Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 8, Issue. 3, March 2019, pg.61 – 71

COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING DIABETES OR PREDIABETES BY RISK FACTORS

Mohammed Layth Zubairi Alkaragole[1]
Department of Information Technologies, Faculty of Engineering and Natural Sciences, Altınbaş University, Turkey
mohammed.alkaragole@ogr.altinbas.edu.tr

Asst. Prof. Sefer Kurnaz
Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Altınbaş University-Turkey
sefer.kurnaz@altinbas.edu.tr

Abstract: Data mining, extraction and analysis techniques play an important role in health and care. Because the analysis and diagnosis of any disease must contain a huge number of data. The important role of extracting and analyzing patterns of medical diagnosis is therefore evident. Diabetes is a group of metabolic diseases that contain a high percentage of blood sugar for a long time. In addition to the challenges of classify and forecasting Diabetes, there is another problem that health data may contain data loss or be incorrect. Because of these problems and circumstances that may hinder the process of processing and overcoming data, many previous studies have provided many automated learning methods for diagnosis, prediction, processing of potential data loss and problem solving. This research will analyze and compare different techniques of data extraction and analysis of diabetes. Recent data mining techniques commonly used in Bayesian, SVM, Decision Tree, etc. This paper represents the proposed framework with hybrid datamining techniques. Results showed that hybrid classification in proposed framework outperforms other classifiers with an accuracy rate of 94%.

Keywords: Data mining, Decision Tree, Naïve Bayes, Support Vector Machine, Diabetes, Electronic Health Records.

Turkish

Özet

Veri madenciliği, çıkarım ve analiz teknikleri sağlık ve bakımda önemli bir rol oynamaktadır.Sebebi ise, hastalığın analiz ve tespitinin önemli bir sayıda bilgi içermesidir.Tıbbi tanı örneklerinin çıkarılması ve analiz edilmesinin önemi ortadadır. Diyabet, uzun süre yüksek oranda kan şekeri içeren bir metabolik hastalık grubudur..Diyabeti sınıflandırma ve tahmin etme zorluklarına ek olarak, sağlık verisinin eksik veya yanlış veri içermesi sorunu da bulunmaktadır.Verilerin işlenmesi ve üstesinden gelinmesi sürecini engelleyebilecek bu sorunlar ve koşullar nedeniyle, önceki çalışmaların birçoğunda tanı, tahmin, potansiyel veri kaybının işlenmesi ve problem çözme için birçok otomatik öğrenme yöntemi sağlanmıştır.Bu çalışma farklı teknikleri analiz etmeyi ve karşılaştırmayı amaçlamıştır. Günümüzde yaygın olarak kullanılan very madenciliği teknikleri Bayesian, SVM ve karar ağaçlarıdır. Bu makale hibrit veri madenciliği teknikleriyle önerilen framework'ü ortaya koymuştur.Sonuçlar, önerilen framework hibrit sınıflandırmasının diğer sınıflandırıcıları% 94'lük bir doğruluk oranı ile geride bıraktığını göstermiştir.

Anahtar Kelimeler: *veri madenciliği, karar ağacı, Naïve Bayes, Support Vector Machine, diyabet, elektronik sağlık kayıtları.*

1. Introduction

Diabetes Mellitus (DM) is a collection of associated conditions in which the body cannot control the volume of sugar in the blood. DM is regarded merely as diabetes which is a constant metabolic sickness that has now become popular and long grown in the world. In a normal person, the blood glucose level is controlled by various hormones, including insulin. Insulin is constructed by the pancreas, a tiny organ separating the stomach and liver. The pancreas hides other necessary proteins that serve to absorb food [1]. Several models of informed diabetes mellitus can be categorized into two sections named as type 1 diabetes and Type2 diabetes [2]. Various different marks can serve to diagnose Type 1 diabetes such as extended thirst, regular urination, appetite, weakness, and blurred vision.

The Type 1 diabetes Strategy intends at supporting healthy blood sugar levels through frequent monitoring, insulin treatment, nutrition, and exercise. The goal of data mining is to obtain valuable information from large databases or data warehouses.

Data mining applications are used for commercial and scientific sides [1]. Data mining is the method of choosing, examining and forming massive volumes of data to identify unknown patterns or relationships which present a clear and valuable outcome to the data interpreter [2].

