**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. 6, Issue. 1, January 2017, pg.43 – 51***

# A Weighted Voting of K-Nearest Neighbor Algorithm for Diabetes Mellitus

**Amal H. Khaleel[1], Ghaida A. Al-Suhail[2], Bushra M. Hussan[3]**

[1]Department of Computer Science, University of Basrah, Iraq

[2]Department of Computer Engineering, University of Basrah, Iraq

[3]Department of Computer Science, University of Basrah, Iraq

[1] amal_albahrany@yahoo.com; [2] ghaida-alsuhail@yahoo.com; [3] bushra5040@yahoo.com

***Abstract— Data mining is recently applied in the medical field to predict and diagnose diseases like Diabetes Mellitus. Such Diabetes is often called a modern-society disease and K-Nearest-Neighbor (KNN) algorithm is one of the best and the most usable classification algorithms used for diabetes diagnoses. However, the main problem in this algorithm is the equality in the effect of all attributes when calculating the distance between the new record and the available records in training dataset. Some attributes are classified with less importance; and others are with more importance. As a consequence, the effect will appear in the misleading of classification process and then decreasing the accuracy of classification algorithm. In this paper, we therefore use One-Attribute-Rule Algorithm to adjust the attributes weights and suggest a new classification algorithm called K-Nearest-Neighbor-Based-OneR (KNNB1R) that improves accuracy of KNN algorithm. The experiments are conducted on diabetes dataset to test and evaluate the performance of the proposed algorithm. The classification accuracy is eventually obtained to approach up to 92.91% for diabetic dataset.***

***Keywords— Diabetes Mellitus, Data Mining, Classification, K- Nearest Neighbor, Attribute Weighting, One-Attribute-Rule Algorithm.***

## I. INTRODUCTION

Nowadays, Diabetes is often called a modern-society disease. It is categorized as one of the major global health problems according to World Health Organization (WHO) reports; for instance, more than 25.8 million people, or 8.3% of the U.S. population, have diabetes. The total cost of health care for diabetes is expected to be $490 billion for 2030, accounting for 11.6% of the total health care expenditure in the world. Diabetes-Mellitus refers to the metabolic disorder that happens from misfunction in insulin secretion and action. It is characterized by hyperglycemia. The persistent hyperglycemia of diabetes leads to damage, malfunction and failure of different organs such as kidneys, eyes, nerves, blood vessels and heart [1-2]. The diagnosis of diabetes is very important; there are so many techniques in Artificial Intelligence (A.I.) that can be effectively used for the prediction and diagnosis of diabetes disease. These algorithms in artificial intelligence prove to be cost-effective and time saving for diabetic patients and doctors. In this paper, we therefore are diagnosing Diabetes Mellitus using a new developed algorithm based on K- Nearest Neighbor algorithm. It can be used as an effective Artificial Intelligence technique compared to traditional KNN [3].

KNN algorithm is one of the best and the most usable classification algorithms which is used largely in different applications. K-Nearest Neighbor is an example of instance-based learning, in which the training

 

dataset is stored, so that a classification for a new unclassified record may be found simply by comparing it to the most similar records in the training set [4]. The distance function is used in this method to determine which member of the training set is closest to an unknown test instance [5]. Also, because of its simplicity, KNN is easy to modify for more complicated classification problems. For instance, KNN is particularly well-suited for the object which has many class labels [6]. In this paper, we used One-Attribute-Rule Algorithm to update the weight attributes and suggest a new classification algorithm called K-Nearest-Neighbor-Based-OneR (KNNB1R). The proposed scheme improves the accuracy of KNN algorithm as the experiments show the performance of the proposed algorithm for diabetes dataset.

The organization of this paper is formed as follows. Related Work is briefed in Section II. The states some basic concepts of classification algorithms have been briefly explained in Section III. Dataset and Attributes is explained in Section IV. Proposed Work is explained in Section V, followed by results and conclusion in Sections VI and VII, respectively.

