An Experimental Study of Various Machine Learning Approaches in Heart Disease Prediction

Md. Shafiul Azam
Dept. of CSE
Pabna University of Science and Technology, Bangladesh

Md. Abu Raihan
Dept. of CSE
Pabna University of Science and Technology, Bangladesh

Humayan Kabir Rana
Dept. of CSE
Green University of Bangladesh, Bangladesh

ABSTRACT

According to recent survey of WHO (World Health Organization) 17.9 million people die each year because of heart related diseases and it is increasing rapidly. With the increasing population and diseases, it has become challenging to diagnosis and treatment diseases at the right time. But there is a light of hope that recent advancements in technology have accelerated the public health sector by advanced functional biomedical solutions. This paper aims to analyze the various machine learning approaches namely Naïve Bayes (NB), Random Forest (RF) Classification, Decision tree (DT), Support Vector Machine (SVM) and Logistic Regression (LR) by employing a qualified dataset for heart disease prediction. This research finds the correlations between the various attributes that are suitable to predict the chances of a heart disease and compares the impact of Principle Component Analysis (PCA) on the accuracy of the above mentioned algorithms.

Keywords

Heart Disease, Machine Learning Algorithms, PCA, Decision Tree, SVM, Random Forest, Logistic Regression, Naïve Bayes

1. INTRODUCTION

The heart is an important organ of human body part. It is nothing more than a pumper, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. The term heart disease refers to the diseases of heart and blood vessel system within it. A number of factors have been shown that increases the risk of heart disease includes, family history, smoking, poor diet, high blood pressure, high blood cholesterol, obesity, physical inactivity, and hyper tension [1,2].

Heart disease is the single largest cause of death world-wide, according to WHO approximately 17.9 million people die each year globally because of Cardiovascular Diseases (CVD) and Four out of five CVD deaths are due to heart attacks and strokes. There are more than 20-fold variations in heart disease mortality rates between countries. Highest mortality rates are in Eastern Europe and Central Asian countries; lowest rate in higher income country. For the working age population, heart disease mortality rates are markedly higher in low and middle income countries rather than higher income countries [3].

Diagnosis and treatment of heart disease is very complex, particularly in developing countries, due to the lack of diagnostic devices and a shortage of physicians and other resources affecting proper prediction and treatment of cardiac patients. With this concern in the recent times computer and machine learning techniques are being used to develop software to assist doctors in making decision of heart disease in the preliminary stage. Early stage detection of the disease and predicting the probability of a person to be at risk of heart disease that can reduce death rate. Medical information has redundancy, multi-attribution, incompleteness and a close relationship with time. The problem of using the massive volumes of data effectively becomes a major problem for the health sector. Machine Learning provides the methodology and technology to convert these data mounds into useful decision-making information. This prediction system for heart disease would facilitate cardiologists in taking quicker decisions so that more patients can receive treatments within a shorter period of time, resulting in saving millions of life [4]. This influenced the research work done in this paper.

This paper intends to adopt five machine learning algorithms (NB, RF, DT, SVM and LR) that can be used to predict heart disease. In this work the impact of PCA on the accuracy of the above mentioned machine learning algorithms are examined.

2. MATERIALS AND METHODS

2.1 Description of the dataset

The Cleveland heart dataset from the UCI machine learning repository has been used for the experiments. The dataset consists of 14 attributes and 303 instances. There are 8 categorical attributes and 6 numeric attributes. The description of the dataset is shown in Table 1 [5].

S.No	Attribute Name	Description	Range of Values
1	Age	Age of the person in years(29-79)	29 to 79
2	Sex	Gender of the person [1: Male, 0: Female]	0, 1
3	Cp	Chest pain type [1-Typical Type 1 Angina 2- Atypical Type Angina 3-Non-angina pain 4-Asymptomatic)	1, 2, 3, 4
4	Trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	Chol	Serum cholesterol in mg/dl	126 to 564
6	Fbs	Fasting Blood Sugar in mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 to 202
9	Exang	Exercise Induced Angina	0, 1
10	OldPeak	ST depression induced by exercise relative to rest	1 to 3

S.No	Attribute Name	Description	Range of Values
11	Slope	Slope of the Peak Exercise ST segment	1, 2, 3
12	Ca	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3, 6, 7
14	Num	Class Attribute	0 or 1

