An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques

Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
emranakabir@gmail.com, szaman22.ruet@gmail.com, rakib063049@gmail.com

***Abstract*—Currently in the healthcare industry different data mining methods are used to mine the interesting pattern of diseases using the statistical medical data with the help of different machine learning techniques. The conventional disease diagnosis system uses the perception and experience of doctor without using the complex clinical data. The proposed system assists doctor to predict disease correctly and the prediction makes patients and medical insurance providers benefited. This research focuses on to diagnosis diabetes disease as it is a great threat to human life worldwide. The system uses the Decision Tree and K-Nearest Neighbor (KNN) Algorithms as supervised classification model. Finally, the proposed system calculates and compares the accuracy of C4.5 and KNN and the experimental result demonstrates that the C4.5 provides better accuracy for diagnosis diabetes. For the clinical database, the Pima Indians Dataset is used in this research.**

***Keywords*—data mining; clinical decision support system; expert application; disease prediction; C4.5; KNN**

I. Introduction

Computational health informatics is an emerging research topic which involving various sciences such as biomedical, medical, nursing, information technology, computer science, and statistics [1]. Data mining techniques are applied to predict the effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among huge clinical and diagnosis data [2]. In medical science, doctor's facilities introduced different data frameworks with a lot of information to manage medical insurance and patient information but unfortunately, data are not mined to discover hidden information for effective decision [2][3]. Clinical test outcomes are regularly made on the basis of doctors' perception and experience rather than on the knowledge enrich data masked in the database and sometimes this procedure prompts inadvertent predispositions, doctor's expertise may not be capable to diagnose it accurately which affects the disease diagnosis system [2][3]. In healthcare sector, the term information mining can mean to analyze the clinical information to predict patient's health status. So discovering interesting pattern from healthcare data, different data mining techniques are applied with statistical analysis, machine learning and database technology.

Fig. 1 describes the framework of health informatics processing pipeline which involves the steps of capturing, storing, sharing, analyzing and decision support. After gathering huge amount of clinical data, different techniques of data mining are applied for analyzing these data. This analyzing procedure starts with data preprocessing, then feature selection and finally machine learning approaches are applied for classification, regression and clustering these healthcare data to explore interesting patterns and meaningful information.

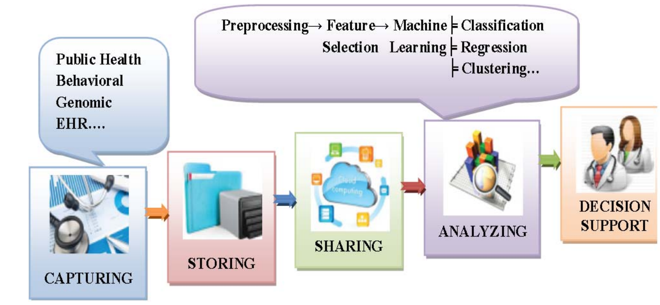

Fig. 1. Pipeline of Health Informatics Processing.

The term diabetes mellitus corresponds to a metabolic disorder like lack of or resistance to insulin, a hormone critical for the regulation of blood sugar and its complications consists of chronic malfunction, strokes, miscarriages, blindness, kidney failure, amputations, damage and failure of several organs [3]. Data mining has been successfully used in knowledge discovery for predictive purposes to make more active and accurate decision. Different data mining techniques i.e. Decision Tree, Bayesian Network, K- Nearest Neighbor, Naïve Bayes, Support Vector Machine etc. are used to predict diabetes in early stage which also helps to avoid the patient's complications.

The objective of this paper is to develop an expert predictive healthcare decision support system using data mining techniques. A common diabetes dataset is trained in this system using C4.5 and KNN algorithm and tested new sample data which predict the patent's outcome.

978-1-5090-5627-9/17/$31.00 ©2017 IEEE

II. Related Works

A number of significant researches that have made use of artificial intelligence and data mining techniques are explained below.

