Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Diabetes Diagnosis and Prediction: An In-Depth Analysis

[1] Dr.R.Suganthi, [2]B.**Keerthana**,[3]**K.S.Keerthika**, [4]**B. Kiruthik**

[1]Professor, Department of Computer Science with Data Analytics, Dr.N.G.P. Arts and Science College, Coimbatore

[2]II B.Sc(CSDA), Department of Computer Science with Data Analytics,Dr.N.G.P. Arts and Science College Coimbatore

[3]II M.Sc(CSDA), Department of Computer Science with Data Analytics,Dr.N.G.P. Arts and Science College Coimbatore

[4]II M.Sc(CSDA), Department of Computer Science with Data Analytics, Dr.N.G.P. Arts and Science College Coimbatore

Abstract Diabetes is a very common disease and beside it causes serious health problems such as fatal kidney damage or blindness; it may lead the patient to death. There is no exact cure for this disease yet but it is manageable with medication and diet. In this manner, importance of correct diagnosis of diabetes is very important to identify the diseases in early stage and take necessary precautions. There is a lot of data accumulated on this subject, as there are so many patients with this condition. This makes it possible for researchers to use data mining techniques on this subject. This study is proposed to classify diabetes by using data mining techniques. The dataset which has been obtained from UCI machine learning depository contains 520 instances, each having 17 attributes. Seven different classification algorithm including Bayes Network, Naïve Bayes, J48, Random Tree, Random Forest, k-NN and SVM have been studied on this dataset. Obtained results indicated that k-NN performed the highest accuracy with 98.07% and this algorithm is the best method to identify and classify diabetes diseases on studies dataset.
Keywords *– Diabetes, Data Mining, Classification, WEKA, Bayesian Network, Naïve Bayes, J48, Random Tree, Random Forest, k-NN, SVM.*

Introduction

Discovering of useful information from large-scale data is called Data Mining. With data mining, it is possible to reveal relationships between data and make accurate predictions for the future. huge amount of data can be studied in this discipline [1]. The main purpose of data mining is to revealing data which can be valuable for decision support systems in institutions after certain methods and processes. During last decade, data mining has been used in various ways in almost every field and today, data mining is an important discipline in which is useful for nearly all kinds of disciplines [2].

Classification algorithms distribute data between different classes defined on a data set Then, algorithm starts to classify the test data and do this process correctly. Labels are

values indicating these classes given on the data set and they are used for determining the class of the data for training and test groups [3].

With computerized technology, content and structure started to change very fast in health sector. Provided health services has to be fast, accurate, qualified and also has to meets the needs. In order to achieve these goals, healthcare professionals need to reach the most accurate and updated information and use this information by help of decision support systems [4]. Healthcare term means detailed processes for treating human injuries, diagnosis of mental disorders, diagnosis, treatment and prevention of diseases [5]. Today, healthcare industry is developing really rapidly and data is constantly accumulating in this sector because large amounts of data are generated and stored, including electronic medical records, reports, and their findings [6]. Effective use of this healthcare data is made possible by data mining, and new and valuable information is extracted from large volumes of data. In healthcare, data mining is used to predict various diseases and to help doctors diagnose [7].

Diabetes is a very common disease which develops when the gland called pancreas stop producing enough insulin hormone in the body or insulin hormone cannot be used effectively by the body. As a result, the person cannot use the glucose that passes into the blood from consumed food and blood sugar rises. With its full name, Diabetes Mellitus is a disease that plays a leading role in the development of many deadly diseases such as lose of kidney or blindness [8].

This study aims to analyze performance of different classification algorithms in data mining for diabetes dataset by using WEKA tool.

Ii Related Work

A study which compared data mining classification techniques for diabetes dataset by using WEKA. In the study, we are able to classified the accuracy of five different algorithms including Naïve Bayes, SMO, J48, REP Tree and random Tree. Results indicated that, Naïve Bayes showed the best performance with 76.3021% accuracy [9].

Followingly we have analyzed performance of 9 different classification techniques with WEKA including Bayesian, Naïve bayes, J48, Random Forest, Random Tree, CART, k-NN and Conjunctive Rule Learning for diabetes dataset. According to study results Random Tree and k-NN showed the best performance in classification with 100% [10].

Ensuingly we applied data mining classification techniques on diabetes dataset. In the study, three data mining algorithms SOM, C4.5 and Random Forest, were applied by using WEKA. According to results, Random Forest achieved the best performance and highest accuracy [11].

subsequently we used WEKA to analyze and predict diabetes disease with data mining classification techniques. J48, REP Tree, Naïve bayes, SMO, and MLP algorithms were used for classification. As a result of the study, SMO gave the highest accuracy with 76.80% accuracy [12].

