EXAMINATION OF UNREMITTING KIDNEY ILLNESS BY UTILIZING MACHINE LEARNING CLASSIFIERS

Fareeha Sarwar[1, 2], Nuno Garrido[1], Pedro Sebastião[1] and Akmal Rehan[2]
[1]*Instituto de Telecomunicações (IT-IUL)*
Ed. Armadas, 1649-026, Lisbon, Portugal
[2]*Faculty of Science, Department of Computer Science*
University of Agriculture, Faisalabad, Pakistan

ABSTRACT

Chronic kidney disease is a rising health issue that affects millions of people worldwide. Early detection and characterization of this disease is essential for effective management and control. This disease is associated with several serious health risks, such as cardiovascular disease, increased risk of stroke, and end-stage renal disease, which can be effectively prevented by early detection and treatment. Medical scientists rely on machine learning algorithms to diagnose the disease accurately at its outset. Recently, adding value to healthcare is being accomplished through the integration of machine learning algorithms into mobile health solution. Considering this, this paper proposes a predictive model of three machine learning classifiers, including Support Vector Machine, Decision Tree, and Multilayer Perceptron for chronic kidney disease prediction. The performance of the model was assessed using confusion matrix and executed in popular machine learning software tools such as WEKA and Rapid Minor. The study found that support vector machine yielded the highest accuracy rate of 98% in predicting chronic kidney disease in WEKA among other standard classifiers by using 10-fold cross validation. In addition, the proposed prediction model has been compared with existing models in terms of accuracy, sensitivity, and specificity. The experimental results indicate that the proposed predictive model shows promising results. These findings could integrate with the development of mobile health solution and other innovative approaches to prevent and treat this debilitating condition.

KEYWORDS

Machine Learning Classifiers, Chronic Kidney Disease, WEKA, Rapid Minor, Mobile Health Solution

1. INTRODUCTION

Worldwide, non-communicable diseases (NCDs) have replaced communicable diseases as the leading cause of morbidity and premature death. In low- and middle-income countries, 80% of the burden occurs, and 25% occurs in people under 60. NCDs are expected to have a significant impact on the global economy: by 2015, just two diseases (cardiovascular disease and diabetes) will account for 5% of global GDP. Heart disease, stroke, and peripheral vascular disease account for approximately half of the economic burden, causing more deaths than HIV/AIDS, malaria, and tuberculosis combined. Among diabetes and cardiovascular disease (including hypertension), kidney disease is an essential determinant of poor health outcomes, and the World Health Organization recommends that national NCD programs focus on preventing kidney disease, particularly at the primary care level (Couser, Remuzzi, Mendis, & Tonelli, 2011). As chronic kidney disease (CKD) does not follow any age limit, it can appear at any age. Further, if one has already developed CKD, one is more likely to experience sudden deterioration of kidney function. Procrastination can cause severe kidney damage. This disease requires early detection in order to be treated successfully. However, CKD doesn't show symptoms in the initial stages, making it impossible to identify it without testing. A person with milder chronic kidney disease may do not show any disease-related symptoms, and that can make it hard to predict. The effected person will show symptoms that are very common, like nausea, high blood pressure and blood in your urine. At this stage if CKD is detected, it can be cured and monitored to prevent prolonged damage to the kidney. Urine and blood test can be conducted to detect it an early stage. When chronic kidney disease worsens or at a high stage, it starts showing symptoms like anemia, low immune response, blood pressure, nerve damage, and weak bones. But it is also observed that some people with chronic kidney disease do not show any symptoms at all. Glomerular filtration rate can be obtained from simple blood tests. The stages of kidney disease are shown in Table 1.

