SCIENCE FORUM (JOURNAL OF PURE AND APPLIED SCIENCES) 25 (2025) 134 - 146

Science Forum (Journal of Pure and Applied Sciences)

Journal Homepage: https://atbuscienceforum.com.ng/index.php/jpas

Enhanced Chronic Kidney Disease Prediction using Boruta Algorithm

Zahraddeen Sufyanu[1], Muhammad U. Yusif[2], Shamsuddeen M. Abubakar[1], Zaharaddeen S. Iro[1]

[1]Department of Computer Science, Faculty of Computing Federal University Dutse, Jigawa State,

[2]I.C.T Department, National Examinations Council (NECO),

Abstract

Chronic Kidney Disease (CKD) is a serious and life-threatening medical condition that impacts millions of individuals worldwide. It arises when both kidneys become damaged, impairing their ability to effectively filter blood. The kidneys play a critical role in maintaining overall health by filtering waste products and excess fluids from the blood, which are then excreted through urine. When kidney function declines, harmful toxins and fluids accumulate in the body, leading to a range of complications, including cardiovascular disease, anemia, bone disorders, and ultimately, kidney failure if left untreated. Recent studies highlight the growing burden of CKD, particularly in Sub-Saharan Africa, where the prevalence of the disease is alarmingly high. For instance, West Africa has reported a CKD prevalence rate of 16%, the highest on the continent. This underscores the urgent need for effective diagnostic and predictive tools to address the rising incidence of CKD in resource-limited settings. In recent years, machine learning techniques have gained significant traction in the field of medical research, particularly for disease prediction and diagnosis. These advanced computational methods offer the potential to improve early detection, enhance diagnostic accuracy, and support clinical decision-making. This paper explores the application of machine learning algorithms to predict CKD, with a focus on improving prediction accuracy. Among the various techniques evaluated, the Boruta algorithm, a feature selection method, demonstrated promising results by achieving satisfactory prediction outcomes even with a limited number of features. This highlights its efficiency in identifying the most relevant predictors of CKD. Furthermore, the Random Forest algorithm emerged as a highly effective model for CKD prediction, delivering exceptional performance metrics. Specifically, it achieved an accuracy of 99%, a precision of 100%, an F1 score of 98%, a Cohen Kappa score of 97%, and a classification error of just 0.0125%. These results underscore the robustness and reliability of Random Forest in accurately classifying CKD cases, making it a valuable tool for early diagnosis and intervention.

Keywords

Chronic Kidney Disease,
Boruta,
Machine Learning,
Classification Error

Corresponding Author: Zahraddeen Sufyanu, sufyanzzzzz@gmail.com Department of Computer Science, Faculty of Computing Federal University Dutse, Jigawa State.

INTRODUCTION

Chronic kidney disease (CKD) occurs when both kidneys are damaged and unable to filter the blood properly and control the levels of salt and minerals in the blood, which include calcium, phosphorus, sodium, and potassium (Priya *et al.*, 2020). CKD is associated with symptoms such as diabetes, age, hypertension, obesity, and primary renal disorders (Nandhini & Aravinth, 2021).

The following are two major tests for chronic kidney disease:

a) eGFR value: This metric measures how well the kidneys filter blood. if the eGFR value is greater than 90 both the kidney are healthy, else the kidney are not healthy(Mula-Abed, 2012).

b) Urine Test: The doctor may order a urine test because the kidneys produce urine, and if the urine contains blood or protein, it means that the kidneys aren't working properly(Levey *et al.*, 2009). CKD is a serious medical condition that affects millions of people worldwide. According to the prevalence of chronic kidney disease in Sub-Saharan Africa was 16%(Abd Elhafeez *et al.*, 2018). CKD is responsible for about 8% to 10% of all hospital admissions in Nigeria (Ulasi & Ijoma, 2010). Additionally, there are no national data on the prevalence of chronic kidney disease, few were discovered in the south and two in the north (Chukwuonye *et al.*, 2018). Due to the rapid growth of African populations, therefore more diagnostic techniques are required to control the disease's prevalence.

