IJCERT

International Journal of Computer Engineering in Research Trends

Multidisciplinary, Open Access, Peer-Reviewed and fully refereed

Research Paper Volume-7, Issue-10, 2020 Regular Edition E-ISSN: 2349-7084

Diagnosis of Chronic Kidney Disease using Naïve Bayes algorithm Supported by Stage Prediction using eGFR

A Victor Ikechukwu[1*], K Nivedha[2], N M Prakruthi[3], Farheen Fathima[4], R Harini[5], L Shamitha[6]

[1, 2, 3, 4, &5]*Department of CSE, Maharaja Institute of Technology Mysore, Visvesvaraya Technological University, Belagavi, India*

E-mail: [1]victora@mitmysore.in, [2]nivedhakjan98@gmail.com, [3]prakruthi.nm1411@gmail.com, [4]farheenfathima0408@gmail.com, [5]harini6498@gmail.com

**Corresponding Author: [1]victora@mitmysore.in*

Available online at: http://www.ijcert.org

Received: 15/10/2020 Revised: 19/10/2020 Accepted: 31/10/2020 Published: 12/11/2020

Abstract: The proliferation of data and availability of open source tools has simplified the diagnosis of diseases such as CKD (Chronic Kidney Disease). As one of the types of kidney disease which results in malfunctioning of kidney, it is paramount to effectively diagnose such diseases to prevent degeneration of vital organs in the body. Despite the advancements in the field of medical imaging, there exists no permanent cure for CKD, but the risk can be mitigated to a larger extent if detected at the early stage. This paper proposes a hybrid approach to early detection of chronic kidney disease by using Naïve Bayes classifier and eGFR (estimated Glomerular Filtration Rate). Naïve Bayes which works on the principle of conditional probability was used to predict whether a patient has CKD or not based on clinical symptoms, and the stage was determined using the eGFR formula. Results were promising as the model was able to predict the prevalence of CKD as well as the stage in which the patient was in. Although we were able to develop a web-based application using machine learning algorithms to aid in the diagnosis of CKD by serving as a "self-diagnostic" tool for medical practitioners, improvements could be made to ensure that the model works according to established ground truth by nephrologists.

Keywords: Naïve Bayes, Random Forest, eGFR, CKD, Medical Diagnosis

1. Introduction

CKD (Chronic Kidney Disease) refers to gradual kidney damage that makes it difficult to filter blood the way they should. According to a recent survey, 10% of the world's population suffer and die due to chronic kidney disease and several millions die each year because doctors are unable to diagnose the disease at the later stages. In India for example, it is estimated that around 800 thousand in a million population has CKD and the incidence of end stage renal disease is around 150-200 thousand, thus it is difficult to state categorically the number of persons affected with CKD. Despite that automation has taken over almost all aspects of medicine, there exists manual treatments and diagnoses of CKD by specialist doctors. So, in order to overcome these challenges we have proposed a system that predicts whether a person is suffering from CKD or otherwise using one of the popular machine learning classification algorithm, i.e. Naïve Bayes algorithm which is

https://doi.org/10.22362/ijcert/2020/v7/i10/v7i1002

supported by predicting the stage using Estimated Glomerular Filtration Rate (EGFR) formula and the proposed system aimed to reduce the time taken to diagnose the prevalence of CKD in patients which subsequently reduces the risk of people entering into End Stage Renal Disease (ESRD) failure. To ensure consistencies in the prediction accuracy, few specialist doctors were contacted and following objectives were agreed upon:

Objectives:

- To survey existing approaches to the diagnosis of CKD in patients above 60 years of age.
- To design a methodology to compute eGFR and compare it with preliminary works carried out by other scholars.
- To develop a web-based portal that will serve as a clinical decision support system.

The paper is organized as follows; Section I contains the introduction of the proposed approach, Section II contains work related to the diagnosis of chronic diseases using machine learning approaches, Section III contain the methodology and relevant figures, Section IV contain the result and discussion and section V concludes the research work with future directions.