Machine learning algorithms are used widely for classification and prediction in the medical field. Because of the dynamic, progressive, and heterogeneous nature of chronic kidney disease, it has become necessary to predict the presence and progression of chronic kidney disease with good accuracy and precision. If the patient has chronic kidney disease, then the severity level is checked by measuring at which stage of chronic kidney disease the patient is. Chronic kidney disease stage measure is very crucial for diagnosis recommendation as treatment decisions are highly influenced by chronic kidney disease stages. Nowadays, electronically availability of medical data makes it easy for researchers to analyze, classify and predict with the help of machine learning algorithms and predictive models. Some mathematical computational models developed in recent years for accurate prediction and support decisions of healthcare providers (Itani, Rossignol, Lecron, & Fortemps, 2019).

Mobile system development in health care system for data gathering and monitoring purpose is an emerging technology that becomes essential especially for emergency situations such as spinal cord and acute trauma. To handle huge amount of data cloud storage-based solutions have been introduced in compliance with mobile application in healthcare. From past years most of the focus was on the development of mHealth Solution, that can manage patients remotely. Many studies focus on the mHealth Solution model, technology, data gathering techniques, and communication. This solution has evolved exponential over time with standard and regulation (Zhang, et al., 2019). The motive of this study is to improve machine learning model performance in predicting chronic kidney disease so that in future it can be integrated with mobile Health Solution.

Table 1. Stages of kidney disease

Stages	GFR Scale	Description
1st	>=90	Kidney function normally working, little bit problem in urinating
2nd	60-89	At this stage kidney function mildly disturbed
3rd	30-59	Moderate level of kidney function working
4th	15-29	At this stage kidney function severely reduced
5th	<15	Depict end stage of kidney functioning, patient will be on dialysis

2. REVIEW OF LITERATURE

Numerous researchers have worked in healthcare domain using Machine Learning and Data Mining classifiers. Various studies delineated that Machine Learning and Data Mining algorithms are potent for prediction and prognosis in medical field. In this research study, author achieved high classification accuracy in prediction using Machine Learning techniques on specified problem. In past few years mHealth System has been introduced that has ability to predict and monitor diseases via mobile applications.

Different Machine Learning classifiers have been applied in prediction of chronic kidney disease.(Jongbo, Adetunmbi, Ogunrinde, & Badeji-Ajisafe, 2020) employed Bagging and Random Subspace approaches on three data mining classifiers including k-Nearest Neighbor, Naïve Bayes, and Decision Tree for improving the performance of these classifiers. The data set used in this research was the data of 400 CKD patients collected from the UCI repository. After preprocessing, data was split into two parts, consisting of 70% data to train the model and the remaining 30% data to test the performance of the model. After that, researchers trained three classifiers (Decision Tree, KNN, and Naïve Bayes) build models and checked their performance. The predictions of the classifiers were gathered utilizing the gathering method (Bagging) with greater part casting a ballot rule and the results of this research explored that 97% accuracy of prediction can be attained by utilizing a random subspace ensemble on k-Nearest Neighbors classifier.

A study conducted by (Islam, Akter, Hossen, Keya, Tisha, & Hossain, 2020) on the risk factor prediction of chronic kidney disease using machine learning classifiers. They collected 1032 patient data through questionnaire and then formatted it into CSV files format. In the preprocessing stage they remove noisy data values and inputted missing values. They proposed model using 6 machine learning algorithms, namely Random Forest, Simple Linear Regression, Naïve Biased, Linear Regression, Simple Logistic Regression and

Decision Stump. The researchers train model with 63% data and test with 37% dataset. The results show that for this real time data set this model gave satisfactory accuracy in prediction of risk of chronic kidney disease. They also concluded that the fundamental driver of kidney ailment is hemoglobin factor.