The experimentation of paper [2] has carried out for diagnosis heart disease using weighted fuzzy rule based system and the performance has compared with neural network based system. For making correct decision about the risk of heart disease, paper [3] has implemented a rule based proposed system with the help of WEKA tool and compared the accuracy with SVM, C4.5, 1-NN, PART, MLP and RBF algorithms. The paper [4] has summarized the classification techniques and compared the accuracy (SVM provided best accuracy- 81.77%) for diagnosis of diabetes disease. A web based application has introduced by paper [5] using Naïve Bayesian algorithm which took symptoms from user and gave the diagnosis result to the user or patient. For predicting diabetes disease, paper [6] has presented a comparison between Decision Tree (accuracy- 76.96%) and Naïve Bayes algorithm (accuracy- 79.56%) with the help of WEKA tool. In paper [7], a comparison has demonstrated among decision tree, KNN and Naïve Bayes algorithm using WEKA, Rapid miner, Tanagra, Orange and Knime tools on Indian Liver Patient Dataset. The paper [8] has proposed a system based on attribute reduction and compared their result (accuracy- 88.3%) with Naïve Bayes (accuracy- 85.2%), support vector machines (accuracy- 81.5%), and artificial neural network (accuracy- 81.5%) to diagnosis heart disease. For the prediction of Dementia, paper [9] has compared the overall classification accuracy, specificity and sensitivity among Multilayer Perceptrons Neural Networks, Radial Basis Function Neural Networks, Support Vector Machines, CART, CHAID and QUEST Classification Trees and Random forests classifiers. The paper [10] has summarized various data mining techniques such as classification, clustering, association to analyze and predict human disease. To predict kidney diseases, paper [11] has demonstrated that the performance of the SVM (accuracy- 76.32%) provided better accuracy than the Naive Bayes classifier algorithm (accuracy- 70.96%). For classification of ECG heartbeat signal, paper [12] has compared performance result and achieved overall accuracy 99.33% against 98.44 and 98.67 % for the C4.5 and CART classifiers, respectively. In paper [13], a multi attribute density estimation based system has proposed for the prediction of diseases. For the prediction of diabetes disease paper [14] has proposed a system with Re-RX with J48graft algorithm and achieved average accuracy 83.33% using 10-fold cross validation technique. To diagnose type II diabetes, paper [15] has compared the overall classification accuracy of Naive Bayes (accuracy- 76.95%), RBF Network (accuracy- 74.35%), and J48 (accuracy- 76.52%) data mining algorithms. For rule based diagnostic classification, paper [16] has compared decision tree, random forest, Naïve Bayes, and support vector machine using five-fold cross validation approach to predict type II diabetes.

III. Proposed Expert System

An expert system is created to predict different diseases. The block diagram of our proposed system is depicted in Fig.2. The workflow of this system is given below.

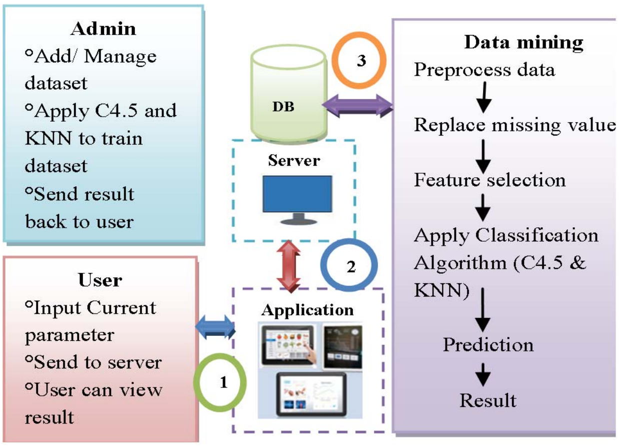

Fig. 2. Block diagram of proposed system.

Step 1: Through the proposed application user (doctor, patient, physician etc.) can input the attribute values of disease and send it to the server with the help of internet. After applying the data mining approach the predicted result can be viewed on the user GUI.

Step 2: On the server, admin can load dataset of different diseases and apply different data mining algorithms to train dataset. Requested user inputs are collected and processed on server to predict the diagnosis result.

Step 3: For analyzing healthcare data, major steps of data mining approaches like preprocess data, replace missing values, feature selection, machine learning and make decision are applied on train dataset. On the server different algorithms, i.e. C4.5 and KNN have executed on train dataset and ready to classify the test dataset.

IV. Methodology

In the proposed system, an expert application has created based on two data mining classification model to predict disease. Decision tree and K-Nearest Neighbor (KNN) classifiers are used to train dataset in this system.