2. Related Work

Several researches have been carried out in tandem with medical practitioners to the early diagnosis of diseases using clinical symptoms and most of them are found in the literature below:-

Hanyu Zhang et al. [1] investigated the performance of Artificial Neural Network (ANN) model while applying the survival expectations on CKD patients. The authors also discussed that the middle endurance time of conclusive stage patients was around 3 years. To aid provide a solution to early diagnosis, the author used two artificial neural network, one of which was a classical Multi-Layer Perceptrons, while the other an integrated LASSO feature selection. Reem A. Alassaf et al. [2] proposed a solution using data mining and supervised machine learning approaches such as ANN, KNN and SVM. The use of recursive feat elimination for feature selection backed up by correlation coefficient was used in selecting appropriate features for the model and it turns out that ANN, SVM and Naïve Bays achieved a testing accuracy of 98% as against KNN which has a precision of 93.9%. Tahira Mahboob et al. [3] proposed a clustering algorithms such as KNN, K-Means and K-Medoids to predict the missing values. To predict the performance of the proposed approach, decision tree and random forest algorithm was used. The results show that the accuracy of KNN and decision tree was 86.67%, and 75.25% for random forest algorithm.

In a similar paper, Dilip Singh Sisodia et al., [4] studied the performance of individual and ensemble learners on CKD patients' dataset sourced from UCI machine learning repository. They evaluated the performance using recall, accuracy, precision values as metrics for comparing performances using the freely available data mining software called WEKA and concluded that selected methods were more accurate to predict the disease.

Veenita Kunwar et al., [5] used ANN and Naïve Bayes in which the former achieved an accuracy of 72.7% and the later an accuracy of 100%. However, the author failed to investigate if the model was overfitted as less number of datasets were used and these were implemented in readily available online tool called Rapid Miner.

Helmie Arif Wibawa et al., [6] proposed a novel approach to CKD prediction using kernel-based extreme learning machine (ELM). Subsequently, other methods such as Linear-ELM, Polynomial-ELM and RBF-ELM were also employed in the study in which results showed that the sensitivity and specificity were 99.38% and 100% respectively, which was indeed a great achievement.

Guneet Kaur et al., [7] used data mining classifiers to predict if the person was suffering from chronic kidney disease. The two techniques used were KNN (K-Nearest Neighbour) and SVM (Support vector Machines). Analysis showed that SVM gave around 78.1% as compared to KNN which was only 70%, the author further compared both accuracies on select datasets and it turned out that the total error was lesser in SVM as compared to KNN, again validated that SVM was a better classifier.

Gunaranthe W.H.S.D et al., [8] carried out a work to predict the patient's status of CKD using the dataset downloaded from UCI repository. The records were trained on various multiclass decision forest algorithm and an accuracy of 99.1% was obtained after the features were reduced to 14 attributes, thus led to the conclusion that the multiclass decision forest algorithm was best because of higher accuracy.

https://doi.org/10.22362/ijcert/2020/v7/i10/v7i1002

M.P.N.M Wickramasinghe et al., [9] investigated factors that led to CKD in patients by considering the level of potassium in their blood using machine learning techniques. Their aim was to develop a healthy and suitable diet plan for CKD patients using machine learning classification algorithms such as multiclass decision jungle, multiclass decision forest, multiclass neural network and multiclass logistic regression and the results obtained showed that multiclass decision forest algorithm returned the highest accuracy of 99.17%.

Devika R et al., [10] investigated CKD prediction using Naïve Bayes and KNN. Although the performance of random forest classifier was better than Naive Bayes and KNN, the authors opined that new classifiers should be used as in the case of destiny paintings to arrive at higher solutions.