Different methods such as extracting and analyzing data, then selecting specific patterns and then cleaning data are the most important pillars and steps of knowledge. The process of discovering knowledge from large data is clearly defined and consists of several major and systematic steps as shown in Figure 1. The process of exploration and data extraction is the main step through which we can discover hidden but useful knowledge of the enormous databases, especially medical databases [4].

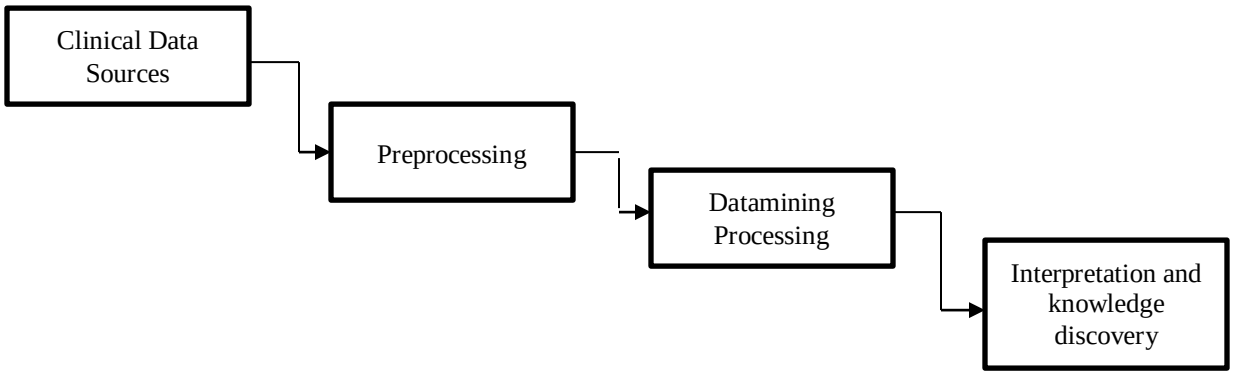

Figure 1 - Levels of actions in the Development of Clinical Data Mining [4]

1.1 Causes of Diabetes and Risk Factors

Hereditary and genetics factors, genders, ages, family history, Insulin deficiency, number of previous medications, High blood pressure, infection, obesity due to excessive food, anxiety and stress, increased cholesterol and triglyceride plus the wrong diet.

1.2 Paper Contributions

Inside our paper, we have focused on data mining classification methods these are capable of forecasting a certain consequence based on a specified input. In particular, we have utilized four classifiers and create comparative study to analyze a medical dataset that recorded previously to diagnosis diabetes disease. The main contribution of this study, apply hybrid techniques (Decision tree and SVM) in a proposed framework which had a highest accuracy to classify diabetes patient records to patient with diabetes in which type or not patient.

A number of trials have been constructed to compare the accuracy of the implemented classifiers on a different size full training dataset with 9 attributes. Results showed that hybrid classification in proposed framework outperforms other classifiers with an accuracy rate of 94.0%, which provided a more effective and comprehensive classification mechanism than other classification techniques.

The purpose of the framework is that the use of context circumstances to realize higher forecast and precision of the information mining technique. The framework structure is found towards medical datasets, and therefore the cases want to see mistreatment the mining, particularly classification rules to develop call support systems associated to the medical domain. So, this paper investigates and implement the following:

- Use standard Medical Dataset (1000 Diabetes's records form UCI standard dataset)
- Classification patient records according to diabetes with different types using hybrid Data mining techniques
- Build comparative study between different datamining techniques to choose the highest two techniques

- Build a proposed framework that cover missing data and use highest two techniques to get highest accuracy

1.3 Paper Structure

The remainder the paper is prearranged as follows. The reviewing of some related works to the proposed approach is presented in section II. Section III discusses the research methodology. Comparative study introduced in section IV. The results and discussions are obtainable into section V. The final conclusions including later works are offered in section VI.

2. Related Work

Mostly, there several data mining techniques that are adopted in health care equivalent to classification, clustering, and association as shown in figure 2.