Table 1. Feature information of the Cleveland dataset

Patients from age 29 to 79 have been selected in this dataset. Male patients are denoted by a gender value 1 and female patients are denoted by gender value 0 [5]. Four types of chest pain can be considered as indicative of heart disease. Type 1 angina is caused by reduced blood flow to the heart muscles because of narrowed coronary arteries. Type 1 Angina is a chest pain that occurs during mental or emotional stress. Nonangina chest pain may be caused due to various reasons and may not often be due to actual heart disease. The fourth type, Asymptomatic, may not be a symptom of heart disease. The next attribute trestbps is the reading of the resting blood pressure. Chol is the cholesterol level. Fbs is the fasting blood sugar level; the value is assigned as 1 if the fasting blood sugar is below 120 mg/dl and 0 if it is above. Restecg is the resting electrocardiographic result, thalach is the maximum heart rate, exang is the exercise induced angina which is recorded as 1 if there is pain and 0 if there is no pain, oldpeak is the ST depression induced by exercise, slope is the slope of the peak exercise ST segment, ca is the number of major vessels colored by fluoroscopy, thal is the duration of the exercise test in minutes, and num is the class attribute. The class attribute has a value of 0 for normal and 1 for patients diagnosed with heart disease.

2.2 Methodologies

This paper intends to adopt five (5) machine learning algorithms include Decision tree, Naïve Bayes, SVM, Logistic Regression and Random Forest for the prediction of effective heart disease. It also compares accuracy and results between them. In this work the above mentioned machine learning algorithms are used with and without Principal component analysis (PCA) to examine the accuracy. The system architecture is given in the following figure 1.

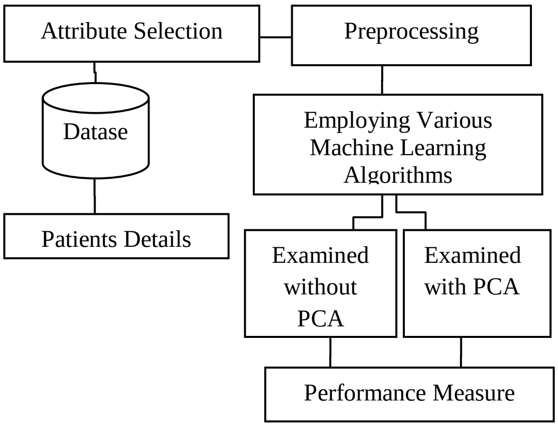

Figure 1: System Architecture used in this study

2.2.1 Principal Component Analysis (PCA)

PCA or Principal Component Analysis is a dimensionality reduction technique that uses orthogonal transformation to reduce a large set of variables to a smaller set of variables while retaining most of the information present. The procedure works by converting the highly correlated variables of the original dataset to a smaller number of uncorrelated linear variables called principal components. These principle components then account for most of the variance of the original dataset [6,7]. PCA is very useful when the data has dimensions of 3 or higher since it becomes extremely hard to make predictions from such huge amount of information. Number of principal components is less than or equal to the smaller of the number of original features or the number of observations. Initially, the dataset comprised 303 observations and 14 useful features, so the maximum number of principal components was 14. Out of those 14, 2 components were taken which attributed to 80% of the original variance as shown in figure 2.

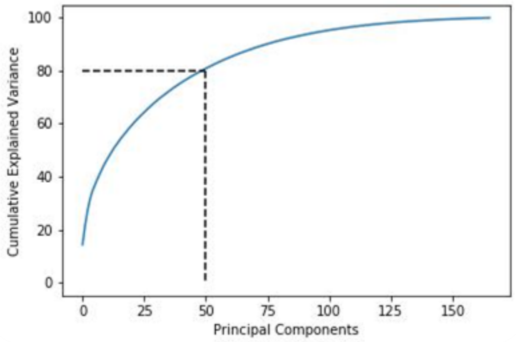

Figure 2: Cumulative Explained Variance against No. of Principal Components

After applying PCA the dataset is reduced to 2 principal components that represent each instance as shown in figure 3.

