Chronic Kidney Disease Prediction in Databricks (Apache Spark) using ML Classifiers

Muhammad Sadiq
Department of Computer Engineering
Baluchistan University of Information Technology, Engineering and Management Sciences (BUITEMS)
Quetta, Pakistan
engrmuhammadsadiq@outlook.com

Sibghat Ullah Bazai
Department of Computer Engineering
Baluchistan University of Information Technology, Engineering and Management Sciences (BUITEMS)
Quetta, Pakistan
sibghat.ullah@buitms.edu.pk

***Abstract*— Despite the strong isolation and increasing prevalence of cardiovascular disease, chronic kidney disease (CKD) is a serious life-threatening condition caused by kidney cancer or kidney failure, and is usually Dialysis or surgery is needed. Early recognition and appropriate treatment are essential to achieve better outcomes. This research examines various machine learning techniques for CKD diagnosis, using big data platforms such as Apache Spark. Hybrid machine learning methods, such as Chi-Squad and Decision Tree, use feature selection and embedding algorithms, proven for accurate diagnostics. This research reduces the initial 25 variables to a subset, identifying the optimal parameters for the diagnosis of CKD. Ten machine learning classifiers were tested, with XgBoost having the highest performance indicators of 0.98 in each case, including accuracy, precision, recall, and F-1 score. The research suggests interesting avenues of action in machine learning and predictive modeling, which are interesting for finding new solutions in predictive accuracy in the fields of kidney disease and beyond.**

Keywords—Chronic Kidney Disease, Machine learning

XgBoost classifier, Classification model.

I. INTRODUCTION

Chronic kidney disease or CKD is a condition in which the kidneys become so damaged that they cannot filter the blood properly. The main function of the kidneys is to remove waste and excess water from the blood, forming urine. CKD means that waste has accumulated in the body. This condition is "further" because the damage increases late. It is a disease that affects people all over the world. CKD can cause various problems with your health. I have observed that various health problems such as diabetes, high blood pressure, and heart disease are just one of three that can be transferred to CKD.

There are usually no symptoms in the early stages of CKD. This is due to the fact that our body normally has the characteristic of completely moderating things that increase in kidney function. CKD is usually not recognized until this stage or if a routine test for another problem, such as a blood or urine test, reveals a possible problem. If it is detected at an early stage, ongoing treatment with medication and routine tests can prevent it from progressing to a more serious condition.

If kidney disease is not caught early or continues to progress after treatment, several symptoms may occur. Kidney failure is the final stage of CKD. It is also called acute renal disease or acute renal failure. At this point dialysis or a kidney transplant may be required. If you have signs or symptoms of kidney disease, you should make an appointment with your doctor. Kidney disease can be caught early if caught early. Over the past two decades, the era of big data has dawned, where digital data is playing an important role in science, healthcare, technology, and society. Assets arise from data volumes, sensor networks and mobile applications, especially in healthcare, and dealing with these different types of data is a challenge without advanced technology.

The challenges that big data presents in healthcare require an advanced processing paradigm. Traditional database management systems struggle with ever-increasing amounts of data. Research offers solutions to such problems, such as Apache Spark, Apache Hadoop, Apache Kafka, and Apache Storm, which focus on solving large and heterogeneous data storage problems, on solving healthcare problems. Addition: Chronic Kidney Disease (CKD) is of intense interest due to its high mortality rate, which is a major threat to new countries. In 2016, CKD claimed the lives of 753 million people worldwide, emphasizing the need for global governments to promote early recognition and treatment. Intelligent machine learning and deep learning, in particular, help in healthcare decision-making, accurate diagnosis, and disease prediction. CKD has generated strong interest in half-breeding, making it a major threat to new countries. In 2016, CKD claimed 753 million lives worldwide, emphasizing the need for global governments in early recognition and treatment. Intelligent machine learning and deep learning, in particular, help in healthcare decision-making, accurate diagnosis, and disease prediction.

Machine learning is the science that studies large volumes of data that contain various variables. Machine learning is mainly formed from the theory of perceptual evaluation and learning-on-the-go. A vital role in helping medical professionals and doctors make accurate and accurate diagnoses, choose the best drugs for patients, recognize increased patient risks, and improve patient recovery at the lowest possible cost. has paid ML has shown prominent roles in various applications, such as speech recognition [13],

computer vision [14], medical diagnosis [15], and engineering [16].

II. Litrerature Review

Several authors have used various machine learning (ML) techniques to diagnose and predict toxic kidney disease, as shown in Table 1. For example, the authors in [27] presented a hybrid model that combines logistic regression (LR) and random forest (RF) to predict CKD disease. They compared their proposed model with six machine learning algorithms (LR, RF, SVM, KNN, Naive Bayes (NB), and Feedforward Neural Network (FNN)), and their model achieved the highest accuracy of 99.83%. Recorded. In [29], NB, K-Star, SVM, and J48 classifiers were used to predict CKD, with the J48 algorithm performing better with 99% accuracy. Some authors have combined ML algorithms with feature selection techniques to help predict CKD. In [22], recursive feature elimination (RFE) synthesis was used, and four classification algorithms (SVM, KNN, DT, and RF) were used in both. The results showed that RF outperformed other algorithms. In [20], the authors used chi-square, CFS, and lasso feature selection to select essential features from the database. They used ANN, C5.0, LR, LSVM, KNN, and RF in both. In [23], five feature selection methods (Random Forest feature selection (RF-FS), Forward Selection (FS), Forward Exhaustion Selection (FES), Backward Selection (BS), and Backward Exhaustion (BE)) were used in the database. Selected the most important features from Four ML algorithms, RF, SVM.