Mobile Health solution is an emerging technology in terms of cost and time saving proposed mHealth solution architecture. (Qureshi, b, Jeon, & Piccialli, 2020) collected data through mobile application, clinics, and healthcare data sets. For handling bulk data cloud-based storage model was incorporated using two tier architecture. Permission framework was used to implement access control system. In the application end security and privacy was achieved through user authentication interface. A dataset of 400 patients containing 25 attributes was taken from the UCI repository. They proposed dynamic and predictive model using 4 machine learning classifiers named Support Vector Machine, Decision Tree, KNN Model, and Naïve Byes Model to predict cardiovascular disease prediction. For this model data was derived from cloud storage and several feature selections techniques were used to obtain dataset. Afterwards dataset was stratified into 75% of training and 25% of testing. This proposed predictive model provided good accuracy with compared to previous predictive models. This model has some limitation also. They suggested some further treatment decisions and early preventive measures about cardiovascular disease.(Estonilo & Festijo, 2022) also advanced mHealth Solution using deep learning classifiers to predict and monitor Diabetes Mellitus remotely. They also improved the quality of mHealth Solution architecture in terms of security and responsiveness. Complied MAG guidance to maintain standard of proposed architecture. A Google Form was used to collect data from November 2021 to December 2021. An online evaluation followed after the data collection. People with diabetes or without diabetes, familiar or not familiar with m-health applications, were the target participants. They named their model DMChecker which gave 93% accuracy in predicting Diabetes Mellitus.

Many research stated that data mining is useful for inconsequential clinical data for discovering hidden patterns from kidney disease patient's test report.(Brito, Esteves, Peixoto, Abelha, & Machado, 2022) worked specifically on serum creatinine value to identify those patients who were at the last stage of the kidney disease. The patients who undergo continuous ambulatory peritoneal dialysis (CAPD) treatment at the last stage of kidney disease were at the subject of this research. As CAPD is not a permanent heal from kidney disease, there is a need for continuous monitoring of these patients' blood tests. Through blood tests, a patient's evolution, therapeutic drug monitoring, and another prognosis can be observed. Creatinine level is used to determine the working of kidney function, through this value kidney disease stages can be determined.

3. MATERIALS AND METHODOLOGY

The main objective of this research is to improve machine learning classifiers accuracy in prediction chronic kidney disease in the proposed architecture of mHealth Solution. For this, dataset was preprocessed and analyzed for generating legitimate results. Proposed methodology specifies three classifiers used to predict chronic kidney disease presence with accuracy shown in Figure 1.

3.1 Data Set Description

Dataset plays a vital role during predictive research. Your results are totally dependent on it so that is really very important to collect accurate data. The data set used in this research was the data of 400 CKD patients collected from the UCI repository (Polat, Mehr, & Cetin1, 2017).The data set contains a total of 14 attributes having 6 nominal values and 8 numeric values. The data were classified into two classes: Yes (containing the data of patients suffering from Chronic Kidney Disease) and No (containing the data of healthy people).A description of each feature can be found in Table 2.