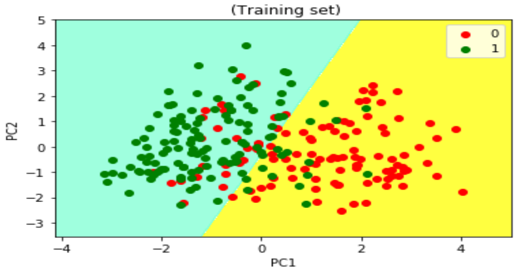

Figure 3: Visualization of Principal Components

2.2.2 Logistic Regression

Logistic Regression is a supervised learning algorithm that trains the model by taking input variables (x) and a target variable (y). In Logistic Regression the output or target variable is a categorical variable, unlike Linear Regression, and is thus a binary classification algorithm that categorizes a data point to one of the classes of the data. The general equation of Logistic Regression is-

$$log(p(X)/(1-p(X)\,) = \beta_0 + \beta_1\, X \dots\dots\dots\dots (1)$$

Where, p(X) is the dependent variable, X is the independent variable, β_0 is the intercept and β_1 is the slope co-efficient [8]. Logistic Regression measures the relationship between the dependent variable, the output, and the independent variables, the input, by estimating probabilities using its underlying logistic function. It uses L2 penalty for regularization. The resultant probabilities are then converted to binary values 0 or 1 by the logistic function, also known as the sigmoid function. The sigmoid function takes any real-valued number and maps it into a value between the ranges 0-1 excluding the limits themselves. Afterwards, a threshold classifier transforms the result to a binary value.

2.2.3 Support Vector Machine (SVM)

Recognition is performed by Support vector machine (SVM) that works on the basis of the principle of structural risk minimization. SVM is treated as binary classifier that separates the two classes of data optimally. The two major aspects of developing SVM as a classifier is: (i) to determine the optimal hyperplane in between two separate classes of data and (ii) to transform the non-linearly separable classification problem into linearly separable problem [9,10,11]. Linearly separable classification problem is shown in figure 4 as for an example.

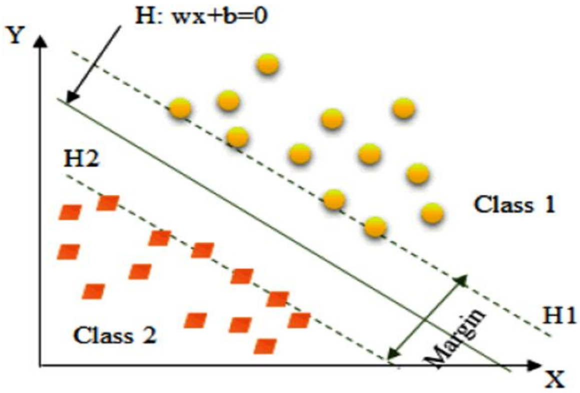

Figure 4: SVM with Linear separable data

Let x is a set of input feature vector and y is the class label. The input feature vectors and the class label can be represented as $\{x_{i,}y_i\}$, where i = 1, 2 . . . N and y = ± 1. The separating hyperplane can be represented as follows,

$$\mathrm{w.x + b = 0} \dots\dots\dots\dots\dots\dots (2)$$

This implies,

$$\mathrm{y_i(w.x_i + b) \geq 1;\ \ i = 1,2,3 \dots N} \dots\dots\dots\dots\dots\dots (3)$$

$\{\mathrm{w, b}\}$ can have numerous possible values which create separating hyperplane. It is believed that points often lie between two data classes in such a way that there is always some margin in between them. SVM maximizes this margin by considering it as a quadratic problem [12]. The SVM is used to make two possible decisions during prediction.

2.2.4 Naive Bayes (NB)

The Naïve Bayes classifier or simply, the Bayesian classifier is based on the Bayes theorem [13]. It is a special case of the Bayesian network, and it is a probability based classifier. In the Naïve Bayes network, all features are conditionally independent. The change in one feature therefore does not affect another feature. The Naïve Bayes algorithm is suitable for classifying high dimensional datasets. The classifier algorithm uses conditional independence. Conditional independence assumes that an attribute value is independent of the values of the other attributes in a class.

Let D be a set of training data and associated class labels. Each tuple in the dataset is defined with n attributes that are represented by $\mathrm{X = \{A1, A2, \dots\dots, A_n)}$. Let there be m classes represented by $\mathrm{C1, C2, \dots C_m}$. For a given tuple X, the classifier predicts that X belongs to the class having the highest posterior probability, conditioned on X. The Naïve Bayes classifier predicts that the tuple X belongs to the class Ci if and only if

$$\mathrm{P(Ci|X)} > P\mathrm{(Cj|X)for1 \leq j \leq m, j \neq i} \dots\dots\dots\dots\dots (4)$$

Thus, P(Ci|X) is maximized. The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. According to Bayes' theorem,

$$\mathrm{(P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}} \dots\dots\dots\dots\dots. (5)$$

If the attribute values are conditionally independent of one another,

$$\mathrm{P(X|Ci) = \prod_{k=1}^{n} P\,(x_k|Ci)} \dots\dots\dots\dots\dots. (6)$$

Where $\mathrm{x_k}$ refers to the value of attribute $\mathrm{A_k}$ for tuple X.