Anusorn Charleonnan et al., [11] applied different machine learning algorithms such as SVM, logistic regression and decision tree classifier to the diagnosis of CKD. Experimental results showed that SVM was the best classifier for predicting Chronic Kidney Disease. Lambodar Jena et al., [12] explored another technique to predict disease accuracy using Naïve Bayes and multilayer perception algorithms. Analysis was done using WEKA; an open-source data mining tool developed by the University of Waikato in New Zealand. Engin AVCI et al., [13] in their work used the CKD dataset from UCI repository and applied Naïve Bayes, K-Star, Support Vector Machine and J48 classifiers. The performance was measured by the values of accuracy and precision. After all the comparisons were done, a conclusion that J48 algorithm gave the highest accuracy among all was drawn.

Anonnya Banerjee et al.[14], carried a research in which food was recommended for CKD patients using machine leaning techniques. Their author had identified five stages, ranging from stage 0 to stage 5. If a person was found to be below stage 2, there is a lower risk of the kidney being completely damaged. Based on the stages, respective food was recommended and analysis was done again using WEKA software.

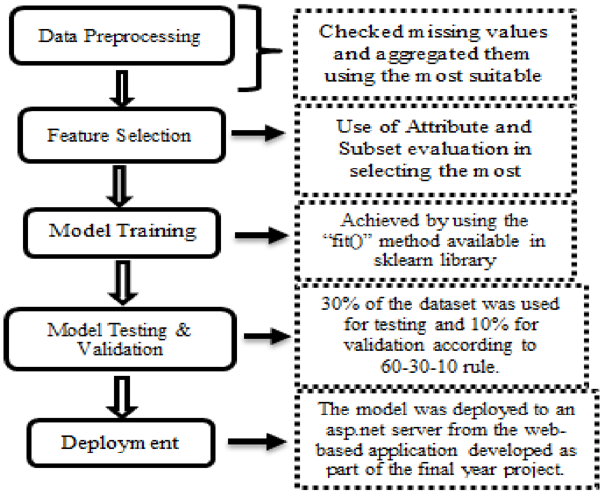
Arif-UI-Islam et al., [15] used Boosting classifiers such as Ada Boost and Logit Boost, and implemented the same in Ant Miner. The research was from a two-fold perspective: firstly, breaking down the exhibition of boosting calculations for distinguishing CKD and secondly, determining rules outlining relationship among the characteristics of CKD. The general results showed that Ada Boost performance was less as compared to Logit Boost by a fraction.

Bilal Khan et al., [16] in their work on empirical evaluation of CKD proposed numerous procedures and models. They utilized experiential examination of machine learning techniques for arranging the kidney patient datasets as CKD or none CKD. The authors then applied several machine learning techniques such as Naïve Bayes, J48, SVM and Composite Hypercube on Iterated Random Projection (CHIRP). The result was promising as it showed that CHIRP performed well in reducing error rates and improved accuracy. Another work that mirrors how machine learning can be applied in diagnosis of heart disease was investigated by Victor Agughasi et al., [17] in which clinical symptoms from patients dataset was used in the diagnosis of chronic heart failure. The author implemented the same using K-Means and Naïve Bayes classifiers and it turned out that Naïve Bayes classifier gave a better accuracy of 90% as compared to K-Means.

It can be observed that numerous research work has been done using Naïve Bayes and other machine learning algorithms based on open source tools. However, most medical practitioners would prefer a personalized decision support system that is tailored towards the needs of the patients and that formed the backbone of our proposed work.

3. Methodology

The process starts by getting the dataset from UCI machine learning repository, followed by feature selection and finally the application of selected supervised machine learning approaches. This can be explained better in a block diagram:


Data Preprocessing
Checked missing values and aggregated them using the most suitable
Feature Selection
Use of Attribute and Subset evaluation in selecting the most
Model Training
Achieved by using the "fit()" method available in sklearn library
Model Testing & Validation
30% of the dataset was used for testing and 10% for validation according to 60-30-10 rule.
Deployment
The model was deployed to an asp.net server from the web-based application developed as part of the final year project.