On deck we used WEKA for early prediction of chronic kidney disease. Researchers compared Random Forest, J48, k-NN, Naïve bayes and SVM. Results showed that Random Forest classifier has the highest accuracy with 100% [13].

By on by we conducted a study in order to detect diabetes at early stage by using classification algorithms including Decision Tree, SVM and Naive Bayes. Results showed that Naive Bayes performed the highest accuracy of 76.30% [14].

Laterly, we used WEKA to analyze a diabetes dataset by using J48, SMO, Naïve Bayes, random forest and k-NN classification algorithms to find the most suitable algorithm for classifying medical data. Results showed that SMO showed the highest performance with 78% accuracy [15].

Posteriorly we study about analyzing performances of decision tree and Naïve Bayes to predict diabetes disease by using WEKA.

Results showed that naïve bayes gave highest accuracy with 98.43% [16].

Subsequently we used classification techniques including Bagging, SVM, MLP, Simple Logistic and Decision Tree by using WEKA, for effective prediction of type 2 diabetes and decision tree performed the highest accuracy with 94% [17].

Finally we used Bayesian Classifier, J48, Naïve Bayes, Multilayer Perceptron, SVM and random forest classification algorithms for prognosis of type 2 diabetes with WEKA. According to results, Bayesian Classifier, J48, Multilayer Perceptron and random forest showed highest accuracies with 100% [18].

I. Proposed Work And Methdology

The study implemented in WEKA and dataset of Sylhet Diabetes Hospital have been used [19]. The following classification algorithms were used in the study:

- Bayes Network
- Naïve Bayes
- Decision tree (J48)
- Random tree
- Random forest
- k-NN
- Support Vector Machines

A. Bayes Network

Bayesian classifiers are defined as statistical classifiers. It is an ideal type of modeling used to describe an event that occurs in everyday life and to predict the probability that any of the possible causes known to cause that event to be a contributing factor. These classifiers always use more than one search algorithms and quality measures which are based on Bayes network classifier [20]. In Bayes Network, there are two types of probabilities. Posterior Probability [P(H/X)] and Prior Probability [P(H)], where X is a data tuple and H is a

hypothesis. According to Bayes' Theorem:

$$P(H/X)= P(X/H)P(H) / P(X) \quad (1)$$

B. Naïve Bayes

Naive Bayes is a probabilistic classifier which is based on Bayes' theorem. This algorithm assumes that there is independence between predictors and the assumed independence assumptions may not have an impact on reality. This is why they are considered as naive. This model is also useful for very large datasets [21].

C. Decision tree (J48)

This statistical classification algorithm is re- implemented version of C4.5 with JAVA for WEKA. J48 chooses the attribute of the data at every node of the tree, which is most effectively splits its arrangement of tests into subsets improved in one class or the other. J48 have some features such as, continuous attribute value ranges, accounting for missing values, decision trees pruning, derivation of rules, etc. This algorithm works for both categorical and continuous dependent variables. [22].

D. Random Tree

Random Tree is a supervised classifier and fast decision tree learner. The algorithm uses bagging idea to create a random set of data to build a decision tree. The algorithm prunes the tree by using reduced-error pruning with back fitting. [23].

E. Random Forest

The random forest classifier is defined as combination of tree classifiers. Aim of random forests algorithm is increasing the classification value. The algorithm achieves this aim by generating more than one decision tree while performing the classification process. Decision trees which are created individually come together to form a decision forest. Each classifier is generated with a random vector sampled independently from input vector [24]. In this algorithm, newly created training sets are built with replacement from the original ones. The tree is created by using a random attribute selection and a new subset. The best split on the random attributes selected is used to split the node [25].

F. K-NN Algorithm

K-nearest neighbors' algorithm (k-NN) is a non- parametric algorithm. It is one of the easiest-to- implement supervised learning algorithms. Usage of this algorithm is suitable for classification and regression problems both, but is mostly used for classification problems. This algorithm is instance- based and it is one of the simplest of all machine learning and data mining algorithms. The algorithm first determines the k parameter. This parameter is the number of nearest neighbors to a given point. Then, by using relevant distance functions, it calculates the distances of the new data that will be included in the sample data set, according to the existing data. The nearest neighbors from the relevant distances are considered. It is assigned to the class of k neighbors according to the attribute values. Finally, the data is labelled [26].

