# Classification of Chronic Kidney Disease Data via Three Algorithms

**Shaimaa Waleed Mahmood[1], Ghalia Twfeek Basheer[2], Zakariya Yahya Algamal[3]**

[1]Department of Statistics and Informatics, University of Mosul, Mosul, Iraq

[2]Department of Operations Research and Intelligent Techniques, University of Mosul, Mosul, Iraq

[3]Department of Statistics and Informatics, University of Mosul, Mosul, Iraq

**Abstract:**

Pattern recognition can be defined as the classification of data based on knowledge already gained or on statistical information extracted from patterns. The classification of objects is an important area for research and application in a variety of fields. In this paper, k-Nearest Neighbor, Fuzzy k-Nearest Neighbor and Modified k-Nearest Neighbor algorithms are used to classify of the chronic kidney disease (CKD) data with different choices of value k. The experiment results prove that the Fuzzy k-Nearest Neighbor and Modified k-Nearest Neighbor algorithms are very effective for classifying CKD data with high classification accuracy.

**Keyword:** Chronic kidney, k-Nearest Neighbor, Fuzzy k-NN, Modified k-NN, Classification.

***Corresponding Author:***

Zakariya Yahya Algamal

Department of Statistics and Informatics

University of Mosul, Mosul, Iraq

E-mail: zakariya.algamal@uomosul.edu.iq

# تصنيف بيانات مرض الكلى المزمن من خلال ثلاث خوارزميات

**م.م.شيماء وليد محمود[1]** **م.م.غالية توفيق بشير[2]** **أ.د.زكريا يحيى الجمال[3]**

[1],[3] جامعة الموصل / قسم الإحصاء والمعلوماتية

[2] جامعة الموصل / قسم بحوث العمليات والتقنيات الذكائية

**المستخلص**

يمكن تعريف تمييز الأنماط كتصنيف بيانات مبنية على معرفة مكتسبة سابقا أو على معلمة إحصائية مأخوذة من أنماط. تعد مواضيع التصنيف مجال مهم للبحث والتطبيق في حقول متنوعة. في هذا البحث ، تم استخدام خوارزميات الجار الأقرب، الجار الأقرب المضببة والجار الأقرب المعدلة لتصنيف بيانات مرض الكلى المزمن مع اختيار قيم k مختلفة. برهنت النتائج التجريبية أن خوارزميات الجار الأقرب المضببة والجار الأقرب المعدلة تكون فعالة جدا لتصنيف بيانات مرض الكلى المزمن مع دقة تصنيف عالي.

**الكلمات المفتاحية: الكلى المزمن، الجار الأقرب ، k-NN المضببة، k-NN المعدلة، التصنيف.**

ISSN (1681- 6870)

## 1. Introduction

Pattern recognition is about assigning labels to objects which are described by a set of values called features or attributes. The current research builds upon foundations laid out in (1960 and 1970) [7]. It has applications in almost every field of human endeavor including astronomy, psychology, geography and geology. More specifically, it is useful in psychological analysis, biometrics, bioinformatics and a set of other applications [8]. There are two major types of pattern recognition problems: supervised and unsupervised. In the supervised category which is also named classification or supervised learning, each object of the data comes with a reassigned class label [7].

Classification of objects is an important area of research and applications in a variety of fields, including artificial intelligence, pattern recognition, vision analysis, statistics, medicine and cognitive psychology [1]. It's one of the most useful techniques in data mining that classify data into structured groups or class. And consists of two phases; the first is learning process phase in a which a huge training data sets are analyzed and then it create the patterns and the rules. The second is testing or evaluation and recording the accuracy of the performance of the classification patterns [13]. The aim of classification technique is to build a model with best generalization capability.

In this paper we implemented the k-Nearest Neighbor, Fuzzy k-Nearest Neighbor and Modified k-Nearest Neighbor algorithms to classify of the CKD data with using different choices of value k, and using the classification accuracy to compare the results between the three algorithms.

## 2. k-Nearest Neighbor Algorithms

The K-Nearest Neighbor (KNN) is one of the most standout classification algorithms in data mining [9]. And is placed in the top ten data mining techniques. It is easy and simple but is a powerful method of classification. The main idea of this classification is that similar things belong to similar layers. It is first introduced by (Fix and Hodges, 1951). This algorithm involves the computation of the distance of the test observation with every observation in the training set. The distance which is computed in this paper is Gower distance. Then the k nearest observations are chosen and class which is assigned to majority out of these k observation is assigned to the test observation. K-NN makes no assumptions regarding the distribution of data, then it is nonparametric [2].

## 3. Fuzzy k-Nearest Neighbor Algorithm

The fuzzy sets were introduced by Zadeh in (1965). Since that, time researchers have found numerous methods to utilize this theory to generalize existing techniques and to develop new algorithms in decision analysis and pattern recognition [5]. The element in the fuzzy set have a membership function attached to them which is in the real unit interval [0,1] [8].

