Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 38-45

DOI: https://doi.org/10.25195/ijci.v50i2.507
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

Chronic Kidney Disease (CKD)Diagnosis using Machine Learning Methodology Classifications

Ahmed Sami Jaddoa
Business Informatics College
University of Information Technology and Communications
Baghdad, Iraq
ahmed.sami@uoitc.edu.iq

ARTICLE INFO
Article History
Received: 01/07/2024
Accepted: 05/08/2024
Published:04/10/2024

ABSTRACT

Early diagnosis of kidney as well as pre-kidney disease is crucial for patients because it allows them to take control of their condition and could potentially avoid or delay more significant consequences that could lower their quality of life. The chance of developing a major disease might be decreased with its assistance. Almost every part of the body could be impacted by chronic kidney disease (CKD). Fluid retention in the lungs, high blood pressure, and swelling of the legs and arms are all potential side effects. This study proposes a model that makes use of machine learning (ML) algorithms for diagnosing kidney disease. The preprocessing dataset, which contains missing values and is preprocessed with the use of mean, delete, and median approaches before data scaling, is the foundation of the suggested model. To achieve the highest classification accuracy, the preprocessing stage receives the results of missing values. Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are the two classification algorithms used to classify whether kidney disease is present or absent. Classify the dataset into testing and training (40% and 60%, respectively). The accuracy, F1-score, recall, and precision have been utilized for evaluating the suggested model. The kidney disease data-set has been used to test the outcomes of the suggested model. Without preprocessing any missing values in the dataset, the algorithms SVM and K-NN obtained maximum accuracy (95% and %89). Through deleting missing values from the dataset, the algorithms SVM and K-NN obtained maximum accuracy (%96 and %93). K-NN and SVM algorithms reached a maximum accuracy of %98 when using a mean technique; when using a median method, such algorithms attained an accuracy ranging from %95 to %98.

Keywords: CKD, KNN, SVM, Mean, Median, Preprocessing, Machine Learning;

1. INTRODUCTION

CKD can be defined as a disability of kidneys to perform their normal function of blood filtering as well as other functions. The disease has been considered as a severe form of kidney failure in which the kidneys are incapable of filtering blood, resulting in a significant buildup of fluid in the body. This causes the body's levels of potassium and calcium salts to rise dangerously. The presence of elevated concentrations of such salts causes the body to experience a number of additional ailments [1]. Permanent dialysis or kidney transplants are frequently required for CKD. High risk of CKD is associated with a family history of kidney disease. Nearly one in three patients who are diagnosed with diabetes also have CKD, according to the available literature. Evidence from the literature suggests that CKD treatment and diagnosis early on could lead to the enhancement of a patient's quality of life. Prediction algorithms in ML could be cleverly applied to predict the development of CKD and offer an early treatment strategy [1]. A computer software that performs a computation and inferring of task-related information and determines properties of matching pattern is referred to as ML. This technology may be one of the potential tools for the diagnosis of CKD because it could produce accurate and affordable disease diagnoses. With information technology advancement, it has taken on a new form as a medical instrument, and the rapid growth of electronic health records expanded its potential applications. ML has previously been applied in the medical field to diagnose a variety of diseases, examine pertinent disease components, and identify the human body state [2][3]. ML is a significant area of study these days. Various

DOI: https://doi.org/10.25195/ijci.v50i2.507
Print ISSN: 2313-190X, Online ISSN: 2520-4912

statistical as well as ML algorithms are used in diverse contexts. We may utilize ML in a variety of sectors, including marketing, socioeconomic behavior research, medical and health issues, and weather forecasting. Various diseases could be identified or predicted through machines in the medical industry through the use of ML algorithms [4]. Diagnosis of Chronic Kidney Disease (CKD) is a risky task since it is directly dependent on people's lives. Accuracy is a major factor because it can be disastrous if not diagnosed accurately. Dataset includes many missing values that will affect the accuracy of the diagnosis. Dataset contains variations in values that affect the operation of the algorithms. The contributions of this paper are illustrated in the following points:

1. Build a model for the diagnosis of chronic kidney disease (CKD) using machine learning algorithms to evaluate its performance.
2. Comparison between Algorithms from the viewpoint of optimization by using a set of standard test metrics such as accuracy.
3. Preprocessing missing values in the dataset to increase the accuracy of the diagnosis model.