Data mining is the process of using specialized software to discover hidden information in large volumes of data. Data mining techniques are interconnected and are used in a wide range of contexts and situations. Through data mining technology we can make predictions, sort data, filter it, and group it. The purpose of the algorithm is to process a training set that contains a set of features and targets, and the plan describes how this should be done. If the data set is very large, data mining is a good way to find patterns in it. However, if the data set is very small, we can reach the same goal with the help of machine learning. Data analysis and pattern recognition are two more capabilities of machine learning. 1,16 Because health data sets have a wide variety, machine learning algorithms are the most suitable method to increase the accuracy of diagnosis. 17,19 Health data Machine learning algorithms are gaining popularity in mining health data sets as a result of the rapidly increasing speed of datasets.9

Information mining techniques have been used in a variety of studies to extract useful information from toxic kidney disease datasets.22 This was done to save more time in analysis, and in addition, With the help of information mining method, the accuracy of prediction would have increased.5 Data mining is also used in the treatment and diagnosis of various diseases and conditions. A variety of work has been done to extract useful information from data sets using information gathering techniques.

III. Methodology

CKD dataset

This research uses the dataset CKD from the UCI Machine Learning Repository11. The CKD dataset contains a total of 24 features and 1 target variable. It can be divided into two categories, yes or no. The data set contains a total of 25 features, of which 11 are numeric and 14 are non-numeric. Machine learning algorithms use the entire data set of 400 samples to make predictions. Of the total 400 quantifications, 250 were classified as having CKD, and the remaining 150 were classified as non-CKD. The features in the data set are shown in Figure 1.

The results of each classifier were tested with different evaluation parameters, and 10-fold cross-validation was used to check the over-fitting of the results. In addition, the nested cross-validation method continued to help improve the basic parameters of the model. The analysis was performed using the Jupyter Notebook web tool and the Python 3.3 programming language, and several Scikit-learn libraries, a free platform for machine learning systems based on the Python programming language, were used.

This analysis included measures of sensitivity, specificity, area under the curve (AUC), and accuracy as measured by the F1 measure. Each model is based on the nature of its parameters, and each model produces different results.

To analyze the CKD data set, different types of machine learning algorithms, such as SVM, KNN, LGBM, and hybrids, were used. In this study, Figure 2 shows the comprehensive structure of the CKD diagnosis process. During the preprocessing step, the mean technique was used to impute missing numerical values, while the mode approach was used to impute missing nominal values. These two techniques combined are called mode method. Recursive feature elimination (RFE) and principal component analysis (PCA) algorithms were used to select features associated with important features for the diagnosis of CKD. These selected features are provided to the disease classifiers to make accurate diagnoses.

	age	blood_pressure	specific_gravity	albumin	sugar	blood_glucose_random	blood_urea	serum_creatinine	sodium	potassium	haemoglobin
count	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000	381.000000	383.000000	313.000000	312.000000	348.000000
mean	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517	57.425722	3.072454	137.528754	4.627244	12.526437
std	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714	50.503006	5.741126	10.408752	3.193904	2.912587
min	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000	1.500000	0.400000	4.500000	2.500000	3.100000
25%	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000	27.000000	0.900000	135.000000	3.800000	10.300000
50%	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000	42.000000	1.300000	138.000000	4.400000	12.650000
75%	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000	66.000000	2.800000	142.000000	4.900000	15.000000
max	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000	391.000000	76.000000	163.000000	47.000000	17.800000