A Fuzzy k-Nearest Neighbor (FKNN) algorithm is one of the most successful techniques for applications due to its simplicity and also because of giving some information about the certainty of the classification decision and was introduced by (Keller *et al*., 1985). The Fuzzy k-NN algorithm classifies the testing data according to two objects, the fuzzy

information taken from the training data and the training data itself. It's explained as follows [6].

1- Find the k- nearest neighbor of a sample z.
2- Evaluate the membership function by using the following equation.

$$\mu_a(z) = \frac{\sum_{b=1}^{k} \mu_{ab} \left( \frac{1}{d_b^{2/m-1}} \right)}{\sum_{b=1}^{k} \left( \frac{1}{d_b^{2/m-1}} \right)} \qquad (1)$$

where $\mu_{ab}$ the membership of training vector $b$ to class $a$ , and calculated from the following relation

$$\mu_{ab} = \begin{cases} 0.51 + 0.49\frac{k_a}{K} & if\ c(z_b) = a \\ 0.49\frac{k_a}{K} & else \end{cases} \qquad (2)$$

such as $k_a$ is the number of points from the original training set among the k nearest neighbors of $z_b$ that belong to the same class as $z_b$ itself. But in some research, each pattern has a membership $\mu_{ab}$ equal 1 to the classes which figure in the k nearest neighbors and equal 0 to the other classes [8]. The $d_b$ is the distance between $z$ and $z_b$ . The parameter $m$ determines how heavily the distance is weighted when calculating each neighbor's contribution to the membership value. The $m$ takes usually the value of 2.
3- The class label, which has highest membership function.

## 4. Modified k-Nearest Neighbor Algorithm

The modified k-Nearest Neighbor (MKNN) algorithm or the distance weight K-NN is identical to the k-nearest neighbor algorithm, because it takes the K-NN into consideration. The difference only is that these k-NN are weighted according to their distance from the test point [8]. The fundamental idea is assigning the class label of the data according to k validated data training points [10, 11]. Each the neighbors is related with the weight $w$, which is defined as [8]

$$w_a = \begin{cases} \frac{d_k - d_a}{d_k - d_1} & if\ d_k \neq d_1 \\ 1 & if\ d_k = d_1 \end{cases} \qquad (3)$$

where $a = 1,2,\ldots,k$, the value of $w_a$ varies from one for the closest pattern to zero for the farthest pattern between the k closest patterns. After calculating the weights $w_a$, the MKNN algorithm assigns the test $P$ to that class for which the weights of the representative between the k-NN sums to the largest value. This modification would mean that outliers patterns will not affect the classification as much as the k-NN algorithm.

ISSN (1681- 6870)

## 5. Evaluation of the three Algorithms

One of the methods to evaluate the performance of classification algorithms is the confusion matrix, which contains information about actual and predicted classification done by the proposed classification system as in Table 1.

**Table 1: Confusion matrix**

| | | Prediction | |
|---|---|---|---|
| Actual | | Positive | Negative |
| | Positive | TP | FN |
| | Negative | FP | TN |

In this paper, we used the classification Accuracy, which is, indicates the ability of classifier algorithm to diagnose of classes of a dataset as the following [3, 12].

$$\text{Classification Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \qquad (4)$$

where TP: Indicates positive cases correctly classified as positive outputs.

TN: Indicates negative cases correctly classified as negative outputs.

FP: Indicates negative cases wrongly classified as positive outputs.

FN: Indicates positive cases wrongly classified as negative outputs.

## 6. Application

In this paper, a dataset of chronic renal disease or chronic kidney disease (CKD) gradually progresses and usually, after months or years, the kidney loses, its functionality from University of California Irvine (UCI) machine learning repository was used. The dataset of CKD contains 24 features. A total of 400 observations are included in 150 not CKD and 250 observations with CKD. Table 2 illustrates the features of CKD data.