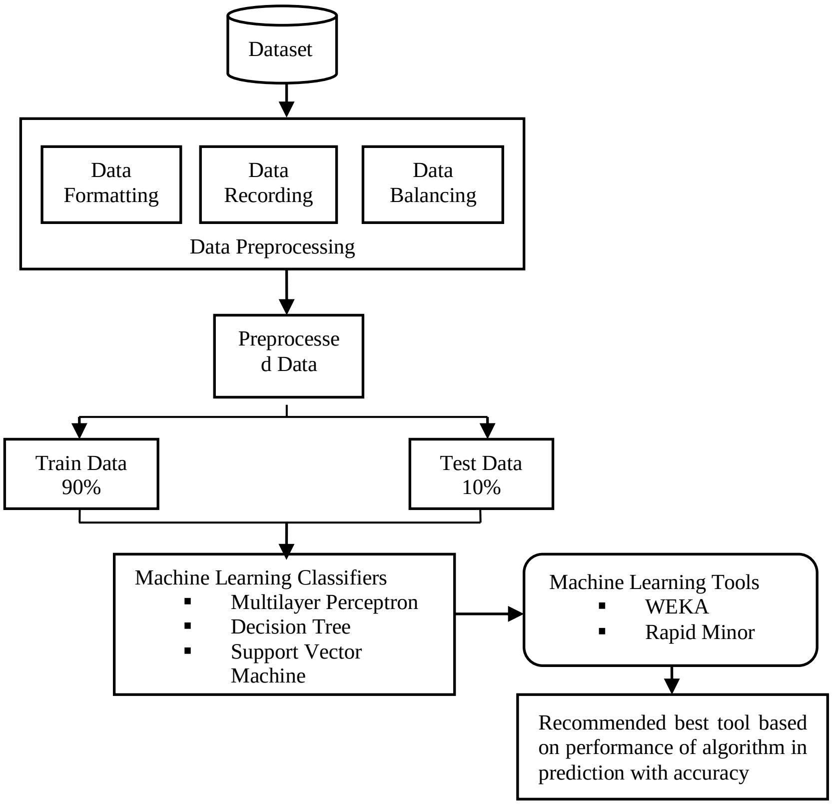

Figure 1. Workflow of proposed methodology

Table 2. Data set description

CKD dataset	Attributes meaning	Category	Scale	Missing values
bp	Blood pressure	Numerical	Mm/Hg	9
sg	Specific gravity	Numerical	1.005 to 1.025 (the higher value means higher risk)	11
al	Albumin	Nominal	0 to 5	47
su	Sugar level	Nominal	0 to 5	49
rbc	Red blood cells	Nominal	Abnormal or Normal	152
bu	Blood urea	Numerical	mgs/dl	19
sc	Serum creatinine	Numerical	mgs/dl	17
sod	Sodium	Numerical	mEq/L	89
pot	Potassium	Numerical	mEq/L	88
hemo	Hemoglobin	Numerical	gms	17
wc	White blood cell count	Numerical	cells/cumm	112
rc	Red blood cell count	Nominal	millions/cmm	131
htn	Hypertension	Nominal	yes, no	2
classification	Class	Nominal	CKD, not CKD	0

3.2 Data Preprocessing

There were outliers and noise in the dataset. Since certain measures may be lost when patients are being tested, the preprocessing stage incorporated the estimation of missing values and noise elimination, including outliers, normalization, and unbalanced data checking. Data was preprocessed in multiple steps. In first step rename the columns in broaden form. Then data types of all columns were corrected. At that moment ablution of categorical columns were performed and to manage the missing values in the data set median value method was used. The noisy values were soothed by using binning technique. The original dataset contained 25 attributes like age, blood pressure (BP), sugar, red blood cells, white blood cell, red blood cells, sodium, potassium, hemoglobin, etc. during the phase of preprocessing, we eliminated few attributes based on importance and change the type of few attributes from discrete values to nominal.

3.3 Mobile Health Solution Architecture

Due to increase support in chronic diseases prediction and monitoring Mobile Health solution has gained a significant attention. Refine diagnostic procedure in outlying and rural areas where medical facilities are not available or at a distant or medical care is not possible, a self-diagnostic application will be developed to predict chronic kidney disease and its severity level. The model will be connected with application using Flask API considered as end points between model and application. Micro services will be used to authenticate user, predict disease, and monitor patients on daily basis. Application will input data through sensors, user interface and clinical history to predict presence of CKD and monitor patient. For huge data storage, cost effective and stability, cloud-based storage solution will be employed. The data from the mobile application and previous history of patient will be stored in this storage model. To ensure privacy and confidentiality of patient data isolation strategies can be implemented. The wearable sensors will monitor the patient and transmit the data to mobile application automatically. Data will go under preprocessing stage where outliers, noisy data will be removed. Missing values will be inputted to enhance the quality of the model. By importing all the libraries feature selection method will be implemented. After preprocessing, three machine learning classifiers named Multilayer Perceptron, Decision Tree and Support Vector Machine will be applied to optimally reduced dataset. The accuracy of each classifier will be evaluated for the prediction of CKD. Based on this prediction, further recommendation, treatments, and continuous monitoring will be suggested through mobile application. With this mobile architecture solution, kidney disease can be detected at an early stage. Patients can be monitored according to their severity level. Figure 2 depicts the system architecture.

