https://doi.org/10.31436/iiumej.v21i1.1206

DIABETES CLASSIFICATION BASED ON *KNN*

AMEER ALI[1]*, MOHAMMED ALRUBEI[2], LAITH FALAH MOHAMMED HASSAN[1], MOHANNAD AL-JA'AFARI[1] AND SAIF ABDULWAHED[2]

[1]*Najaf Technical Institute, Al-Furat Al-Awsat Technical University, 31001 Al-Najaf, Iraq.*
[2]*Al-Furat Al-Awsat Technical University, 31001 Al-Najaf, Iraq,*
[3]*Al-Furat Al-Awsat Technical University, 31001 Al-Najaf, Iraq.*

**Corresponding author: inj.ame7@atu.edu.iq*

(Received: 26th July 2019; Accepted: 7th December 2019; Published on-line: 20th January 2020)

ABSTRACT: Diabetes is a life-threatening syndrome occurring around the world; it can have huge complications and is documented by large amounts of medical data. Therefore, attempts at early detection of this disease took a large area of research and many methods were used to deal with diabetes. In this paper, different types of KNN algorithm have been used to classify diabetes disease using Matlab. The dataset was generated by the criteria of the American diabetes association. For the training stage, 4900 samples have been used by the classifier learner tool to observe the results. Then, 100 of the data samples were used for the test. The results show that the KNN types (Fine, Weighted, Medium and Cubic) give high accuracy over the Coarse and the Cosine methods. Fine KNN is considered the most suitable according to its accuracy of classified samples.

ABSTRAK: Penyakit kencing manis adalah sindrom penyakit ancaman nyawa yang berlaku di seluruh dunia dan ia mempunyai data perubatan yang besar serta komplikasi tinggi. Oleh itu, cubaan dalam mengesan awal penyakit ini mempunyai potensi luas dalam kajian dan banyak kaedah telah digunakan bagi mengkaji penyakit kencing manis. Dalam kajian ini, pelbagai jenis algoritma KNN telah digunakan bagi mengelas penyakit kencing manis menggunakan Matlab. Setdata dihasilkan berdasarkan kriteria Kesatuan Kencing Manis Amerika. Pada peringkat latihan, sebanyak 4900 sampel telah digunakan oleh pelatih alat pengelasan bagi memantau dapatan kajian. Kemudian, 100 daripada sampel data telah digunakan bagi ujian. Keputusan menunjukkan jenis KNN (Halus, Berat, Sederhana dan Kubik) lebih tepat berbanding kaedah Kasar dan Kosinus. KNN Halus di dapati lebih sesuai berdasarkan ketepatan sampel pengelasan.

KEYWORDS: *diabetes; KNN; classification machine learning*

1. INTRODUCTION

Diabetes disease is chronic and widespread in the world. It occurs by disorder in insulin secretion that causes an irregular increase in glucose level. Spread of this disease has been observed in due to unhealthy diets [1]. Generally, a higher probability of diabetes infection correlates to various factors such as female gender, age over 35, and persons with unhealthy weight [2]. Many methods have been produced for diagnosis of different types of diseases, which used intelligent algorithms to classify, cluster, and diagnose these diseases. Support Vector Machine (SVM) used as a classifier to diagnose diabetes based on a medical dataset [3,4,5]. Bayes theorem used to classify prediction accuracy in diabetes data with a data set delivered from Diabetes 130-US hospitals [6]. Decision Tree and Naïve Bayes algorithms presented to analyse and classify the patterns in order to

https://doi.org/10.31436/iiumej.v21i1.1206

diagnose the diabetes disease [7]. Coupling methods: likelihood ratio test, joint clustering, and classification based data presented from Boston Medical Center, in order to diagnose diabetes [8]. Fasting plasma glucose used to predict and diagnose type 2 diabetes based on two machine-learning algorithms: logistic regression and naive Bayes classifier [9]. Fuzzy logic also presented system to diagnose diabetes based on five layers [10]. Convolution neural network used to diagnose many diseases belonging to the diabetes family using [11]. In this paper, six types of KNN algorithms have been used to investigate its ability to classify glucose levels.