A. Decision Tree

Decision tree is a powerful classification technique and ID3, C4.5, C5, J48, CART and CHAID algorithms are available to predict the data. In this work, the dataset contains continuous values so we have selected C4.5 as our classifier. The decision nodes are found by calculating the highest information gain from all attributes. The information gain is calculated by (1) which is useful to classify unknown records.

$$\mathrm{Gain(p)} = \mathrm{F(Info\ (T)} - \mathrm{Info\ (p,T))} \qquad (1)$$

$$\text{Where, } \mathrm{Info(T)} = \mathrm{Entropie(p)} = -\sum_{i=1}^{n} \mathrm{pi} \times \log(\mathrm{pi})$$

$$\text{and } \mathrm{Info(p,T)} = \sum_{j=1}^{n} (\mathrm{pj} \times \mathrm{Entropie(pi)})$$

F = number of known sample / total number of sample in the dataset for a given attribute, pi = the set of probability distribution, T= Test, pj= the set of all possible values for attribute T

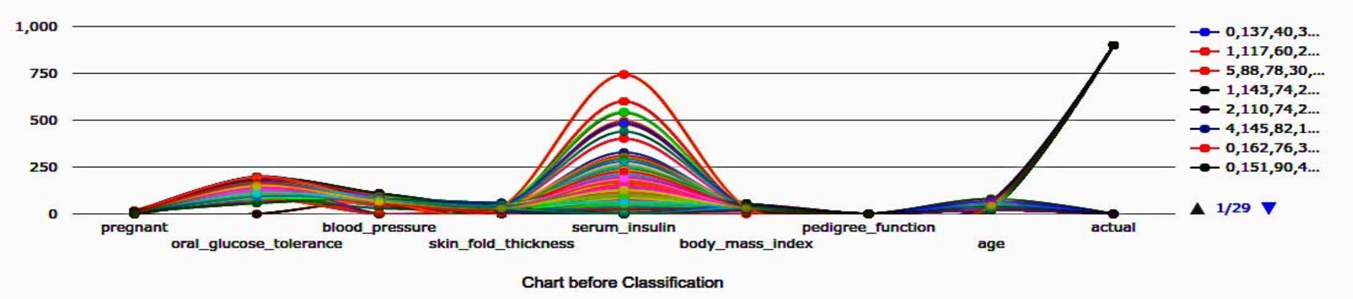

Fig. 3. Line chart of attribute values before classification.

Pseudo-code for the DecisionTree (T) [17] is stated below:

```
Step 1: ComputeClassFrequency(T);
Step 2: if OneClass or FewCases
        Return a leaf;
        Create a decision node N;
Step 3: ForEach Attribute A
        ComputeGain(A);
Step 4: N.test = AttributeWithBestGain;
Step 5: if N.test is continuous
        Find Threshold;
Step 6: ForEach T' in the splitting of T
Step 7: if T' is Empty
            Clild of N is a leaf
        else Child of N = DecisionTree(T');
Step 8: ComputeErrors of N;
        Return N
```

B. K-Nearest Neighbor (KNN)

KNN is a supervised learning algorithm which classifies new data based on minimum distance from the new data to the K nearest neighbor. The proposed work has used Euclidean Distance (2) to define the closeness.

$$d(X,Y) = \sqrt{\sum_{i=0}^{n} (\mathrm{Xi} - \mathrm{Yi})^2} \qquad (2)$$

Where, X=(x1, x2……,xn) and Y=(y1,y2…..yn)

Pseudo-code for the KNN classifier [18] is stated below:

```
Step 1: Input: D= {($x_1$, $c_1$),….($x_n$, $c_n$)}
        x=($x_1$,…..,$x_n$) new instance to be classified
Step 2: For each labeled instance ($x_i$,$c_i$)
        *Calculate* d($x_i$, x)
Step 3: *Order* d($x_i$, x) from lowest to highest, (i=1,….,N)
Step 4: *Select* the K nearest instances to x : $D_x^K$
Step 5: *Assign* to x the most frequent class in $D_x^K$
```

V. Implementation

A. Dataset

The Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases [6] [14] [15] has used for prediction the diabetes disease. This dataset consists of 768 samples with 8 numerical valued attribute where 500 are tested negative and 268 are tested positive instances. The selected attributes have been described as details for diabetes data analysis in TABLE I.

TABLE I. Attributes Description

Attribute name(all numeric valued)	**Mean**	**Standard deviation**
Number of times pregnant	3.8	3.4
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	120.9	32.0
Diastolic blood pressure (mm Hg)	69.1	19.4
Triceps skin fold thickness (mm)	20.5	16.0
2-Hour serum insulin (mu U/ml)	79.8	115.2
Body mass index (weight in kg/(height in m)^2)	32.0	7.9
Diabetes pedigree function	0.5	0.3
Age (years)	33.2	11.8

Fig. 3 describes the line chart of all attribute values for 768 instances before classification.