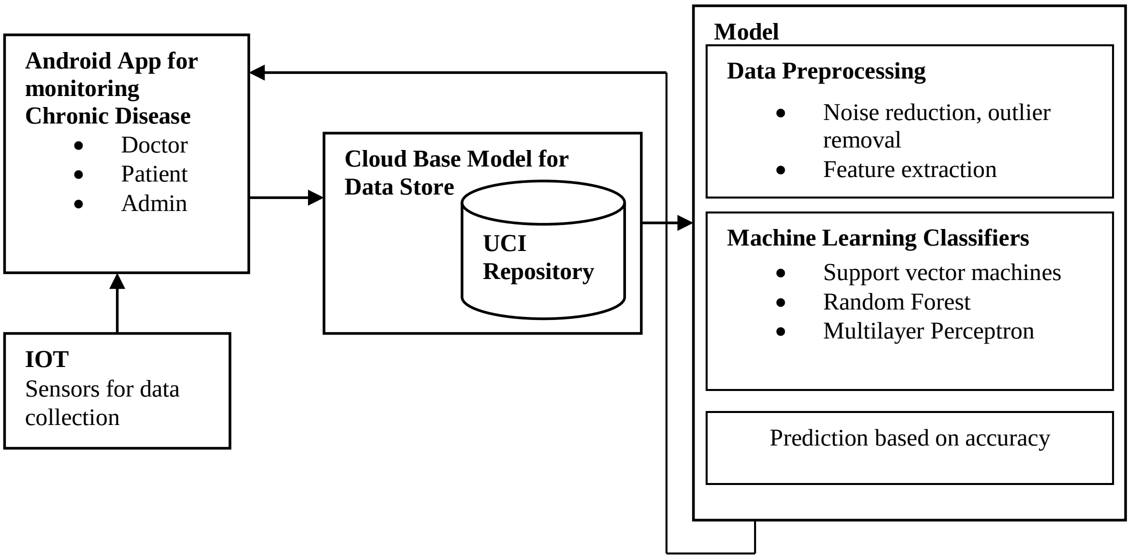

Figure 2. System architecture

3.4 Evaluating the Model

In this peculiar analysis, value of CKD marked as positive and value of not CKD marked as negative. With the help of this confusion matrix the performance of machine learning classifiers was assessed. With the help of this confusion matrix the specific results were displayed. From the CKD samples the true positive (TP) specifies that the disease was accurately identified whereas false negative (FN) specifies that the disease was incorrectly identified. For the non-CKD samples false positive (FP) indicates that the disease was wrongly diagnosed. While on the contrary, true Negative (TN) highlights the disease was accurately diagnosed. Model performance was measured using precision, recall, sensitivity, accuracy, F1 score, and specificity.

4. EXPERIMENTAL RESULTS

Using two machine-learning tools, three classifiers were applied to a dataset. In this research, two phases of machine learning will be employed. First is the training phase, where a test dataset is trained to fit the classifier parameters. Second is the Validation Phase, which estimates how well the model has been introduced. In order to train and test the data, we used 10 cross-validation methods. Machine learning algorithms Decision Tree, Support Vector Machine, and Multilayer Perceptron will be compared. Machine learning tools WEKA and Rapid Miner will be used to train and test the dataset. The comparative analysis will be performed among the selected classifiers based on accuracy and precision in the prediction of kidney disease (Rady & Anwar, 2019).