## II. RELATED WORK

There is a variety of research work which has been carried out by many researchers based on the observed medical diabetes data. Some of such works are discussed hereafter. Yu and Zhengguo (2007) [7] have presented the investigational result which shows the classification using traditional KNN algorithm and produce normal evaluation value, with fulfilment rate of 75%. In [8], HardikManiya *et al.* (2011) have done the comparative study of Naïve Bayes Classifier and KNN for Tuberculosis and they were able to justify the effectiveness of results using kNN to get further improvement. This was obtained by increasing the number of data sets and for Naïve Bayesian classifier and by increasing attributes or by selecting weighted features. Also, Jianping Gou *et al.* (2011) [9] have proposed classifier mainly employs the dual weighted voting function to reduce the effect of the outliers in the k nearest neighbors of each query object. In 2012, Karegowda *et al.* [10] have also used cascading K mean and K nearest Neighbor algorithm for categorization of diabetic patients in their paper. They classified diabetic patients by proposing results using KNN and K mean. The accuracy achieved by the proposed system was up to 82%.

Christobel *et al.* (2013) [11] have proposed a new class-wise K-Nearest Neighbor (CKNN) classification algorithm for classification of diabetes data. They have used diabetes data set for testing the CKNN algorithm and compared the various cases. The proposed CKNN model gives better classification accuracy of 78.16% compared to simple KNN. Also, S. Peter (2014) [12] has present an analytical study of numerous algorithms which includes clustering, classification, vector machines and neural networks. An analytical result has been validated for the approaches such as clustering, neural network, vector machines and hybrid approaches. It is observed that the hybrid approaches are observed to produce significant results in terms of the classification accuracy, processing time, etc.

Farahmandian *et al.* (2015) [13] have applied diabetes data set on various classification algorithms like SVM, KNN, Naïve bayes,ID3, CART and C5.0 to classify the diabetes data. They have compared the classification accuracy of these models. The findings indicate that SVM gives best classification accuracy of 81.77% compare to other schemes. Finally, T. Daghistani and Alshammari (2016) [14] have applied adult population data from Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia on three data mining algorithms, namely Self-Organizing Map (SOM), C4.5 and Random Forest, to predict diabetic patients using 18 risk factors. Random Forest achieved the best performance compared to other data mining classifiers.

## III. BACKGROUND AND EXISTING METHOD

Data mining (DM) is the method of identifying, exploring and modeling huge amounts of data that discover unidentified Patterns or relationships that produce a correct result. Thus data mining tools can be successfully applied in various fields in order to find patterns automatically with least amount of user input and efforts. Many Organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis and evaluate various trends and pattern of market and to produce quick and effective market trend analysis [15]. The tools and methods are mainly categorized as follows: Online Analytical Processing (OLAP), Classification [16], Clustering [15], Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining, Web Mining [17], Text Mining etc. These methods use different algorithms and can be implemented with different data and types.

More specifically, there are various data classification algorithms available in DM. Among these, K-Nearest Neighbor algorithm (KNN) used for this research is discussed hereafter.

 

**Algorithm: K-Nearest Neighbor** [16]:

1. Determine K means of the quantity of nearest neighbors. This value is all up to you.
2. Compute the distance between the query instance and all the training samples. You can use any distance algorithm.
3. Sort the distances for all the training samples and determine the nearest neighbor based on the K-th minimum distance.
4. Since this is supervised learning, get all the categories of your training data for the sorted value which fall under K.
5. Use the majority of nearest neighbors as the prediction value.

*A. The K-Nearest Neighbor Algorithm*

Algorithm of K-Nearest Neighbor (K-NN) is defined as a supervised learning algorithm used for classifying objects based on closest training examples in the feature space. KNN is the most basic type of instance-based learning or lazy learning. It assumes all instances are points in n-dimensional space. A distance measure is needed to determine the "closeness" of instances [3]. KNN algorithm is a simple technique that stores all available cases and classifies new cases based on a similarity measure. It is a type of lethargic knowledge where the function is only approximated nearby and the entire working out is deferred until classification. An entity is classified by the best part of its neighbors. K is always a positive integer. The correct classification is known because the neighbors are selected from a set of objects [16].