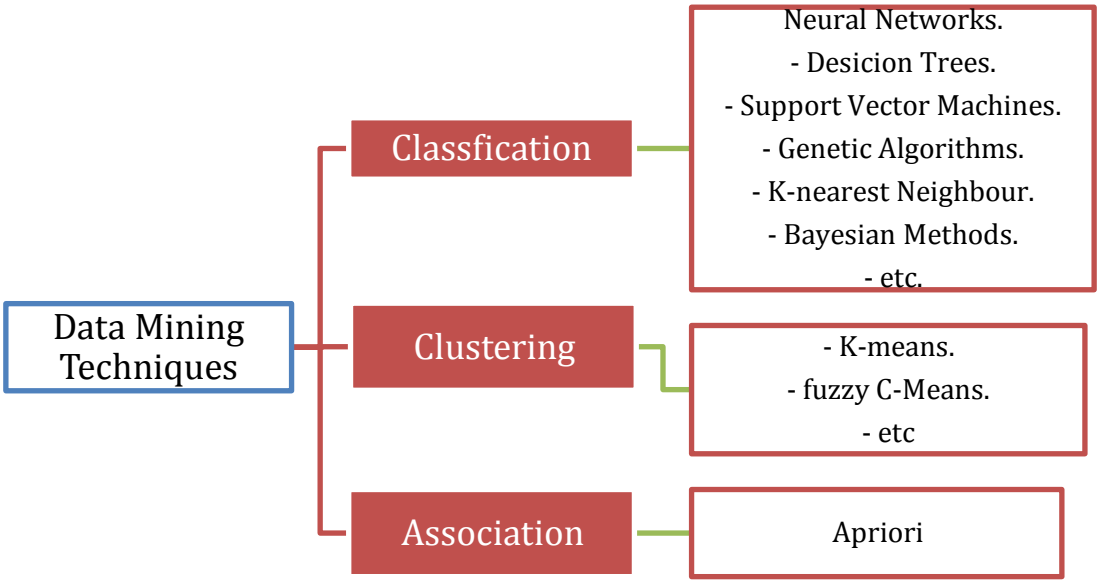

Figure 2- The various data mining methods applied in healthcare control [12].

In [5], they proposed and realized a method to verify and observance coronary artery illness. Authors need to use the standard dataset of heart disease called Cleveland. They need to be enforcement their applied to 303 cases with thirteen features choose from seventy-six features obtainable within the dataset. Throughout the primary experience, SVM produced 88.3% accuracy whereas, throughout the second experience, Bayes Net and SVM both produced 83.8% efficiency and FT produced 81.5% accuracy.

In [6], heart diseases are diagnosed apply Naïve Bayes algorithm. The utilized dataset is received by one amongst this leading diabetic study academy in Chennai. In their trials, they applied the WEKA application of dimension separation to achieve the classification process. The outcomes have proved that Naive Bayes has 86% efficiency.

In [7], the new method applies a comparative of Genetic Algorithm (GA) and decision trees for the diagnosis of diabetes was given. A comparative of the C4.5 and GA models were applied to improve the accuracy, rate, and the diagnosis of diabetes. In the decision tree, feature identification and choice in every node states mean that the feature is extra effective than others in data choice. Accordingly, the user can get the final judgment more accurate and quicker. In this research, Pima data of decision tree including 768 people by 8 features were assessed. The suggested method has produced 89.7% identification accuracy.

In [8], they attempted to predict the liver infection applying Naïve Bayes and SVM categorization methods. They have applied the standard dataset of ILPD that involves 560 status and 10 features. The review from the couple methods is included in expressions of both efficiency and performance time. MATLAB application is applied as an essential instrument. The test events produce that SVM has an efficiency 79.66% while Naive Bayes has 61.28 %.

In [9], they suggested a technique for analyzing liver cases applying a dataset received of UCI. Authors have applied decision tree, RF, SVM, perceptron and NB classifiers by the WEKA instrument. Following feature choice, the test events have given that the accuracies are 69.1252%, 70.669%, 70.8405%, 70.8405%, and 71.8696% for J48, SVM, MLP, Bayesian Network, and Random Forest, respectively.

Chaudhari et al [10] Illness analysis is an example of the several significant importance of such a scheme as it is one of the principal conditions of mortality all over the world. Foretell the personal control from complicated experiments carried in labs and further foretell the infection based on risk circumstances such as tobacco family history, age, smoking, diabetes, alcohol, high cholesterol, obesity. This document examines implementing KNN to improve healthcare specialists in the analysis of disease particularly heart disease. The outcomes show that KNN has good accuracy in the analysis of heart disease.

Ahmed et al. [11] Heart illness is a critical determinant of morbidity and death in the contemporary community. The therapeutic analysis is a necessary but complex responsibility that should be completed correctly and efficiently. This study article purposed to get out the heart illness within data mining, SVM, Genetic Algorithm, Neural Networks, association rules, and rough set. So it is mentioned that data mining could assist in the classification or the forecast of raised or low-risk heart illness.

