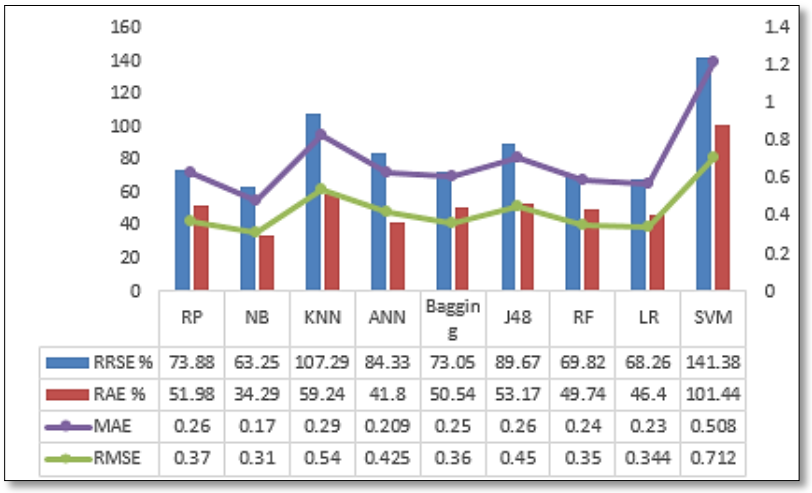

Figure 1 Graphical Representation of Experimental Results

Table 3 show the percentage difference of error rate of NB with respect to another classifier. That is also shown graphically in Figure 2.

Table 3 Percentage Difference of NB as compare to Rest of used Classifiers

S No	Techniques	Diff in RRSE	Diff in RAE	Diff in RMSE	Diff in MAE
1	NB and LR	7.61%	30%	10.40%	30%
2	NB and RF	9.87%	36.78%	12.12%	34.14%
3	NB and Bagging	14.40%	38.31%	14.92%	38.09%
4	NB and RT	15.50%	41.01%	17.64%	41.86%
5	NB and ANN	28.60%	19.73%	31.30%	20.60%
6	NB and J48	34.55%	43.20%	36.84%	41.90%
7	NB and KNN	51.64%	53.35%	54.11%	52.17%
8	NB and SVM	76.36%	98.94%	78.70%	99.70%

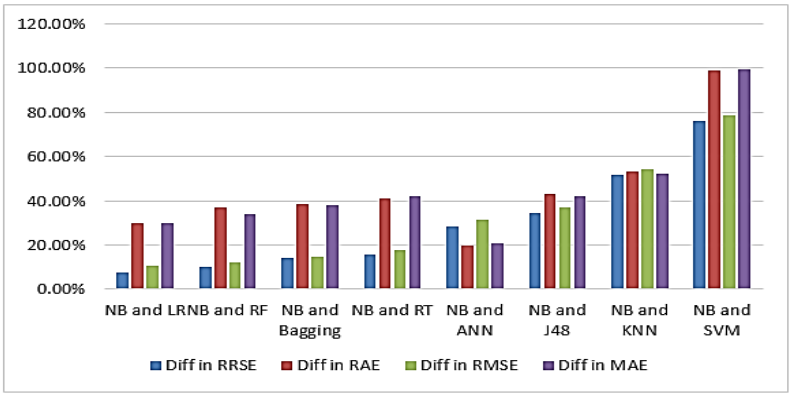

Figure 2 Percentage Difference of NB and other Classifiers

6. Conclusion

This paper focuses on the comparison of different classification techniques on heart disease dataset selected for UCI machine learning repository. Going through comparative analysis, nine classification algorithms are employed that are listed in Table 2. The evaluation results show that using the selected dataset, NB performance is achievable as compare to the rest of used classification algorithms. The difference between the error rates of NB as compare to others for all evaluation measures are significant which is listed in Table 3 also graphically shown in Figure 2 for easy consideration.