G. *Support Vector Machines*

Support vector machine (SVM) is a bi-classification algorithm which is used for classifying class based on decision boundary. SVM can also be defined as a vector space-based machine learning method that finds a decision boundary between two classes that are the furthest from any point in the training data. SVM is mostly used to separate data that consists of two classes (binary classification), for example separating each data in a data set into female or male. However, the data can sometimes belong to more than two classes, in such cases the basic SVM algorithm becomes dysfunctional. The objective of SVM algorithm is finding a hyper plane in n number of features which distinctly classifies the data points. SVM is finds a space with the maximum margin, which means the maximum distance between data points of both classes. Maximizing the margin distance provides reinforcement. As a result, future data points can be classified with higher accuracy [27].

II. TOOLS USED

A. *WEKA Engine*

With full name, Waikato Environment for Knowledge Analysis (WEKA), can be defined as a collection of machine learning algorithms for data mining and machine learning tasks. It developed at the University of Waikato, New Zealand. WEKA contains tools for tasks such as data pre-processing, clustering, classification, association rules, regression and visualization. In this paper, WEKA has been used as a data mining engine [28].

B. *Dataset*

The dataset includes names of attributes and the explanation of these attributes shown in Table I. The dataset contains 520 instances, each having 17 attributes. Also the dataset includes one class with two possibilities as tested positive and tested negative.

Table 1. Dataset Description

Attribute Numbers	Attribute Names	output
1	Age_1	20-65
2	Gender_2	1. Male 2.Female
3	Polyuria_3	1.Yes 2.Nos
4	Polydipsia_4	1.Yes 2.Nos
5	Sudden weight loss_5	1.Yes 2.Nos.
6	Weakness_6	1.Yes 2.Nos.
7	Polyphagia_7	1.Yes 2.Nos.
8	Genital thrush_8	1.Yes 2.Nos.
9	Visual blurring_9	1.Yes 2.Nos.

10	Itching_10	1.Yes 2.Nos.
11	Irritability_11	1.Yes 2.Nos.
12	Delayed healing_12	1.Yes 2.Nos.
13	Partial paresis_13	1.Yes 2.Nos.
14	Muscle stiffness_14	1.Yes 2.Nos.
15	Alopecia_15	1.Yes 2.Nos.
16	Obesity_16	1.Yes 2.Nos.
17	Class_17	1.Positive, 2.Negative.

I. GAIN RATIO CALCULATION

Gain Ratio is used for reducing the bias resulting effect which causes from the use of information gain. The information gain measure is biased toward tests with many outcomes. This situation means that, Gain Ratio prefers to select attributes which have a large number of values. Gain Ratio is for adjusting the information gain for each attribute to allow for the breadth and uniformity of the attribute values [29]. The gain ratio can be calculated as follows:

Gain Ratio = Information Gain / Split Information (2)

where the split information is a value-based and it is on the column sums of the frequency table [30].

Search Method: Attribute Ranking
Attribute Evaluator (Supervised, Class(Nominal): 17 Class): Gain Ratio feature evaluator
9 Class: tested_positive or tested_negative

Table Ii. Ranked Attributes

Attribute Rank	Attribute-nominal	Attributes
0.4623	3	Polyuria
0.3619	4	Polydipsia
0.172	2	Gender
0.1518	5	Sudden weight loss
0.1457	13	Partial paresis
0.0812	11	Irritability
0.0883	7	Polyphagia
0.0651	15	Alopecia

0.047	9	Visual blurring

Selected attributes: 3,4,2,5,13,11,7,15,9 : 9

II. RESULTS OF DIFFERENT CLASSIFICATION METHODS

I. Bayes Network

In first experiment, Bayes Network classifier used to classify the diabetes dataset. In this experiment 10 fold cross validation technique were used to split training and testing dataset. Bayes Network classifier could classify 86.92% of instances correctly.

Table Iii. Results Of Bayes Network

Proper Classified Instances	500	65.1042%
Improper Classified Instances	268	34.8958%
Total Number of Instances	768	

1. Naïve Bayes:

In second experiment, Naïve Bayes classifier used to classify the diabetes dataset. In this experiment 10 fold cross validation technique were used to split training and testing dataset. Naïve Bayes could classify 87.11% of instances correctly.