https://doi.org/10.22362/ijcert/2020/v7/i10/v7i1002

4. Results and Discussion

The dataset used in this study was obtained from the UCI repository (https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease) containing around 400 instances and 25 attributes (with the 25th column being the predictor class), and analysis was done using Naive Bayes (NB) algorithm supported by stage prediction using eGFR. The following information is given about the classifier: As a classification technique based on Bayes' Theorem with an assumption of independence among predictors, Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature and owing to its simplicity, it is known to outperform even the most sophisticated classification methods. According to the equation below:

$$P(S_i|X) = \frac{P(X|S_i) * P(S_i)}{P(X)} \qquad (i)$$

Where:

$P(S_i|X) =$ The probability of occurrence of the event, *Si* when event *X* occurs

$P(X|S_i) =$ The probability of occurrence of the event, *X* when the event *Si* occurs.

$P(S_i), P(X) =$ The prior probability of events *Si* and *X* respectively.

Furhermore, the estimated Glomerular Filtrate Rate (eGFR) was calculated using the formula:

$$186 * \left(\frac{Creatine}{88.4}\right)^{-1.154} * (Age)^{-0.203} * (0.742\ if\ female) * (1.210\ if\ black) \qquad (ii)$$

The above equation is commonly used by nephrologists in local clinical laboratories, however, another useful formula employed in the project is as shown beow:

$$eGFR = (140 - age) * \frac{weight(kg)}{72 * P_{cr}} \qquad (iii)$$

Where:

$P_{Cr} =$ The creatinine level in *mg/dl*

Weight = Weight of the individual in *kg*

For instance, the eGFR value for a male of 57 years, PCr value of 3.36 was approximately 19 as shown in Figure 2. It is worthy to mention that other authors have considered "race" as a determining factor in which the value changes from 19 to 24 for black races. Confusion matrix was used to validate the performance of the system by computing the accuracy, precision, sensitivity and F-Score:

Table I: Confussion Matrix

		Predicted		
		Yes	***No***	
Actual	***Yes***	TP	FN	TP + FN
	No	FP	TN	FP + TN
	Total	**TP + FP**	**FN + TN**	

Table II: Naïve Bayes Confussion Matrix

		Predicted	
		CKD	***NonCKD***
Actual	***CKD***	346	10
	NonCKD	22	22
	Total	**368**	**32**

a) The accuracy of the model shows the total classification ratio and calculated using the formula below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$= \frac{346 + 22}{346 + 22 + 22 + 10} = \frac{368}{400} = 0.92 => 92\%$$

b) Precision determined the positive features of the whole classification algorithm, and expressed as:

$$Precision = \frac{TP}{TP + FP} = \frac{346}{368} = 0.94 => 94\%$$

c) Sensitivity (also known as recall), refers to correctly classified positive and falsely classified negative values, viz:

$$Sensitivity = \frac{TP}{TP+FN} = \frac{346}{346+10} = \frac{346}{356} = 0.97 => 97\%$$

https://doi.org/10.22362/ijcert/2020/v7/i10/v7i1002

d) F-Measure was used to calculate the weighted average of both the sensitivity and precision as shown below:

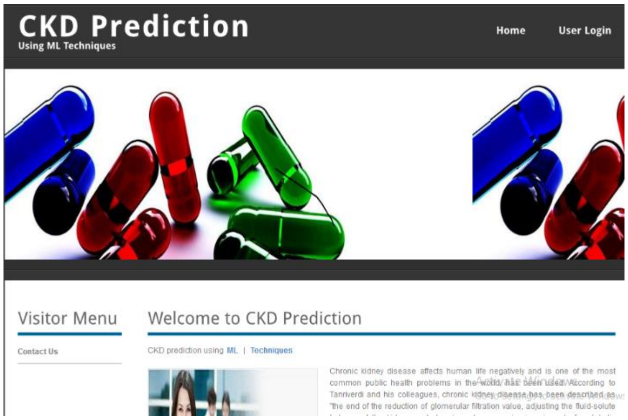
$$F - Measure = \frac{2 * Sensitivity * Precision}{Sensitivity + Precision}$$

$$= \frac{2 * 0.97 * 0.94}{0.97 + 0.94} = \frac{1.824}{1.91} = 0.95 => 95\%$$

TABLE III: Summary of the Performance Metrics used

Algorithm Used	Performance Measures			
	Accuracy	Precision	Sensitivity	F-Measure
Naïve Bayes	92%	94%	97%	95%