2. RELATED WORKS

Xiao et al., 2019 [5], the aim of study to prediction of chronic kidney disease progression using Logistic regression, elastic Net, SVM, random forest (RF), ridge regression, neural network (NN), XG Boost, and KNN were used in the suggested prediction of CKD. They came to the conclusion that, with an accuracy of 0.87, logistic regression performed better. **Kilvia L et al., 2020 [6],** in this study, SVM were utilized along with Decision Tree (DT), RF, and SVM with RBF, sigmoid, and polynomial functions. They made use of the MIMIC database in their investigation. With prediction accuracy of 0.80 and 0.87, they came to the conclusion that RF and DT produced the best outcomes. **Hamida Ilyas et.,2021[7],** the aim of this study is to predict the various stages of CKD using machine learning classification algorithms on the dataset obtained from the medical records of affected people. Specifically, we have used the Random Forest and J48 algorithms to obtain a sustainable and practicable model to detect various stages of CKD with comprehensive medical accuracy. Where Random Forest give accuracy 78.25, While J48 give an accuracy 0.85. **Saurabh Pal 2023, [8],** the objective of this research is creating a model that uses SVM, ANN, and RF to diagnose CKD early on. With an accuracy of 0.92, RF out performed SVM, which had an accuracy of 0.88, and ANN, which had an accuracy of 0.80.

3. Methodology

The outcome of this study is the detection of the absence or presence of kidney disease. A kidney disease model is proposed depending on CKD extracted from the publicly accessible UCI ML Repository. The two main stages of the suggested diagnosis cooperate to achieve the diagnosis' objectives. The classification and preprocessing stages are the main steps of the diagnosis. When learning a dependence from data, to avoid overfitting, it is important to divide the data into the training set and the testing set. We first train our model on the training set, and then we use the data from the testing set to gauge the accuracy of the resulting model. Empirical studies show that the best results are obtained if we use 40% of the data for testing, and the remaining 60% of the data for training. A general diagram of the proposed model has been depicted in Figure1.