Figure 1 Dataset Features Description

Table 1 The Description of Chronic Kidney Disease Dataset

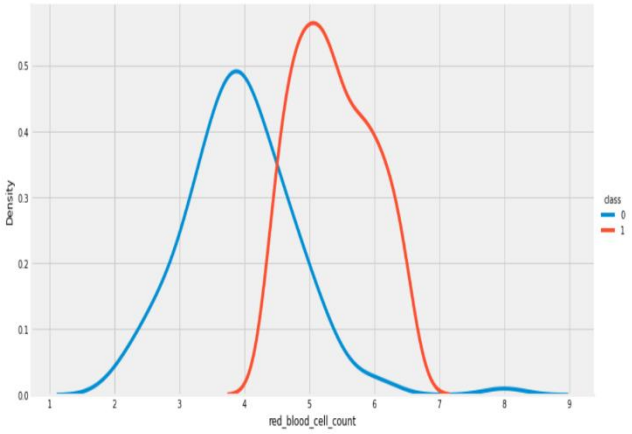
Features	*Features Meaning*	*Scale*	*Missing*	*Category*
age	Age	Years	9	Numerical
bp	Blood pressure	mm/Hg	12	Numerical
sg	Specific gravity	1.005 to 1.025	47	Categorical
al	Albumin	0 to 5	46	Categorical
su	Sugar	0 to 5	49	Categorical
rbc	Red blood cells	Abnormal, Normal	152	Categorical
pc	Pus cell	Abnormal, Normal	65	Categorical
pcc	Pus cell clumps	Not present, Present	4	Categorical
ba	Bacteria	Not present, Present	4	Categorical
bgr	Blood glucose random	mgs/dl	44	Numerical
bu	Blood urea	mgs/dl	19	Numerical
sc	Serum creatinine	mgs/dl	17	Numerical
sod	Sodium Numerical	mEq/L	87	Numerical
pot	Potassium	mEq/L	88	Numerical
hemo	Hemoglobin	gms	52	Numerical
pcv	Packed cell volume	P cv	71	Numerical
wc	White blood cell count	cells/cumm	106	Numerical
rc	Red blood cell count	millions/cmm	131	Categorical
htn	Hypertension	No, Yes	2	Categorical
dm	Diabetes mellitus	No, Yes	2	Categorical
cad	Coronary artery disease	No, Yes	2	Categorical
appet	Appetite	Poor, Good	1	Categorical
pe	Peda edema	No, Yes	1	Categorical
ane	Anemia	No, Yes	1	Categorical
classification	Class	Not CKD, CKD	0	Categorical

In this work, CKD was diagnosed using SVM, KNN, LGBM, Xg, Cat Boost, Ada, Hybrid, and several other classifiers. The presented machine learning model is expected to achieve good classification performance with a limited number of features and the best performance is achieved by using PCA.

Data preprocessing

It is important that the quality of the data is high so that reasonable performance in data mining is possible. The CKD data set must be fully populated with variables that are missing from the database. When continuous properties exist, it is possible to adapt the techniques to create discrete properties under certain conditions. Each of them has smart prices and some are missing. The process of professionally organizing the original data to restore it to health is called "data preprocessing". Through this process, the raw data is organized and made suitable for use in a machine learning model.

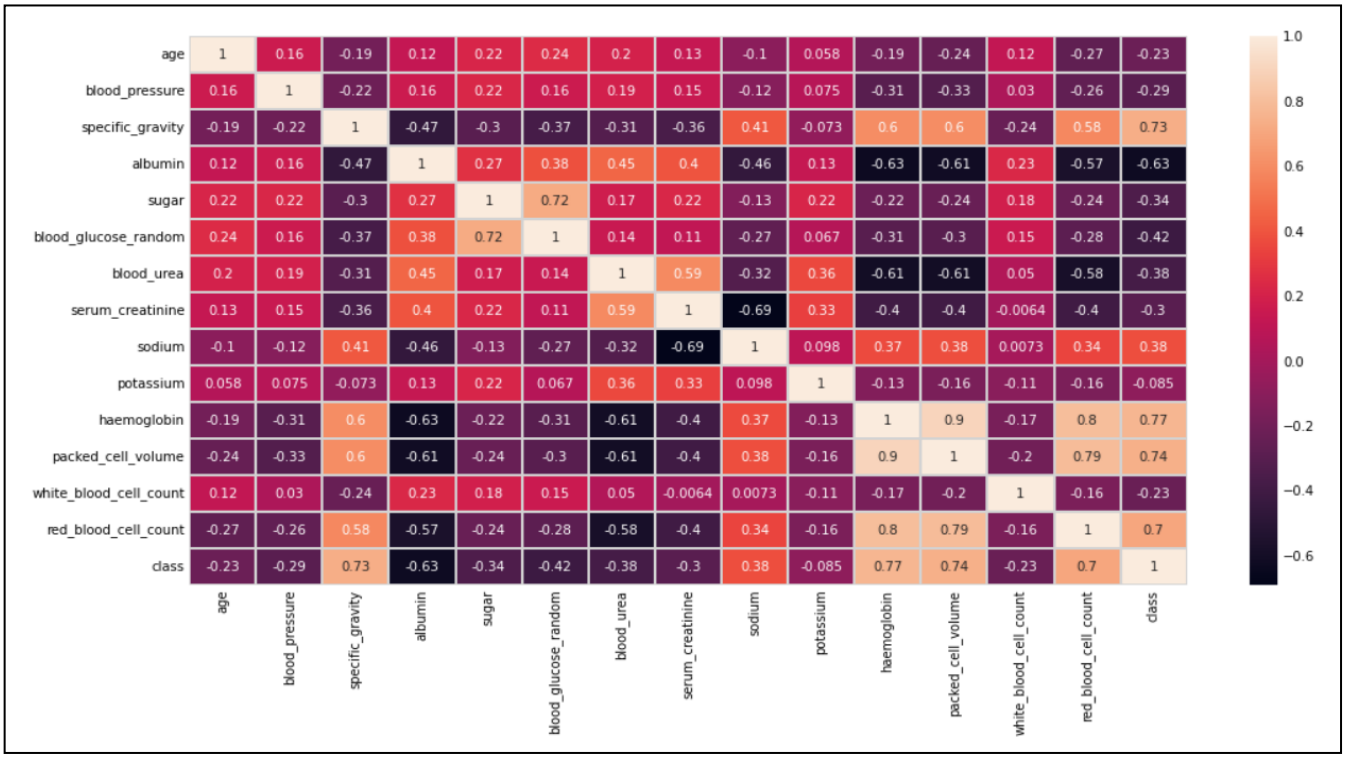
The data preprocessing steps are shown in Figure 3. It is important to consider that there are many intangibles in the data. Details of each variable are shown in Figure 1 which can be found here. 3 The datasets mentioned should be cleaned because they contain "NaNs" (undefined numbers), and numeric properties should be converted to flats. In plain language, we were instructed to remove all rows containing "NaNs", and there was no limit to this process; That is, any array that contains a single NaN should be removed.