If $\mathrm{x_k}$ is categorical, then $\mathrm{P(x_k|C_i)}$ is the number of tuples of class $\mathrm{C_i}$in D having the value $\mathrm{x_k}$ for $\mathrm{A_k}$, divided by |$\mathrm{C_i}$,D|, the number of tuples of class $\mathrm{C_i}$ in D. The classifier predicts the class label of X is the class $\mathrm{C_i}$ if and only if,

$$\mathrm{P(X|Ci)P(Ci)} > P\mathrm{(X|Cj)P(Cj)for1 \leq j \leq m, j \neq i} \dots\dots\dots (7)$$

Bayesian classifiers are effective in the sense that they have the minimum error rate for classification [14,15].

2.2.5 Decision Tree (DT)

Decision Tree is a supervised learning algorithm that is used for classification and regression. It works by splitting the data into two or more subsets based on the values of the input variables. A cost function or splitting criterion is used to determine the best split (one with the lowest cost) among all the split points [16]. The data is split recursively into groups until the leaves contain only one sample. In this model, an optimized version of the CART (Classification and Regression Trees) algorithm is used to implement the Decision Tree classifier using Scikit-Learn [16]. Gini impurity is used as the splitting criterion to measure the uncertainty. Decision Trees can handle both categorical and numerical variables as input so it is appropriate for this model, since the dataset contains both variable types. In this model, the relationship between the feature variable and target variable is complex and highly non-linear. So a Decision Tree has a greater chance of outperforming linear models like Logistic Regression [17].

2.2.6 Random Forest Classifier (RF)

Random forest is a tree based classification algorithm. As the name indicates, the algorithm creates a forest with a large number of trees. It is an ensemble algorithm which combines multiple algorithms. It creates a set of decision trees from a random sample of the training set [18]. It repeats the process with multiple random samples and makes a final decision based on majority voting. The Random forest algorithm is effective in handling missing values but it is prone to over fitting. Appropriate parameter tuning can be applied to avoid over fitting [19]. The algorithm for Random forest is shown in table 2.

Let D be a training set, $D = \{(x_{1,}y_1), \dots\dots (x_{n,}y_n)\}$
Let $h = h_1(x), h_2(x), \dots\dots h_k(x)$ *an ensemble of weak classifiers*
If each h_k *is a decision tree, the parameters of the tree are defined as* $\theta = (\theta_{k1,}\theta_{k1}, \dots\dots \theta_{kp,})$
Each decision tree k leads to as classifier $h_k(X) = h_k(X|\theta_k)$
Final Classification $f(x) = \text{Majority of } h_k(X)$

Table 2: Random forest algorithm [18, 19]

3. EXPERIMENT SETTING & RESULTS

After employing the different Machine Learning algorithms, the next step is to find out how the models performed. This is done by running the models on the test dataset which was set aside earlier. The test dataset comprised of 30% of the original data for heart disease. To determine and compare the performance of different algorithms, several performance metrics were used.

3.1 Performances Metrics

The performance of Machine Learning algorithms is evaluated using several performance metrics. Performance metrics relating to classifications are discussed here since the paper only deals with classification problems. For heart disease, if the target variable (risk of heart disease) is 1 then it is a positive instance, meaning the patient has heart disease. And if the target variable is 0, then it a negative instance, meaning the patient does not have heart disease.

3.1.1 Confusion Matrix

Confusion Matrix is the easiest way to determine the performance of a classification model by comparing how many positive instances were correctly/incorrectly classified and how many negative instances were correctly/incorrectly classified. In a Confusion Matrix, the rows represent the actual labels and the columns represent the predicted labels shown in the following table 3.

	Predicted Positive(has heart disease)	Predicted Negative(no heart disease)
Actual Positive (has heart disease)	TP	FN
Actual Negative (no heart disease)	FP	TN

Table 3. Confusion Matrix

True Positives (TP): True positives are the instance where both the predicted class and actual class is True (1), i.e., when patient actually has complications and is also classified by the model to have complications.