LITERATURE REVIEWS

Its prevalence has increased partly due to rising of incidences of hypertension and diabetes mellitus, and it is predicted to become major cause of death worldwide by 2040. Nishat *et al.* (2018) used seven machine learning algorithms with University of California Irvine machine learning repository. The dataset contains 400 instances and 25 attributes. The model was evaluated using various metrics such as accuracy, precision, sensitivity, specificity, F1-score, and ROC. Random Forest produced the highest accuracy of 99.75%.

Padmanaban and Parthiban (2016) proposed machine learning algorithms to develop a model for early detection of CKD in diabetic patients using Naïve Bayes and Decision Tree with Weka Tool. The data collected from diabetes research center in Chennai, Naive Bayes archieved the highest accuracy of 91%. Desai (2019) used Boruta algorithm to examine the factors that increase the likelihood of a patient having CKD. The findings revealed that hypertension, blood pressure, uric acid, albumin, age, and serum creatinine are some of the leading causes of CKD. Only seven of the total 24 attributes were confirmed.

Haq *et al.*(2020) suggested a machine learning model for early diabetic illness prediction. Additionally, it was determined that machine learning has the potential to be a significant factor in healthcare. Akmal *et al.* (2020) proposed J48 and Random Forest to detect mobile malware, and Boruta Algorithm was used as feature selection techniques. Pal and Parija (2021)used the Random Forest algorithm to create a heart prediction Model using Kaggle dataset. The model had an accuracy of 86%, a sensitivity of 90.6%, and a specificity of 82.7%.

Almeida *et al.* (2020) created a model for detecting kidney failure using Decision Tree, RandomForest, and Support Vector Machine (SVM). Random Forest and Decision Tree reported the best prediction accuracy of 80% and 87% respectively. Meghana *et al.* (2021) used three algorithms Random Forest, Support Vector Machine and Hybrid Neural Network. The dataset has 400 rows and 26 columns including the id column. The accuracy of the Hybrid neural network was the highest at 97.82 Mtsweni *et al.* (2020) used a Neural Network Model and 10-fold cross validation to propose a model for classifying patients with CKD. Support Vector Machine, K-nearest neighbors, Decision Tree, and Gradient Boost the classifiers used. The experiment revealed that the performance of the Neural Network with 10-fold cross validation was 98.25%.

In the work of Abhijitbet *al.* (2021), Support Vector Machine and Naïve Bayes were used to predict chronic kidney disease. The dataset collected from kaggle.com with 400 records with 25 features. Support Vector Machine reported the accuracy of 61%.

Priya *et al.* (2020) used a multilayer perceptron on 400 different patients with 25 different attributes. The data was obtained from the UCI repository, but only 13 attributes were used in their work after removing irrelevant features. The accuracy of the model is 95.36% and the error rate is 98%, the Cohen Kappa was reported 96% and the MSE was 0.0175.
In another research by Iliyas *et al.* (2020), Deep Neural Network used detect the presence and absence of CKD with 400 patients from Bade General Hospital in Yobe State, Nigeria. The model achieved accuracy 98%, precision 100%, recall 96%, F1-score 98%, Cohen kappa 96.6%, ROC 100%, sensitivity 96%, and specificity 100%.

Majority of research in the literature focuses on full features, while others select some features without using any selection technique. in addition, using feature selection techniques helps eliminate unimportant features and can enhance model performance. In this study, the Boruta algorithm was applied to identify the most relevant features for chronic kidney disease prediction,

effectively reducing the feature space while retaining essential information. By selecting only, the most significant features, the model's efficiency improved, leading to promising accuracy in classification.