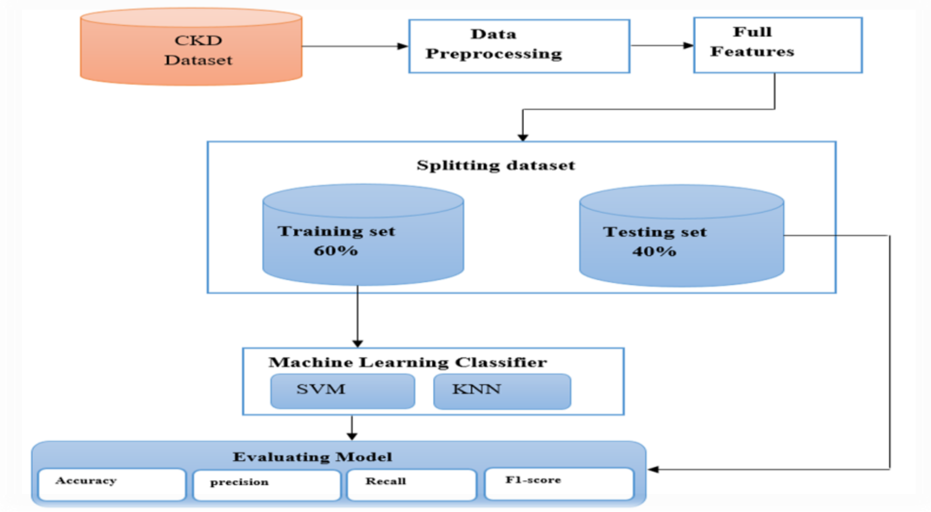

Fig. 1. Diagnosis Model of CKD

3.1 Chronic Kidney Disease Dataset

UCI ML Repository, which is accessible to the public, provided the dataset that was employed in this study to diagnose CKD [9]. There were 400 records in the dataset at first, comprising 24 features and a class feature. Of the twenty-four features, 11 are numeric and 14 are nominal. The class feature establishes whether the instance is CKD or not. This dataset belongs to a hospital in Karaikudi, Tamil Nadu, India and is intended for classification problems. Therefore, this dataset helps in training supervised algorithms. The dataset contains a large proportion of missing values, which people can decide to fill with either mean or median values, or leave them blank to allow the algorithm to learn without noisy data, or delete the missing values. This makes it a big challenge to process these missing values and compare all the results before and after processing. Table 1 displays the dataset's details.

TABLE I. CKD Dataset

No	Feature	Description	Value/Range
1	Age	Age	Years
2	BP	A Blood Pressure	Mm/Hg
3	Al	Albumin	0,1,2,3,4,5
4	Sg	Specific Gravity	1.005,1.01,1.015,1.02,1.025
5	PC	Pus Cell	Normal, Abnormal
6	Su	Sugar	0,1,2,3,4,5
7	RBC	Red Blood Cells	Normal, Abnormal
8	Ba	Bacteria	Present, Not present
9	PCC	Pus Cell Clumps	Present, Not present
10	SC	Serum Creatinine	Mgs/Dl
11	BGR	Blood Glucose Random	Mgs/Dl
12	Bu	Blood Urea	Mgs/Dl
13	Hemo	Hemoglobin	Gms
14	Sod	Sodium	Meq/L
15	Pot	Potassium	Meq/L
16	RBCC	Red Blood Cell Count	Millions/Cmm
17	PCV	Packed Cell Volume	
18	WBCC	White Blood Cell Count	Cells/Cumm
19	DM	Diabetes Mellitus	Yes, No
20	Htn	Hypertension	Yes, No
21	PE	Pedal Edema	Yes, No
22	Cad	Coronary Artery Disease	Yes, No
23	Applet	Appetite	Good, Poor
24	Ane	Anemia	Yes, No
25	Class	Class	Ckd, NotCkd

3.2 Data Preprocessing

In order to transform a raw dataset into a cleaned dataset set that could be useful for applying various ML algorithms, data pre-processing is crucial in ML. Because of the volume or the method, the data was converted from manual to computerized, the majority of acquired data is very likely to be missing, insufficient, and unreliable. Several techniques will be employed in this section: first, clean up the data for removing inconsistent and null values. Second, data reduction eliminates features that are redundant or have no bearing on prediction outcomes. Third, to help increase efficiency and accuracy, data transformation or normalization through the scaling and classifying high-range numeric properties [10].

DOI: https://doi.org/10.25195/ijci.v50i2.507
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Data Cleaning: A highly important and crucial task in pre-processing is data cleaning, which involves routines for either filling in or removing null values. Additionally, more uncleaned or unreliable data might result in confusion and unreliable results. The utilized dataset contained some missing values. This article describes three methods to impute missing values: either eliminate the records containing the missing values, replace them with random values, or, and this is the most advised approach, replace missing values with the average of other accessible values. Thus, this method is selected to be utilized in order to eliminate missing values from the dataset to obtain more accurate results [11] [12].

3.3 Classification Algorithms

In order to create classification templates, understandable and new patterns were defined using ML algorithms. In order to perform classification and regression in clinical as well as medical diagnostics, unsupervised and supervised learning methods both call for the creation of models depending upon prior analyses. Two stages are used by ML algorithms to create predictive/classification models: the training phase creates a model from a group of the training data with expected outputs, and validation stage assesses quality of trained models by using a validation dataset that does not contain the expected output.