The organization of this paper is arranged as follows: the description of KNN algorithms and their distance equations are presented in section two. Section three presents the dataset criteria for diabetes. The discussion of results for KNN algorithms has been explained in section four to identify which type of KNN is more suitable. Finally, the conclusion of this experiment has been introduced in section five.

2. K-NEAREST NEIGHBOURS (*KNN*)

The classification is a type of supervised machine learning. The KNN is one of the classification techniques that is commonly used to classify data input into pre-defined classes (k) [12]. It was proposed by Cover and Hart in 1968. The straightforward mechanism of the KNN algorithm is to compute the Euclidean distance function between pre-defined classes and each varying sample. After that, the KNN algorithm chooses the minimum nearest neighbours according to each category. The samples are assigned to their category based on the nearest k neighbours. There are many versions of distance function between the samples. In this paper, the most commonly used is the Euclidean distance, expressed in Eq. (1) [13].

$$d = \sqrt{\sum_{k=1}^{n}(X_{1k} - X_{2k})^2} \tag{1}$$

Where k is the number of values in each sample vector, and X_1, $X_2$are input samples.

Six types of KNN have been chosen to classify the dataset. Their details are described as follows [14]:

Fine KNN takes one neighbour to distinguish the sample data, while the Medium KNN takes more neighbours than Fine KNN for distinction. This type will cause a low distinction feature to the algorithm. The Coarse KNN takes more neighbours than Medium KNN, which leads to the lowest distinction feature amongst the three types.

The Cosine KNN uses a Cosine distance metric as in Eq(2). The Cubic KNN uses a cubic distance metric as in Eq(3). The weight KNN uses a distance weight as in Eq(4). The last three types have the same number of neighbours as Medium KNN [15].

$$d = \left(1 - \frac{x_1 x_2'}{\sqrt{(x_1 x_1')(x_2 x_2')}}\right) \tag{2}$$

$$d = \sqrt[3]{\sum_{k=1}^{n}|x_{1k} - x_{2k}|^3} \tag{3}$$

$$d = \sqrt{\sum_{k=1}^{n} w_i (x_{1k} - x_{2k})^2} \tag{4}$$

When the numbers of neighbours are decreased, the accuracy of the classifier increases. This will increase the complexity of the classifier model, but does not guarantee that the out-of-samples will be classified correctly [16].

https://doi.org/10.31436/iiumej.v21i1.1206

3. DIABETES CRITERIA

The dataset presents amounts for each of the variables based on the criteria, such as glucose before food and glucose after food of an object. Each value or amount is called as a datum. The criteria used in this research are summarized in Table 1. The objective of these criteria is to generate a dataset to diagnose diabetes in humans. Based on a personal dataset, such as (HB A1C test, Fasting test, and Random test), it tries to decide if a human subject has diabetes, based on whether its values are normal or not. Diabetes criteria are already in public domain, thanks to the American diabetes association standards [17].

Table 1: Diabetes criteria [17]

HB A1C **A**	**Fasting (FPG)** **B**	**Random (PG)** **C**	**Response** **Y**
5-5.9 A0	90-119 B0	140-199 C0	Normal (0)
6-6.5 A1	120-140 B1	200-250 C1	Neuropath (0.25)
6.6-7 A2	141-180 B2	251-300 C2	Retinopathy (0.5)
7.1-7.5 A3	181-250 B3	301-350 C3	Nephropathy (0.75)
7.6-8 A4	251-600 B4	351-500 C4	Heart disease (1)

The diagnosis of diabetes is based on glucose and plasma; either the 2-h Plasma Glucose (2-h PG) value after a 75-g oral glucose tolerance test or the Fasting Plasma Glucose (FPG). A1C (threshold $\geq$6.5%) is added as a third option in the diagnosis. The A1C test uses a Diabetes Control that is certified by the National Glycohemoglobin Standardization Program.

