International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Performance Analysis of Chronic Kidney Disease Detection Based on *K*-Nearest Neighbors Data Mining

**Mohtady Ehab Barakat[1], Chung Gwo Chin*[2], Lee It Ee[3]**

**Submitted**: 22/04/2023 **Revised**: 16/06/2023 **Accepted**: 24/06/2023

**Abstract:** Kidney diseases are a leading cause of death in the United States. According to the Centers for Disease Control and Prevention (CDC), in 2021, approximately 37 million US adults, or 1 in 7, are estimated to have chronic kidney disease (CKD), and most are undiagnosed. Moreover, Medicare costs for people with CKD were $87.2 billion in 2019. Thus, data mining has been used in the healthcare industry to assist authorities in providing patients with health information as well as identifying patients earlier. In this paper, data mining is implemented for the classification of laboratory data from CKD patients. The K-Nearest Neighbors (KNN) algorithm is proposed to train the machine learning model to detect CKD based on blood test lab results such as sugar count, white blood cell count, red blood cell count, hemoglobin, albumin, etc. The model also includes general factors such as age and blood pressure. From the obtained results, other machine learning methods produce inferior accuracy, such as linear regression and decision tree. By training the model on a dataset containing 400 different anonymous patients using KNN, the accuracy reaches 99%. Based on the prediction, around 40% of the patients are fully healthy. This paper aims to detect whether the patient has CKD or not, depending on lab results and general information about the patient.

***Keywords:*** *Chronic kidney disease, data mining, K-Nearest Neighbors, linear regression, decision tree*

## 1. Introduction

The sixteenth cause of death worldwide is chronic kidney disease (CKD) [1]. In order to avoid adverse CKD-related outcomes, such as end-stage renal disease, cardiovascular disease, and death, major care clinicians must perform proper screening, diagnosis, and therapy. Around the world, 8% to 16% of people have CKD. The most typical causes of CKD in developed nations are diabetes and hypertension [2]. However, less than 5% of patients with early chronic kidney disease reported that they were aware of their disease. This is very crucial since CKD will lead to more severe sickness, as stated above, if it is not detected earlier.

Recent technical advancements have made it possible to detect various diseases via laboratory data. The value of this data is determined by how well it can be interpreted and analyzed, as well as how well it can be used to inform future policies and choices. One of the latest and best-known techniques used for this purpose is data mining [3]. Data mining is frequently described as the automated or convenient extraction of patterns indicating knowledge implicitly stored or collectible in huge databases, data warehouses, the Web, other big information repositories, or data streams. Usually, this data is not collected with the goal of analysis or processing but rather as a by-product of an operational system.

Data mining was commonly applied to find anomalies, trends, and correlations within huge data sets in order to forecast outcomes. In the business track, data mining can be utilized to improve customer interactions, decrease risks, save expenses, and generate income using a variety of approaches. On the other hand, data mining is useful in the healthcare industry for grouping patients with similar diseases or health problems in order to provide them with appropriate treatments and to provide patient hospitality information. [4, 5].

There are a lot of techniques applied in data mining using machine learning for classification. Three of the most commonly used machine learning techniques are linear regression, decision tree, and *K*-Nearest Neighbors (KNN). In [6], Dr. Parul Sinha projected work to predict CKD using machine learning algorithms. A comparison of support vector machines (SVM), Bayesian, and KNN was made. The result of the dataset showed that Bayesian provided the most accuracy, proving that Bayesian was the best algorithm. However, it is noticeable that the *K* number in the KNN was poorly chosen. There are also several classification techniques reported in [7] for CKD prediction. Meanwhile, Dr. El-Houssainy A. Rady [8] compared a few algorithms, which are probabilistic neural

[1] *Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia.*
[2] *Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia.*
*ORCID ID : 0000-0002-3262-3451*
[3] *Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia.*
*ORCID ID : 0000-0002-0922-8859*
** Corresponding Author Email: gcchung@mmu.edu.my*

networks (PNN), multilayer perceptrons (MLP), SVM, and radial basis functions (RBF). PNN produced the most accuracy. However, it was only able to achieve more than 99% accuracy for one type of CKD.

With the growing demand for treatment and diagnosis, predicting disorders of the heart, liver, and kidney at the earliest opportunity is of significant importance. In order to compare the disease with other varying parameters, classification methods were proposed in [9], including SVM and Random Forest (RF). This effectively advances illness analysis and disease prediction at various stages. In 2015, Dr. Parul Sinha [6] compared SVM and KNN, which significantly showed KNN as the best option. MATLAB was used as the primary tool to read the dataset and train the users. The dataset consisted of 400 patients. Nonetheless, the accuracy was only around 80%.