To facilitate computer processing, each type of variable, also known as a nominal variable, is assigned a code. When it comes to rbc and pc values, the coding system has common and non-common. were expressed as the numbers 1 and 0, respectively. Pay and not pay are assigned the numbers 1 and 0 respectively for the pcc and ba values. Answers "yes" and "no" were assigned the values "1" and "0" for the variables htn, dm, cad, pe, and ane, respectively. The value of high happiness was also, fortunately, assigned values of 1 and 0 for good and poor, respectively. Although the initial data specification defines three variables - sg, al, and su - as a specific type, the values of these three variables are based on more numerical information Therefore, these variables are treated as numerical variables. Each categorical variable was converted into a factor. A non-adjacent number from 1 to 400 was assigned to each sample. A large number of values are missing in the data set, and only 158 instances are complete. Many values are missing. Usually, before a diagnosis is made, patients have few measurements for a number of reasons. Accordingly, if the diagnostic characteristics of samples are unknown and must be imputed as a probability, missing values will be included in the data and will amount to an imputation.

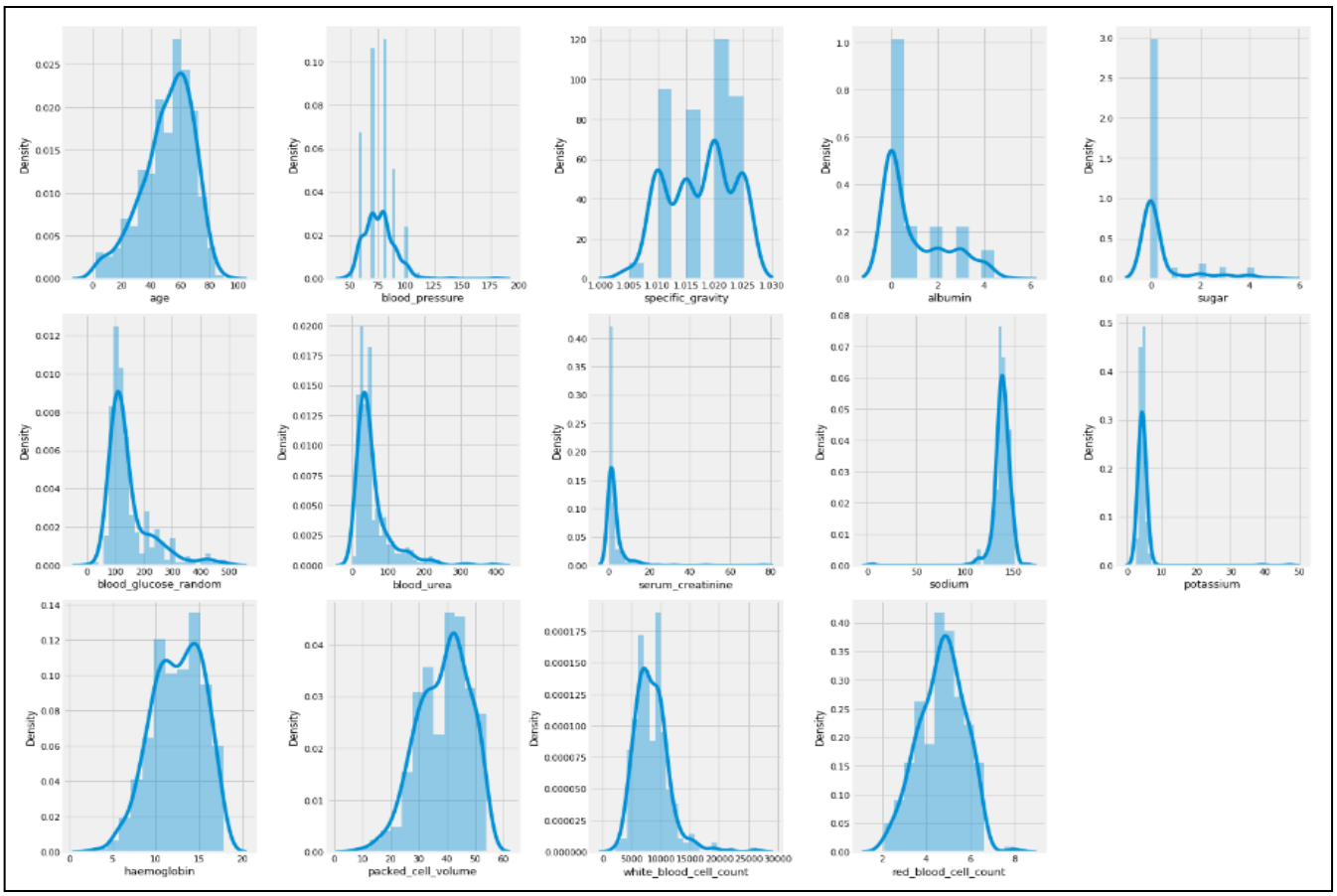
Just as the physical conditions of people with functional states are comparable, it is logical that physical disabilities are comparable. To fill in the missing numbers, a kNN-based synthesis is used, because physical disabilities are expected to be consistent physical disabilities for individuals in similar physical conditions.