METHODOLOGY

Figure 1 below described the general methodology followed to achieve the proposed method. The methodology began with data collection, preprocessing, feature selection, model development and performance evaluation to identify the best algorithm among Random Forest, Support Vector machine and Naïve Bayes for chronic kidney prediction

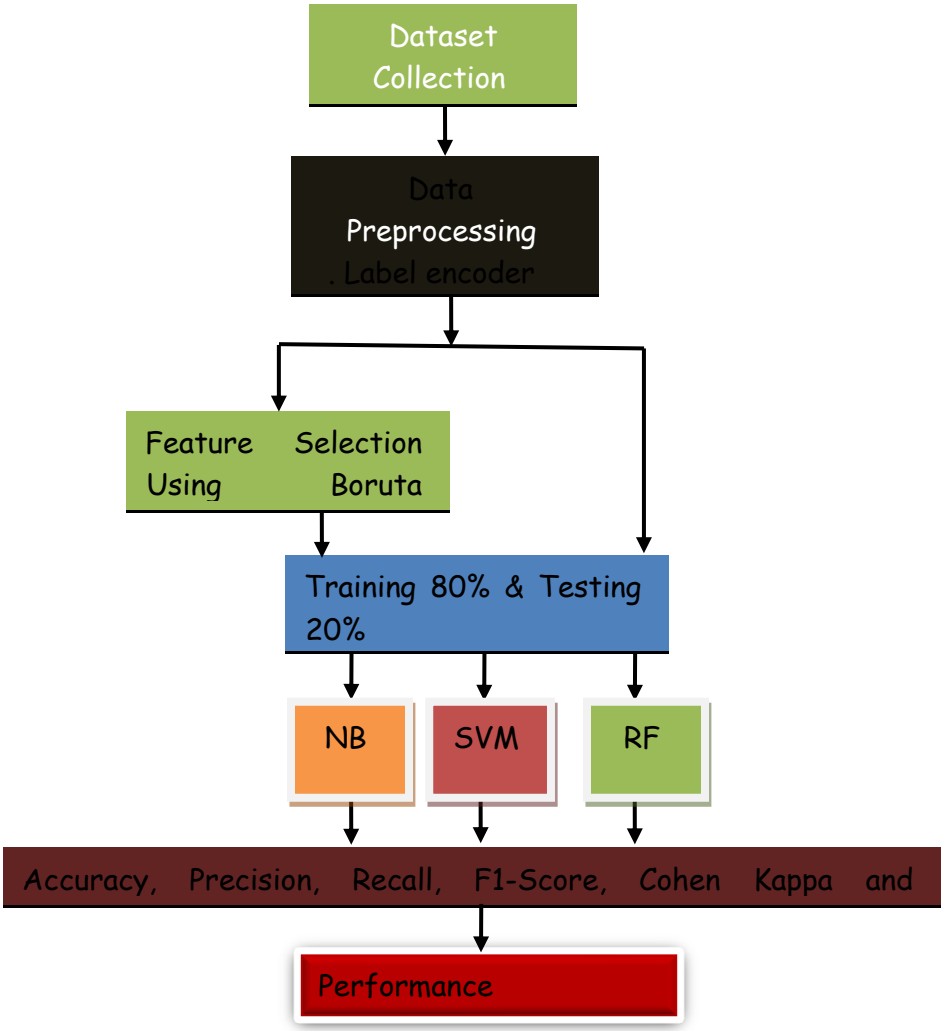

Figure 1: The Proposed Work

Data Collection

To process the chronic kidney disease dataset, Python programming language and Anaconda IDE (Integrated Development Environment) was used, along with the Scikit Learn library.

Table 1: Kaggle Dataset Description

S/N	Representation	Features	Type	Number of missing values	Percentage of missing values
1	Age	Patients age	Discrete Integer Values	9	2.25%
2	Bp	Blood pressure	Discrete Integer Values	12	3.00%
3	Sg	Specific gravity	Nominal Values	47	11.75%
4	Al	Albumin	Nominal values	46	11.50%
5	Su	Sugar	Norminal Values	49	12.25%
6	Rbc	Red Blood Cells	Nominal values	152	38.00%
7	Pc	Puss cell	Nominal Values	65	16.26%
8	Pcc	Puss cell clumps	Nominal values	4	1.00%
9	Ba	Bacteria	Nominal Values	4	1.00%
10	Bgr	Blood glucose in random	Numerical Values	44	11.00%
11	Bu	Blood urea	Numerical	19	4.75%
12	Sc	Serum creatinine	Numerical Values	17	4.25%
13	Sod	Sodium	Numerical Values	87	21.75
14	Pot	Potassium	Numerical Values	88	22.00
15	Hemo	Haemoglobin	Numerical values	52	13.00%