A more advanced method has been proposed, as R. Subhashini [10] introduced a methodology called the Optimal Fuzzy-K (OF) Nearest Neighbor technique in his article. The optimum performance of fuzzy was achieved by altering the membership functions using the Bat optimization algorithm. Then, the OF was utilized to measure the similarity in the KNN for the classification of diseases. The performance of the proposed technique was analyzed by comparison with conventional methods. Accuracy was considered the primary metric for evaluating performance.

Recently, V. Manoranjithem [11] has implemented KNN to detect CKD and compared distance measurement methods such as Sjaccard, S3wjaccard, Sczekanowski, Srogertanimoto, Ssokalmichener, Srussellrao, and DEuclid (Euclidean distance. The result proved that Euclidean distance has the lowest error rate when $K = 190$. Meanwhile, the research published by C. Priyadharshini [12] has the aim of detecting CKD but the dataset has a terrific number of missing characteristics. It was later solved in the research by using KNN and utilizing its attribution to fill in the missing qualities. In order to address the missing information in fragmented cases, a few complete examples with the most comparative estimations were chosen.

Nevertheless, this paper intends to analyze the performance of CKD detection based on data mining. Several machine learning algorithms, such as linear regression, decision tree, and KNN, are implemented to classify the data for CKD detection. Laboratory data such as sugar count, white blood cell count, red blood cell count, hemoglobin, albumin, etc. will be used for categorizing the patients with and without CKD diseases. General factors such as age and blood pressure are also considered during the training process. Performance analysis is then performed to compare the accuracy of each of the machine learning algorithms. Further analysis on choosing the suitable $K$ number for the KNN algorithm will also be conducted to increase the accuracy of the prediction. Henceforth, data mining techniques are used to detect CKD, as shown in the later section.

## 2. Machine Learning Algorithms

Machine learning algorithms are widely used in data classification. Classification can start with two groups that have one difference, such as filtering e-mails as spam or not spam; henceforth, they can be plotted on a 2D plane, a more complicated 3D plane or something larger. There are various machine learning algorithms [13]. Each one of them is most suitable for a certain case. For example, two groups of circles are green or red. Since the classification is based on color, a simple linear regression model will usually be the best. But in more complicated situations, other algorithms can produce higher accuracy in prediction. In this paper, we will discuss the most popular three algorithms used for classification in data mining that have a lot of variables to detect CKD diseases.

### 2.1. Linear Regression

Simple linear regression is useful for figuring out the relationship between two continuous variables [14], as seen in Fig. 1. A prediction is made by an independent variable, whereas a response is made by a dependent variable. Instead of deterministic correlation, it attempts statistical correlation. What is deemed deterministic can be accurately expressed by a relationship between two variables. For example, it is likely to forecast a pound using a weight in grammes accurately. Statistical correlation cannot be used to accurately determine the relationship between two variables. such as the relationship between weight and height.

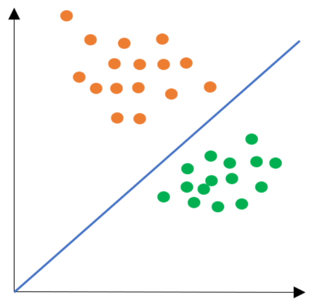

**Fig. 1.** Example of a linear regression model recognizing $N$ samples from two different classes.

The primary goal of regression is to find the line that best fits the data. The best-fit line has the lowest potential total prediction error across all data points, which can be described as the distance between the regression line and the point. In order to predict with 2-D linear regression, a normal linear function is used [14].

$$Y(\text{predict}) = b + mX$$

(1)

where $X$ is the predictor (independent) variable, $Y$ is the response (dependent) variable, $m$ is the estimated slope, and $b$ is the estimated intercept.

The error can be minimized as much as possible using the squared mean error [14].

$$\sum_{i=1}^{n}(\text{actual output - predicted output})^{2}$$

(2)

The mean error is calculated to update the $X$ and $Y$ values, as shown in Fig. 1.

### 2.2. Decision Tree

A decision tree is a map with a node representing every possible decision. The probability of each decision is then calculated, and the best decision for the next step is chosen. It is one of the few algorithms that are considered supervised learning but work for both classification and regression problems [15].

A decision tree starts with a root node, and then a process called splitting happens where some nodes are split into two or more sub-nodes, as shown in Fig. 2. When a sub-node is divided, it's called a decision node. If not, it is called a leaf or terminal node. Whereas a branch or sub-tree is a subsection of the tree.

### 2.3. *K*-Nearest Neighbor

KNN is a classification algorithm that organizes the data using the nearest neighbor/point [16]. KNN is also known as the lazy learning algorithm. When you submit the training data, KNN saves the data during training, instead of performing any training or calculations. It will only build a model once a query is executed on the dataset. Thus, KNN has a fast computation time and it is ideal for data mining.