For example, the physiological parameters of healthy people should be proportional within a given range. When comparing diseased individuals, the physical disabilities of the



diseased individuals should be comparable to the physical disabilities of individuals with a similar level of disease. In particular, there should be no major differences in physical disabilities between individuals whose conditions are comparable. This strategy, which has been used in the field of hyperuricemia, should also be applied to diagnostic data for other diseases.

Research was conducted using the CKD dataset. [^11] This dataset has 400 rows and 14 columns. The output column of the "class" column has a value of "yes" or "no". "yes" and "no" responses are given values of "1" and "0" respectively. A value of "1" indicates that the patient is a CKD patient, while



a value of "0" indicates that the patient is not a CKD patient. Figure 4 shows the unoptimized columns view of the data set without PCA. Figure 5 shows a PCA with unsupervised columns view of the data set where diabetes mellitus is represented in over 250 occasions that are not CKD patients and in 140 occasions that are CKD patients. In the target class distribution (Figure 6 there are approximately 250 rows of CKD patients and 150 non-CKD patients.

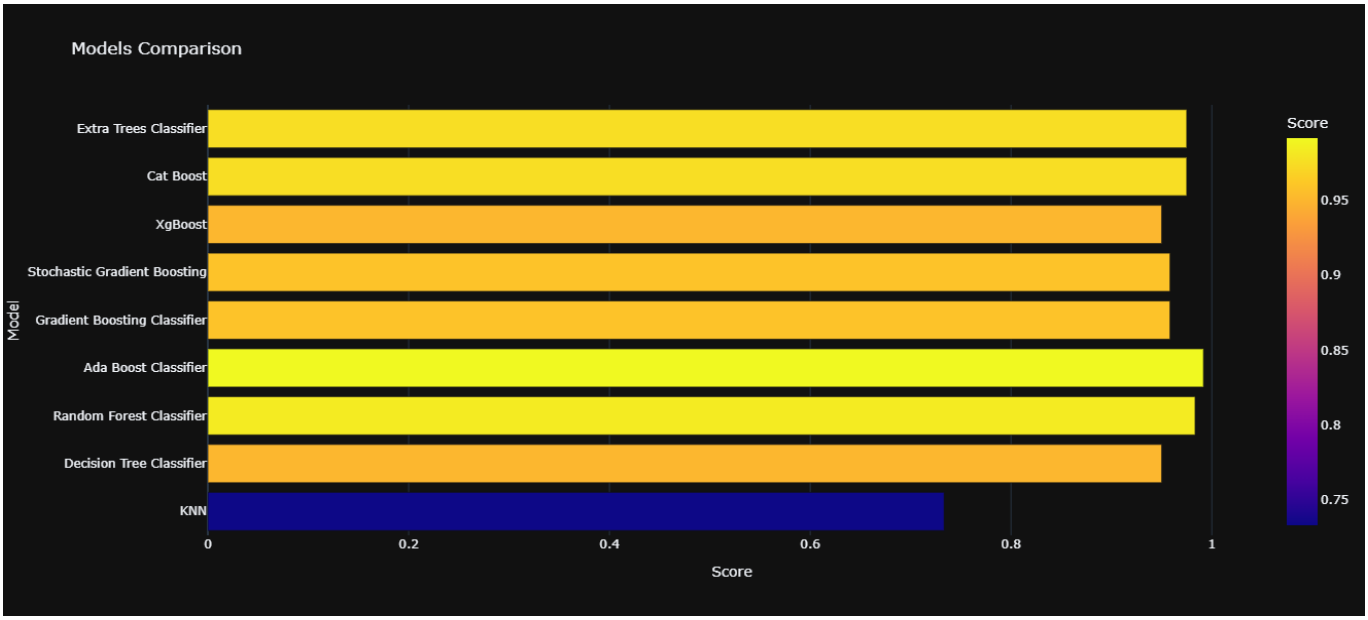
Some features have some categories that are not valid. A check if there is any mismatch between the classes. Further, the sample of patients in the more psychotic segment: 62.5%. Percentage of Lusaka specimens that are not Lusaka kidney disease: 37.5%. Clearly there is not much comparison between the classes. In the Heatmap (Figure 7), the absolute values of the relationships between class labels and features show that blood pressure, specific gravity, albumin, sugar, blood urea, serum creatinine, blood glucose random, and sodium all have positive relationships. are While hemoglobin, potassium, white blood cell count, and red blood cell count have negative relationships. Figure 7 can be observed to see the correlations of the data given in the Heatmap.

