Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 6s (2024)

Detection of CKD Status using Feature Selection Approach based on Random Forest Classifier

Shibi Mathai[1], Dr. K.S. Thirunavukkarasu[2]

[1]Ph.D., Scholar (Part-Time), Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, India, shibimathai@gmail.com

[2]Assistant Professor and Research Supervisor, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Pallavaram,Chennai, India, thirukst@gmail.com

Article History:

Received: *15-06-2024*

Revised: *24-07-2024*

Accepted: *06-08-2024*

Abstract:
Chronic kidney disease (CKD) is a serious condition that can last a lifetime occurs on by either impaired kidney function or kidney cancer. With CKD, the kidneys are unable to filter blood properly or have totally stopped functioning, which results in the accumulation of toxins in the bloodstream and ultimately killing the patient. Early detection of CKD is probably impossible, and saving a patient's life in the last stages of CKD is extremely challenging. High variance frequently impedes clinical decision-making in the prognosis of chronic disorders, resulting in ambiguity and unfavorable outcomes, particularly in situations like CKD. Early identification and suitable treatment can raise this risk. Techniques for Machine Learning (ML) have become important instruments for improving clinical decision-making and lowering unpredictability. However, because they rely on a small number of biological characteristics, current approaches for CKD identification frequently lack accuracy. This study investigates creatinine levels by assessing estimated Glomerular Filtration Rate (eGFR) and Blood Urea Nitrogen (BUN) derived from serum creatinine through indirect measurement using the dataset's available attributes. The dataset comprises 28 attributes, including "Gender," and encompasses a total of 523 records. This research work suggests label encoder, hot encoder, standard scaler, and iterative imputation for missing values as preprocessing techniques to solve issues in medical datasets. The Boruta method is used for feature selection, and ML algorithms are utilized to create the model. This prediction analysis involved eGFR and BUN computing values for assist in providing an accurate classification of CKD and non-CKD status using proposed Boruta Feature Selection (BFS) technique with Ensemble Based Random Forest Classifier (ERFC). Moreover, the performance evaluation of proposed BFS-ERFC model is compared with Radom Weighted Optimization (RWO) with Neural Network (NN), RWO with Logistic Regression (LR) and ERFC for evaluating the patient's medical records to determine the improved classification status of CKD.

Keywords: Chronic Kidney Disease (CKD), Boruta Feature Selection, Random Forest Classifier, Detection, Stages

1. Introduction

CKD is a serious threat to a person's general quality of life as well as their physical health. Treatment for the resulting consequences is achievable when the GFR declines, which can lower the risk of cardiovascular disease development and increase survival rates. Regularly scheduled laboratory testing can serve the dual purpose of diagnosing and treating chronic renal disease. Treatments for reduced GFR and associated consequences can postpone, halt, or even completely avoid the disease. A number

of risk factors, including smoking, adopting bad eating habits, not getting enough sleep, and not managing one's weight, can contribute to the development of CKD. Over 700 million individuals worldwide were afflicted by this illness in 2016. As the illness progresses, kidney failure could occur. The current diagnostic procedure involves assessing creatinine levels through serum and urine analysis. This is achieved by employing a variety of medical techniques, some of which include ultrasound and screening techniques [1]. Before undergoing any testing, a patient is examined for prior history, hypertension, current condition, and history of renal disease. Using this method, GFR may be estimated from the model's serum creatinine level and the Albumin to Creatinine Ratio (ACR) taken from a morning urine sample.