Some of the main advantages of KNN are: (*i*) it is very too simple to implement and easy to justify the outcome of KNN (*ii*) Robust to noisy training data (especially if we use Inverse Square of weighted distance as the "distance"), and (*iii*) Effective if the training data is large. Although KNN has those advantages, it has some disadvantages such as: (*i*) There is no thumb rule to determine value of parameter K, (*ii*) A high computation cost since it depends on computing the distance of each test instance to all training samples, and finally (*iii*) Low accuracy rate in multidimensional data sets with irrelevant features [7].

*B. The Performance of Nearest Neighbor Classification*

There are several key elements that may affect the performance of Nearest Neighbor classification [5]:

*1) Choosing Factor K*: One of the parameters to choose is the value of K. The value for K is pre-selected and the optimal value of K depends on the size and nature of the data [18]. Because all K nearest neighbors are considered equally important with respect to the classification, the choice of K is crucial [17]. Since using relatively large K may include too many points from other classes and on the other hand, using very small K may make the result sensitive to noise points. In both cases the classification accuracy will decrease [6]. The data analyst needs to balance these considerations when choosing the value of K that minimizes the classification or estimation error and highest accuracy [4].

*2) Choice of Distance Metric*: A distance metric measures the dissimilarity between two data points in terms of some numerical value. It also measures similarity; we can say that more distance is the less similar the data points, and less distance is the more similar the data points [18].

The choice of the distance measure is another important consideration. Commonly, Euclidean or Manhattan distance measures are used. For two points x and y, with n attributes, these distances are given by the following formulas:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad \text{Euclidean distance} \quad (1)$$

$$d(x,y) = \sqrt{\sum_{k=1}^{n}|x_k - y_k|} \quad \text{Manhattan distance} \quad (2)$$

where $x_k$ and $y_k$ are the $k^{th}$ attributes (components) of x and y, respectively [6]. Although there are other possible choices, most instance-based learners special k nearest neighbors (KNN) classification use Euclidean distance function [5].

 

*3) Weighted Voting*: The problem, when using the Euclidean distance between two points in KNN algorithm is the weighting of the contributions of the different attributes [19], where the KNN algorithm uses all the record attributes equally however; all attributes might not have the same role in the classification process. Perhaps, some of these attributes are irrelevant to the classification, these irrelevant attributes can lead to distinguish two near records so far from each other and so the classification cannot be done correctly. Idiomatically, this problem is called as a curse of dimensionality [20].

Generally, there are two types of combination functions: un weighted voting and weighted voting. In the un-weighted voting combination function, the class label which has the majority between neighbors of new record is selected as the class label of the new record without considering the preference of each neighbor. But, in the weighted voting more weight is given to some neighbors that are so close to the new record. In other words, the ones which are more similar to the new record [20], where each objects weights vote by its distance and the influence of a particular record is inversely proportional to the distance of the record from the new record to be classified; but when the distance is zero, the inverse would be undefined.

In this case the algorithm should choose the majority classification of all records whose distance is zero from the new record. The weight factor is often taken to be the reciprocal of the squared distance [4,6]:

$$P_{wi} = 1/(d(x_1, x_2))^2 \quad (3)$$

Where $x_1$, $x_2$ denotes the two records.

*C. One-Attribute-Rule Algorithm*

OneR is basic algorithm used to find classification rules. OneR is a simple, cheap method that often comes up with quite good rules for characterizing the structure in data. It generates a one-level decision tree expressed in the form of a set of rules that all test one particular attribute [21].

The basic idea is that rules are constructed to test a single attribute and branch for every value of that attribute. For each branch, the class with the best classification is the one occurring most often in the training data. The error rate of the rules is then determined by counting the number of instances that do not have the majority class in the training data. Finally, the error rate for each attribute's rule set is evaluated, and the rule set with the minimum error rate is chosen [22].