KNN calculates outliers, and it's not efficient for memory as it calculates the distance of each point individually, as shown in Fig. 3. So, choosing the dataset for training significantly impacts accuracy. In order to calculate the distance, KNN uses the *Euclidean* distance, $d$ [17]:

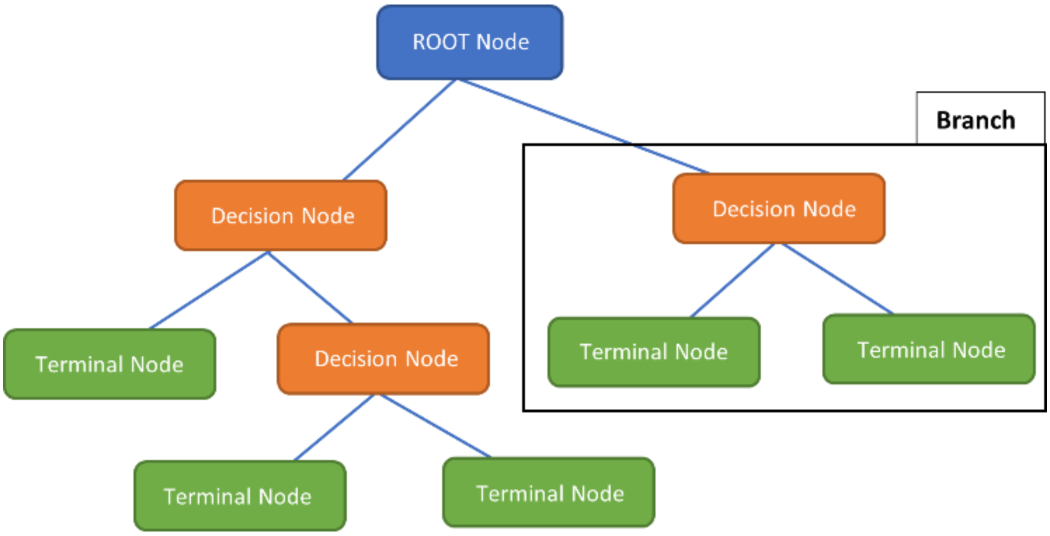

**Fig. 2.** Example of a decision tree architecture.

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

(3)

In KNN, $K$ is the number of neighbours to calculate the distance from the node.

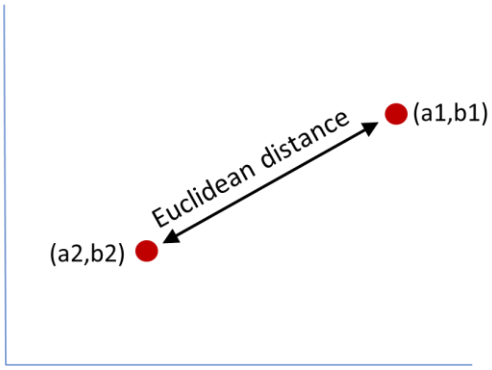

**Fig. 3.** *Euclidean* distance between two points for KNN.

## 3. Research Methodology

Fig. 4 presents the flow chart of the proposed methodology in this paper. The following illustrates the detail of each steps for analyzing the performance of the CKD detection using several machine learning algorithms.

a. Dataset:

In the first stage, pre-processing has to be done on the dataset to improve its readability during the model training later. For example, all the incorrect inputs are replaced with the right values. Then, to deal with numeric numbers, the Pandas library is used to convert all needed columns to be numeric, such as blood sugar, and patients that have CKD will have a value of 0, and patients that don't have CKD will have a value of 1. Besides that, missing values will be detected and filled with a question mark ("?") so that the data can still be used for training. A simple function is used for this purpose: df = data.replace(np.nan, "?"). Scikit Learn is then used to make and convert a one-hot encoding of categorical variables into a comma-separated values (CSV) Excel file. Data can be visualized using the Matplotlib library, as shown in Fig. 5.