16	Pcv	Packed cell volume	Numerical values	70	17.50%
17	Wc	White blood cell count	Numerical values	105	26.25%
18	Rc	Red blood cell count	Numerical values	130	32.40%
19	Htn	Hypertension	Nominal values	2	0.50%
20	Dm	Diabetes mellitus	Nominal values	0	0.00%
21	Cad	Coronary artery disease	Nominal values	0	0.00%
22	Appet	Appetite	Nominal values	1	0.25%
23	Pe	Pedel edema	Nominal values	1	0.25%
24	Ane	Anaemia	Nominal values	1	0.25%
25	Class	Patients is having CKD or Not	Nominal	0	0.00%

Data Preprocessing

Data preprocessing is a method of transforming raw data into a clean dataset. In this work, data is preprocessed in two ways. Features that have more than 20% missing values was removed. Red Blood Cells, Sodium, Potassium, White Blood Cell Count, and Red Blood Cell Count are completely excluded from the work (Imesh & Damayanthi, 2020). Missing values are filled with the mean, median, or mode values of the respective features; however, missing values are filled with the median value of all attribute columns. The categorical values were converted into binary numbers 1 and 0 using a label encoder.

Boruta Feature Selection Algorithm

Feature Selection is the process of selecting a subset of the most relevant

features in a dataset to describe the target variable, which becomes prominent especially in datasets with a large number of variables and features (Akmal *et al.*, 2020).

The Boruta algorithm is a feature selection method that improves the Random Forest algorithm by determining the most important features in a dataset. It works by generating jumbled copies of the original features, known as shadow features, then training a Random Forest model on them. The value of each original characteristic is then compared to the greatest relevance of the shadow features.

(Maithili, 2019)

Random Forest

Random Forest was trained using a chronic kidney disease dataset by randomly selecting features to evaluate at each node. The process was repeated until the entire tree was built. The CKD data was fed into the trees, and the predictions were made by averaging all N predictions to obtain the overall aggregate predictions. As a result, the forest's final prediction is the most common.

Support Vector Machine

Support Vector Machine (SVM) performs classification by making hyper plane in a dimensional space, which divided the cases of different class labels. SVM can handle both continues and categorical values (Bano & Muhammad, 2016).

Naïve Bayes

Naive Bayes is a classification Algorithm that states that all features are independent and unrelated to one another. The dataset of the present study contains some features such as age, creatinine, blood pressure, sugar, and potassium, among others. The goal is to predict whether a person has CKD or not.

According to naive bayes, each observation determines whether it belongs to class 1 or class 2. Given the same conditions as shown in the dataset (see Table…), the probability that the first person has or does not have chronic kidney disease is calculated in my dataset. This is repeated throughout the 400 records, with the class prediction based on the features or condition.

Data Splitting

Splitting data is essential in machine learning and is widely used. The algorithm divided the data for training and testing. The training test fits the model, while the testing set evaluates it. For training and testing, the data in this work is divided into 80% and20% for better accuracy.

Performance Evaluation

a) Accuracy

This is defined as a model's ability to detect the ratio of true predictions to total number of instances.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

Accuracy is not a sufficient metric for assessing model performance in the medical field. Because the incorrectly predicted cases were not taken into account, other metrics such as Precision, F1-measrue, Kohen Kappa, and Classification Error were adopted.

b) Precision

A metric for correctness that can be defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

c) F1-Score

This is calculated as the harmonic mean of recall and precision. It can also be written as

$$\text{F1-measure} = \frac{2*Precision + Recall}{\text{precision} + \text{Recall}} \quad (3)$$

d) Kohen Kappa

Kohen Kappa is a Statistical method for determining inter-rate reliability for quantitative items in a classifier.