True Negatives (TN):

True negatives are the instances where both the predicted class and actual class is False (0), i.e., when a patient does not have complications and is also classified by the model as not having complications.

False Negatives (FN):

False negatives are the instances where the predicted class is False (0) but actual class is True (1), i.e., when a patient is classified by the model as not having complications even though in reality, they do.

False Positives (FP):

False positives are the instances where the predicted class is True (1) while actual class is False (0), i.e., when a patient is classified by the model as having complications even though in reality, they do not.

3.1.2 Accuracy (ACC)

Accuracy determines the number of correct predictions over the total number of predictions made by the model. Even though it is widely used, it is not a very good measure of performance especially when the dataset is imbalanced like in this case. The formula for Accuracy is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

3.1.3 Precision

Precision is a measure of the proportion of patients that actually had complications among those classified to have complications by the system. The formula for Precision is:

$$Precision = \frac{TP}{TP + FP}$$

3.1.4 Recall/ Sensitivity

Recall or sensitivity is a measure of the proportion of patients that were predicted to have complications among those patients that actually had the complications. The formula is:

$$Recall = \frac{TP}{TP + FN}$$

3.1.5 Specificity

Specificity is the opposite of Recall. It is a measure of the number of patients who were classified as not having complications among those who actually did not have the complications. The formula is:

$$Specificity = \frac{TN}{TN + FP}$$

3.1.6 F1 Score

F1 Score is the harmonic mean of the Recall and Precision that is used to test for Accuracy. The formula is:

$$F1\ 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.1.7 Experimental Result

In this study, the five classification algorithms were used, namely, Logistic Regression (LR), Support Vector Machines (SVM), Naïve Bayes (NB), Decision Tree (DT) and Random Forest (RF) were applied on the dataset using with and without PCA and shown in the table 4. In each experiment, the performance was measured using Accuracy, Precision, Recall, Specificity, F1 Score and Sensitivity. The following table 4 illustrates the results of different performance metrics for the algorithms to detect risk of heart disease with PCA and without PCA.

Methods	Accuracy	Precision	Recall	F1 Score	Specificity	Sensitivity
LR without PCA	81.33	86	88	87	81	42
LR with PCA	80.29	82	82	87	78	79
NB without PCA	85.25	84	91	87	70	94
NB with PCA	83.21	82	94	88	71	91
SVM without PCA	81.97	81	88	85	77	91
SVM with PCA	78.2	83	85	84	74	88
DT without PCA	78.69	84	76	80	70	73
DT with PCA	73.77	78	74	76	69	70
RF without PCA	88.52	91	88	90	77	91
RF with PCA	77.05	81	76	79	71	86

Table 4. Performance Analysis of various approaches

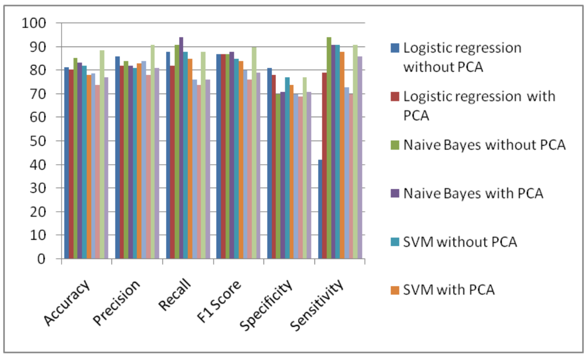

Figure 5: Performance Analysis of various approaches

In terms of accuracy, the best result is produced by Random Forest without PCA with a score of 88.52%. Naïve Bayes is close behind, with a score of 85.25%. However, Naïve Bayes has a better than Random Forest. Logistic regression also has a recall score of .88, so even though the accuracy is low, Random Forest is considered to be the best algorithm for the prediction of Heart Disease without PCA. When PCA is introduced, the accuracy of most of the algorithms remains change except Naïve Bayes. Naïve Bayes accuracy remains to 0.83. Comparing all the results, the best combination to predict the risk of heart disease is Random Forest without PCA.

4. CONCLUSION

In this research, we have tried to compare different machine learning algorithms and predict if a certain person, given various personal characteristics and symptoms, will get heart disease or not. The main motive of our research was to employing various machine learning algorithms with and without applying PCA and comparing the accuracy and analyzing the reasons behind the variation of different algorithms.