applying PCA for XgBoost classifier (Table 3). The performance of KNN and MLP is very weak. The performance of ANN classifiers for both datasets is 60% because of the low number of data. The table summarizes the results of experiments conducted on each model, including its testing and training accuracy, F-1 measure, precision, recall, and confusion metrics. According to the results of the analysis, all the models have good performance in terms of accuracy in identifying CKD including hemoglobin, specific gravity, and albumin, sugar, blood glucose random, serum creatinine, potassium, packed cell volume, white and red blood cells.

However, the data set we used has certain limitations for this development:

First of all, it is expected that the size of the grant will contain only 400 examples, which may affect the validity of the research.

Secondly, the problem identification data set is an additional data set that has the same characteristics that will be compared for performance evaluation.



IV. Results and Discussion

The results of each classifier were evaluated with different diagnostic criteria, and a 10-fold cross-validation method was used to verify the results against overfitting. 23 In addition, the different parameter settings of the models They have gone for the purpose of fine-tuning. A Google Collab web application written in Python using this programming language was used to run the project's experiments. This project uses Scikit-learn23, an open-source software library for machine learning in Python. Validity, F-1 score, precision, and recall were evaluated for this study. Different values for each model's parameters produce unique outputs from each other as shown in Table 2. The best performance for XgBoost for the original CKD data set was 0.9833 accuracy. An accuracy of 0.9916 was obtained for the same classifier when PCA was applied to the data set. adaBoost, random dorest, gradient boosting, LGBM, and Extra Tree showed an accuracy of 0.9833 for the original CKD dataset. For this data set, the highest accuracy is 0.9833, which is increased to 0.9916 after

Table 4 and Figure 11 show a comparison of previous research and the proposed research. Chittora et al.9 used kNN, SVM, and ANN classifiers to predict CKD and obtained a high accuracy of 96.5% for SVM. Almasoud and Ward 3 achieved an accuracy of 98.5% with the RF classifier. Islam et al.18 obtained slightly higher accuracy with the RF

	Model	Score
3	Ada Boost Classifier	0.991667
2	Random Forest Classifier	0.983333
7	Cat Boost	0.975000
8	Extra Trees Classifier	0.975000
4	Gradient Boosting Classifier	0.958333
5	Stochastic Gradient Boosting	0.958333
1	Decision Tree Classifier	0.950000
6	XgBoost	0.950000
0	KNN	0.733333

classifier. According to Gudeti et al.14, SVM failed in other classifiers. Aljaaf et al.2 obtained the general accuracy shown in figure 11. It can be said that previous researches were based on kNN, RF, NB, GB, SVM, and ANN algorithms for prediction of CKD data set which is publicly available. We obtained an accuracy of 0.990 with the GB algorithm and 0.983 for the CKD dataset and hybrid algorithms for the CKD dataset. But the XgBoost algorithm showed high accuracy, which is 0.992. Our proposed model shows the best accuracy.

V. Conclusion

This study has presented various machine learning algorithms that allow the diagnosis of CKD in the first stage. The models that use CKD patients are trained and validated using previously approved input parameters. Studies have been done on the relationship between various factors so that the number of features can be reduced and the information can be completed. When the remaining features were applied to the feature selection filter, it was found that hemoglobin, albumin, and special gravity have the greatest effect on the diagnosis of CKD. This method was used. This allows for early stage diagnosis of CKD. The original CKD dataset was first processed to validate the machine learning based detection model. After that, PCA was performed in order to identify the most dominant features, thus making a diagnosis of CKD. The models that use CKD patients are trained and validated using previously approved input parameters.

In this article, hybrid machine learning techniques based on big data platforms (IPACHI SPARK) were used for Chinese candidate (CKD) prediction. Relief-F and chi-squared feature selection techniques were used to select important features from the data set. Machine learning algorithms, decision tree (DT), logistic regression (LR), neural network (NB), random forest (RF), support vector machine (SVM), and gradient boosting tree (GBT) classifiers such as ensemble learning algorithms for CKD data. Used to compare sets. They were also applied to full features and selected features. Grid search and cross-validation were used to optimize ML parameters. In addition, four types of comparisons, namely accuracy, precision, recall, and F-one measure, were used to verify the results and cross-validation and testing data were registered. The results showed that SVM, DT, and GBT classifiers achieved the best performance with selected features. In general, the performance of Relief-F feature selection is better than that of tea-squared feature selection.

References

[1] G. Manogaran and D. Lopez, "Health data analytics using scalable logistic regression with stochastic gradient descent," *International Journal of Advanced Intelligence Paradigms*, vol. 10, no. 1-2, pp. 118–132, 2018.

[2] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: a technology tutorial," *IEEE access*, vol. 2, pp. 652–687, 2014.

[3] "Apache Hadoop," 2021, https://hadoop.apache.org/.

