Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 4 (2023)

Improved Detection and Prediction of Chronic Renal Disease by Evaluating Machine Learning Algorithms with Predominantly Reduced Features

[1*]Vasanthakumar G U, [2]Shakuntala Inamati, [3]Akash Patil KM, [4]Aishwarya M, [5]Impana B S

[1] [2] [3] [4]Nitte Meenakshi Institute of Technology, Bengaluru, India.
[5]CGI Inc., Bengaluru, India.

E-mail: vasanth.gu@nmit.ac.in

Abstract: The kidneys play as an important organ which help in removal of toxic waste from the body. Their malfunctioning may lead to Chronic Renal Disease (CRD) if not attended and treated appropriately at the right time. This chronic situation expedite kidney failure and in-turn death if not diagnosed and attended on time. This work depicts the appropriate, relevant and correlated attributes among all the attributes and reduction of features in the dataset using Chi-square test on to the patients' dataset for extraction of predominant features for better detection and prediction of CRD. The DPCRD algorithm is implemented and the results are predominantly used in Logistic Regression, K-Nearest Neighbour classification and Random Forest classification technique to enhance and improve its prediction accuracy on CRD. The results revealed that the Random Forest algorithm achieved an improved accuracy of 100.00% after the selection and reduction of attributes using Chi-square test when compared to 99.33% and 99.17% of Logistic Regression and K-Nearest Neighbour classifiers respectively.
Keywords: Chronic Renal Disease, Classification Technique, Kidney Disease, K-Nearest Neighbour, Logistic Regression, Random Forest.

1. Introduction

Chronic Renal Disease (CRD) is unyielding disease. Kidney failure causes death as it does not remove the waste which is created in our bodies. Chronic renal illness affects 10% of the global population. There are approximately 61.7 million women and 48.3 million men among them. According to the Renal Association, out of 18 million individuals in Bangladesh, 35000 to 40000 suffer each year from chronic kidney failure [1]. The role of kidneys is to filtering your blood and the extra water. If kidneys are not functioning properly, experiencing one or more symptoms such as nausea, puking, diarrhea, nosebleeds, back pain, abdominal discomfort, and fever may arise. There can be many conditions like Covid-19 as well which may affect and lead to malfunctioning of kidneys [2].

The kidney is one of the most crucial organs for both humans and animals, as we all know. Osmoregulation and excretion are among the primary functions of the kidney. It plays a crucial part in cleansing the blood and getting rid of harmful toxins from the body. Since it impairs kidney function, CRD is a serious condition that poses a risk to society. Permanent renal failure may result from this illness. Since the kidneys begin to work incorrectly, the amount of waste that is excreted from the body will also be small. Therefore, it's crucial to identify the disease early on, yet some people don't exhibit any symptoms. In order to determine whether a person has CRD or not, machine learning can be useful.

The prediction along with detection of Chronic Kidney Disease (CKD), nothing but CRD is performed with numerous unsupervised machine learning techniques, including Auto-encoder, I-Forest, DB-SCAN and K-means with an accuracy of 91% when all 24 features were considered. Four types of feature selection methods were employed to speed up and save money on the CKD diagnosis process [3]. The prominent two illnesses related to CKD are as follows: Diabetes and excessive Blood Pressure. Asia region has the highest rate of CKD patients in the world, led by Japan and followed by Taiwan [4]. CKD is a highly common condition now-a-days, and two life-threatening conditions that can result from it are cardiovascular and final-stage renal disease. These might be avoided by detecting its conditions little early and timely treatment of those who are in danger.

By eliminating toxins from the body through urination through the bladder, the kidneys filter the blood. Renal failure may cause death if the kidneys are unable to eliminate waste that is tainted with toxins. Acute or chronic kidney problems can be categorized. For those with renal disease condition, wastes can accumulate to the highest amounts in our blood vessels, which can cause issues like high blood pressure, brittle bones, anemia, poor nutrition, and damage to our nerve cells. Renal disease also increases the chance of Cardio-vascular issues. The authors [5] have provided an outline of several tools available for early detection of Alzheimer's disease. This has opened a new door for researchers by providing outline of already available techniques which can be utilized to further develop the improved and advanced tools and this can be achieved through proper handling of the input image data obtained from different modalities. The advanced tools could make more clear representations so that it helps experts to decide quickly and accurately to make decision for diagnosing patients without or with minimal manual interventions. One of the most significant issues in the medical areas is the prediction of this disease.