**Algorithm: One-attribute-Rule(OneR)**

```
For each attribute A:
    For each value V of that attribute, create a rule:
        Count how often each class appears
        Find the most frequent class, c
        Make a rule "if A=V then C=c"
    Calculate the error rate of this rule.
Pick the attribute whose rules produce the lowest error rate.
```

*D. Performance Measures*

For the calculation of the predicted positive cases the below mentioned formulas are used [3]:

- True positive (TP): Those Sick people who are correctly diagnosed as sick
- False positive (FP): The Healthy people who are incorrectly identified as sick
- True negative (TN): The Healthy people who are correctly identified as healthy
- False negative (FN): The Sick people who are incorrectly identified as healthy

Various performance measures like sensitivity, specificity, accuracy and F-Measure are calculated using this matrix as depicted in Table 1:

 

TABLE I

THE PERFORMANCE MEASURE FORMULAS [23]

| **Performance Measure** | **Formulas** |
|---|---|
| Precision | TP / (TP+ FP) |
| Recall (Sensitivity) | TP / (TP+ FN) |
| Accuracy | (TP + TN) / (TP+ TN +FP +FN) |
| Specificity | TN / ( TN+ FP) |
| F-Measure | (2* Recall* Precision) / ( Precision + Recall) |

## IV. DATA SET AND ATTRIBUTES

The Indian Diabetes dataset, the dataset consists of 8 attributes plus class (Table 2). The dataset was collected from 768 females. The diagnosis can be carried out depending on personal data (age, number of times pregnant) and results of medical examination (blood pressure, body mass index, result of glucose tolerance test, triceps skin fold thickness, serum insulin, pedigree function). There are 500 samples of class 1 (diabetes) and 268 of class 2 (not diabetes). The original source of the data in Indian is the National Institute of Diabetes, and we have used in our work is taken from - http://mlearn.ics.uci.edu/MLRepository.html. [21].

To improve the performance of K- nearest neighbor technique for classification dataset, we used *Min–Max Normalization* [25] *to prevent attributes with initially large ranges from outweighing attributes with initially smaller ranges, and* filling the missing value by used *hybrid k-nearest neighbor imputation (KNNI)-based diabetes* method to impute missing values [26]. This algorithm is very important because it increases the plausibility and accuracy of the forecasts.

TABLE II

THE CHARACTERISTICS USED FOR DIABETES TYPE II DIAGNOSE [24]

| **No. of Feature** | **Feature** | **Descriptions and Feature Values** |
|---|---|---|
| 1 | Number of times Pregnant | Numerical values |
| 2 | Plasma Glucose Concentration | Numerical values |
| 3 | Diastolic Blood Pressure | Numerical values in (mm Hg) |
| 4 | Triceps Skin Fold Thickness | Numerical values in mm |
| 5 | 2-Hour Serum Insulin | Numerical values in (mu U/ml) |
| 6 | Body Mass Index (BMI) | Numerical values in (weight in kg/(height in m)^2) |
| 7 | Diabetes Pedigree Function (DPF) | Numerical values\ |
| 8 | Age | Numerical values |
| 9 | Diagnosis of type 2 diabetes disease | sick=1 , Normal=0 |

## V. THE PROPOSED ALGORITHM KNNB1R-BASED DIABETES

We present a new classification algorithm named K-Nearest-Neighbor-Based-OneR (KNNB1R) that improves accuracy of KNN algorithm. In our paper, we used to weigh attributes association rules (OneR) algorithm. The advantage of this method of figuring out of the distance, not only, the quantity of attributes that is considered but also, the quality of attributes that is emphasized, so it increases the classification accuracy. It's obvious that how accurate weights may be the classification accuracy increase but, if the weights are not selected accurately, the classification accuracy even decreases than before.

 

The following steps describe our proposed (KNNB1R) algorithm for diabetes classification:

**The Proposed (KNNB1R) Algorithm:**

1. Input: The Diabetes Dataset
2. For each attribute A:
   For each value V of that attribute:
   - Count how often each class appears.
   - Find the most frequent class(C), c.
   - Make a rule "if A=V then C=c".

   Calculate the succeed rate of this rule.
   Calculate the succeed rate of this attribute.
3. Weighted Voting of attribute ($w_i$) is equal the succeed rate of this attribute.
4. Determine the parameter K number of nearest neighbors.
5. Calculate the distance (d) between the query-instance and all the training samples, use Euclidean distance with weighted voting (w).

   $d(x_1,x_2) = \sqrt{\sum_{i=1}^{n}\left(w_i \times (x_{1i} - x_{2i})\right)^2}$, where n denotes the number of attributes.
6. Sort the distances for all the training samples and determine the nearest neighbor based on the K-th minimum distance.
7. Since this is supervised learning, get all the categories of your training data for the sorted value which fall under K.
8. Use the higher weight among weights of nearest neighbors as the prediction value. Where, the weight of neighbor ($p_{wi}$)= $\frac{1}{(d(x_1,x_2))^2}$ , but if the distances for all the samples are equal zero then use the majority of nearest neighbors as the prediction value.