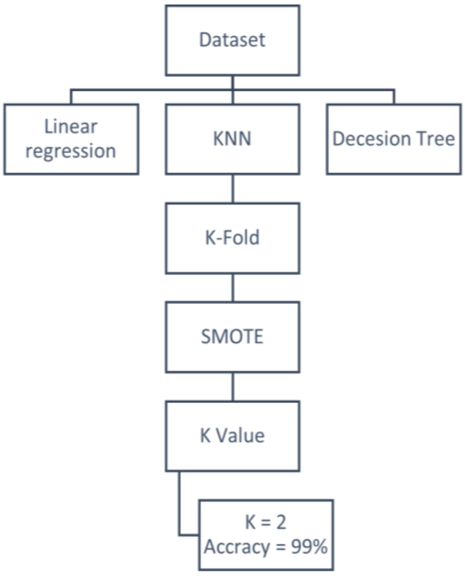

**Fig. 4.** Flow chart of the proposed methodology.

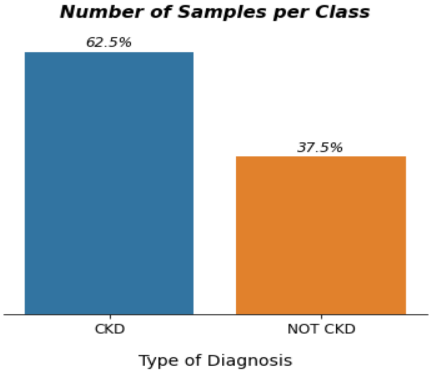

**Fig. 5.** Distribution of the dataset with or without CKD.

Fig. 6 and Fig. 7 presents the statistical distribution of the dataset for all the laboratory data such as age, blood pressure, blood glucose, blood urea, hemoglobin, packed cell volume, specific gravity, albumin, serum creatinine, sodium, white blood cell count, sugar, and potassium. This is to show the overall distributed values of the data used in this paper to train the machine learning model for CKD prediction.

For instance, the age of the patients is concentrated around 40 to 70 years old since these are the common ages of patients that are diagnosed with CKD. Other more common factors, such as the blood pressure, glucose level, and substances such as sodium, sugar, and potassium in the patients, are also considered for the detection of CKD cases. The graphs show that the values are uniformly distributed in their normal range.

On the other hand, more specific dataset that related to CKD disease are used in training the proposed machine learning model [1, 2], as shown in Fig. 6 and Fig. 7 respectively. The amount of solutes in the urine is determined by the specific gravity of the urine (SG). It determines how much urine is concentrated relative to water by calculating the ratio of urine density to water density. It typically falls between 1.005 and 1.030. Human blood's level of urea nitrogen is known as blood urea. It is a waste product produced by the breakdown of protein by your liver. Urea nitrogen levels in blood or serum should be between 5 and 20 mg/dl, or 1.8 and 7.1 mmol of urea per litre. The packed cell volume (PCV) is a measurement of the proportion of blood that is made up of cells. Typically, it ranges from 40.7 to 50.3% for men and 36. to 44.3% for women.

Besides that, the amount of creatinine in your blood is measured by serum creatinine. For men, a normal result ranges from 0.7 to 1.3 mg/dL (61.9 to 114.9 mol/L) and for women, it ranges from 0.6 to 1.1 mg/dL (53 to 97.2 mol/L). A white blood count measures the number of white cells in your blood. The immune system is made up of white blood cells. 4,000 to 11,000 white blood cells per microliter are considered typical. A red blood cell (RBC) count is a blood test that tells you how many red blood cells you have. Whereas haemoglobin is a substance in red blood cells.

Last but not least, the Scikit library is again used to replace all values with scaled values falling between 0 and 1. Using the method "model_selection," the data is split into training and testing sets.

b. Machine learning algorithms:

Three machine learning algorithms, such as linear regression, decision tree, and KNN, are applied in this paper for performance comparison. The Scikit library is used to train all these models. The accuracy of detecting

CKD is then calculated and stored at the end of the process.