[4] A. Kafka, "Apache Kafka," 2021, https://kafka.apache.org/.

[5] A. Storm, "Apache Storm," 2021, https://storm.apache.org/.

[6] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *JAMA*, vol. 319, no. 13, pp. 1317-1318, 2018.

[7] L. R. Nair, S. D. Shetty, and S. D. Shetty, "Applying spark based machine learning model on streaming big data for health status prediction," *Computers & Electrical Engineering*, vol. 65, pp. 393–399, 2018.

[8] A. A. Ali, "Stroke prediction using distributed machine learning based on Apache spark," *Stroke*, vol. 28, no. 15, pp. 89–97, 2019.

[9] World Health Organization, Public Health Agency of Canada, and Canada. Public Health Agency of Canada, *Preventing Chronic Diseases: A Vital Investment*, World Health Organization, Geneva, Switzerland, 2005.

[10] B. Bikbov, N. Perico, and G. Remuzzi, "Disparities in chronic kidney disease prevalence among males and females in 195 countries: analysis of the global burden of disease 2016 study," *Nephron*, vol. 139, no. 4, pp. 313–318, 2018.

[11] K. Disease, "Improving global outcomes (kdigo) transplant work group. kdigo clinical practice guideline for the care of kidney transplant recipients," *American Journal of Transplantation*, vol. 9, no. 3, pp. S1–S155, 2009.

[12] Cdc, "Chronic Kidney Disease in the united states," 2021, https://www.cdc.gov/kidneydisease/publications-resources/ckd-national-facts.html.

[13] L. Deng and X. Li, "Machine learning paradigms for speech recognition: an overview," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.

[14] M. Q. Huang, J. Nini´c, and Q. B. Zhang, "Bim, machine learning and computer vision techniques in underground construction: current status and future perspectives," *Tunnelling and Underground Space Technology*, vol. 108, Article ID 103677, 2021.

[15] P. Oza, P. Sharma, and S. Patel, "Machine learning applications for computer-aided medical diagnostics," in *Proceedings of the Second International Conference on Computing, Communications, and Cyber-Security*Springer, New York, NY, USA, 2021.

[16] T. Bikmukhametov and J. J¨aschke, "Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models," *Computers & Chemical Engineering*, vol. 138, Article ID 106834, 2020.

[17] R. J. Palma-Mendoza, D. Rodriguez, and L. D. Marcos, "Distributed relieff-based feature selection in spark," *Knowledge and Information Systems*, vol. 57, no. 1, pp. 1–20, 2018.

[18] M. Nassar, H. Safa, A. A. Mutawa, A. Helal, and I. Gaba, "Chi squared feature selection over Apache spark," in *Proceedings of the 23rd International Database Applications & Engineering Symposium*, pp. 1–5, Athens Greece, June 2019.

[19] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *Proceedings of the 2016 management and innovation technology international conference (MITicon)*, Bang-San, +ailand, October 2016.

[20] P. Chittora, S. Chaurasia, P. Chakrabarti et al., "Prediction of chronic kidney disease-a machine learning perspective," *IEEE Access*, vol. 9, no. 17312, p. 17334, 2021.

[21] A. Spark, *Apache Spark*, 2021 https://spark.apache.org/.

[22] E. M. Senan, M. H. A. Adhaileh, F. W. Alsaade et al., "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," *Journal of Healthcare Engineering*, vol. 2021, Article ID 1004767, 10 pages, 2021.

[23] A. A. Abdullah, S. A. Hafidz, and W. Khairunizam, "Performance comparison of machine learning algorithms for classification of chronic kidney disease (ckd)," *Journal of Physics: Conference Series*, vol. 1529, no. 5, Article ID 052077, 2020.

[24] S. I. Ali, B. Ali, J. Hussain et al., "Cost-sensitive ensemble feature ranking and automatic threshold selection for chronic kidney disease diagnosis," *Applied Sciences*, vol. 10, no. 16, p. 5663, 2020.

[25] O. A. Jongbo, A. O. Adetunmbi, R. B. Ogunrinde, and B. B. Ajisafe, "Development of an ensemble approach to chronic kidney disease diagnosis," *Scientific African*, vol. 8, Article ID e00456, 2020.

[26] L. Jena, B. Patra, S. Nayak, S. Mishra, and S. Tripathy, "Risk prediction of kidney disease using machine learning strategies," in *Intelligent and Cloud Computing*, pp. 485–494, Springer, New York, NY, USA, 2021.

[27] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidneydisease," *IEEE Access*, vol. 8, Article ID 20991, 2019.

[28] N. A. Almansour, H. F. Syed, N. R. Khayat et al., "Neural network and support vector machine for the prediction of chronic kidney disease: a comparative study," *Computers in Biology and Medicine*, vol. 109, pp. 101–111, 2019.

[29] E. Avci, S. Karakus, O. Ozmen, and D. Avci, "Performance comparison of some classifiers on chronic kidney disease data," in *Proceedings of the 2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, pp. 1–4, Antalya, Turkey, March 2018.