## VI. Experimental Results And Discussion

### *A. Configurations*

The framework of project work is designed by (Delphi Ver.7) on a PC with the following configurations Core i7 laptop with 4GB of RAM for implementing the algorithms, running under Microsoft Windows 8_64 bits.

### *B. Testing Results*

We experimentally tested KNNB1R accuracy using the Indian Diabetes dataset. It was divided into 70% of training and 30% of testing data. In this paper, it is found that when using relatively large K, it may include too many points from other classes and on the other hand, using very small K may make the result more sensitive to noise points. Thus to reduce the execution time in program, we test values of K (3, 5, 7 and 10) on training sets to estimate the error rate of the classifier. Table 3 shows the chosen values of K for K-nearest neighbor method. For example, in Fig.1 we chose K=5 because it achieves the best accuracy.

It is evident from Table 4 that the performance results with weighted voting are much better than those without weighted voting in k-nearest neighbor methods. For example, the accuracy of dataset without missing achieves (73.88%) in case of KNN method without weight; meanwhile the accuracy of this method achieves (92.91 %) when we use weight. Figure 2 clearly shows results of Performance Measures for K = 5 between the existing KNN without weight and the proposed K-NNB1R methods.

 

TABLE III

THE PERFORMANCE COMPARISON BETWEEN PROPOSED K-NN METHODS (EFFECT OF K)

| KNN Method | Accuracy % | | | |
|---|---|---|---|---|
| | K=3 | K=5 | K=7 | K=10 |
| K-Nearest Neighbor without weight | 69.40 % | 73.88% | 70.89 % | 67.16 % |
| Proposed K-Nearest Neighbor with weight | 92.53 % | 92.91 % | 86.94 % | 89.55 % |

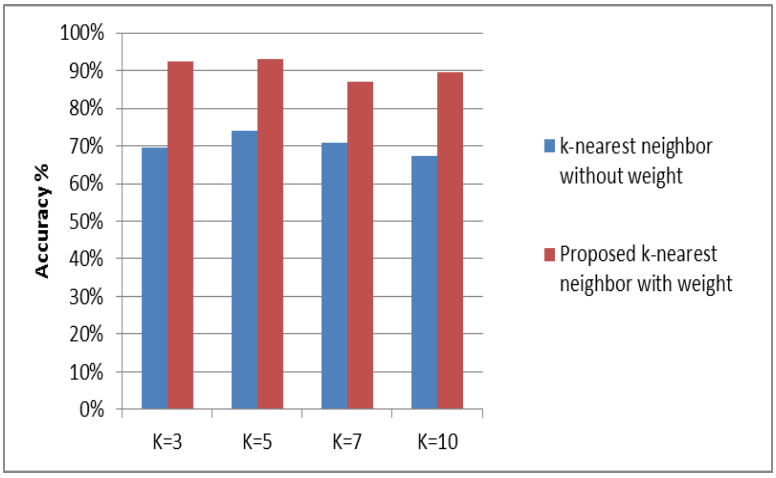


Fig. 1 The Performance comparison between KNN and proposed KNNB1R method to describe the effect of K on accuracy

TABLE IV

THE PERFORMANCE MEASURES BETWEEN KNN AND PROPOSED KNNB1R FOR K = 5

| Performance | KNN without weight | Proposed KNN with weight (KNNB1R) |
|---|---|---|
| Precision | 0.83598 | 0.89171 |
| Recall | 0.80203 | 0.98591 |
| Specificity | 0.56338 | 0.86507 |
| F-Measure | 0.81865 | 0.93645 |
| Accuracy % | 73.88059% | 92.91044% |