Following are the five stages of CKD:

• First Stage where GFR > 90 ml/min, either normal or high

• Mild CKD second stage where GFR is between 60–89 ml/min

• Moderate CKD third stage where GFR is between 30-59 ml/min

• Severe CKD fourth stage where GFR is between 15–29 ml/min

• Final Phase: Final Stage where GFR<15 ml/min

The condition known as CKD causes the kidneys' ability to function to steadily deteriorate over time. An estimated 14% of people are thought to be affected by chronic renal disease [2]. This is true as more than 1.5 million people need dialysis due to renal failure or kidney transplants. Numerous body processes, such as the blood pressure, synthesis of red blood cells and calcium metabolism, are regulated by a hormone that the kidneys manufacture. A number of variables can influence eGFR which include gender, age and creatinine levels. The main indicator for CKD stage classification is eGFR. The kidneys go through five different phases of function. Still, the great majority of instances fall into stage 3. There is a moderate reduction in each of the two stages' performance abilities. The research on ML algorithms for predicting renal disease has utilized various samples.

A part of AI called ML allows computers to carry out particular tasks without direct guidance. When it comes to reliable techniques, ML techniques can be trained to recognize the patterns in model data and forecast results for fresh data using the collected data. Compared to traditional research, ML incorporates more complex mathematical operations and generally produces better results when forecasting outcomes affected by a wide range of factors with complex, non-linear interactions [3]. Recent research has shown that ML can perform at a level that is higher than both human and traditional statistical capabilities [4].

The research that is currently available on classifying and predicting CKD using a variety of data mining techniques, as well as assessing BCR and eGFR to ascertain the degree of CKD, is summarized in this paper. The CKD-EPI expression is created utilizing a collective data set derived from ten investigations of patients with and without the aid of CKD, every one of those have assessed GFR utilizing urine iothalamate clearance and blood creatinine traceable in accordance to international resource criteria. Nevertheless, the BUN-CR and EGFR are conducted using these factors to ascertain the precise Sc concentration in the patients. Hence, this paper focuses on generating BCR and EGFR values to reliably classifying CKD levels utilizing Feature selection of Boruta approach for rapid and

reliable diagnoses. This procedure improves computing speed as well as forecast accuracy [5]. Boruta is a method that uses the RF algorithm to carry out feature selection and ranking. Determining the significance of each variable and assisting in the statistical selection of important variables are two benefits of using Boruta. Furthermore, increasing the p-value to 0.01 could improve the method's resilience. The term "estimator" is used to indicate how many times the algorithm has been run. The variables should be picked with greater selectivity the larger the nEstimator.100 is the standard value. This strategy offers a top-down method by contrasting the initial features with pertinent attributes. The study also tackles the challenges associated with evaluating CKD data in the presence of missing values. This is accomplished through the utilization of a novel approach and the evaluation of various techniques employing the UCI dataset.

The primary contributions of this work include:

- To categorize and identify earlier possibilities of CKD, justification may be provided utilizing EGFR and BCR.
- The input data from the kidney disease patient's dataset is divided into training and testing subsets as part of the data preprocessing procedure.
- Implementing BFS-ERFC model for import train_test_split, LabelEncoder, OneHotEncoder and StandardScalar in sklearn package.
- After the model has been trained, this research work utilizes it to anticipate the labels for both our training and test datasets. Subsequently, these predictions allow us to examine the performance accuracy of training and test methods.

Literature Review

Numerous researchers have shown a keen interest in predicting CKD, employing diverse classification methods to develop effective and specific prediction systems.

Shahinda et al. [6] introduces a sophisticated model for classifying and predicting kidney-related diseases. The proposed model employs a customized Deep Belief Network (DBN) as the classification algorithm, utilizing Softmax as the activation function and Categorical Cross-entropy as the loss function. Poonia et al. [7] utilizes different ML approaches whereas the results indicated that LR method, optimized with attributes selected through the Chi-Square method, achieved the peak precision at 98.75%. Debabrata et al. [8] focusses research with the objective of creating a ML technique for the early recognition of CKD by utilizing the UCI CKD dataset. The investigators incorporated imputation methods, data normalization and data balancing sampling technique in their approach. Nine features were chosen through the SVM and chi-square test were employed for categorization. Despite its contributions, the research had constraints, including the omission of progressive imputation techniques and the potential loss of information resulting from minimizing the variable set. Z. Ullah et al [9] discusses to estimate the progression of CKD uutilizing the implementation of a Decision Tree-based imputation method for handling missing values. The researchers carried out the KNN for categorization and feature selection through the utilization of filter method. Notably, the research did not incorporate data scaling techniques or employ hyperparameter tuning methods. Farjana et al [10] emphasis on predicting CKD through ML methods utilizing the