[30] M. S. Wibawa, I. M. D. Maysanjaya, and I. M. A. W. Putra, "Boosted classifier and features selection for enhancing chronic kidney disease diagnose," in *Proceedings of the 2017 5th international conference on cyber and IT service management (CITSM)*, pp. 1–6, Denpasar, Indonesia, August 2017.

[31] U. M. L. Repository, "Chronic kidney disease data set," 2021, https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease.

[32] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Proceedings of the European conference on machine learning*, pp. 171–182, Catania, Italy, April 1994.

[33] V. B. Canedo and B. Remeseiro, "Sanchez-maro no, n., and alonso-betanzos," *Transactions on Computational Collective Intelligence XX*, vol. 9420, pp. 78–98, 2017.

[34] V. Chaurasia, S. Pal, and B. Tiwari, "Chronic kidney disease: a predictive model using decision tree," *International Journal of Engineering Research and Technology*, vol. 11, no. 11, pp. 1781–1794, 2018.

[35] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, Article ID 3860146, 21 pages, 2018.

[36] H. Shi, H. Wang, Y. Huang, L. Zhao, C. Qin, and C. Liu, "A hierarchical method based on weighted extreme gradient boosting in ecg heartbeat classification," *Computer Methods and Programs in Biomedicine*, vol. 171, pp. 1–10, 2019.

[37] Akhter T, Islam MA, Islam S. Artificial neural network based covid-19 suspected area identification. J Eng Adv 2020;1:188‑194.

[38] Aljaaf AJ, Al-Jumeily D, Haglan HM, et al. Early prediction of chronic kidney disease using machine learning supported by predictive analytics. 2018 IEEE Congress on Evolutionary Computation (CEC). IEEE; 2018. p. 1‑9.

[39] Almasoud M,Ward TE. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int J Soft Comput Appl 2019;10.

[40] Ani R, Sasi G, Sankar UR, Deepa O. Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE; 2016. p. 1287‑1292.Arora, M., Sharma, E.A., 2016. Chronic kidney disease detection by analyzing medical datasets in weka. International Journal of Computer of machine learning algorithms. New advances in machine learning 3, 19‑48.

[41] Arora M, Sharma EA. Chronic kidney disease detection by analyzing medical datasets in weka. Int J Comput Mach Learn Algor New Adv Mach Learn 2016;3:19‑48.

[42] Banik S, Ghosh A. Prevalence of chronic kidney disease in Bangladesh: a systematic review and meta-analysis. Int Urol Nephrol 2021;53:713‑718.

[43] Charleonnan A, Fufaung T, Niyomwong T, Chokchueypattanakit W, Suwannawach S, Ninchawee N. Predictive analytics for chronic kidney disease using machine learning techniques. 2016 Management and Innovation Technology International Conference (MITicon). IEEE; 2016.pp. MIT‑80.

[44] Chen Z, Zhang X, Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. Int Urol Nephrol 2016;48:2069‑2075.

[45] Chittora P, Chaurasia S, Chakrabarti P, et al. Prediction of chronic kidney disease-a machine learning perspective. IEEE Access 2021;9:17312‑17334.

[46] Cueto-Manzano AM, Cortés-Sanabria L, Martínez-Ramírez HR, Rojas-Campos E, Gómez- Navarro B, Castillero-Manzano M. Prevalence of chronic kidney disease in an adult population. Arch Med Res 2014;45:507‑513.

[47] Dua D, Graff C. UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml 2017.

[48] Eroğlu K, Palabaş T. The impact on the classification performance of the combined use of different classification methods and different ensemble algorithms in chronic kidney disease detection. 2016 National Conference on Electrical, Electronics and Biomedical Engineering (ELECO). IEEE; 2016. p. 512‑516.

[49] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intel Learn Syst Appl 2017;9(01):1.

[50] Gudeti B, Mishra S, Malik S, Fernandez TF, Tyagi AK, Kumari S. A novel approach to predict chronic kidney disease using machine learning algorithms. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE; 2020. p. 1630‑1635.

[51] Heung M, Chawla LS. Predicting progression to chronic kidney disease after recovery from acute kidney injury. Curr Opin Nephrol Hypertens 2012;21:628‑634.

[52] Islam M, Shampa M, Alim T, et al. Convolutional neural network based marine cetaceans detection around the swatch of no ground in the bay of bengal. Int J Comput Digit Syst 2021;12:877‑893.

[53] Islam, M.A., Akhter, T., Begum, A., Hasan, M.R., Rafi, F.S. Brain tumor detection from MRI images using image processing.

[54] Islam MA, Akter S, Hossen MS, Keya SA, Tisha SA, Hossain S. Risk factor prediction of chronic kidney disease based on machine learning algorithms. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE; 2020. p. 952‑957.

[55] Islam MA, Hasan MR, Begum A. Improvement of the handover performance and channel allocation scheme using fuzzy logic, artificial neural network and neuro-fuzzy system to reduce call drop in cellular network. J Eng Adv 2020;1:130‑138.

[56] Mahesh B. Machine learning algorithms-a review. Int J Sci Res (IJSR)[Internet] 2020;9: 381‑386.

[57] Polat H, Danaei Mehr H, Cetin A. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. J Med Syst 2017;41:1-11.