Furthermore, a web-based application was developed to facilitate model deployment as shown below:



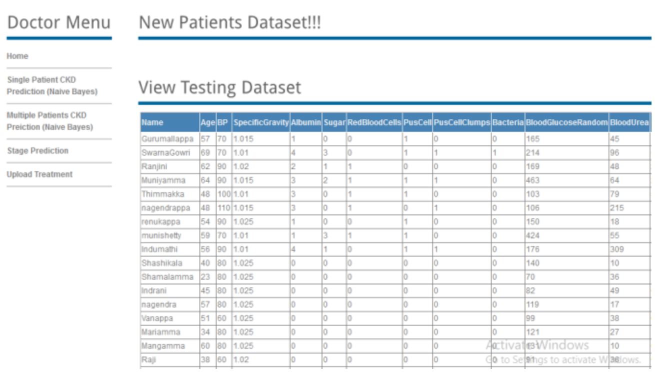
Welcome to CKD Prediction

CKD prediction using ML | Techniques

Chronic kidney disease affects human life negatively and is one of the most common public health problems in the world has been used. According to Tanriverdi and his colleagues, chronic kidney disease has been described as "the end of the reduction of glomerular filtration value, adjusting the fluid-solute

Figure 1: Homepage of the application

The dataset was split into training and testing data which is visible to the doctor to ensure inconsistencies were reduced to the barest minimum



Doctor Menu
Home
Single Patient CKD Prediction (Naive Bayes)
Multiple Patients CKD Preiction (Naive Bayes)
Stage Prediction
Upload Treatment

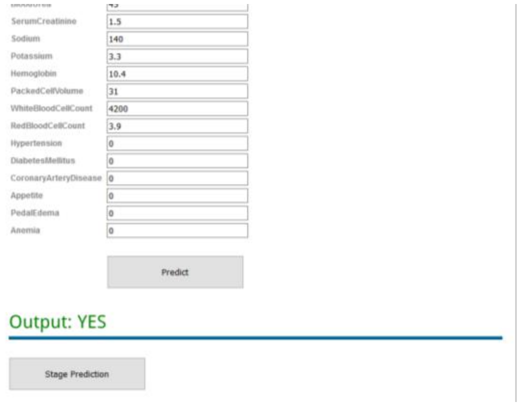
New Patients Dataset!!!

View Testing Dataset

Name	Age	BP	SpecificGravity	Albumin	Sugar	RedBloodCells	PusCell	PusCellClumps	Bacteria	BloodGlucoseRandom	BloodUrea
Gurumallappa	57	70	1.015	1	0	0	1	0	0	165	45
SwarnaGowri	69	70	1.01	4	3	0	1	1	1	214	96
Ranjini	62	90	1.02	2	1	1	0	0	0	169	48
Muniyamma	64	90	1.015	3	2	1	1	1	0	463	64
Thimmakka	48	100	1.01	3	0	1	1	0	0	103	79
nagendrappa	48	110	1.015	3	0	1	0	1	0	106	215
renukappa	54	90	1.025	1	0	0	1	0	0	150	18
munishetty	59	70	1.01	1	3	1	1	0	0	424	55
Indumathi	56	90	1.01	4	1	0	1	1	0	176	309
Shashikala	40	80	1.025	0	0	0	0	0	0	140	10
Shamalamma	23	80	1.025	0	0	0	0	0	0	70	36
Indrani	45	80	1.025	0	0	0	0	0	0	82	49
nagendra	57	80	1.025	0	0	0	0	0	0	119	17
Vanappa	51	60	1.025	0	0	0	0	0	0	99	38
Mariamma	34	80	1.025	0	0	0	0	0	0	121	27
Mangamma	60	80	1.025	0	0	0	0	0	0	131	10
Raji	38	60	1.02	0	0	0	0	0	0	91	36