3. Proposed Framework

In this part, the block design of the recommended mechanism is presented in Illustration 3. We crossed over two stages to developing the suggested tool: data preprocessing and data classification processing. A separate subsection is applied to a specific stage.

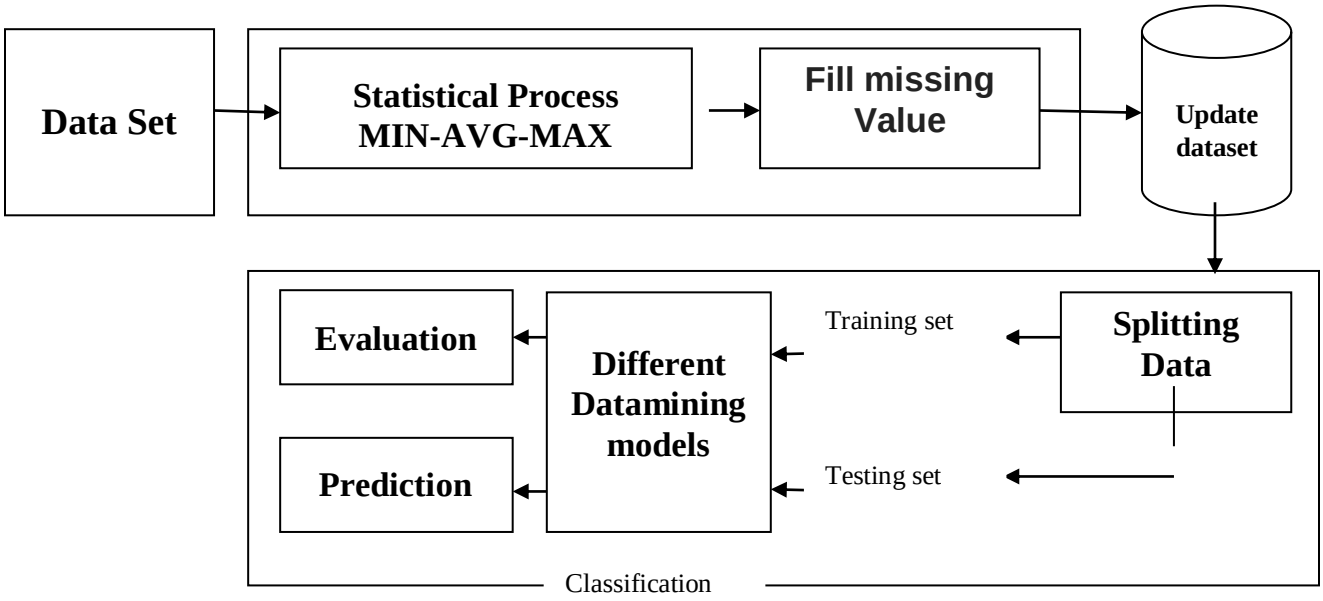

Figure 3 - The block design of the recommended tool

3.1.1 Pre-processing Stage

A. Statistical Processing

- Compute Min, Max, Mean and Standard Deviation

B. Filling Missing value

- Any missing values would be replaced with Mean value, or archived value

3.1.2 Hybrid Classification processing Stage

A. Sampling and voting

- Numerous classifiers elective includes separating the training data in reduced equal subsections from data and construct a Decision Tree structure for each subsection of data. Election is based on variety or majority voting. Every discrete classifier donates a solo vote.

B. Decision Tree

- In the decision tree method, we need to pick the excruciating feature that reduces the value from entropy and exploiting the Information Gain. To recognize an excruciating feature from the Decision Tree, should compute the Information Gain to every feature also then choose a feature that exploits an Information Gain.

$$E = \sum_{i=1}^{k} -P_i \, log_2 \, P_i \qquad 1)$$

Where

- k is that value from classes of this objective feature
- Pi is a value of incidences from class i separated via the whole value from occurrences

4. Comparative Study

4.1 Dataset

The dataset describes ten years (2006-2016) of clinical responsibility at one-hundred thirty US clinics and combined distribution channels. It covers over 9 characteristics describing patient and clinic results.