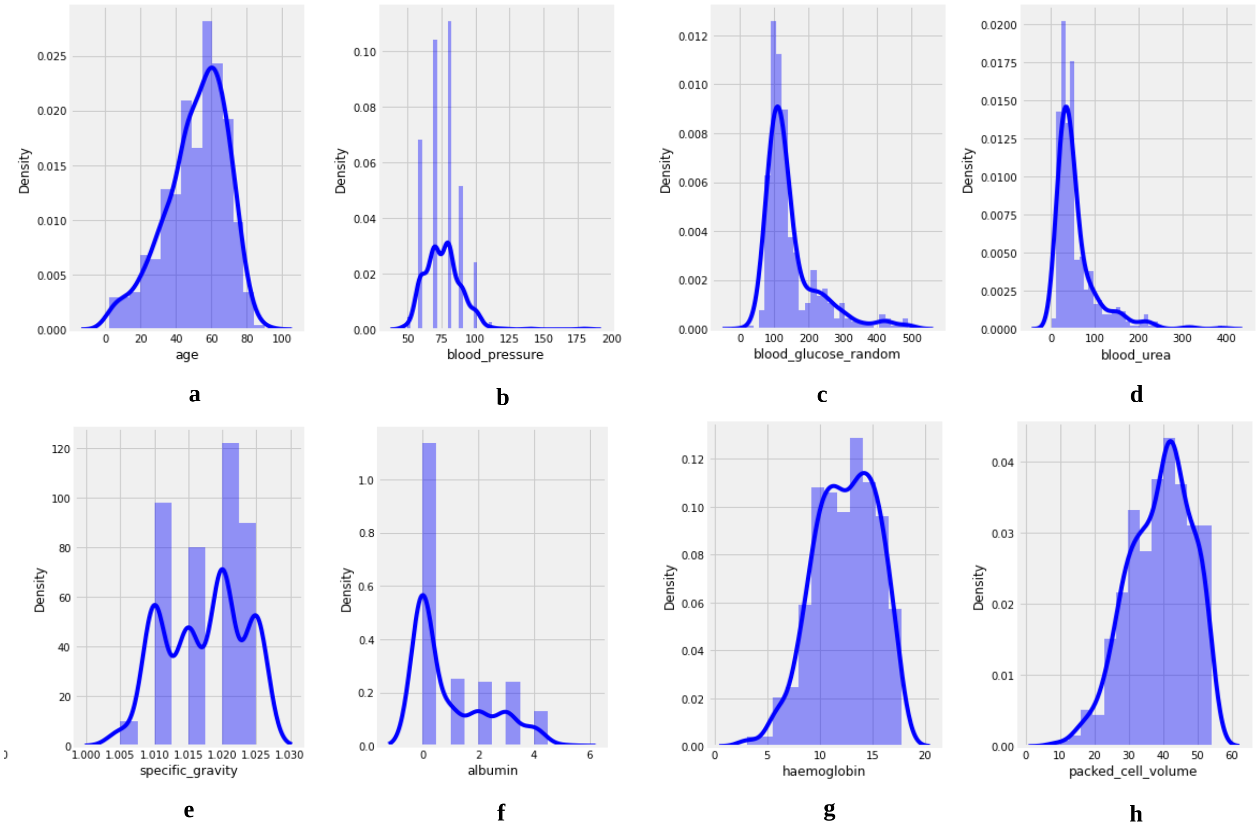

**Fig. 6.** Distribution of dataset for (a) age, (b) blood pressure, (c) blood glucose, (d) blood urea, (e) specific gravity, (f) albumin, (g) hemoglobin, and (h) packed cell volume.