The relationship between the A1C and the risk of retinopathy, as with FPG and 2-h PG, were shown in epidemiological data. The compatibility between the 2-h PG tests with FPG and between A1C with glucose-based tests are <100%. The A1C and oral glucose tolerance have several benefits to the FPG, including greater convenience and less daily disturbances during stress and sickness [17,18].

4. RESULTS

In this experiment, we will train the classifier using the sample data generated from Table 1. Each training sample has four values. The first three values are input samples that indicate HB A1C, Fasting (FPG), and Random (PG). The fourth one is the response. Then the classifier was tested using out-of-sample data to calculate the accuracy of the distinction. Table (2) summarizes the parameters of the experiment.

Table 3 shows the accuracy prediction speed and training time for all KNN types using 4900 input samples. For the accuracy, Fine KNN, Weighted KNN, Medium KNN, Cubic KNN, Cosine KNN, and Coarse KNN have been arranged according to their performance. In addition, according to the mathematical computation distance, the maximum training time was taken by the Cosine KNN while the Medium KNN takes the least training time. Thus, the prediction speed (observation/second) is maximum for Fine KNN and minimum for Cosine KNN.

Table 2: Simulation parameters

No.	Coefficient	Description
1	**Fine KNN**	K=1
2	**Medium KNN**	K=10
3	**Coarse KNN**	K=100
4	**Cosine KNN**	K=10
5	**Cubic KNN**	K=10
6	**Weighted KNN**	K=10
7	**Number of training samples**	4900
8	**Number of testing samples**	100
9	**Training tool**	Classification learner using Matlab

Table 3: Experiment outcome information

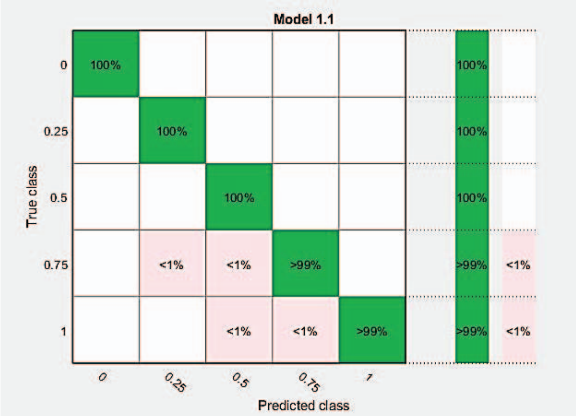
Preset	Accuracy	Training time	Prediction speed
Fine KNN	99.9%	0.54411 sec	200000 obs/sec
Medium KNN	98.4%	0.26525 sec	100000 obs/sec
Coarse KNN	74.3%	0.46863 sec	43000 obs/sec
Cosine KNN	85.6%	0.69376 sec	25000 obs/sec
Cubic KNN	98.2%	0.47105 sec	52000 obs/sec
Weighted KNN	99.8%	0.27984 sec	110000 obs/sec

The pre-trained classifier tested 100 random out-of-sample data to check the distinction capability, which it is enough to cover the reliability of the classifier over all classes. As expected, the numbers of classification errors are 0, 0, 19, 16, 2, and 0 for Fine KNN, Medium KNN, Coarse KNN, Cosine KNN, Cubic KNN, and Weighted KNN, respectively.

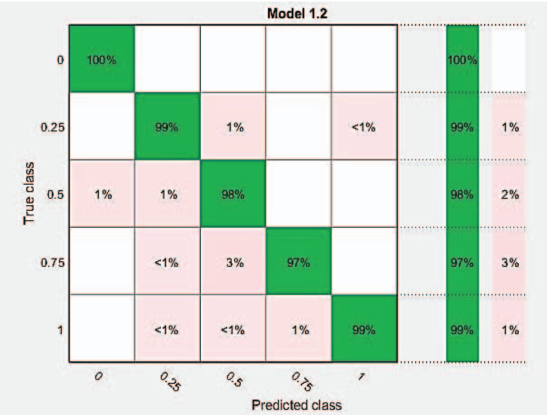
Figure 1 shows the confusion matrix for each of the KNN types. In each model in Fig. 1, the diagonal of green squares represents the correct prediction ratio of the predict classes over true classes while the red squares give the incorrect class ratio. The ratio of each square is associated with the disparity of the colours. The white squares are empty for samples from the training stage.