In the application, there is an option to input sample diabetes data for new prediction is shown in Fig. 4. Medical practitioners can test sample data through this application.

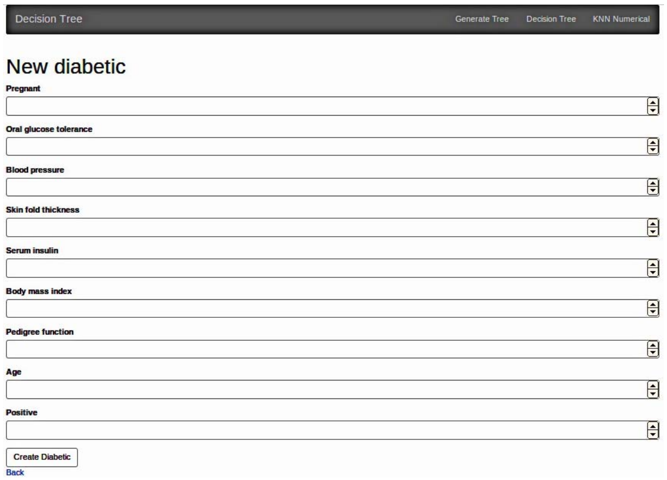

Fig. 4. Sample diabetes data for new prediction.

B. Performance Evaluation

Performance evaluation is carried out by accuracy calculation as (3) which is the ratio of the number of correctly classified instances to the total number of instances of the test data [8].

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{TN}+\text{FN})} \times 100\% \qquad (3)$$

Where,
TP, FP, TN and FN are the number of true positive, false positive, true negative and false negative respectively.

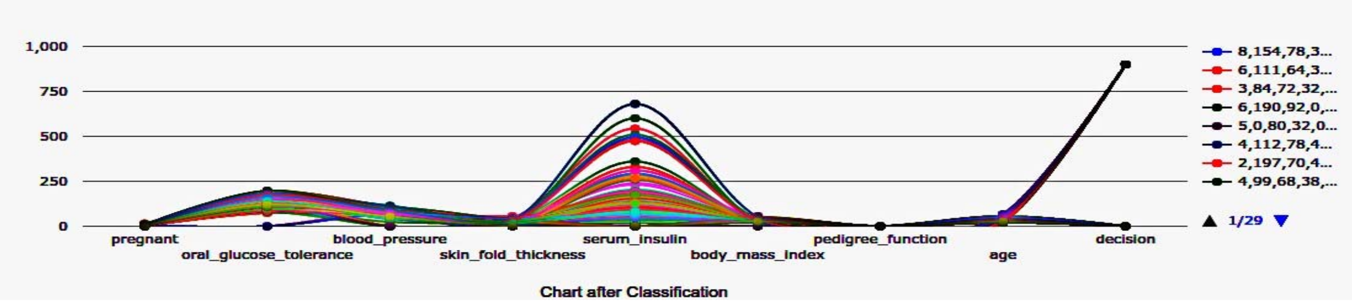

Fig. 5. Line chart of attribute values after classification.

C. Result and Discussion

The outcomes of clinical decision support system for diabetes disease prediction are presented in this section. It trains and tests this expert system using 70:30 percentage split method to evaluate the classification result. Here randomly chosen 538 instances (about 70%) of original dataset are used for training the system and 230 instances (about 30%) are used for testing purpose. This prediction system has been developed using two well known algorithm i.e. C4.5 and KNN algorithm. The performance of these two algorithms is described below.

1) Performance evaluation of KNN algorithm: The performance of proposed system has been experimented using KNN classification algorithm. To guarantee the validity of result is carried out by assigning various values of K. After applying KNN classifier with K=7, Fig. 5, TABLE II and Fig. 6 have found as the experimental result of the system. The Fig. 5 shows the line chart of attribute values with decision after classification. In TABLE II, the prediction result shows that KNN classifier has correctly classified 177 instances and incorrectly classified 53 instances. The accuracy of correctly classified instance is 76.96% and incorrectly classified instance is 23.04%. The pie chart of accuracy is shown in Fig. 6.