Is define as $$\text{K} = \frac{\text{P(o)} - \text{P(E)}}{1 - \text{P(E)}}$$ (4)

Where Po indicate relative observed agreement among the raters

Where P(e) is the probability of agreement by chance.

e) Classification Error

Classification error displays the incorrect rate of prediction results derived from the confusion matrix. It is defined as $$\text{Error} = \frac{\text{FP+FN}}{\text{TP +TN+FP+FN}} \quad (5)$$

Results

This section displays the output of three different algorithms, Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes, on the Dataset.

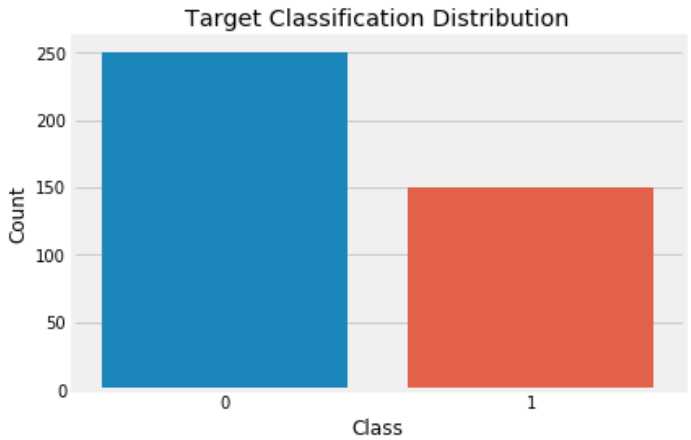

Figure 2: Dataset Distributions

The distributions of the chronic kidney disease dataset, which included 250 CKD and 150 non-CKD patients, are depicted in figure 2 above.

Table 2 shows the features selected by ranking on Kaggle dataset.

Table 2: Ranking significant features of kaggle Dataset

Serial Number	Feature	Ranking
1	Age	1
16	Appetite	1
14	Diabetes_Mellitus	1
13.	Hypertension	1
12	Packed_Cell_Volume	1
11	Haemoglobin	1
10	Serum_Creatinine	1
17	Peda_Edema	1
9	Blood_Urea	1
5	Puss_Cell	1
4	Sugar	1
3	Alumin	1
2	Specific_Gravity	1
1	Blood_Pressure	1
8	Blood_Glucose_Random	1
18	Anaemia	2
15	Coronary_Artery_Disease	3
7	Bacteria	4
6	Puss_Cell_Clumps	5

Table 2 deficts Features significantly more important than shadow features are assigned a ranking of 1, indicating their selection, while less important ones receive rankings greater than 1, signifying their rejection. Uncertain features are re-evaluated in multiple iterations until they are confirmed or rejected. In this process, 15 features were chosen, each representing a significant aspect of the dataset, ensuring improved model

Table 3: Results of the Classifiers without Boruta Algorithm

S/N	Classifiers	Accuracy (%)	Precision (%)	F1-Measure (%)	Cohen kappa (%)	Classification Error (%)	Time (s)
1	Random Forest	96	96	98	96	0.037	0.6
2	Support Vector Machine	97	100	96	96	0.025	0.4
3	Naïve Bayes	96	99	96	95	0.0037	0.3

The table 3, above shows the performance of classifiers without feature selection using Precision, Recall, F1 Measure, Cohen Kappa and Classification Error on the Dataset. Random Forest and Naïve Bayes measured the same accuracy of 96%, while Support Vector Machine reported the accuracy of 97%.

According to the below table, the Random Forest reported the highest accuracy of 99%, while the Support Vector Machine and Nave Bayes reported 97% and 95% accuracies respectively. The Random Forest model has the minimal classification error.