In order to resolve or mitigate the human error done by medical professionals, predicting the risk of CRD and automating the process its identification is done using the best Machine Learning algorithm. The following are the objectives of the proposed work based on the research gaps identified:

- To clean and handle the missing data in the dataset for improved detection.
- To select the appropriate, relevant and correlated attributes among all the attributes and reduction of features in the dataset using Chi-square test for better detection of CRD.
- To improve the prediction accuracy of CRD using Logistic Regression, K-Nearest Neighbour and Random Forest.

Rest of the paper is consciously organized in an effort to hold the readers' interest. Section-2 impart a summary of the related work in the literature and relevant research done to date by various researchers. The Section-3 defines the problem whereas the proposed system and the mathematical model of feature selection technique is represented in Section-4. The system model is described in Section-5, Results are discussed in Section-6 along with the entire work summarization and conclusion in Section-7.

2. Related Work

The main cause of death in poorer countries, the CKD is the silent killer in wealthy countries [6]. CKD is one of the greatest threats to public health due to its increased cases. It is clear that CKD may be detected early, which can reduce the amount of maturity-related damage. The patient needs to go to a diagnostic facility in order to speak with a doctor. CKD is characterised by a slow, months or years-long decline in kidney functioning [7]. One of the most significant issues in the medical areas is the prediction of this disease. Despite modifications in treatment and surgical care, CKD remains a health issue [8]. Researchers from all around the world are highly motivated in prospering gassed-up methods for the diagnosis, treating, and preventative therapy of CKD despite its recent rises. Learning features that are important for the problem may result in improving the performance. Therefore, automated tools and techniques are useful in helping doctors predict CKD for timely and better treatment saving lives. But these tools will always have a limitations of having detection mechanisms only to those known features in the data set.

The authors [9] offered a method to detect CKD along with risk variables that are essential for its early identification to avoid the prognosis of the disease to end-stage using seven deep learning algorithms. The study shows a comprehensive evaluation of deep learning methods on CKD. Random forest algorithm accuracy is 97.12%, while ANN accuracy is 94.5%. The early detection of chronic renal disorders will be assisted by this technology. The authors [10] carried out feasibility study of a distributed approach for the management of alarms related to the monitoring of CKD patients. For this purpose and from the methodological point of view, the predictions are limited to alarms and prioritization issues which are addressed methodologically according to the definitions provided by the ISO IEC/CD 60601. If the main criteria that develops CKD are known, even those without a diagnosis of the disease may learn something about the state of their kidneys from a medical test they took for a different purpose. Following that, they may carry out a proper CKD check [11]. This approach is limited to only with those known features leading to CKD.

The authors [12] presented and examined a novel sensing paradigm for rapid and precise CKD diagnosis. The deep learning algorithm is directly fed with raw sensor signal in order to make a prediction. The suggested 1-

D CNN-SVM process correctly defined the samples with an accuracy of 98.04% by removing features from the unprocessed signal. In this method, feature selection is limited to prominently known ones. The physician checks and analyses the suggested sensing strategy. The combined model would provide appropriate accuracy after KNN imputation is used without supervision [13]. They also think that using this methodology to really diagnose CKD would have a positive outcome. Medical records of 491 UAE patients who had CKD and were at risk for cardiovascular disease were examined [14], and the machine learning techniques that could accurately predict the possibility that these patients will acquire CKD at stages 3-5 is developed. This approach is unable to detect if the CKD is at stages 1 and 2.

The study [15] has found a trustworthy method for categorizing CKD and choosing attributions that is also more cost-effective and simpler to use. Training and choosing the best classifiers, calculating the feature importance using SHAP values were performed, and obtaining a smaller dataset based on the pathological tests and determined feature importance. Second, they used these smaller data sets to train the classifiers, then tested those classifiers using the test datasets. The findings showed that the crucial elements that SHAP had discovered were in line with what clinical thought was currently thinking but are limited to smaller data sets.

In order to determine whether a person is at risk of acquiring CKD, the authors propose a classifier powered by neural networks [16]. The model is trained on two groups with and without CKD in Colombia during the year 2018 and the results were quite impressive with 95% accuracy on the dataset but the accuracy of the model slightly gets reduced with samples from other country patients.