TABLE II. CLASSIFICATION SUMMARY OF KNN CLASSIFIER

KNN	**Prediction Result**		**Accuracy**	
	Correctly Classified instances	***Incorrectly Classified instances***	***Correctly Classified instances***	***Incorrectly Classified instances***
Training instances: 538 Testing instances: 230	177	53	76.96%	23.04%

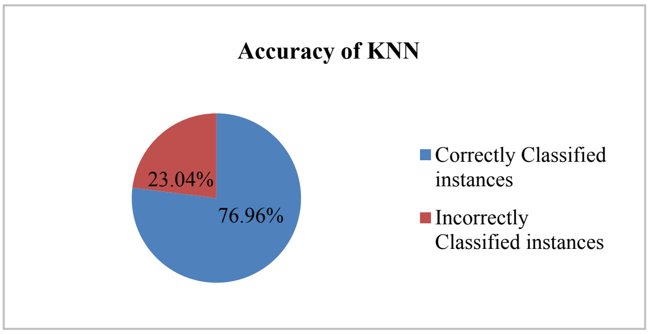

Fig. 6. Accuracy chart of KNN classifier.

2) Performance evaluation of Decision Tree algorithm: After executing the system with C4.5 classifier, our application has correctly classified 208 instances and incorrectly classified 22 instances. The accuracy of correctly classified instance is 90.43% and incorrectly classified instance is 9.57%. TABLE III and Fig. 7 have been found as the experimental result of decision tree algorithm.

TABLE III. CLASSIFICATION SUMMARY OF C4.5 CLASSIFIER

C4.5	**Prediction Result**		**Accuracy**	
	Correctly Classified instances	***Incorrectly Classified instances***	***Correctly Classified instances***	***Incorrectly Classified instances***
Training instances: 538 Testing instances: 230	208	22	90.43%	9.57%

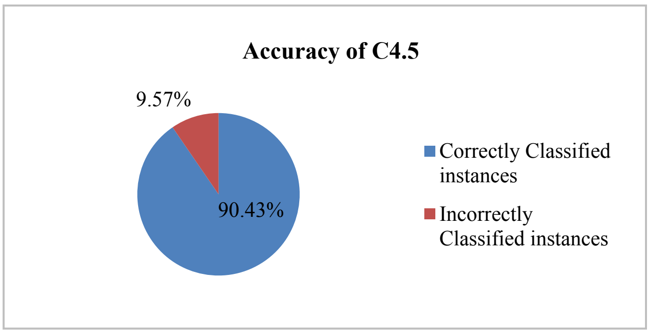

Fig. 7. Accuracy chart of C4.5 classifier.

3) Comparison of classification accuracy: The performance of proposed clinical expert system was analyzed with diabetes dataset using KNN and C4.5. Acoording to experimental results, TABLE IV represents the performance comparison of KNN and C4.5 classifier based on percentage split (70:30) technique.

TABLE IV. PERFORMANCE COMPARISON OF KNN AND C4.5

Classifier	**No. of instances**		**Accuracy**
KNN	Correctly classified	177	76.96%
	Incorrectly classified	53	23.04%
C4.5	Correctly classified	208	**90.43%**
	Incorrectly classified	22	9.57%

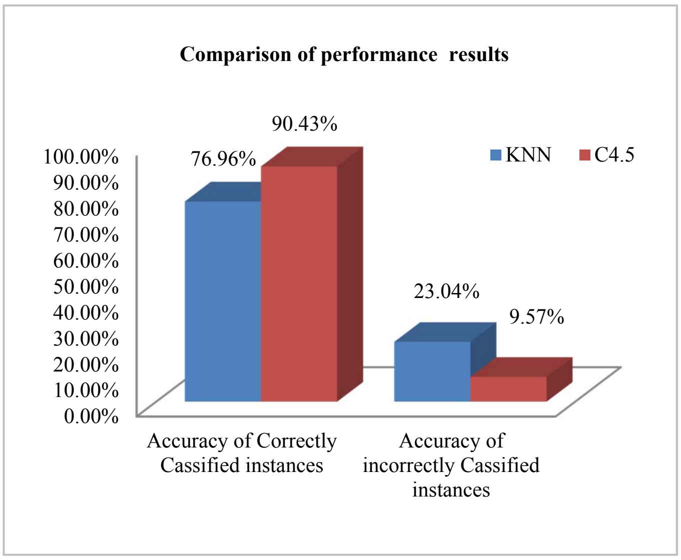

Fig. 8. Graphical representation of performance comparison.

In Fig. 8, the graphical representation of performance comparison is described. From the graphical model it is observed that C4.5 algorithm performs better with greater accuracy.

The KNN based classifier determines neighborhoods directly from training observations and doesn't build any classification model. It only works with numeric feature vector. On the other hand, decision tree predicts a class using predefined classification tree with contains both numerical and categorical feature vector. So if it is needed to use any medical dataset with numerical and nominal features, decision tree is preferred than KNN.