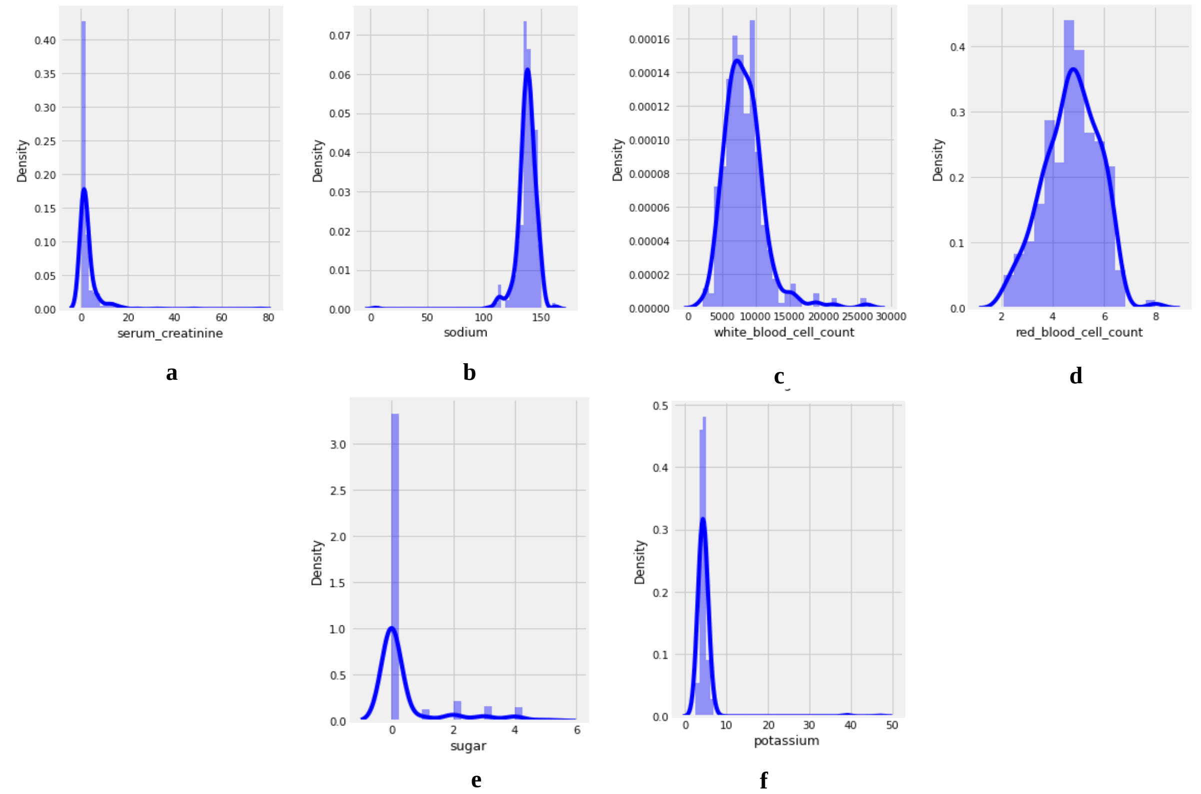

**Fig. 7.** Distribution of dataset for (a) serum creatinine, (b) sodium, (c) white blood cells count, (d) red blood cell count, (e) sugar, and (f) potassium.

c. *K*-fold:

Typically, a linear model's *K*-fold cross-validation technique divides the data set into *K* pieces evenly and at random (if possible). A candidate model is constructed based on a training set, which is the *K* - 1 part of the dataset. A test set containing the data from the hold-out part is then used to evaluate the candidate model's prediction accuracy [18]. *K*-fold is used to increase the accuracy of the model.

d. SMOTE:

When employing class-imbalanced data for classification, the majority class is favored. The bias is significantly stronger for high-dimensional data, when the number of variables greatly exceeds the number of samples. Oversampling, which generates class-balanced data, can mitigate the issue. A highly popular oversampling technique called Synthetic Minority Oversampling Technique (SMOTE) is introduced to enhance random oversampling. Thus, SMOTE is applied to the whole dataset [19].

e. *K*-Value:

The KNN algorithm's *K* parameter, which determines how many neighbors will be chosen. The choice of *K* has a considerable impact on the diagnostic performance of the KNN algorithm. When a large *K* lessens the effect of variance carried on by random error, it also runs the danger of obscuring minute but significant patterns. Finding a balance between overfitting and underfitting is crucial for selecting an optimal *K* value. SMOTE is a statistical method for increasing the number of cases in your dataset in a balanced way. The component creates new instances using minority cases that you supply as input from previously existing instances. [20].

## 4. Results and Discussion

In this paper, there are 400 samples of anonymous patients collected to train the prediction model. For choosing the appropriate *K* value for the KNN algorithm, we try to first simulate the *K* range from 1 to 200 and calculate the prediction accuracy individually. The accuracy starts dropping when the *K* increases, as depicted in Fig. 8. Thus, we repeat the simulation for the *K* range from 1 to 20, as shown in Fig. 9. It is noticeable that *K* = 2 is the best value, as it produced an accuracy of 99%, which is the highest among all the other *K* values.

Fig. 10 presents the performance comparison for the linear regression, decision tree, and KNN algorithms. Matplotlib is used to draw the figure. The value of *K* is by default set to 5, but most of the previous research papers used *K* = 1. However, as shown in Fig. 8, KNN with a *K* value of 2 achieves the highest prediction accuracy. This shows the importance of choosing the most accurate value for *K*, as it is noticeable that the linear regression has a better performance as compared to KNN when the *K* value is poorly determined (*K* = 5). On the other hand, linear regression and decision tree produce maximum accuracy of 97% and 96%, respectively, which makes KNN the best option as it has a performance accuracy of 98% at *K* = 1 and 99% at *K* = 2. From the obtained results, around 40% of the patients are predicted to be healthy, whereas 60% of them are suspected to have CKD diseases. This statistic is almost in line with the information shown in Fig. 5. Hence, the proposed KNN model with *K* = 2 is able to detect CKD with very high accuracy.