datasets. The scholars addressed missing information by imputing average numerical data and implemented withhold authentication. Light GBM showcased highest proficiency; however, the research did not incorporate advanced outlier handling, imputation techniques, feature selection, data scaling, or model optimization. Islam et al [11] describes ML methods were employed to estimate CKD. The study utilized mode and mean methods to handle missing information and applied principal component analysis for feature selection and recursive feature elimination. Hassan et al. [12] focuses by using ML to predict CKD by analyzing patients' clinical records. Predictive mean matching was employed to impute missing data, followed by K-means clustering analysis on the dataset. The researchers adopted the Xtreme Gradient Boosting (XGB) approach alongside SHapley Additive exPlanations(SHAP) value evaluation for feature selection. Nonetheless, scaling methods or hyperparameter optimization were not integrated into the study.

Kaur et al. [13] discusses ML methods for CKD prediction where MCAR test was performed to investigate the absence of data, while the feature selection process employed the Ant Colony Optimization algorithm. Ensemble methods were applied, with bagging yielding the most promising outcomes. Nonetheless, the research did not incorporate cross-validation, scaling techniques or hyperparameter optimization methods. P. Chittora et al. [14] describes rapid expansion of electronic healthcare databases is directly contributing to the increasing prevalence of ML methods in the healthcare sector. A further approach based on Recursive Feature Elimination (RFE) is presented by E. M. Senan et al. [15]. The RFE approach was utilized to select the CKD traits that were most highly representative. To categorize the characteristics, KNN, SVM, RF and decision trees were employed. All techniques yielded promising results, and all classifier factors were adjusted to deliver the best classification outcomes. The RF technique outperformed all other approaches by all parameters. The framework underwent scrutiny and assessment via multiclass statistical methods. Observed outcomes from Decision Tree (DT), KNN and SVM methods disclosed significant performance measurement of 99.17%, 98.53%, and 96.67%. Hossain, M. [16] describes several feature optimization strategies were detailed to examine their impact on the efficiency of the ML algorithms, which was evaluated on five prominent classification methods, in order to properly select the feature subset. Experiments using XGB, RF, KNN, SVM and Logistic Regression have demonstrated that the model's accuracy can be improved by employing a Linear Discriminant Analysis (LDA) feature selection tool that yields the best possible outcome. Hussianzadah [17] introduces a diagnostic method for CKD employing four distinct classifiers: SVM, DT classifier, Naive Bayes classifiers and Multi-Layer Perception (MLP). These classifiers were employed to three individual datasets, each containing corresponding values 14, 12, and 13. The SVM classifier, which achieved an accuracy of 91%, was of particular interest in the study. Krishnamurthy et al. [18] discusses several AI models for CKD prediction. The LightGBM approach identified crucial attributes for CKD forecast, including age, gout, angiotensins and sulfonamides. Convolutional Neural Networks exhibited the most superior Proficiency, achieving the maximum AUROC metric of 0.954 evaluated to all other methods. Singh et al. [19] introduces a novel hybrid technique for diagnosing CKD and attained a prediction accuracy of 92.5%. Ifraz et al. [20] focusses on the dataset and employed DT, KNN and LR methods to train three distinct methods for CKD forecast. LR attained a superior performance rate of 97% in contrast to DT with 96.25% and k-NN with 71.25%.

Research Methodology

The research employed the CKD dataset [21]. This dataset has 523 rows and 27 columns is shown in figure.1. The result of the column "class" has two distinct potential values: "1" and "0." A number of "0" signifies that the patient does not have CKD, whereas a value of "1" suggests that CKD. Prior to pre-processing, there were 380 total CKD data and 143 total non-CKD data points. Figure 2. illustrates how age is grouped according to gender.