7. Acknowledgements

I would like to thanks Mr. Bilal Khan for his support, also I would like to thank USJ for their cooperation.

References

[1] B. Venkatalakshmi and M. V Shivsankar, "Heart Disease Diagnosis using Predictive DataMining," vol. 3, no. 3, 2014.

[2] V. Chaurasia and S. Pal, "Data Mining Approach to Detect Heart Diseases," *Int. J. Adv. Comput. Sci. Inf. Technol.*, vol. 2, no. 4, pp. 56–66, 2014.

[3] U. Shafique and L. Campus, "Data Mining in Healthcare for Heart Diseases," *Int. J. Innov. Appl. Stud.*, vol. 10, no. 4, pp. 1312–1322, 2015.

[4] D. Chandna, "Diagnosis of Heart Disease Using Data Mining Algorithm," vol. 5, no. 2, pp. 1678–1680, 2014.

[5] M. A. Nishara Banu and B. Gomathy, "Disease forecasting system using data mining methods," *Proc. - 2014 Int. Conf. Intell. Comput. Appl. ICICA 2014*, pp. 130–133, 2014.

[6] K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.

[7] M. A. M. Hlaudi Daniel Masethe, "Prediction of Heart Disease using Classification Algorithms," vol. 1, pp. XIII–XIV, 2014.

[8] S. Kodati, R. Vivekanandam, and G. Ravi, *Comparative Analysis of Clustering Algorithms with Heart Disease Datasets Using Data Mining Weka Tool*, vol. 898. Springer Singapore, 2019.

[9] J. Kim, J. Lee, and Y. Lee, "Hir-21-167," vol. 21, no. 3, pp. 167–174, 2015.

[10] A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," *2015 Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2015*, pp. 704–706, 2015.

[11] L. Verma, S. Srivastava, and P. C. Negi, "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data," *J. Med. Syst.*, vol. 40, no. 7, 2016.
[12] T. P. R and J. Thomas, "Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2016," *Proc. IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT 2016*, 2016.
[13] N. B. Muppalaneni, M. Ma, and S. Gurumoorthy, *Soft Computing and Medical Bioinformatics*. Springer Singapore, 2019.
[14] G. P. Diller *et al.*, "Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: Data from a single tertiary centre including 10 019 patients," *Eur. Heart J.*, vol. 40, no. 13, pp. 1069–1077, 2019.
[15] B. C. S. Science, "B. Clinical Sleep Science and Practice VII. Pediatrics," vol. 41, pp. 292–293, 2018.
[16] M. Shamsollahi, A. Badiee, and M. Ghazanfari, "Using Combined Descriptive and Predictive Methods of Data Mining for Coronary Artery Disease Prediction: a Case Study Approach," *J. AI Data Min.*, vol. 7, no. 1, pp. 47–58, 2019.
[17] D. O. Makumba, W. Cheruiyot, and K. Ogada, "A Model for Coronary Heart Disease Prediction Using Data Mining Classification Techniques," vol. 3, no. 4, pp. 1–19, 2019.
[18] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim. Res.*, vol. 30, no. 1, pp. 79–82, 2005.
[19] F. Collopy and J. Armstrong, ""Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons," *Int. J. Forecast.*, vol. 8, pp. 69–80, 1992.
[20] Y. Gultepe, "The Use of Data Mining Techniques in Heart Disease Prediction," vol. 8, no. 4, pp. 136–141, 2019.
[21] M. Majumder, *Artificial Neural Network*. 2015.
[22] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics and Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.
[23] J. P. A. Jebamalar, S. Paul, and D. P. P. Latha, *Mining Classification Algorithms : A Comparative Study*. Springer Singapore, 1882.
[24] S. Moral-García, C. J. Mantas, J. G. Castellano, M. D. Benítez, and J. Abellán, "Bagging of credal decision trees for imprecise classification," *Expert Syst. Appl.*, vol. 141, p. 112944, 2020.