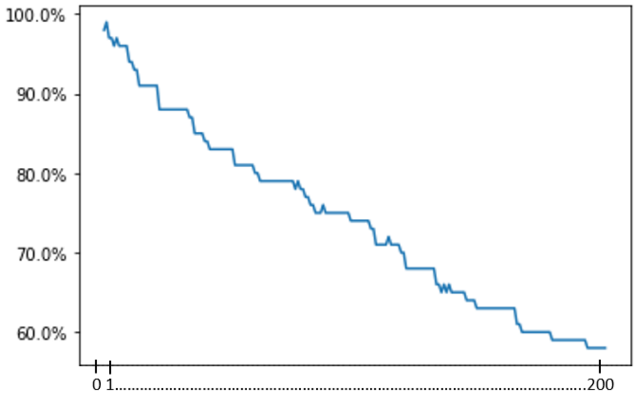

**Fig. 8.** Prediction accuracy for *K* range from 1 to 200.

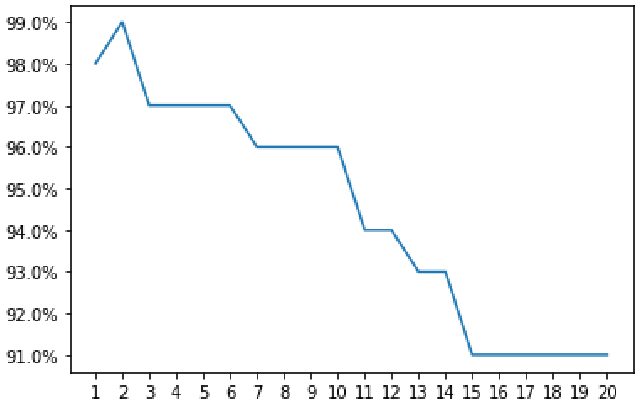

**Fig. 9.** Prediction accuracy for *K* range from 1 to 20.

## 5. Conclusion

In conclusion, a performance analysis has been conducted in this paper to detect CKD based on laboratory data using machine learning algorithms such as linear regression, decision tree, and KNN. Results show that choosing an appropriate *K* value for the KNN improves prediction accuracy. Moreover, KNN has achieved better performance accuracy over linear regression and decision tree.

Despite proving that KNN is the best machine learning algorithm to be used for detection, KNN is capped at an accuracy of 99%. In order to increase the prediction accuracy, additional data will need to be added to the dataset, such as Computed Tomography (CT) scan images, as reported in [19]. Providing the model with such data along with more blood test results will improve its accuracy significantly. As the model runs fast and needs low processing power, it can be easily implemented as software to help doctors or patients monitor their health frequently.

## Acknowledgements

This research project is fully sponsored by Internal Research Fund (MMUI/220061), Multimedia University.

## Author contributions

**Mohtady Ehab Barakat:**, Methodology, Software, Field study, Data curation **Gwo Chin Chung:** Conceptualization, Writing-Original draft preparation, Investigation, Validation **It Ee Lee:** Visualization, Validation, Writing-Reviewing and Editing.

## Conflicts of interest

The authors declare no conflicts of interest.

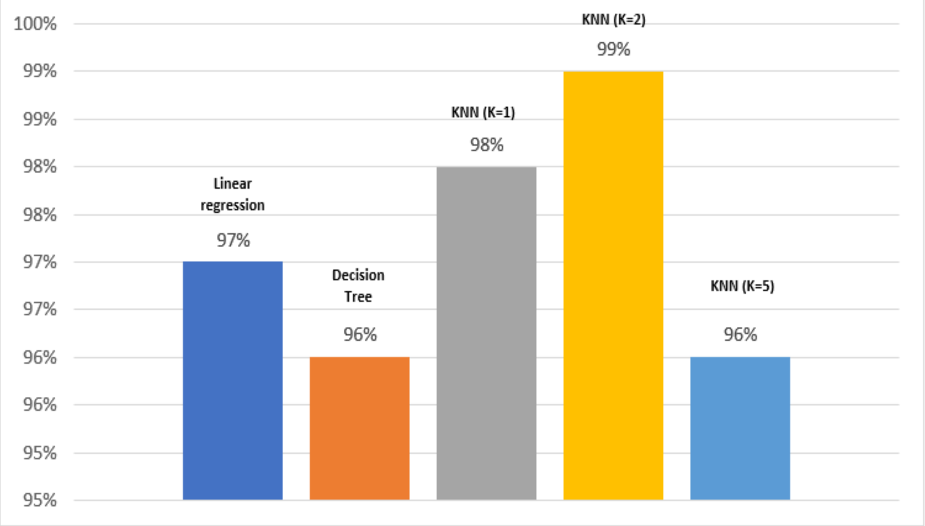

**Fig. 10.** Accuracy comparison for linear regression, decision tree, and KNN.

## References

[1] T. K. Chen, D. H. Knicely, D. H. and M. E. Grams, “Chronic kidney disease diagnosis and management,” *The Journal of the American Medical Association (JAMA)*, vol. 322, no. 13, pp. 1294, 2019. https://doi.org/ 10.1001/jama.2019.14745

[2] C. P. Kovesdy, “Epidemiology of chronic kidney disease: an update 2022,” *Kidney International Supplements*, vol. 12, no. 1, pp. 7-11, 2022. https://doi.org/10.1016/j.kisu.2021.11.003

[3] T. Calders and B. Custers, “What is data mining and how does it work?,” *Studies in Applied Philosophy, Epistemology and Rational Ethics*, pp. 27–42, 2013. https://doi.org/10.1007/978-3-642-30487-3_2