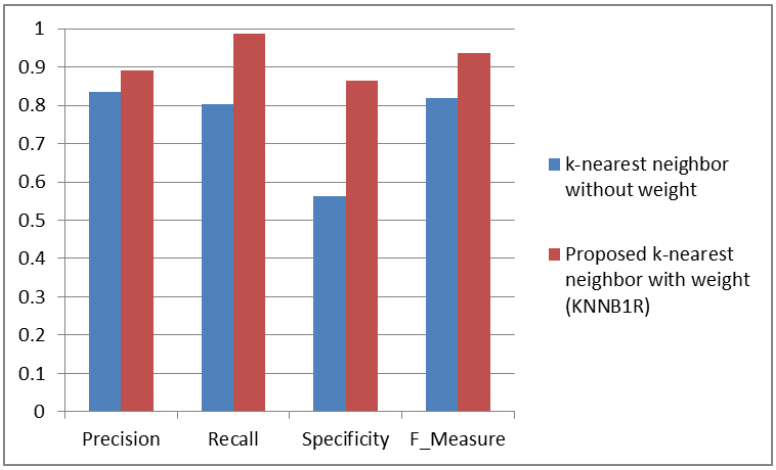


Fig. 2 A Performance comparison between KNN and the proposed KNNB1R method for K = 5

 

*C. Comparison of Proposed Work vs. Existing Works*

The implementation of our proposed work in Section V is compared with other existing works as in Table 5. It is found that our proposed KNNB1R algorithm outperforms the other existing methods in [27], [11] and [28] and it becomes slightly superior to the accuracy obtained by Sarwar and Sharma [29]. Obviously, the proposed KNNB1R achieves 92.91%.

TABLE V

A COMPARATIVE ACCURACY OF PROPOSED WORK WITH VARIOUS EXISTING WORKS

| **Authors /Methods** | **Accuracy** |
|---|---|
| Y. A. Christobel and P. Sivaprakasa [27] | 73.38% |
| Christobel *et al.* [11] | 78.16% |
| G.Visalatchi *et al.* [28] | 78 % |
| A. Sarwar and V. Sharma [29] | 91 % |
| Proposed work | 92.91% |

## VII. CONCLUSION

A number of algorithms were proposed for the prediction and diagnosis of diabetes. These algorithms provide more accuracy than the available traditional systems. However, in this paper, we have considered a new classification algorithm KNNB1R which uses the One-attribute-Rule algorithm in the KNNB1R algorithm in order to increase the classification accuracy of the KNN algorithm. In this algorithm, each of attributes is given weight by One-attribute-Rule algorithm. The performance of classification is then measured with respect to sensitivity, specificity and accuracy. It is found that accuracy has increased significantly in the case of proposed KNNB1R algorithm. Future works may address hybrid classification model using KNN with other data mining techniques.

# REFERENCES

[1] A. K. Dewangan, and P. Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," *International Journal of Engineering and Applied Sciences (IJEAS)*, vol. 2, no. 5, May 2015.

[2] M. Marinov, A.S. M. Mosa,I. Yoo and S. A. Boren, "Data-Mining Technologies for Diabetes: A Systematic Review,"*Journal of Diabetes Science and Technology*, vpl. 5, no. 6, Nov. 2011.

[3] K. Saxena, Z. Khan and S. Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm," *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 2 , no. 4, July-Aug 2014.

[4] D. T. Larose, *Discovering Knowledge in Data*, John Wiley & Sons, United States of America, 2005.

[5] I. H. Witten and E. Frank, "*Data Mining Practical Machine Learning Tools and Techniques*", 2ed Elsevier, United States of America, 2005.

[6] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*, Taylor & Francis Group, USA 2009.

[7] W. Yu and W. A. Zhengguo, "*Fast KNN algorithm for Text Categorization*," *in Proc. of the 6th International Conference on Machine Learning and Cybernetics*, Hong Kong, 2007.

[8] H. Maniya, M. I. Hasan and K. P. Patel, "Comparative Study of Naïve Bayes Classifier and KNN for Tuberculosis," *International Journal of Computer Applications* (*IJCA*), pp. 22-26, 2011.

[9] J. Gou, T. Xiong and Y. Kuang, "A Novel Weighted Voting for K-Nearest Neighbor Rule," *Journal of Computers*, vol. 6, no. 5, May 2011.

[10] A. G. Karegowda, M.A. Jayaram and A.S. Manjunath,"Cascading K-means Clustering and K-NearestNeighbor Classifier for Categorization of Diabetic Patients," *International Journal of Engineering and Advanced Technology* (*IJEAT),* vol.1, no.3, pp. 147-151, 2012.