3.3.1. K-Nearest Neighbors Classifier (KNN)

KNN algorithm classifies a new test point to the most comparable class amongst available classes based on how comparable the stored and new data points (training points) are. The lazy learning algorithm, or KNN, is nonparametric and keeps training data-set instead of learning from it. The new dataset (test data) is classified according to the value of k, and the distance from the new point to the training points that are stored is calculated using Euclidean distance. The maximum number of neighbors is used to classify a new point. The nearest neighbor in features vector was located by applying Euclidean distance function (Di) as indicated in Eq. 1 [13].

$$\text{E(d,p)} = \sqrt{\sum_{i=1}^{n}(d_i - p_i)^2} \qquad (1)$$

Where E is the Euclidean distance, d is the data point from the dataset, p is the new data point to be predicated and n is the number of dimensions.

3.3.2. Support Vector Machine (SVM)

SVM serves as learning mechanism with a feature space of significant dimensions and a hypothesis space based upon the linear functions. Learning algorithms based on the ideas of optimization theory help its training. Kernel function and parameters employed throughout training phase have a significant impact on accuracy level that SVM model achieves. SVM could be classified into two categories, which are: Linear and Nonlinear SVMs, depending on its features. Through dividing data into classes with a soft margin, a hyperplane is used in linear SVMs. Fig (2) is an illustration of the linear SVM. However, Non-linear SVM maps data into higher-dimensional space in order to use the kernel approach. SVM is essentially the process of locating the best separator on the hyperplane, as Fig (3) illustrates. Finding value of f(x) on hyperplane margin yields the optimal separating hyperplane [14].

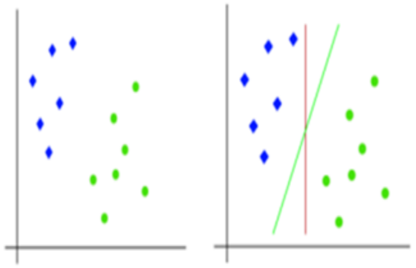

Fig 2. Linear SVM

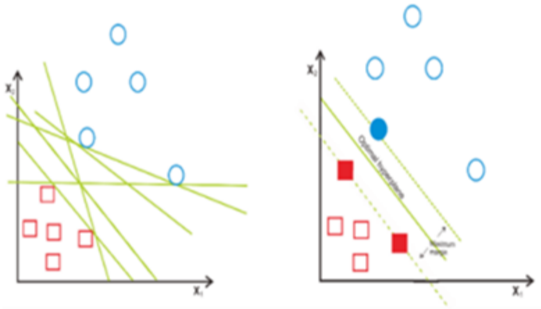

Fig 3. Find the best hyperplane

3.4. Evaluating the Model

Four common metrics have been utilized in order to assess the model: accuracy, F1-score, recall, and precision [15].

- Accuracy: The capacity of classification algorithm for accurately predicting classes in the data-set is implied by accuracy. It is a measurement of the degree to which the theoretical or actual value and the predicted value. The ratio of correct predictions to total number of occurrences is typically used to define accuracy. Eq. 2 displays the accuracy equation.

DOI: https://doi.org/10.25195/ijci.v50i2.507
Print ISSN: 2313-190X, Online ISSN: 2520-4912

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \qquad (2)$$

- Precision: From all of the predicted values in the actual class, the actual values that were successfully predicted are measured as precision. The capacity of classifiers to not categorize a negative example as positive is measured by precision. Eq. 3 displays precision equation.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (3)$$

- Recall: The rate of positive values which are accurately classified is measured by recall. How much of actual positives are correctly classified is answered by recall. Eq. 4 displays the recall equation.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (4)$$

- F1-score: The harmonic mean of recall and precision is the F1-score, which is commonly known as F-measure. The F1-score equation is displayed in Eq. 5.

$$\text{F1-score} = \frac{2*precision*recall}{precision+recall} \qquad (5)$$

4. RESULTS OF THE PROPOSED MODEL

Applying classifiers (SVM and KNN). Subsequently, accuracy, recall, precision, and f1-score are used to assess performance. The comparison results between SVM and KNN applied to the dataset without any preprocessing are shown in Table 2 and Fig 4.