5. FUTURE SCOPE

In future we will try to use the primary data as more as possible in Bangladesh and employing deep learning with some ensemble classification techniques find out the risk of heart diseases and try to conclude the reasons and to make some recommendation.

6. ACKNOWLEDGMENTS

We are grateful to the dept. of Computer Science and Engineering, Pabna University of Science and Technology for permitting us to use the Advanced Computing Lab (ACL) for our experiments.

7. REFERENCES

[1] Meijers, Wouter C., and Rudolf A. de Boer. "Common risk factors for heart failure and cancer." Cardiovascular research vol.115, no.5, pp. 844-853, 2019.

[2] NK Podder, HK Rana and et al., "A system biological approach to investigate the genetic profiling and comorbidities of type 2 diabetes", Gene Reports, pp. 100830, https://doi.org/10.1016/j.genrep.2020.100830, 2020.

[3] WHO, "Cardiovascular diseases (CVDs)", 2017. [Online].who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed: 22- Aug- 2020].

[4] Sa, S., "Intelligent heart disease prediction system using data mining techniques.", International Journal of healthcare & biomedical Research, vol. 1, pp. 94-101, 2013.

[5] Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." Informatics in Medicine Unlocked 16 (2019): 100203.

[6] Rana, H. K., Azam, M. S., &Akhtar, M. R., "Iris recognition system using PCA based on DWT". SM Journal of Biometrics & Biostatistics, vol. 2, no. 3, p. 1015, 2017.

[7] Rana, H. K., Azam, M. S., Akhtar, M. R., Quinn, J. M., &Moni, M. A.,"A fast iris recognition system through optimum feature extraction.",PeerJ Computer Science, vol. 5, p. e184, 2019.

[8] HosmerJr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. "Applied logistic regression". vol. 398. John Wiley & Sons, 2013.

[9] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., &Scholkopf, B., "Support Vector Machines", IEEE Intelligent Systems and their applications, vol. 13, no. 4, pp. 18-28, 1998.

[10] Jony, M. H., Johora, F. T., Khatun, P., &Rana, H. K., "Detection of Lung Cancer from CT Scan Images using GLCM and SVM", In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1-6, IEEE, 2019.

[11] Johora, F. T., Jony, M. H., Hossain, M. S., &Rana, H. K. "Lung Cancer Detection Using Marker Controlled Watershed with SVM", GUB Journal of Science and Engineering, vol. 5, no. 1, pp. 24-30, 2018.

[12] Azam, M.S. &Rana H.K., "Iris Recognition using Convolutional Neural Network", International Journal of Computer Applications, vol. 175, no. 12, pp. 24-28, 2020.

[13] Hossen, M. R., Azam, M. S., &Rana, H. K., "Performance evaluation of various DNA pattern matching algorithms using different genome datasets", Pabna University of Science and Technology Studies, vol. 3, no. 1, pp. 14-8, 2018.

[14] Mujtaba, M. A., Azam, M. S., &Rana, H. K., "Performance evaluation of various data mining classification techniques that correctly classify banking transaction as fraudulent", GUB Journal of Science and Engineering, vol. 4, no. 1, pp. 59-63, 2017.

[15] Maheswari, S., &Pitchai, R. "Heart Disease Prediction System Using Decision Tree and Naive Bayes Algorithm", Current Medical Imaging, vol. 15, no. 8, pp. 712-717, 2019.

[16] Iliyas, M. M. K., &Shaikh, M. I. S. "Prediction of Heart Disease Using Decision Tree", AllanaInst of Management Sciences, Pune, vol. 9, pp. 1-5, 2019.

[17] Joloudari, J. H., HassannatajJoloudari, E., Saadatfar, H., GhasemiGol, M., Razavi, S. M., Mosavi, A., &Nadai, L. "Coronary artery disease diagnosis; ranking the significant features using a random trees model", International journal of environmental research and public health, vol. 17, no. 3, p. 731, 2020.

[18] Vallée, A., Petruescu, L., Kretz, S., Safar, M. E., &Blacher, J. "Added value of aortic pulse wave velocity index in a predictive diagnosis decision tree of coronary heart disease", American journal of hypertension, vol. 32, no. 4, pp. 375-383, 2019.

[19] Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." Informatics in Medicine Unlocked 16 (2019): 100203.