4.2 Dataset description

Table 1 - Dataset Statistical Information

Patient Records	1000 Records
Genders	476 Males 524 Females
Race	713 Caucasian 251 African American 7 Asian 8 Hispanic 21 Others
Range of ages	20 – 90 years

4.3 Properties of dataset

- Age
- Family History
- Gender
- Weight
- Blood Sugar
- cigarette smoking
- alcohol drinking
- admission type
- number of medications
- emergency visits in the year

4.4 Naïve Bayes

Table 2 - Naïve Bayes Result

Result	Values
Correctly Classified Instances	829

In Correct Classified Instances	171
Precision	86 %
Recall	82 %
ROC Area	90 %
Time	0.5 second

In the table2, we discuss the Naïve Bayes algorithm's results. We applied 10-fold cross validation for splitting records. We noticed that corrected classified records 829 and incorrect classified records 171 from total 1000 records. Also, we noticed that the accuracy ratio is low with 82% but it's fast. It is applied in 0.5 second.

4.5 Decision Tree

Table 3 – Decision Tree Result

Result	Values
Correctly Classified Instances	832
In Correct Classified Instances	168
Precision	83 %
Recall	83 %
ROC Area	86 %
Time	0.52 second

In the table3, we discuss the Decision tree algorithm's results. We applied 10-fold cross validation for splitting records. We noticed that corrected classified records 832 and incorrect classified records 171 from total 1000 records. Also, we noticed that the accuracy ratio is low with 83% but it's fast. It is applied in 0.52 second.

4.6 SVM

Table 4 – SVM Result

Result	Values
Correctly Classified Instances	860

In Correct Classified Instances	140
Precision	86 %
Recall	85 %
ROC Area	91 %
Time	0.3 second

In the table2, we discuss the SVM algorithm's results. We applied 10-fold cross validation for splitting records. We noticed that corrected classified records 860 and incorrect classified records 140 from total 1000 records. Also, we noticed that the accuracy ratio is good with 86% but it's very fast. It is applied in 0.3 second.

5. Experiments and Results

Table 4 - Tools and device used to preform proposed framework

Metric	**Values**
CPU	Intel core i7
RAM	4G
Operating system	Windows 10
Programming Language	PHP v4
Server Platform	Apache server

Table 5 - Confusion Matrix

	Predicted patient with diabetes disease (positive)	Predicted Healthy Persons (negative)
Actual Patient with diabetes disease	True Predicted Patient as (TP)	False Predicted Person as (FN)
Actual Healthy Persons	False Predicted Patient as (FP)	True Predicted Person as (TN)

For training and testing the data sets, we use ten-fold cross validation technique. This technique splits the dataset to 10 portions 9 portions are then applied to training and that tenth fragment is applied for testing. This is recurring, applying the alternative portion to the test section. Individually the data portion is utilized 1 for testing and 9 events for training. This is recurrent 10 events, including a novel portion doing the testing part. The average outcome is produced from the 10 runs.

The accuracy of the applied procedures must be evaluated applying under titles about rightly classified instances, wrongly classified instances, recall, precision, processing Time and accuracy.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad 2)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \times 100\% \quad 3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad 4)$$

Table 6 - Comparative Result between Proposed hybrid algorithm and other algorithms

Classifier	**Sensitivity**	**Specificity**	**Accuracy**
Decision Tree	83%	83%	86%
Naïve Bayes	86%	82%	90%
SVM	86%	85%	91%
Proposed Ensemble SVM+ Decision Tree (iteration=100)	91%	91%	94%