TABLE II. RESULTS OF CLASSIFIERS WITHOUT PREPROCESSING DATASET.

Algorithms	Precision	Recall	F1-score	Accuracy
KNN	%100	%86	%92	%89
SVM	%100	%93	%96	%95

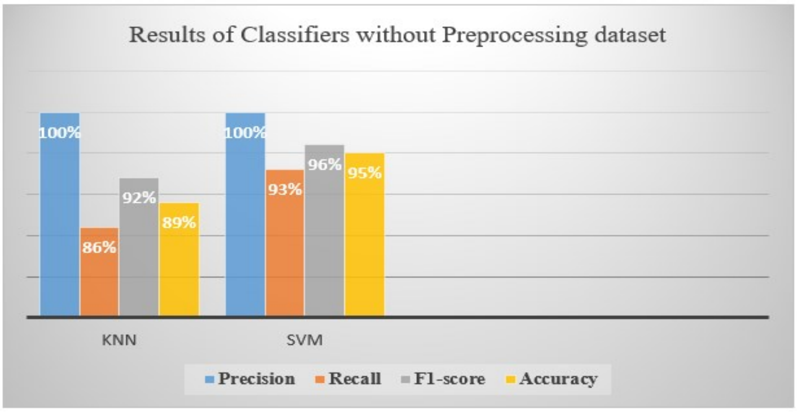

Fig 4. Show results of Classifiers without Preprocessing dataset

Applying classifiers (SVM and KNN). Subsequently, accuracy, recall, precision, and f1-score are used for the assessment of performance. Table 3 and Fig 5 presents the comparative results between SVM and KNN while preprocessing the dataset by deleting missing values.

TABLE III. RESULTS OF CLASSIFIERS WITH PREPROCESSING MISSING VALUES USING DELETING MISSING VALUES IN THE DATASET

Algorithm	Precision	Recall	F1-score	Accuracy
KNN	%100	%92	%95	%93

DOI: https://doi.org/10.25195/ijci.v50i2.507
Print ISSN: 2313-190X, Online ISSN: 2520-4912

SVM	%100	%95	%97	%96

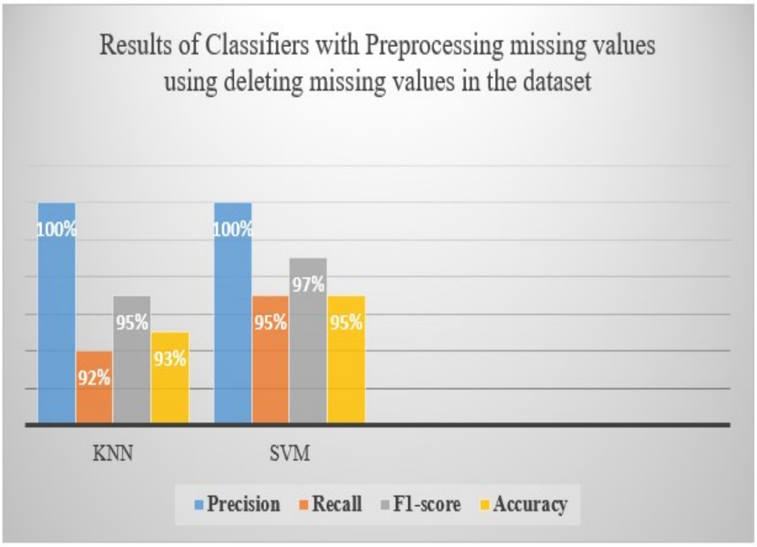

Fig 5. Show Results of Classifiers with Preprocessing missing values using deleting missing values in the dataset

Applying classifiers (SVM and KNN). Subsequently, accuracy, recall, precision, and f1-score are utilized in order to assess performance. The comparison between SVM and KNN applied to the data-set with preprocessing missing values with the use of mean approach is shown in Table 4 and Fig 6.