Figure 2: Test Dataset with prominent features

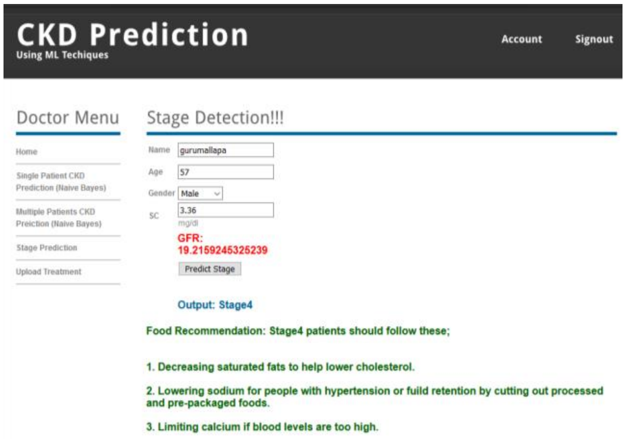
After the doctor validates the dataset and clicks on "Predict" button, the model then predicts the presence or absence of CKD, followed by the stage prediction using eGFR.



SerumCreatinine	1.5
Sodium	140
Potassium	3.3
Hemoglobin	10.4
PackedCellVolume	31
WhiteBloodCellCount	4200
RedBloodCellCount	3.9
Hypertension	0
DiabetesMellitus	0
CoronaryArteryDisease	0
Appetite	0
PedalEdema	0
Anemia	0

Output: YES

Figure 3: Doctors Prediction



Stage Detection!!!

Name: gurumallapa
Age: 57
Gender: Male
SC: 3.36 mg/dl

GFR: 19.2159245325239

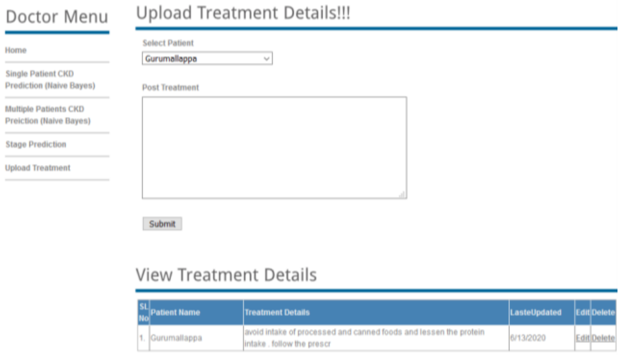
Output: Stage4

Food Recommendation: Stage4 patients should follow these;

1. Decreasing saturated fats to help lower cholesterol.
2. Lowering sodium for people with hypertension or fuild retention by cutting out processed and pre-packaged foods.
3. Limiting calcium if blood levels are too high.

Figure 4: Stage Prediction

Finally, appropriate treatment details would be presented to the patient with the help of the model.


Doctor Menu
Home
Single Patient CKD Prediction (Naive Bayes)
Multiple Patients CKD Preiction (Naive Bayes)
Stage Prediction
Upload Treatment
Submit

Upload Treatment Details!!!

Select Patient: Gurumallappa

Post Treatment

View Treatment Details

SL No	Patient Name	Treatment Details	LasteUpdated	Edit	Delete
1.	Gurumallappa	avoid intake of processed and canned foods and lessen the protein intake . follow the prescr	6/13/2020	Edit	Delete

Figure 5: Recommendation

5. Conclusion and Future Scope

This paper proposed a method for predicting whether a patient is suffering from CKD or not using Naïve Bayes algorithm with an accuracy of 92%. An attempt was made at predicting the stage in which the patient is suffering from, using the eGFR method with greater accuracy. The result was promising as it will help doctors in predicting the stage in which the patient is suffering, with the option to recommend foods that should be taken, and which should be avoided if need be.