Table Iv. Results Of Naïve Bayes Classification

Proper Classified Instances	586	76.3021%
Improper Classified Instances	182	23.6979%
Total Number of Instances	768	

2. Decision Tree (J48):

In third experiment, J48 Decision Tree used to classify the diabetes dataset. In this experiment 10 fold cross validation technique were used to split training and testing dataset. J48 could classify 95.96% of instances correctly.

Table V. Results Of J48 Classification

Proper Classified Instances	552	71.875%
Improper Classified Instances	216	28.125%
Total Number of Instances	768	

3. Random Tree

In fourth experiment, Random Tree used to classify the diabetes dataset. In this experiment 10 fold cross validation technique were used to split training and testing dataset. Random

Tree could classify 96.15% of instances correctly.

Table Vi: Results Of Random Tree Classification

Proper Classified Instances	523	68.099%
Improper Classified Instances	245	31.901%
Total Number of Instances	768	

1. Random Forest

In fifth experiment, Random Forest used to classify the diabetes dataset. In this experiment 10 fold cross validation technique were used to split training and testing dataset. Random Forest could classify 97.5% of instances correctly.

Table Vii. Results Of Random Forest Classification

Proper Classified Instances	584	75.7813%
Improper Classified Instances	187	24.2188%
Total Number of Instances		768

I. **Results**

Algorithm names	Properly Classified instances	Improper Classified Instances
Bayses Net	65.1042%	34.8958%
Naive Bayes	76.3021%	23.6979 %
Decision Tree (J48)	71.875%	28.125 %
Random Tree	68.099%	31.901%
Random Forest	75.7813%	24.2188%

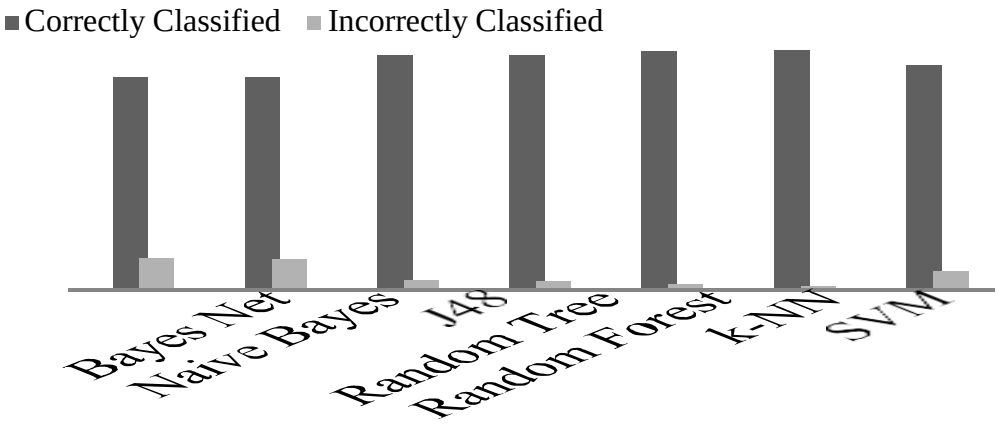


Fig 2. Results of Classification Algorithms

TABLE VII. Results of Random Forest Classification

	Positive Precision	**Negative Precision**	**Error Rate**
Bayes Net	0.1438	0.1100	0.1308
Naive Bayes	0.1438	0.1050	0.1288
J48	0.0500	0.0250	0.0404
Random Tree	0.0406	0.0350	0.0385
Random Forest	0.0219	0.0300	0.0250
k-NN	0.0250	0.0100	0.0192
SVM	0.0688	0.0950	0.0788

The results of tested classification algorithms on diabetes dataset are represented in Table X and Figure 1. According to obtained results, k-NN algorithm performed the highest accuracy with 98,07% while Bayes Net performed the lowest accuracy with 86.92%.

A. Evaluation Measures

Formulas which are used for evaluation measures for classification techniques are as follow [31]. In formulas; TP represents True Positive, TN represents True Negative, FP represents False Positive and FN represents False Negative.

a) Sensitivity = >TP/(TP+FN) (3)
b) Specificity= >TN/(FP+TN) (4)
c) Accuracy= > (TP+TN)/(TP+FP+TN+FN) (5)
d) Positive Precision= >FP/(TP+FP) (6)
e) Negative Precision= >FN/(TN+FN) (7)
f) Error Rate= >FP+FN)/(TP+FP+TN+FN) (8)

Table Xi. Results Of Evaluation Measures For Sensitivity, Specificity And Accuracy

	Sensitivity	**Specificity**	**Accuracy**
Bayes Net Algorithm	0.9267	0.7956	0.8693