3. Problem Definition

Given the dataset pertaining to the patients with CRD, the problem is to select the appropriate, relevant and correlated attributes among all the attributes and reduction of features in the dataset using Chi-square test for better detection and prediction of CRD.

4. Proposed System

In this work, a methodology for improving the prediction of CRD status based on clinical data is proposed. It includes data preparation, a mechanism for addressing missing values, clustering, and attribute values selection and reduction. Once the data is prepared, it is tested for accuracy on Logistic Regression, K-Nearest Neighbour and Random Forest classifiers.

4.1 Data Collection and Preprocessing

The dataset for analysis of the proposed model is considered form the UCI online repository which consists of 400 instances and 24 attributes. During preprocessing of the dataset, identifying and handling of missing values, outliers, and inconsistencies in the dataset if performed. Techniques such as imputation for missing values or removal of extreme outliers, are applied for ensuring the quality of the dataset.

4.2 Feature Selection

Analysing the collected features and selecting the relevant ones that have a significant impact on improving the detection of CRD is of prime importance. Feature selection methods like correlation analysis, mutual information, or domain knowledge-based approaches can be employed for improving the accuracy of the classifier. The statistical analysis that evaluates the appropriateness of the observed real data is the chi-square statistics and measures the dependence between two categorical variables as depicted in Equation (1).

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad \text{Equation… (1)}$$

where,
x^2 = Chi-square
O_i = observed value
E_i = expected value

Chi-square testing can be particularly useful for feature selection when dealing with categorical target variables or when working with datasets that have a mix of categorical and continuous features. As in this work, the role in feature selection and reduction of features is significant, thus it is considered for the same here. Equation (2) provides the technique for Chi-square test.

$$Chi2.pdf(x,a) \frac{1}{(2 * gamma\left(\frac{a}{2}\right) * \left(\frac{x}{2}\right)^{\left(\frac{a}{2}-1\right)} * \exp\left(\frac{-x}{2}\right))}$$

Equation… (2)

After performing the chi-square test on each attribute, a chi-square statistics and a corresponding score for each attribute is obtained. The chi-square statistics quantifies the extent of correlation between the attribute and the target variable, while the scores indicate the statistical significance of that association. Based on the results of the chi-square test, decisions are made for selecting or excluding attributes for further analysis or modeling in this work.

5 System Model

5.1 System Architecture

System architecture diagram is depicted in the Fig. 1, which shows how the dataset considered from University of California Irvine for analysis which is preprocessed to select and reduce the appropriate and relevant correlated attributes using chi-squared testing model. The data is pre-processed to extract meaningful insights like: Serum creatinine, Random Blood glucose, potassium, albumin, packet cell volume, age, sugar, hypertension, diabetes mellitus, blood pressure and categorized to perform whether the patient is suffering from CRD or is a non-CRD. Here, 10 attributes are selected and reduced from 24 attributes which are relevant and significant for further testing.