The results show the magnitude of decision ratio between the true value and the prediction value. Fine, Medium, Cubic, and Weighted give the best decision while Coarse and Cosine give high failure ratios. Figure 1 shows that the Fine KNN classifier is the most preferable because it gives 100% accuracy for Normal, Neuropath, and Retinopathy classes and <1% for Nephropathy and Heart disease.

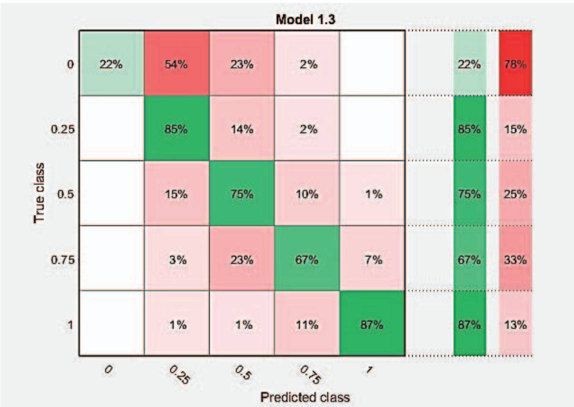
https://doi.org/10.31436/iiumej.v21i1.1206



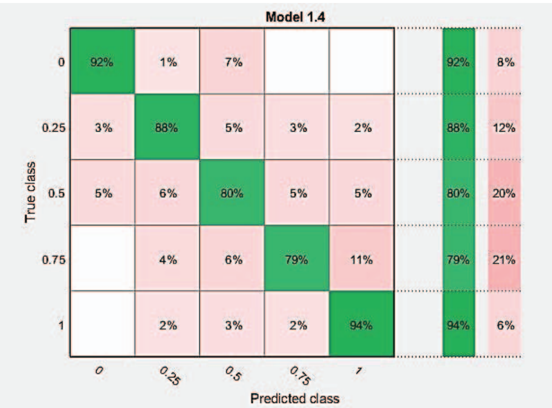
Fine KNN



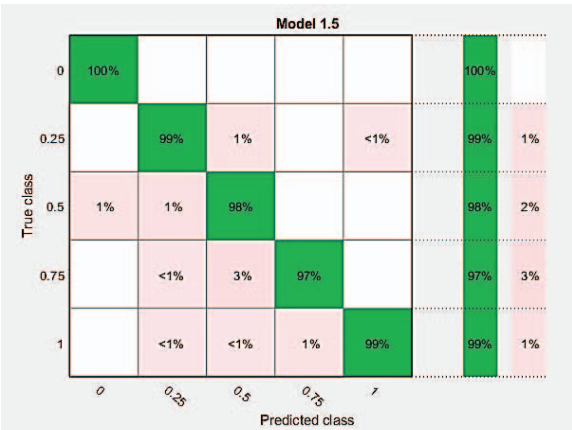
Medium KNN



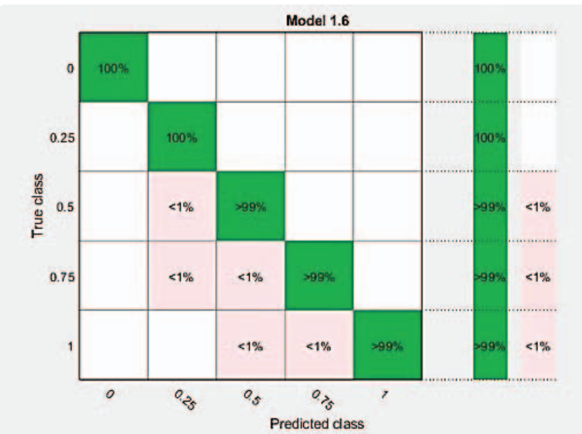
Coarse KNN



Cosine KNN



Cubic KNN



Weighted KNN

Fig. 1: The confusion matrix for different types of KNN.