TABLE IV. RESULTS OF CLASSIFIERS WITH PREPROCESSING MISSING VALUES USING THE MEAN METHOD

Algorithm	Precision	Recall	F1-score	Accuracy
KNN	%100	%97	%98	%98
SVM	%100	%97	%98	%98

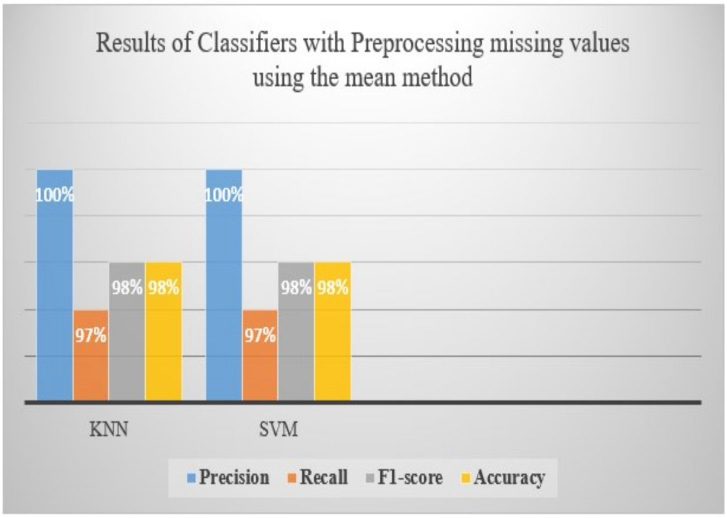

Fig 6. Show Results of Classifiers with Preprocessing missing values using the mean method

DOI: https://doi.org/10.25195/ijci.v50i2.507
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Applying classifiers (KNN and SVM). Subsequently, accuracy, recall, precision, and f1-score are utilized in order to assess performance. The comparison between SVM and KNN applied to the dataset after preprocessing the missing values with the use of median approach is shown in Table 5 and Fig 7.

TABLE V. RESULTS OF CLASSIFIERS WITH PREPROCESSING MISSING VALUES USING THE MEDIAN METHOD

Algorithm	Precision	Recall	F1-score	Accuracy
KNN	%100	%94	%96	%95
SVM	%100	%97	%98	%98

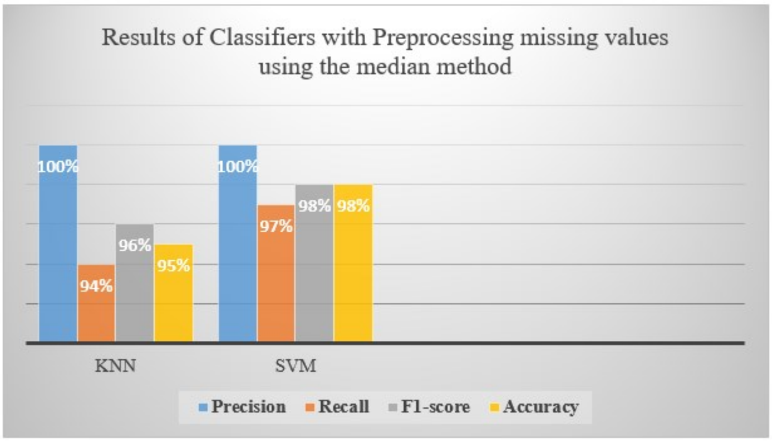

Fig 7. Show Results of Classifiers with Preprocessing missing values using the median method