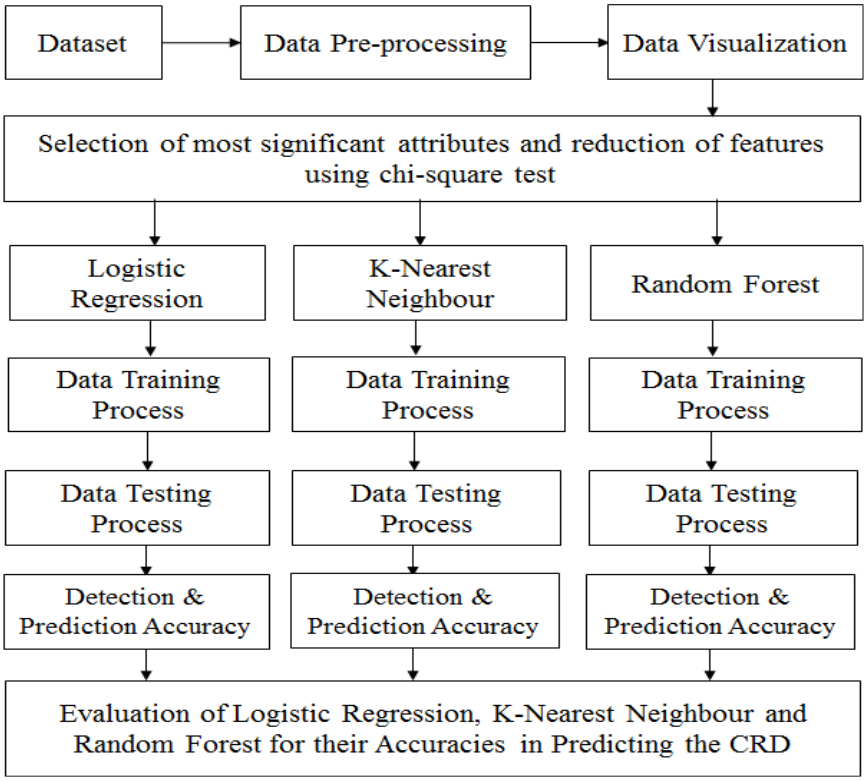

Fig 1: System Architecture Diagram

Further, 70% of data is trained and remaining 30% of data is used for testing Logistic Regression, K-Nearest Neighbour and Random Forest machine learning algorithms for evaluation of their performances in predicting CRD.

5.2 DPCRD Algorithm

Algorithm-1, depicts the proposed DPCRD algorithm, showing how data is acquired from each patient's file and is examined by conducting Explanatory Data Analysis (EDA), feature engineering, chi-square test to get cleaned and useful data further to perform and predict the CRD on the useful data considering Logistic Regression, K-Nearest Neighbour and Random Forest machine algorithms. At first, the data set is considered and analyzed for any missing values and the data types. After performing the data cleaning, the data set is evaluated by utilizing mean method and filling the missing values, as well as utilizing the mode method to handle null values.

Algorithm 1: *Detection and Prediction of Chronic Renal Disease (DPCRD) Algorithm*

```
while (True) do
for (Raw Data) do
    Perform EDA
    Perform Data Cleaning
    Perform Feature Engineering to select most significant attributes using Chi-square Test
    Perform Feature Reduction to Select Predominantly Significant Attributes
    Get Preprocessed Data
end for
for (Preprocessed Data) do
                Apply LR and Compute its Prediction Accuracy
                if (Prediction Accuracy < Threshold) then
                        Perform Hyper-parameter tuning to classify the Patient as CRD/Non-CRD
                else
                        Classify the Patient using Binary Classification as CRD/Non-CRD
                Apply KNN and Compute its Prediction Accuracy
                if (Prediction Accuracy < Threshold) then
                        Perform Hyper-parameter tuning to classify the Patient as CRD/Non-CRD
                else
                        Classify the Patient using Binary Classification as CRD/Non-CRD
                Apply RF and Compute its Prediction Accuracy
                if (Prediction Accuracy < Threshold) then
                        Perform Hyper-parameter tuning to classify the Patient as CRD/Non-CRD
                else
                        Classify the Patient using Binary Classification as CRD/Non-CRD
end for
    Evaluate LR, KNN and RF to determine the best method for detection and prediction of CRD
end while
```

Further, proceeded to conduct a correlation analysis on the data set using chi-square testing. Consequently, identified the most correlated data variables that are well suited for subsequent analysis and implementation within various machine learning models. As in the proposed algorithm, the results of prediction accuracy is achieved better for CKD, these methods are thus employed.

6 Results and Analysis

The dataset is provided by University of California Irvine, which is then stored according to date. 24 attributes in the data set describe things like age, blood pressure, sugar, RBC, hypertension, and more. The CRDP algorithm is implemented on an Intel Pentium i7 processor with 4GB RAM and built in Python using Jupyter notebook, an open-source software development tool, where Windows 8 supports these setups.