[11] Y. A. Christobel and P. Sivaprakasam, "A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset,"*IJEAT*, vol. 2, no. 3, pp. 396-400, February 2013.

[12] S. Peter, "An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus," Bonfring *International Journal of Data Mining*, Vol. 4, No. 2, June 2014.

[13] M. Farahmandian, Y. Lotfi and I. Maleki, "Data Mining Algorithms Application in Diabetes Diseases Diagnosis: A Case Study," MAGNT Research Report. vol. 3, pp. 989-997, 2015.

[14] T. Daghistani and R. Alshammari, "Diagnosis of Diabetes by Applying Data Mining Classification Techniques," *International J. of Advanced Computer Science & Applications (IJACSA)*, vol. 7, no. 7, 2016.

[15] R. Naik and N. Deepika, "Data Mining System and Applications: A Study", *International Journal of Computer Science and Mobile Computing (IJCSMC)*, vol. 5, no. 12, pp.103 – 110. Dec. 2016,

 

[16] K. Saravananathan and T. Velmurugan, "Analyzing Diabetic Data using Classification Algorithms in Data Mining," *In proceeding of Indian Journal of Science and Technology*, vol 9, no. 43, Nov 2016.
[17] Z. Markov and D. T. Larose, *Data Mining The Web*, John Wiley & Sons, United States of America, 2007.
[18] M. A. M. Khan, "Fast Distance Metric Based Data Mining Techniques Using Ptrees: K-Nearest-Neighbor Classification and k-Clustering," M.Sc. Thesis, North Dakota State University, North Dakota, USA, 2001.
[19] M. Bramer, *Principles of Data Mining*, London Limited, England, 2007.
[20] M. Moradian and A. Baraani, "KNNBA: K-Nearest-Neighbor-Based-Association Algorithm," *Journal of Theoretical and Applied Information Technology*, vol. 6, no.1, pp.123 - 129, 2009.
[21] H. A. Nguyen and D. Choi, "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model," *In proceeding of APNOMS*, pp. 399–408, 2008.
[22] M. Berry and M. Browne, *Lecture Notes in Data Mining*, World Scientific, USA, 2006.
[23] A. K. Dewangan and Pragati Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," *International Journal of Engineering and Applied Sciences (IJEAS)*, vol. 2, no. 5, May 2015.
[24] S. Sa'di1, A. Maleki1, R. Hashemi, Zahra Panbechi1 and Kamal Chalabi1, "Comparison Of Data Mining Algorithms In The Diagnosis Of Type II Diabetes," *International Journal on Computational Science & Applications (IJCSA)*, vol. 5, no.5, Oct 2015.
[25] L. Al Shalabi, Z. Shaaban, and B. Kasasbeh, "Data Mining: A Pre-processing Engine," *Journal of Computer Science*, vol. 2, no. 9, pp. 735-739, 2006.
[26] Bushra M. Hussan, Ghaida Al-Suhail, and Amal Hameed Khaleel, "Studying the Impact of Handling the Missing Values on the Dataset On the Efficiency of Data Mining Techniques," *Basrah Journal of Science (A)*, vol. 30, no. 2, pp.128-141, 2012.
[27] Y. A. Christobel and P. Sivaprakasa, "Improving the Performance of K-nearest Neighbor Algorithm for the Classification of Diabetes Dataset With Missing Values," *International Journal of Computer Engineering and Technology (IJCET)*, vol. 3, no. 3, pp. 155-167, Oct- Dec 2012.
[28] G.Visalatchi, S. J Gnanasoundhari, and M. Balamurugan, "A Survey on Data Mining Methods and Techniques for Diabetes Mellitus," *International Journal of Computer Science and Mobile Applications* (IJCSMA), vol.2, no 2, pg. 100-105, Feb 2014.
[29] A. Sarwar and V. Sharma, "Comparative Analysis of Machine Learning Techniques in Prognosis of Type II Diabetes," AI & SOCIETY *Journal of Knowledge, Culture and Communication*, Springer, vol. 29, pp 123–129, Feb 2014.

 