As with every medical research, the application has to be further enhanced based on the ground truth recommended by leading medical practitioners to ensure that our application worked as intended and detailed study of other ensemble methods will be investigated further to improve the overall accuracy of the system.

6. References

[1] H. Zhang, C. L. Hung, W. C. C. Chu, P. F. Chiu, and C. Y. Tang, "Chronic Kidney Disease Survival Prediction with Artificial Neural Networks," Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018, pp. 1351–1356, 2019.

[2] R. A. Alassaf et al., "Preemptive Diagnosis of Chronic Kidney Disease Using Machine Learning Techniques," Proc. 2018 13th Int. Conf. Innov. Inf. Technol. IIT 2018, pp. 99–104, 2019.

[3] T. Mahboob, A. Ijaz, A. Shahzad, and M. Kalsoom, "Handling Missing Values in Chronic Kidney Disease Datasets Using KNN, K-Means and K-Medoids Algorithms," ICOSST 2018 - 2018 Int. Conf. Open Source Syst. Technol. Proc., pp. 76–81, 2019.

[4] D. S. Sisodia and A. Verma, "Prediction performance of individual and ensemble learners for chronic kidney disease," Proc. Int. Conf. Inven. Comput. Informatics, ICICI 2017, no. Icici, pp. 1027–1031, 2018.

[5] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," Proc. 2016 6th Int. Conf. - Cloud Syst. Big Data Eng. Conflu. 2016, pp. 300–305, 2016.

[6] H. A. Wibawa, I. Malik, and N. Bahtiar, "Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease," 2018 2nd Int. Conf. Informatics Comput. Sci. ICICoS 2018, no. x, pp. 33–36, 2019.

[7] G. Kaur and A. Sharma, "Mining Algorithms In Hadoop," no. Icici, 2017.

[8] G. W.H.S.D, "Performance Evaluation on Machine Learning Classification Techniques for Disease (CKD)," Ieee, pp. 291–296, 2017.

[9] M. P. N. M. Wickramasinghe, D. M. Perera, and K. A. D. C. P. Kahandawaarachchi, "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," 2017 IEEE Life Sci. Conf. LSC 2017, vol. 2018-Janua, pp. 300–303, 2018.

[10] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, "Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest," Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. Iccmc, pp. 679–684, 2019.

[11] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," 2016 Manag. Innov. Technol. Int. Conf. MITiCON 2016, pp. MIT80–MIT83, 2017.

[12] L. Jena and R. Swain, "Work-in-Progress: Chronic Disease Risk Prediction Using Distributed Machine Learning Classifiers," Proc. - 2017 Int. Conf. Inf. Technol. ICIT 2017, pp. 170–173, 2018.

[13] E. Avci, S. Karakus, O. Ozmen, and D. Avci, "Performance comparison of some classifiers on Chronic Kidney Disease data," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018-Janua, pp. 1–4, 2018.

[14] A. Banerjee, A. Noor, N. Siddiqua, and M. N. Uddin, "Food Recommendation using Machine Learning for Chronic Kidney Disease Patients," 2019 Int. Conf. Comput. Commun. Informatics, ICCCI 2019, pp. 1–5, 2019.

[15] Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019, pp. 7–9, 2019.

[16] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," IEEE Access, vol. 8, pp. 55012–55022, 2020.

[17] V. I. Agughasi, Y. Dk, and S. Das M, "Early Prognosis of Heart Failure from Clinical Symptoms using K-Means and Naïve Bayes Algorithms," vol. 9, no. 7, pp. 55–61, 2020.

Authors Profile

Victor Ikechukwu A., is a research scholar with interest in Medical Imaging and Deep Learning who works in proximity with clinical experts to understand and develop prognostic tools that will assist doctors in early diagnosis of chronic medical conditions.