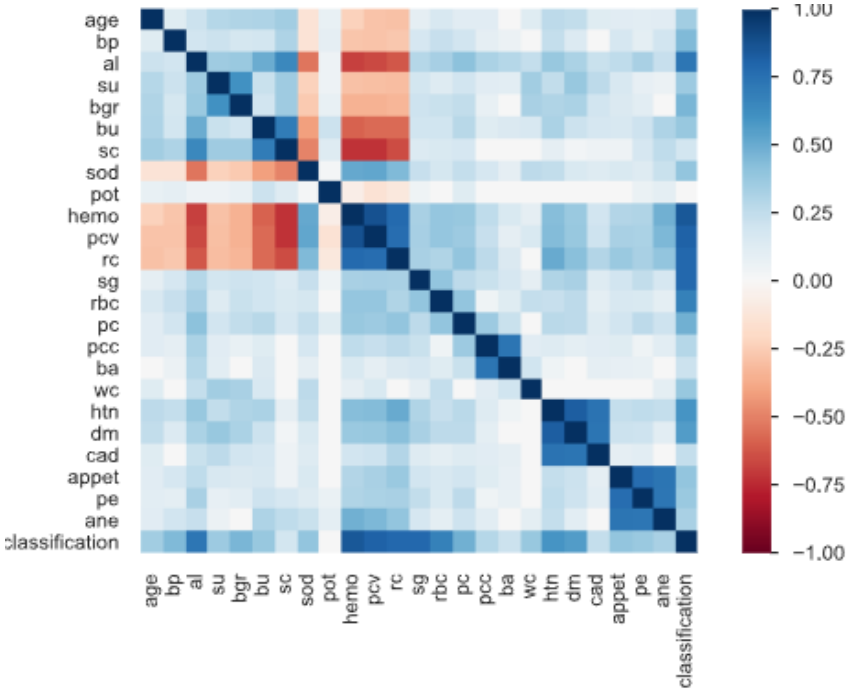

Fig 2: Correlation Matrix

Chi-square testing is used to analyze the correlation matrix displayed in Fig 2. After performing all correlation methodology like Spearman, Pearson, Kendall, Cramer, Phik, it is observed that there are some independent and dependent variables available in the dataset. The correlation scores of each and every attributes is then computed and is as depicted in Table.1. It is observed from the table that the correlation scores of every attribute is not same and only few attributes have higher scores.

As per the scores of relationships between dependent and independent columns by implementing this correlation, the highest scored correlated features considering only 10 high score attributes out of 24 attributes are obtained and are as in Table.2, which helps in getting less number attributes for further testing and training Logistic Regression and K-Nearest Neighbour and Random Forest models so as to analyze the accuracy of the algorithms and a faster way to predict the two possibilities like CRD or non-CRD.

Table 1: Attribute Scores of all 24 attributes after Chi-squared testing

Attribute ID	Features	Score
1	age	115.859940
2	bp	81.786701
3	sg	0.005035
4	al	216.000000
5	su	94.800000
6	rbc	3.754674
7	pc	10.696296
8	pcc	25.200000
9	ba	13.200000
10	bgr	2241.651289
11	bu	2343.097145
12	sc	357.792101

Attribute ID	Features	Score
13	sod	27.558749
14	pot	2.951339
15	hemo	123.856342
16	pcv	308.181415
17	wc	80.604980
18	rc	19.113025
19	htn	88.200000
20	dm	82.200000
21	cad	20.400000
22	appet	49.200000
23	pe	45.600000
24	ane	36.000000

Table 2: Attribute Scores of top 10 attributes

Attribute ID	Features	Score
11	bu	2343.097145
10	bgr	2241.651289
12	sc	357.792101
16	pcv	308.181415
4	al	216.000000
15	hemo	123.856342
1	age	115.859940
5	su	94.800000
19	htn	88.200000
20	dm	82.200000

The 10 chosen attributes are the most relevant and informative features for the task of early detection of CRD and the scores distribution is as shown in Fig. 3. These attributes shown a significant association with CRD, in indicating the potential importance in predicting the disease.

According to the depicted results indicate that individuals in the 60-64 age group have a higher likelihood of developing CRD. While kidney disease can occur at any age, the risk increases significantly after the age of 60. This observation can be attributed to the fact that as people age, their kidneys also undergo natural changes and may become more susceptible to kidney-related issues.