Table 4: Result of the Classifier with Boruta Algorithm

S/N	Classifiers	Accuracy (%)	Precision (%)	F1-Measure (%)	Cohen kappa (%)	Classification Error (%)	Time (s)
1	Random Forest	99	100	98	97	0.012	0.2
2	Support Vector Machine	97	96	96	94	0.025	0.1
3	Naïve Bayes	97	96	96	96	0.025	0.04

The research work was carried out in two phases; in the first phase, we focused on full features to train and test machine learning algorithms and all the records were recorded. In the second phase, the Boruta algorithm was used to remove irrelevant features from the dataset, which helps to improve the model's prediction performance. Therefore, various metrics were used to assess the model's performance. The Random Forest model produced a promising accuracy of 99% with a low classification error of 0.0125% and a relatively short running time of 0.2(s), compared the work of (Abhijitbet *al.*,2021) conducted on the same dataset. This demonstrates that, Boruta Feature Selection Algorithm can improve the prediction of CKD most significantly using Random Forest.

Conclusion

Boruta feature selection was used to enhance the prediction of chronic kidney disease and proved to be effective. This wrapper method identified the most relevant features in the Kaggle dataset, selecting 15 features out of 19. The dataset was split into 80% training and 20% testing. The entire experiment was conducted and documented, demonstrating the impact of feature selection on model performance. Using the Boruta feature selection wrapper method, the most important features were identified, and Random Forest, Support Vector Machine, and Naïve Bayes were applied to the dataset. Among these models, Random Forest achieved the highest accuracy of 99%, showing its effectiveness in classification. In the future, techniques for dataset balancing should be explored, and alternative feature selection methods could be considered to further enhance the model's performance.

References

Abd Elhafeez, S., Bolignano, D., D'Arrigo, G., Dounousi, E., Tripepi, G., & Zoccali, C. (2018). Prevalence and burden of chronic kidney disease among the general population and high-risk groups in Africa: A systematic review. *BMJ Open*, *8*(1). https://doi.org/10.1136/bmjopen-2016-

015069

Abhijitb pathak, Most. Asma Gani, Abras, Hossain Tasin, Sanjida Nusrat Sania, Md. Adil, SuraiyaAkter. (2021).Chronic Kidney Disease (CKD) Prediction using Data Mining Techniques. Available at http//www.researchgate.net/publication/34642247.

Akmal, C., Yahaya, C., Firdaus, A., Mohamad, S., Ernawan, F., Faizal, M., & Razak, A. (2020). Automated Feature Selection using Boruta Algorithm to Detect Mobile Malware. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(5), 9029–9036. https://doi.org/10.30534/ijatcse/2020/307952020

Anantha Padmanaban, K. R., & Parthiban, G. (2016). Applying machine learning techniques for predicting the risk of chronic kidney disease. *Indian Journal of Science and Technology*, *9*(29). https://doi.org/10.17485/ijst/2016/v9i29/93880

Bano Saima, Muhammad Naeem and Ahmed Khan.(2016). A Frame work to improve Diabetes Prediction using K-NN and SVM. *International Journal of Computer Science and Information Security*. 14 (11).

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00327-4

Chukwuonye, I. I., Samuel Ogah, O., Ndukaife Anyabolu, E., Arinze Ohagwu, K., Collins Nwabuko, O., Onwuchekwa, U., Erinma Chukwuonye, M., Chukwuebuka Obi, E., & Oviasu, E. (2018). Prevalence of chronic kidney disease in Nigeria: Systematic review of population-based studies. *International Journal of Nephrology and Renovascular Disease*, *11*, 165–172. https://doi.org/10.2147/IJNRD.S162230

De Almeida, K. L., Lessa, L., Peixoto, A., Gomes, R., Celestino, J., Kidney, J. C., De, K. L., Almeida, A., Peixoto, A. B. S., & Gomes, R. L. (2020). Kidney Failure Detection Using Machine Learning Techniques. *… on Advances in …*, 1–8. https://hal.archives-ouvertes.fr/hal-02495264

Desai, M. (2019). Early detection and prevention of chronic kidney disease. *Proceedings - 2019 5th International Conference on Computing, Communication Control and Automation, ICCUBEA 2019*. https://doi.org/10.1109/ICCUBEA47591.2019.9128424

Iliyas, I. I., Rambo Saidu, I., Dauda, A. B., & Tasiu, S. (2020). *Prediction of Chronic Kidney Disease Using Deep Neural Network*.