Naive Bayes Algorithm	0.9288	0.7956	0.8712
J48 Algorithm	0.9838	0.9242	0.9596
Random Tree Algorithm	0.9778	0.9379	0.9616
Random Forest Algorithm	0.9812	0.9652	0.9750
k-NN Algorithm	0.9937	0.9612	0.9808
SVM Algorithm	0.9401	0.8916	0.9212

Table Xii. Results Of Evaluation Measures For Positive Precision, Negative Precision And Error Rate

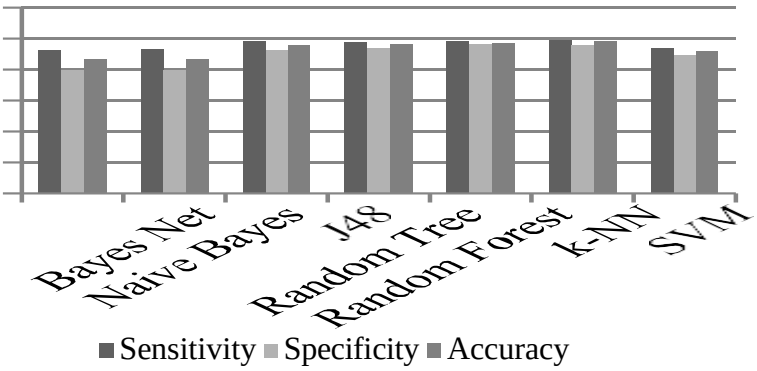
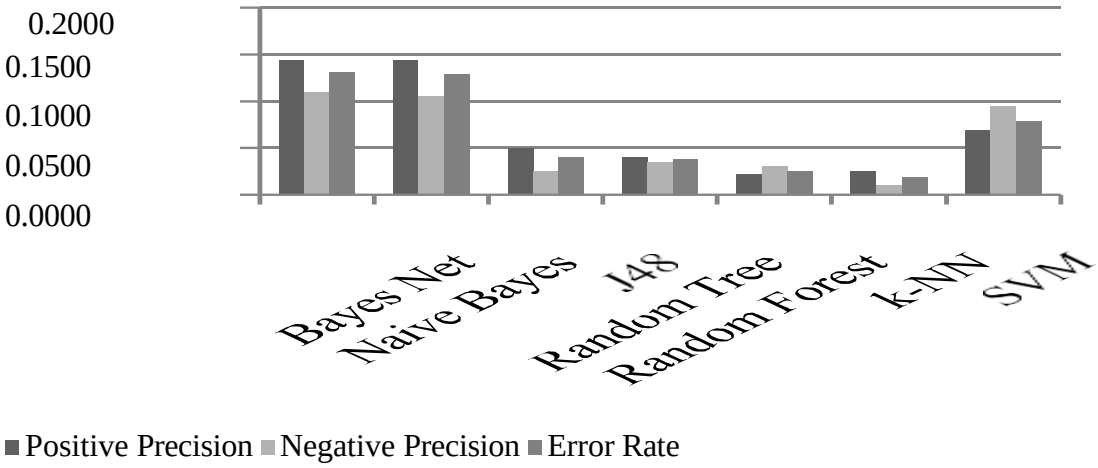


Fig 3. Results Sensitivity, Specificity and Accuracy



VII. **Conclusion**

Applying data mining techniques on medical datasets is an research as there are lots of health

issues and cases to investigate. On the other hand, classification is a very useful technique for knowledge discovery because it can accurately and efficiently classifies the data. there are lots of conducted studies by using classification technique on diabetes datasets in order to help medical workers to identify diabetes based on different datasets and attributes because diabetes mellitus is a challenging disease. So, its classification is significant in medical field [32]. In this study, popular and frequently used classification techniques namely Bayesian Network, Naïve Bayesian, J48, Random Tree, Random Forest, k-NN and SVM have been studied and experiments conducted on the dataset which has been obtained from UCI machine learning depository, in order to find the best classification technique for diagnosis of diabetes diseases. Results indicated that all techniques works above 86% accuracy and k-NN performed the highest accuracy with 98.07%. As a result, k-NN is an effective classification technique to identifying diabetes disease and helps to medical workers for faster decision making about this disease, based on the attributes.

References

[1] Friedman, J. H. (1998). Data Mining and Statistics: What's the connection. *Computing science and statistics*, *29*(1), 3-9.