5. CONCLUSION

The dataset of diabetes has been classified using different types of KNN algorithms. The simulation results show that Fine, Medium, Cubic, and Weighted KNN types have a superior performance over Coarse and Cosine. The classifier model of all KNN types required less than 0.7 sec to predict the target. In addition, it can be said that the Fine KNN

https://doi.org/10.31436/iiumej.v21i1.1206

algorithm is suitable to solve the diabetes classification problem with higher accuracy than other types.

REFERENCES

[1] T.Karthikeyan, K.Vembandadsamy. (2015). An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus. International Journal of Computer Application,7:2250-1797.

[2] Catherine C. Cowie, Keith F. Rust,Danita D. Byrd-Holt, Edward W. Gregg, Earl S. Ford, Linda S. Geiss, Kathleen E. Bainbridge, Judith E. Fradkin. (2010). Prevalence of diabetes and high risk for diabetes using A1C criteria in the US population in 1988–2006. Diabetes care, 33: 562-568.

[3] V. Anuja Kumari, R.Chitra. (2013). Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications,3:1797-1801.

[4] Longfei Han, Beijing, Senlin Luo, Jianmin Yu, Limin Pan, Songjing Chen. (2013) Rule Extraction from Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes. IEEE,19:2168-2194.

[5] Nahla H. Barakat, Andrew P. Bradley, Mohamed Nabil H. Barakat (2010). Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. IEEE,40:2168-2194.

[6] Subhankar Manna, Malathi G. (2017). Performance Analysis Of Classification Algorithm On Diabetes Healthcare Dataset. International Journal of Research - Granthaalayah,5:260-266.

[7] Aiswarya Iyer, S. Jeyalatha, Ronak Sumbaly (2015). Diagnosis Of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP),5:1-14.

[8] Theodora S. Brisimi, Tingting Xu, Taiyao Wang, Wuyang Dai, William G. Adams , Ioannis Ch. Pasc halidis(2018). Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach. IEEE,106:1-18.

[9] Bum Ju Lee, Boncho Ku, Jiho Nam, Duong Duc Pham, Jong Yeol Kim (2014). Prediction of Fasting Plasma Glucose Status Using Anthropometric Measures for Diagnosing Type 2 Diabetes. IEEE,18:555-561.

[10] Chang-Shing Lee, Senior Member, Mei-Hui Wang (2011) A Fuzzy Expert System for Diabetes Decision Support Application. IEEE,41:139-153.

[11] Yuliang Liu, Quan Zhang, Geng Zhao, Zhigang Qu, Guohua Liu, Zhiang Liu, Yang An (2018). Detecting Diseases by Human-Physiological- Parameter-Based Deep Learning. IEEE,7: 2169-3536

[12] Shichao Zhang, Xuelong Li,Ming Zong, Xiaofeng Zhu, Ruili Wang (2017). Efficient kNN Classification With Different Numbers of Nearest Neighbors. IEEE,29:1-12.

[13] Mehdi Zekriyapanah gashti (2018). A Modified Model Based on Flower Pollination Algorithm and K-Nearest Neighbor for Diagnosing Diseases. IIUM Engineering Journal, 19:144-157.

[14] Jenifer Mariam Johnson and Anamika Yadav(2016). Fault detection and classification technique for HVDC transmission lines using KNN, Second International Conference on ICT for Sustainable Development, Goa, 1-2 July 2016.

[15] Jaime Vitola , Francesc Pozo, Diego A. Tibaduiza, Maribel Anaya (2017). A Sensor Data Fusion System Based on k-Nearest Neighbor Pattern Classification for Structural Health Monitoring Applications. Sensors, 17.

[16] Asad Hussain, Sajjad Ahmed Ghauri, M. Farhan Sohail,Sheraz A. Khan, Ijaz Mansoor Qureshi. (2016). KNN BASED CLASSIFICATION OF DIGITAL MODULATED SIGNALS. IIUM Engineering Journal, 17: 71-82.

[17] American Diabetes Association (2014). Standards of Medical Care in Diabetes 2014 ,Diabetes Care 37: S14-S80.