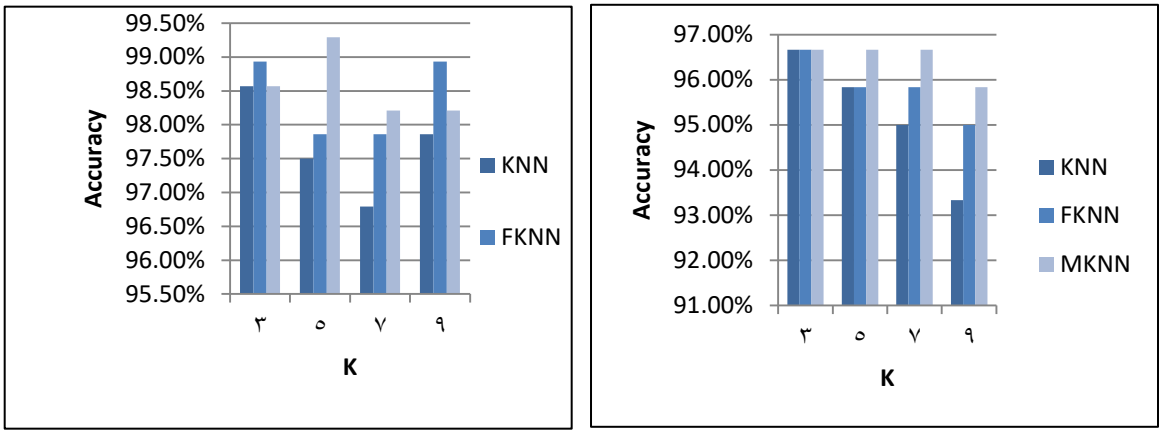
**Table 2: Features of CKD**

| Features | Features | Features |
|---|---|---|
| Age | Bacteria (ba) | White blood cell count (wbcc) |
| Blood pressure (bp) | Blood glucose random (bgr) | Red blood cell count (rbcc) |
| Specific gravity (sg) | Blood urea (bu) | Hypertension (htn) |
| Albumin (al) | Serum creatinine (sc) | Diabetes mellitus (dm) |
| Sugar (su) | Sodium (sod) | Coronary artery disease (cad) |
| Red blood cells (rbc) | Potassium (pot) | Appetite (appet) |
| Pus cell (pc) | Hemoglobin (hem) | Pedal edema (pe) |
| Pus cell clumps (pcc) | Paked cell volume (pcv) | Anemia (ane) |

To classify the CKD dataset, the data are converted to a standard format because of the different units of measurement. The data were randomly divided into a training set with 70% of the original size and a testing set with 30%. Using Gower distance, we found the distance matrix for two sets. Table 3 and Figure 1 explain the results of the classification performance of the three algorithms used with different k values.

**Table 3: comparison between three algorithms**

| K | Training set | | | Testing set | | |
|---|---|---|---|---|---|---|
| | KNN | FKNN | MKNN | KNN | FKNN | MKNN |
| 3 | 98.5714 | 98.9286 | 98.5714 | 96.6667 | 96.6667 | 96.6667 |
| 5 | 97.5 | 97.8571 | 99.2857 | 95.8333 | 95.8333 | 96.6667 |
| 7 | 96.7857 | 97.8571 | 98.2143 | 95 | 95.8333 | 96.6667 |
| 9 | 97.8571 | 98.9286 | 98.2143 | 93.3333 | 95 | 95.8333 |




a) Training set b) Testing set

**Figure 1: classification accuracy of the two sets**

As can be seen from Table 3 that the FKNN and MKNN algorithms have the best results in terms of the classification accuracy for both the training and testing datasets, where the FKNN has the largest classification accuracy of 98.93% when k=3,9 and the MKNN has the largest classification accuracy of 99.29%, 98.21% when k=5,7. In contrast, the KNN algorithm provided less classification accuracy in all k values. This indicted that FKNN and MKNN algorithms were better than KNN algorithm.

## 7. Conclusions

Presence of irrelevant information in the dataset makes the learning and prediction process difficult and inaccurate. In the present paper, the implementation of the three algorithms on CKD dataset demonstrated that the FKNN and MKNN algorithms in both the training and the test in datasets having the capacity to produce higher classification accuracy, therefore, the fuzzy KNN and modified KNN algorithms were better than KNN algorithm in the classification data.

ISSN (1681- 6870)

## References

[1] Batchelor, B. G., "Pattern Recognition," *Plenum Press, New York*, 1978.

[2] Ertuğrul, Ö. F., & Tağluk, M. E, "A novel version of k nearest neighbor: Dependent nearest neighbor," *Applied Soft Computing, 55*, 480-490, 2017.

[3] Fawcett, T, "An introduction to ROC analysis," *Pattern recognition letters, 27*(8), 861-874, 2006.

[4] Fix, E, "Discriminatory analysis: nonparametric discrimination, consistency properties," USAF school of Aviation Medicine, 1951.

[5] Kandel, A, "Fuzzy techniques in pattern recognition," John Wiley & Sons, 1982.

[6] Keller, J. M., Gray, M. R., & Givens, J. A, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, (4), 580-585, 1985.

[7] L. I. Kuncheva, "Combining Pattern Classifiers, Methods and Algorithms," New York: Wiley, 2005.

[8] Murty, M. N., & Devi, V. S, "Pattern recognition: An algorithmic approach," Springer Science & Business Media, 1988.

[9] Pakize, S. R., & Gandomi, A, "Comparative study of classification algorithms based on Map Reduce model," *International Journal of Innovative Research in Advanced Engineering (IJIRAE), 1*(7), 251-254, 2014.

[10] Parvin, H., Alizadeh, H., & Minaei-Bidgoli, B, "MKNN: Modified k-nearest neighbor," In *Proceedings of the world congress on engineering and computer science*, Vol.1, News wood Limited, 2008.

[11] Parvin, H., Alizadeh, H., & Minati, B, "A modification on k-nearest neighbor classifier," *Global Journal of Computer Science and Technology*, 2010.

[12] Santra, A. K., & Christy, C. J, "Genetic algorithm and confusion matrix for document clustering," *International Journal of Computer Science Issues (IJCSI), 9*(1), 322, 2012.

[13] Smitha, T., & Kumar, V. S, "Applications of big data in data mining," *International journal of emerging technology and advanced engineering, 7*(3), 2013.

[14] Zadeh, L. A, "Fuzzy sets." Information and control, 8(3): 338-353, 1965.