[2] Horita, F. E., de Albuquerque, J. P., Marchezini, V., & Mendiondo, E. M. (2017). Bridging the gap between decision- making and emerging big data sources: An application of a model-based framework to disaster management in Brazil. *Decision Support Systems*, *97*, 12-22.

[3] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

[4] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, *2*(1), 3.

[5] Sun, J., & Reddy, C. K. (2013, August). Big data analytics for healthcare. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1525-1525).

[6] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*(02), 250-255.

[7] Vispute, N. J., Sahu, D. K., & Rajput, A. (2015). An empirical comparison by data mining classification techniques for diabetes data set. *International Journal of Computer Applications*, *131*(2), 6-11.

[8] Joshi, S., & PriyankaShetty, S. R. (2015). Performance analysis of different classification methods in data mining for diabetes dataset using WEKA tool. *International Journal on Recent and Innovation Trends in Computing and Communication*, *3*(3), 1168-1173.

[9] Daghistani, T., &Alshammari, R. (2016).Diagnosis of diabetes by applying data mining classification techniques. *Int. J. Adv. Comput. Sci. Appl*, *7*(7), 329-332.

[10] Verma, D., & Mishra, N. (2017, December).Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques.In *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 533-538).IEEE.

[11] Kumar, N., & Khatri, S. (2017, February). Implementing WEKA for medical data classification and early disease prediction. In *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)* (pp. 1-6).

[12] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, *132*, 1578-1585.

[13] Nass, L., Swift, S., & Al Dallal, A. (2019). Indepth analysis of medical dataset mining: a comparative analysis on a diabetes dataset before and after preprocessing. *KnE Social Sciences*, 45- 63.

[14] Karthikeyan, R., Geetha, P., & Ramaraj, E. (2019, February). Rule Based System for Better Prediction of Diabetes. In *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)* (pp. 195-203). IEEE.

[15] Shuja, M., Mittal, S., & Zaman, M. (2020). Effective prediction of type ii diabetes mellitus using data mining classifiers and SMOTE. In *Advances in Computing and Intelligent Systems* (pp. 195-211). Springer, Singapore.

[16] Bhatti, S. (2020). Prognosis of Diabetes by Performing Data Mining of HbA1c. *International Journal of Computer Science and Information Security (IJCSIS)*, *18*(1).

[17] Islam, M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis* (pp. 113-125). Springer, Singapore.

[18] Tang, B., He, H., Baggenstoss, P. M., & Kay, S. (2016). A Bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *28*(6), 1602-1606.

[19] Flach, P. A., & Lachiche, N. (2004). Naive Bayesian classification of structured data. *Machine Learning*, *57*(3), 233- 269.

[20] Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(6).

[21] Cutler, A., & Zhao, G. (2001). Pert-perfect random tree ensembles. *Computing Science and Statistics*, *33*, 490-497.

[22] Sridharan, K., & Komarasamy, G. (2020). Sentiment classification using harmony random forest and harmony gradient boosting machine. *Soft Computing*, *24*(10), 7451-7458.

[23] Wang, Z., Chegdani, F., Yalamarti, N., Takabi, B., Tai, B., El Mansori, M., & Bukkapatnam, S. (2020). Acoustic Emission Characterization of Natural Fiber Reinforced Plastic Composite Machining Using a Random Forest Machine Learning Model. *Journal of Manufacturing Science and Engineering*, *142*(3).

[24] Li, S., Zhang, K., Chen, Q., Wang, S., & Zhang, S. (2020). Feature Selection for High Dimensional Data Using Weighted K-Nearest Neighbors and Genetic Algorithm. *IEEE Access*.

[25] Fung, G., & Mangasarian, O. L. (2002, April). Incremental support vector machine classification. In *Proceedings of the 2002 SIAM International Conference on Data Mining* (pp. 247- 260). Society for Industrial and Applied Mathematics.

[26] Garner, S. R. (1995, April). Weka: The waikato environment for knowledge analysis. In *Proceedings of the New Zealand computer science research students conference* (Vol. 1995, pp. 57-64).

[27] Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271-277.

[28] Shouman, M., Turner, T., & Stocker, R. (2011, December). Using decision tree for diagnosing heart disease patients. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121* (pp. 23-30).

[29] Sahin, H., & Subasi, A. (2015). Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques. *Applied Soft Computing*, *33*, 231-238.

[30] Olaniyi, E. O., & Adnan, K. (2014). Onset diabetes diagnosis using artificial neural network. *Int J Sci Eng Res*, *5*(10), 754-759.