[4] M. L. Kolling, L. B. Furstenau, M. K. Sott, B. Rabaioli, O\P. H. Ulmi, N. L. Bragazzi and L. P. Tedesco, “Data mining in healthcare: Applying strategic intelligence techniques to depict 25 years of research development,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, pp. 3099, 2021. https://doi.org/10.3390/ijerph18063099

[5] A. Garg and V. Mago, “Role of machine learning in medical research: A survey,” *Computer Science Review*, vol. 40, pp. 100370, 2021. https://doi.org/10.1016/j.cosrev.2021.100370

[6] P. Sinha and P. Sinha, “Comparative study of chronic kidney disease prediction using KNN and SVM,” *International Journal of Engineering Research and Technology (IJERT)*, vol. 4, no. 12, 2015. https://doi.org/ 10.17577/IJERTV4IS120622

[7] P. Tikariha and P. Richhariya, “Comparative study of chronic kidney disease prediction using different classification techniques,” presented at the *Proceedings of International Conference on Recent Advancement on Computer and Communication (ICRAC)*, pp. 195-203), Springer Singapore, 2018. https://doi.org/10.1007/978-981-10-8198-9_20

[8] E.-H. A. Rady and A. S. Anwar, “Prediction of kidney disease stages using data mining algorithms,” *Informatics in Medicine Unlocked*, vol. 15, pp. 100178, 2019. https://doi.org/ 10.1016/j.imu.2019.100178

[9] A. AhmedK, S. Aljahdali and S. Naimatullah

Hussain, “Comparative prediction performance with support vector machine and random forest classification techniques,” *International Journal of Computer Applications*, vol. 69, no. 11, pp. 12–16, 2013. https://doi.org/10.5120/11885-7922

[10] R. Subhashini, M. Jeyakumar and N. Islam, “OF-KNN technique: An approach for chronic kidney disease prediction,” *Computer Science*, vol. 116, no. 24, 2017.

[11] V. Manoranjithem and M. Venkatesulu, “KNN classification in chronic kidney disease dataset, *International Journal of Mathematics and* Computer Science (IJMCS), vol. 15, no. 4, pp. 1337–1343, 2020.

[12] C. Priyadharshini, K. Sanjeev, M. Vignesh, N. Saravanan and M. Somu, “KNN based detection and diagnosis of chronic kidney disease,” Annals of the Romanian Society for Cell Biology, vol. 25, no. 4, pp. 2870, 2021.

[13] S. Suthaharan, Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, Springer, 2015.

[14] A. Schneider, G. Hommel and M. Blettner, “Linear regression analysis,” Deutsches Ärzteblatt International, vol. 107, no. 44, pp. 776-782, 2010. https://doi.org/10.3238/arztebl.2010.0776

[15] I. Jenhani, N. B. Amor and Z. Elouedi, “Decision trees as possibilistic classifiers,” International Journal of Approximate Reasoning, vol. 48, no. 3, pp. 784–807, 2008. https://doi.org/10.1016/j.ijar.2007.12.002

[16] Z. Zhang, “Introduction to machine learning: K-Nearest Neighbors,” Annals of Translational Medicine, vol. 4, no. 11, pp. 218–218, 2016. https://doi.org/10.21037/atm.2016.03.37

[17] H. Rajaguru and S. K. Prabhakar, “KNN Classifier and K-Means Clustering for Robust Classification of Epilepsy from EEG Signals. A Detailed Analysis,” Anchor Academic Publishing, 2017.

[18] Y. Jung and J. Hu, “A K-fold averaging cross-validation procedure,” Journal of Nonparametric Statistics, vol. 27, no. 2, pp. 167–179, 2015. https://doi.org/10.1080/10485252.2015.1010532

[19] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” BMC Bioinformatics 14, no. 106, 2013. https://doi.org/ 10.1186/1471-2105-14-106

[20] G. S. K. G. Prasad, A. A. Chowdari, K. P. Jona and R. Senapati, “Detection of CKD from CT Scan images using KNN algorithm and using edge detection,” presented at the 2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET), pp. 1-4, 2022. https://doi.org/10.1109/icefeet51821.2022.9848173

[21] M, T. ., & K, P. . (2023). An Enhanced Expectation Maximization Text Document Clustering Algorithm for E-Content Analysis. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 12–19. https://doi.org/10.17762/ijritcc.v11i1.5982

[22] Dr. Bhushan Bandre. (2013). Design and Analysis of Low Power Energy Efficient Braun Multiplier. International Journal of New Practices in Management and Engineering, 2(01), 08 - 16. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/12