Imesh Udara Ekanayeke and Damayanthi Herath(2020). Chronic Kidney Disease prediction using machine learning method. *ResearchGate*. Https/: DOI: 10.1109/MERCon50084.2020.9185249

Levey, A. S., Cattran, D., Friedman, A., Miller, W. G., Sedor, J., Tuttle, K., Kasiske, B., & Hostetter, T. (2009). Proteinuria as a Surrogate Outcome in CKD: Report of a Scientific Workshop Sponsored by the National Kidney Foundation and the US Food and Drug Administration. *American Journal of Kidney Diseases*, *54*(2), 205–226.

https://doi.org/10.1053/j.ajkd.2009.04.029

Meghana L, H., Kuber, V. S., S, Y. B., L, V. T., & M, V. B. (2021). *Chronic Kidney Disease Prediction using Neural Network and ML Models*. www.ijert.org

Maithili Desai .(2019) Early Detection and Prevention of Chronic Kidney Disease. *International conference of computing communication control and automation,* University of Rochester.

Mtsweni, E. S., Hörne, T., Poll, J. A. van der, Rosli, M., Tempero, E., Luxton-reilly, A., Sukhoo, A., Barnard, A., M. Eloff, M., A. Van Der Poll, J., Motah, M., Boyatzis, R. E., Kusumasari, T. F., Trilaksono, B. R., Nur Aisha, A., Fitria, -, Moustroufas, E., Stamelos, I., Angelis, L., … Khan, A. I. (2020). No 主観的健康感を中心とした在宅高齢者における健康関連指標に関する共分散構造分析 Title. *Engineering, Construction and Architectural Management*, *25*(1), 1–9. http://dx.doi.org/10.1016/j.jss.2014.12.010%0Ahttp://dx.doi.org/10.1016/j.sbspro.2013.03.034%0Ahttps://www.iiste.org/Journals/index.php/JPID/article/viewFile/19288/19711%0Ahttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.678.6911&rep=rep1&type=pdf%0Ahtt

Mula-Abed, W. A. S. (2012). Estimated Glomerular Filtration Rate (eGFR): A serum creatinine-based test for the detection of chronic kidney disease and its impact on clinical practice. *Oman Medical Journal*, *27*(4), 339–340. https://doi.org/10.5001/omj.2012.87

Nandhini, G., & Aravinth, J. (2021). Chronic kidney disease prediction using machine learning techniques. *2021 6th International Conference on Recent Trends on Electronics, Information, Communication and Technology, RTEICT 2021*, *September*, 227–232. https://doi.org/10.1109/RTEICT52294.2021.9573971

Nishat, M., Faisal, F., Dip, R., Nasrullah, S., Ahsan, R., Shikder, F., Asif, M., & Hoque, M. (2018). A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, 170671. https://doi.org/10.4108/eai.13-8-2021.170671

Pal, M., & Parija, S. (2021). Prediction of Heart Diseases using Random Forest. *Journal of Physics: Conference Series*, *1817*(1). https://doi.org/10.1088/1742-6596/1817/1/012009

Priya, S., S.Nirmal, K., G.Sibi, S., E., P., & Ashuthosh, K. P. (2020). *Chronic Kidney Disease Prediction Using Data Mining Algorithms*. 92–111. https://doi.org/10.4018/978-1-7998-2742-9.ch006

Shaji, S., Ajina, S. R., & Priya R, V. S. (n.d.). *Chronic Kidney Disease Prediction using Machine Learning*. www.ijert.org

Ulasi, I. I., & Ijoma, C. K. (2010). The enormity of chronic kidney disease in Nigeria: The situation in a teaching hospital in south-east Nigeria. *Journal of Tropical Medicine*, *2010*. https://doi.org/10.1155/2010/501957