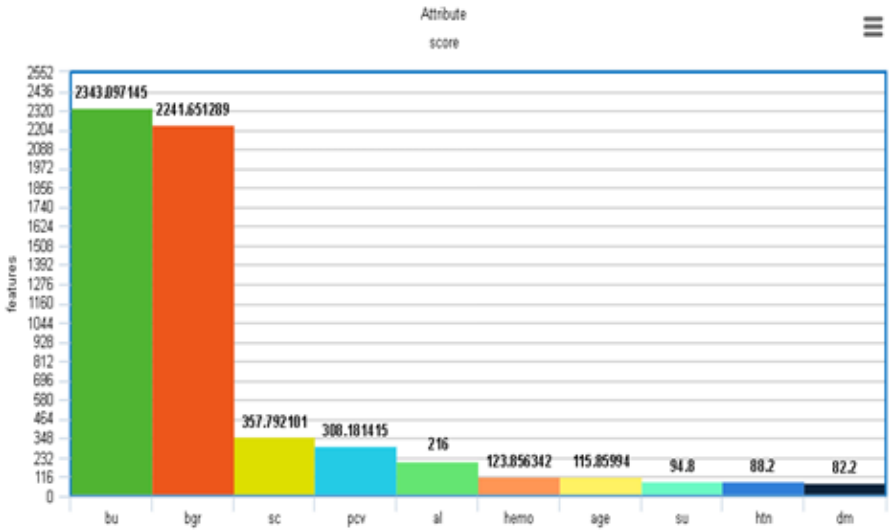

Fig 3: Top 10 Attribute Scores

Including all available attributes in the model can lead to a high-dimensional feature space, especially if the dataset contains a large number of features. This will introduce challenges in terms of computational resources, model complexity, and potential over fitting. Choosing a subset of 10 attributes helps to simplify the model and reduce the dimensionality while retaining the most important features.

By selecting a smaller set of attributes, the resulting model is more interpretable and easier to understand. This is particularly important in healthcare applications where clinicians and domain experts need to trust and comprehend the model's decision-making process. The chosen 10 attributes might align with known risk factors, biomarkers, or indicators of CRD, enhancing the model's interpretability. This also enhances the speed of detection, in-turn early detection of the disease.

With a limited number of attributes, the model training process is faster and more efficient. It also reduces the risk of over fitting, where the model performs poorly on unseen data and becomes too specific to the training data. By focusing on a subset of 10 attributes, the model can generalize better and have a higher chance of performing well on new, unseen instances.

The performances of the models can be penetrated with different aspects through these metrics and can help evaluate its effectiveness in classification.

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

Equation… (3)

$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}$$

Equation… (4)

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$$

Equation… (5)

In the context of a Logistic Regression, K-Nearest Neighbour and the Random Forest classifiers, the confusion matrix can be generated by evaluating the predictions of the ensemble on a set of labeled test data. By examining the confusion matrix, various performance metrics can be derived to assess the capabilities of Logistic Regression, K-Nearest Neighbour and the Random Forest classifiers, such as accuracy, precision, and recall as shown in Equations (3), (4) and (5) respectively.

Table 3: Results of Evaluation Parameters of Logistic Regression, K-Nearest Neighbour and Random Forest

Evaluation Parameters	Logistic Regression	K-Nearest Neighbour	Random Forest
Accuracy	98.33 %	99.17 %	100.00 %
Precision	0.9833	0.9778	1.000
Recall	0.9833	1.0000	1.000

From the results obtained as shown in Table 3, it is observed that the accuracy obtained for predicting CRD through Random Forest classifier 100% which is improved and better than Logistic Regression and K-Nearest Neighbour classifiers after reduction of attributes using Chi-square test and selection of predominantly appropriate attributes for improved detection and prediction of CRD.

7. Conclusion

The most important stage in illness prevention is disease detection. Today, the number of CRD patients is steadily rising. The primary focus of this work, the disease detection, which plays a crucial role in chronic illness prevention. CRD is on the rise, emphasizing the need for accurately detecting the disease which is achieved in this work. The main objective of this work, i.e to accurately identify CRD utilizing a number of tests and indicators is achieved. Out of 24, only 10 relevant attributes are appropriately selected and reduced, which are used to conduct testing and produce reliable findings. Lesser input attributes are frequently used, but the aim here is to predict with more attributes and with greater speed the likelihood of developing CRD.

The results revealed that the Random Forest algorithm achieved an improved accuracy of 100.00% after the selection and reduction of attributes using Chi-square test when compared to 99.33% and 99.17% of Logistic Regression and K-Nearest Neighbour classifiers respectively. The work can be extended to compute the detection and prediction accuracy for various other machine learning algorithms and the best can be chosen to better improve the detection of CRD in less time.

References

[1] Yashfi, Shanila Yunus, Md Ashikul Islam, Nazmus Sakib, Tanzila Islam, Mohammad Shahbaaz, and Sadaf Salman Pantho. "Risk prediction of chronic kidney disease using machine learning algorithms." In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-5. IEEE, 2020.