5. CONCLUSIONS

The number of missing data was very large, which affected the accuracy of the model results. The accuracy reached (%95) without preprocessing the dataset indicates that the studies and findings conducted indicate that the purity and clarity of the dataset have an impact on the diagnosis's accuracy. Since the diagnostic accuracy rate drops to (%89) without a preprocessing stage, kidney diagnosis is necessary. The suggested work is significantly impacted by the preprocessing of the dataset. The outcomes demonstrate that the dataset's missing values were caused by preprocessing. As evidence, the mean and median approaches for preprocessing of the missing values in the dataset yielded diagnosis accuracy of 98% for all suggested classification approaches. When compared to earlier relevant efforts, the performance of the suggested model was good. After preprocessing the dataset with median and mean approaches to fill in missing values, two classification algorithms produced the best accuracy. Difference results can also be obtained using other ML Algorithms like K mean, decision trees, ANN etc. Future works may hybrid classification models using KNN or SVM with other techniques of ML.

References

[1] S. Revathy, B. Bharathi, P. Jeyanthi, and M. Ramesh, "chronic kidney disease prediction using machine learning models," Int. J. Eng. Adv. Technol., vol. 9, no. 1, pp. 6364–6367, 2019, doi: 10.35940/ijeat. A2213.109119.

[2] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," IEEE Access, vol. 8, pp. 20991–21002, 2020, doi: 10.1109/ACCESS.2019.2963053.

[3] Ahmed Sami Jaddoa, "Heart Disease Prediction System Using (SMOTE Technique)," vol. 050006, 2023.

[4] M. U. Emon, A. M. Imran, R. Islam, M. S. Keya, R. Zannat, and Ohidujjaman, "Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches," Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021, no. February, pp. 713–719, 2021, doi 10.1109/ICICT50816.2021.9358491.

[5] J. Xiao et al., “Comparison and development of machine learning tools in the prediction of chronic kidney disease progression,” J. Transl. Med., vol. 17, no. 1, pp. 1–13, 2019, doi: 10.1186/s12967-019-1860-0.

[6] K. L. De Almeida et al., “Kidney Failure Detection Using Machine Learning Techniques,” pp. 1–8, 2020, [Online]. Available: https://hal.archives-ouvertes.fr/hal-02495264.

[7] H. Ilyas et al., “chronic kidney disease diagnosis using decision tree algorithms,” BMC Nephrol., vol. 22, no. 1, pp. 1–11, 2021, doi: 10.1186/s12882-021-02474-z.

[8] S. Pal, “Prediction for chronic kidney disease by categorical and non_categorical attributes using different machine learning algorithms,” Multimed. Tools Appl., pp. 41253–41266, 2023, doi: 10.1007/s11042-023-15188-1.

[9] The chronic kidney disease dataset is downloaded from https:// archive.ics.uci.edu/ml/datasets/chronic_kidney_disease.

[10] K. M. Orabi, Y. M. Kamal, and T. M. Rabah, “Early predictive system for diabetes mellitus disease,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9728, pp. 420–427, 2016, doi: 10.1007/978-3-319-41561-1_31.

[11] S. A. Aboalnaser and H. R. Almohammadi, “Comprehensive study of diabetes miletus prediction using different classification algorithms,” Proc. - Int. Conf. Dev. eSystems Eng. DeSE, vol. October-20, pp. 128–133, 2019, doi 10.1109/DeSE.2019.00033.

[12] M. S. Arif, A. Mukheimer, and D. Asif, “Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model,” Big Data Cogn. Comput., vol. 7, no. 3, 2023, doi: 10.3390/bdcc7030144.

[13] E. M. Senan et al., “Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques,” J. Healthc. Eng., vol. 2021, 2021, doi: 10.1155/2021/1004767.

[14] A. Syarif, O. D. Riana, D. A. Shofiana, and A. Junaidi, “A Comprehensive Comparative Study of Machine Learning Methods for Chronic Kidney Disease Classification: Decision Tree, Support Vector Machine, and Naive Bayes,” Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 10, pp. 597–603, 2023, doi: 10.14569/IJACSA.2023.0141063.

[15] M. A. Abdel-Fattah, N. A. Othman, and N. Goher, “Predicting Chronic Kidney Disease Using Hybrid Machine Learning Based on Apache Spark,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/9898831.