As a result of previous research, the Multilayer Perceptron provided an accuracy of 92.5% when detecting kidney disease. In order to train and test their model, researchers used a dataset containing 25 attributes. Based on multilayer perceptrons applied to a reduced dataset, this study achieved 96.5% accuracy on a 10-fold validation with an error rate of 3.5%. A total of 386 instances are correctly classified, while 14 instances are incorrectly classified. A 10-fold validation divides the dataset into ten parts, with one part being used to train the model and the other part being used to evaluate whether the model is correct (Vashisth, Dhall, & Saraswat, 2020). Previously, SVM was applied to a dataset with 25 attributes in order to detect kidney disease with an accuracy of 70%. In this study, Support Vector Machines were applied to a chronic kidney disease dataset with 14 attributes and gave 98% accuracy with a 2.5% error rate on a 20-fold validation. In a set of 400 instances, 391 are correctly classified while 9 are incorrectly classified. 20-fold validations allow the dataset to be divided into 20 parts, one of which is used to train the model and the other part is used to evaluate whether or not the model generates correct results (Vashisth, Dhall, & Saraswat, 2020). Previous research has found that decision trees are 95% accurate when there are 25 attributes in the dataset. On a dataset with 14 attributes, Decision Tree provided 97.75% accuracy with an error rate of 2.25% after 10-fold validation. In a 10-fold validation, 400 instances are divided into 10 parts and one part is used to train the model, and the other part is used to determine which model is reliable. 391 instances have been correctly classified while 9 have been incorrectly classified. Results can be found in Table 3 (Baidya, Umaima, & Islam, 2022).

Table 3. Comparison of algorithms in weka

Classifiers	Correctly Classified %	Precision	F measure	Error Rate
Multilayer Perceptron	96.5	0.967	0.977	3.5
Support Vector Machine	98	0.981	0.98	2
Decision Tree	97.75	0.978	0.978	2.25

As in Rapid Miner, three classifiers are applied to the dataset and the results are presented separately in Table 4. First, the results are presented one by one, and then they are presented collectively in a tabular format. The Decision Tree provided 94% accuracy when applied to the dataset. The accuracy of the Support Vector Machine on the chronic kidney disease dataset was 94.50% while the accuracy of Multilayer Perceptron on the chronic kidney disease dataset was 63.50%.

Table 4. Comparison of algorithms in rapid miner

Classifiers	Accuracy %
Decision Tree	94.17
Support Vector Machine	94.50
Multilayer Perceptron	63.75

Table 5 compares the performance of the same algorithm on two different machine learning tools. Three machine learning algorithms were applied to a kidney disease dataset, and the results indicate that Support Vector Machine is the most accurate in predicting chronic kidney disease, with 98% accuracy in WEKA. The multilayer perceptron did not perform well in Rapid Miner, but the accuracy in WEKA was satisfactory in predicting chronic kidney disease. Compared to previous research, this predictive model shows good accuracy with a reduced dimension dataset. There is evidence that most previous work shows good accuracy with a dataset with 25 attributes, but this accuracy can be achieved using a dimensionally reduced dataset with 14 features, which will result in a reduction of processing time and space.

Table 5. Comparison of algorithms in both tools

Tools	**Algorithm**	**Accuracy %**
WEKA	Decision Tree	97.75
	Multilayer Perceptron	96.50
	Support Vector Machine	98.00
Rapid Minor	Decision Tree	94.17
	Multilayer Perceptron	63.75
	Support Vector Machine	94.50

5. CONCLUSION

The advancement of technology makes healthcare services more advanced, and M-Health is also integrated with such technologies to monitor, process, and store patient data. Further, machine learning techniques are being employed to accurately predict disease analysis and classify disorders. In this study, three machine learning classifiers were used to predict chronic kidney disease, including support vector machine, decision tree, and multilayer perceptron. These classifiers were executed in machine learning software tools such as WEKA and Rapid Miner. The results of this study showed that support vector machine had the highest accuracy rate of 98% in predicting kidney disease in WEKA, while multilayer perceptron did not show good accuracy in Rapid Miner, but showed good accuracy in WEKA. It seems that the proposed machine learning model would be an effective tool for predicting kidney disease, based on an analysis of patient features from a dataset. Compared to existing state-of-the-art models, it outperformed them in accuracy, sensitivity, and specificity. In conclusion, the results of this study demonstrate the potential of machine learning for improving the accuracy of kidney disease prediction that can be integrated with mobile health solution architecture. Future research in this area could focus on the development of hybrid machine learning approaches for predicting chronic diseases, including kidney disease with mobile health solution integration. Additionally, dimension reduction techniques could be applied to minimize computational resources and time in the prediction of kidney disease and other chronic diseases.