[2] Vasanthakumar, G. U., N. Ramu, and M. N. Thippeswamy. "PRGR-C19: Profiling Rapid Growth Regions of COVID-19 Pandemic, A Data-Driven Knowledge Discovery Approach." *In International Conference on Information Processing*, Springer International Publishing, pp. 366-379, 2021.

[3] Antony, Linta, Sami Azam, Eva Ignatious, Ryana Quadir, Abhijith Reddy Beeravolu, Mirjam Jonkman, and Friso De Boer. "A comprehensive unsupervised framework for chronic kidney disease prediction." *IEEE Access* 9 (2021): 126481-126501.

[4] Islam, Md Ashiqul, Shamima Akter, Md Sagar Hossen, Sadia Ahmed Keya, Sadia Afrin Tisha, and Shahed Hossain. "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, 2020.

[5] Vinutha, N., G. U. Vasanthakumar, P. Deepa Shenoy, and K. R. Venugopal. "A Comprehensive Survey on Tools for Effective Alzheimer's Disease Detection." *Neuroscience International*, vol. 9, no. 1 pp. 1-10, 2018.

[6] Samet, Sarra, Mohamed Ridda Laouar, and Issam Bendib. "Predicting and Staging Chronic Kidney Disease using Optimized Random Forest Algorithm." In *2021 International Conference on Information Systems and Advanced Technologies (ICISAT)*, pp. 1-8. IEEE, 2021.

[7] Maurya, Akash, Rahul Wable, Rasika Shinde, Sebin John, Rahul Jadhav, and R. Dakshayani. "Chronic kidney disease prediction and recommendation of suitable diet plan by using machine learning." In *2019 International Conference on Nascent Technologies in Engineering (ICNTE)*, pp. 1-4. IEEE, 2019.

[8] Elkholy, Shahinda Mohamed Mostafa, Amira Rezk, and Ahmed Abo El Fetoh Saleh. "Early prediction of chronic kidney disease using deep belief network." *IEEE Access* 9 (2021): 135542-135549.

[9] Akter, Shamima, Ahsan Habib, Md Ashiqul Islam, Md Sagar Hossen, Wasik Ahmmed Fahim, Puza Rani Sarkar, and Manik Ahmed. "Comprehensive performance assessment of deep learning models in early prediction and risk identification of chronic kidney disease." *IEEE Access* 9 (2021): 165184-165206.

[10] Estudillo-Valderrama, Miguel A., Alejandro Talaminos-Barroso, Laura M. Roa, David Naranjo-Hernandez, Javier Reina-Tosina, Nuria Areste-Fosalba, and Jose A. Milan-Martin. "A distributed approach to alarm management in chronic kidney disease." *IEEE journal of biomedical and health informatics* 18, no. 6 (2014): 1796-1803.

[11] Nishanth, Anandanadarajah, and Tharmarajah Thiruvaran. "Identifying important attributes for early detection of chronic kidney disease." *IEEE reviews in biomedical engineering* 11 (2017): 208-216.

[12] Bhaskar, Navaneeth, and Suchetha Manikandan. "A deep-learning-based system for automated sensing of chronic kidney disease." *IEEE Sensors Letters* 3, no. 10 (2019): 1-4.

[13] Qin, Jiongming, Lin Chen, Yuhua Liu, Chuanjun Liu, Changhao Feng, and Bin Chen. "A machine learning methodology for diagnosing chronic kidney disease." *IEEE Access* 8 (2019): 20991-21002.

[14] Chicco, Davide, Christopher A. Lovejoy, and Luca Oneto. "A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease." *IEEE Access* 9 (2021): 165132-165144.

[15] Rashed-Al-Mahfuz, Md, Abedul Haque, Akm Azad, Salem A. Alyami, Julian MW Quinn, and Mohammad Ali Moni. "Clinically applicable machine learning approaches to identify attributes of Chronic Kidney Disease (CKD) for use in low-cost diagnostic screening." *IEEE Journal of Translational Engineering in Health and Medicine* 9 (2021): 1-11.

[16] Vásquez-Morales, Gabriel R., Sergio M. Martinez-Monterrubio, Pablo Moreno-Ger, and Juan A. Recio-Garcia. "Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning." *Ieee Access* 7 (2019): 152900-152910.