VI. Conclusion

In this research, an expert system is proposed for predicting the diseases like diabetes using data mining classification technique. The system gives benefit to the doctors, physicians, medical students and patients to make decision regarding the diagnosis of the diseases. In this application 70:30 percentage ratio is used to train and test the database. The system finds 100% accuracy for training phase using decision tree algorithm. For the testing phase C4.5 and KNN provide 90.43% and 76.96% accuracy respectively. In paper [6], [14], [15] and [16], decision tree was applied as a classification technique on same dataset and 76.96%, 83.83%, 76.52% and 62.17% highest accuracy have been found respectively. Using the same dataset for diabetes disease prediction this system has achieved best accuracy (90.43%) for decision tree algorithm. The application can be used by anybody specially for medical practitioners via internet for diagnosis purpose.

References

[1] R. Fang, S. Pouyanfar, Y. Yang, S. Chen and S. Iyengar, “Computational health informatics in the big data age: a survey,” ACM Comput. Surv., New York, Vol. 49, pp. 12-47, June 2016.

[2] P. K. Anooj, “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,” J. of King Saud Uni. Comput. and Inform. Sci., ELSEVIER, Vol. 24, pp. 27-40, 2012.

[3] Purushottam, K.Saxena and R. Sharma, “Efficient Heart Disease Prediction System,” Proced. Comput. Sci., ELSEVIER, Vol. 85, pp. 962 – 969, 2016.

[4] P. Agrawal and A. Dewangan, “A brief survey on the techniques used for the diagnosis of diabetes-mellitus,” Int. Res. J. of Eng. and Tech. IRJET, Vol. 02, pp. 1039-1043, June-2015.

[5] P. Bhandari, S. Yadav, S. Mote and D.Rankhambe, “Predictive system for medical diagnosis with expertise analysis,” Int. J. of Eng. Sci. and Comput., IJESC, Vol. 6, pp. 4652-4656, April 2016.

[6] A. Iyer, S. Jeyalatha and R. Sumbaly, “Diagnosis of diabetes using classification mining techniques,” Int. J. of Data M. & Know. Manag. Process, IJDKP, United Arab Emirates, vol. 5, pp. 1-14, January 2015.

[7] A. Naik and L. Samant. “Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer , Tanagra ,Orange and Knime,” Int. Con. on Computa. Mod. and Sec., ELSEVIER, Vol. 85, pp. 662-668, 2016.

[8] N. Long, P. Meesad and H. Unger, “A highly accurate firefly based algorithm for heart disease prediction,” Expert Syst. with App., ELSEVIER, Vol. 42, pp. 8221–8231, 2015.

[9] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana and A. Mendonca, “Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests,” Maroco et al. BMC, Vol. 4, pp. 299-313, 2011.

[10] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain,” Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March 2016.

[11] S. Vijayarani and S.Dhayanand, “Data mining classification algorithms for kidney disease prediction,” Int. J. on Cyber. & Informatics, Vol. 4, pp. 13-25, August 2015.

[12] E. Alickovic and A. Subasi, “Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier,” Patient Facing System, J Med Syst, Vol. 40, pp. 108-120, 2016.

[13] M. Inbavalli and T. Arasu, *“*Multi-attribute density estimation based location selection approach in multi-agent disease prediction model for decision support system using diagnosis pattern and data mining,” Middle-East J. of Scien. Res., Vol. 24, pp. 65-72, 2016.

[14] Y. Hayashi and S. Yukita, “Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset,” Informatics in Medicine Unlocked, ELSEVIER, Vol. 2, pp. 92-104, 2016.

[15] S. Sa’di, A. Maleki, R. Hashemi, Z. Panbechi and K. Chalabi. “Comparison of data mining algorithms in the diagnosis of type II diabetes,” Int. J. on Comput. Sci. & App., Vol. 5, pp.1-12, October 2015.

[16] G.Huang, K.Huang, T.Lee, J. Tzu-Ya and Weng, “An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients,” Huang et al. BMC Bioinformatics, Vol. 16, pp.55-65, 2015.

[17] S. Ruggieri, "Efficient C4.5 [classification algorithm]," IEEE trans. on know. and data eng., Vol. 14, pp. 438-444, 2002.

[18] M. Jabbar, B. Deekshatulua and P. Chandra, “Classification of heart disease using k-nearest neighbor and genetic algorithm," Procedia Tech., ELSEVIER, Vol. 10, p.85-94, 2013.