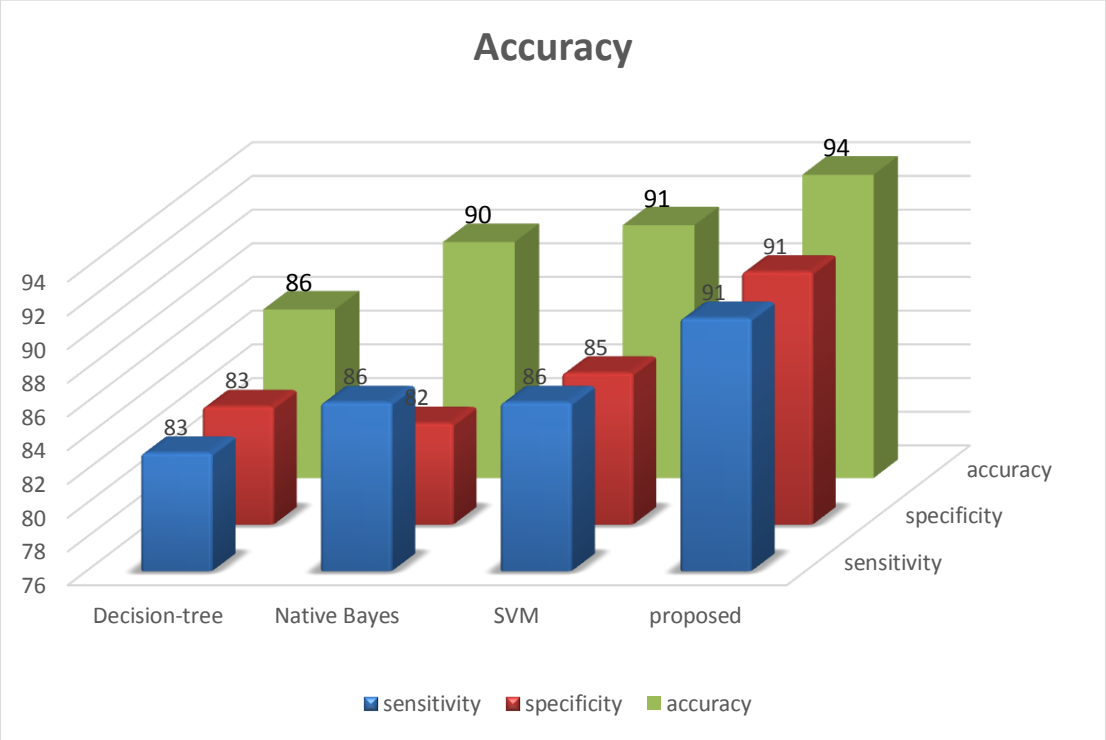

Figure 4- Accuracy Diagram of Comparative Study between proposed hybrid classification algorithm and other algorithms

6. Conclusion

Nowadays, data mining methods are playing an essential role in healthcare<management. In this document, we perform an overview of some uses of data mining procedures in identifying, diagnosing, and foretelling various conditions and syndromes. Lastly, a set of operation was led to assess this accuracy regarding a set of data mining procedures including Decision Trees (j48), Naïve Bayes, and hybrid proposed method of decision-tree and SVM into diabetes disease diagnosis. The trial outcomes have shown that a hybrid classifier is presenting the greatest achievement concerning accuracy by the large dataset.

References

[1] Assal, J. P., and L. Groop. "Definition, diagnosis and classification of diabetes mellitus and its complications." World Health Organization (2016): 1-65.
[2] National Diabetes Data Group. "Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance." Diabetes 28.12 (2016): 1039-1057.
[3] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare."Journal of healthcare information management 19.2 (2011): 65.
[4] SA, S. "Intelligent heart disease prediction system using data mining techniques." International Journal of Healthcare & Biomedical Research 1 (2013): 94-101.
[5] Otoom, A.F., E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour. Effective Diagnosis and Monitoring of Heart Disease. International Journal of Software Engineering and Its Applications. 9 (2015): 143-156.
[6] Vembandasamy, K., Sasipriya, R. And Deepa, E. Heart Diseases Detection Using Naive Bayes Algorithm. IJISET International Journal of Innovative Science, Engineering & Technology, 2(2015): 441-444.
[7] Afshari, A. A., and S. M. Mirhosseini, "A New Approach in Diabetes Diagnosis by Hybrid of Genetic Algorithm and Decision Tree." International Journal of Science Volume-5, Issue-1, pp. 805-814, 2016.
[8] Vijayarani, S. And S. Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms. International Journal of Science." Engineering and Technology Research (IJSETR), 4(2015): 816-820.
[9] Gulia, A., R.Vohra, and P. Rani, "Liver Patient Classification Using Intelligent Techniques." (IJCSIT) International Journal of Computer Science and Information Technologies, 5(2014): 5110-5115.
[10] Anand A. Chaudhari, Prof.S.P.Akarte, " Fuzzy and Data Mining based Disease Predection using K-NN Algorithm", International Journal of Innovations in Engineering and Technology, Vol. 3, Issue No. 4, April 2014
[11] Aqueel Ahmed, Shaikh Abdul Hannan, " Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering, Vol. 1, Issue No. 4, September 2012
[12] Ahmad, Parvez and Saqib Qamar. "Techniques of Data Mining In Healthcare: A Review." (2015).