ACKNOWLEDGEMENT

This work was supported by Instituto de Telecomunicações, Portugal.

REFERENCES

Baidya, D., Umaima, U., Islam, M. N., Shamrat, F. J. M., Pramanik, A., & Rahman, M. S. (2022, April). A deep prediction of chronic kidney disease by employing machine learning method. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1305-1310). IEEE.

Brito, C., Esteves, M., Peixoto, H., Abelha, A. and Machado, J., 2022. A data mining approach to classify serum creatinine values in patients undergoing continuous ambulatory peritoneal dialysis. *Wireless Networks*, pp.1-9.

Couser, W. G., Remuzzi, G., Mendis, S., & Tonelli, M. (2011). The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. *Kidney international*, *80*(12), 1258-1270.

Estonilo, C.G. and Festijo, E.D., 2022, January. Evaluation of the Deep Learning-Based M-Health Application Using Mobile App Development and Assessment Guide. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0435-0440). IEEE.

Islam, M.A., Akter, S., Hossen, M.S., Keya, S.A., Tisha, S.A. and Hossain, S., 2020, December. Risk factor prediction of chronic kidney disease based on machine learning algorithms. In *2020 3rd international conference on intelligent sustainable systems (ICISS)* (pp. 952-957). IEEE.

Itani, S., Rossignol, M., Lecron, F. and Fortemps, P., 2019. Towards interpretable machine learning models for diagnosis aid: a case study on attention deficit/hyperactivity disorder. *PloS one*, *14*(4), p.e0215720.

Jongbo, O.A., Adetunmbi, A.O., Ogunrinde, R.B. and Badeji-Ajisafe, B., 2020. Development of an ensemble approach to chronic kidney disease diagnosis. *Scientific African*, *8*, p.e00456.

Kulkarni, E.G. and Kulkarni, R.B., 2016. Weka powerful tool in data mining. *International Journal of Computer Applications*, *975*, p.8887.

Pandit, A. K. (2020). Chronic Kidney Disease Prediction by Machine Learning.

Polat, H., Danaei Mehr, H. and Cetin, A., 2017. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. *Journal of medical systems*, *41*, pp.1-11.

Qureshi, K.N., Din, S., Jeon, G. and Piccialli, F., 2020. An accurate and dynamic predictive model for a smart M-Health system using machine learning. *Information Sciences*, *538*, pp.486-502.

Rady, E.H.A. and Anwar, A.S., 2019. Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, *15*, p.100178.

Tahmasebian, S., Ghazisaeedi, M., Langarizadeh, M., Mokhtaran, M., Mahdavi-Mazdeh, M. and Javadian, P., 2017. Applying data mining techniques to determine important parameters in chronic kidney disease and the relations of these parameters to each other. *Journal of renal injury prevention*, *6*(2), p.83.

Vashisth, S., Dhall, I., & Saraswat, S. (2020, January). Chronic kidney disease (CKD) diagnosis using multi-layer perceptron classifier. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 346-350). IEEE.

Webster, A.C., Nagler, E.V., Morton, R.L. and Masson, P., 2017. Chronic kidney disease. *The lancet*, *389*(10075), pp.1238-1252.

Zhang, Y., Liu, C., Luo, S., Xie, Y., Liu, F., Li, X. and Zhou, Z., 2019. Factors influencing patients' intentions to use diabetes management apps based on an extended unified theory of acceptance and use of technology model: web-based survey. *Journal of medical Internet research*, *21*(8), p.e15023.