	ID	age	Gender	bp	sg	al	su	rbc	pc	pcc	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	1	48	Male	80	1.020	1.0	0.0	normal	normal	notpresent	...	44.0	7800.0	5.2	yes	yes	no	good	no	no	ckd
2	3	62	Male	80	1.010	2.0	3.0	normal	normal	notpresent	...	31.0	7500.0	4.8	no	yes	no	poor	no	yes	ckd
3	4	48	Female	70	1.005	4.0	0.0	normal	abnormal	present	...	32.0	6700.0	4.7	yes	no	no	poor	yes	yes	ckd
4	5	51	Female	80	1.010	2.0	0.0	normal	normal	notpresent	...	35.0	7300.0	4.6	no	no	no	good	no	no	ckd
5	6	60	Female	90	1.015	3.0	0.0	normal	normal	notpresent	...	39.0	7800.0	4.4	yes	yes	no	good	yes	no	ckd
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
605	606	55	Male	80	1.020	0.0	0.0	normal	normal	notpresent	...	47.0	6700.0	4.9	no	no	no	good	no	no	notckd
606	607	42	Male	70	1.025	0.0	0.0	normal	normal	notpresent	...	54.0	7800.0	6.2	no	no	no	good	no	no	notckd
607	608	12	Male	80	1.020	0.0	0.0	normal	normal	notpresent	...	49.0	6600.0	5.4	no	no	no	good	no	no	notckd
608	609	17	Male	60	1.025	0.0	0.0	normal	normal	notpresent	...	51.0	7200.0	5.9	no	no	no	good	no	no	notckd
609	610	58	Male	80	1.025	0.0	0.0	normal	normal	notpresent	...	53.0	6800.0	6.1	no	no	no	good	no	no	notckd

523 rows × 27 columns

Figure.1 CKD dataset of 523 rows and 27 columns

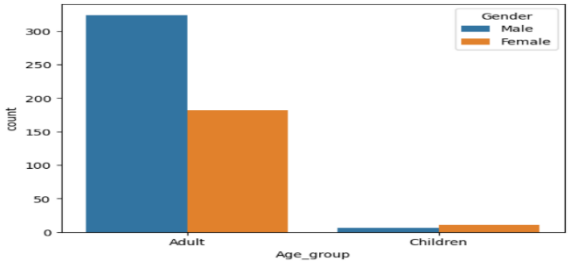

Figure 2 Grouping the age by children and adult

Measurement of creatinine levels to classify the status of CKD

EGFR: A quick and easy method for labs to assist healthcare professionals in identifying CKD serves to use a traceable equation to calculate the EGFR using serum creatinine. Providers may be able to determine that CKD is present even in cases when a patient's blood creatinine concentration seems to be within or slightly above the standard baseline interval using the attributes of age, gender, and racial factors included in CKD-EPI equations. The recommendations for figuring out the adult GFR equation depending on whether a patient has CKD or not. A two-slope "spline" is used in the CKD-EPI equation to represent a link between serum creatinine (Sc) and GFR. The equation also takes into account a number of other factors, such as ethnicity, sex, and age. Consequently, the formulas (1) and (2) that have been stated using different parameters to compute eGFR are provided below.

$$eGFR = 141 \times (Sc/S, 1)^{\alpha} \times (0.993)^{Age} \quad \text{..... (1)}$$

$$eGFR = 144 \times (Sc/S, 1)^{\alpha} \times (0.993)^{Age}$$ ..... (2)

Where,

S represents the kappa factor is 0.9 for men and 0.7 for women and α represents Sex alpha factor

BUN-CR: It uses two variables, creatinine and blood urea nitrogen, both of which are tested in serum. This can be dependent on kidney function. The BUN-CR formula is represented in equation 3 which is expressed in (mg/dL).

BUN-CR = BUN / serum creatinine .....(3)

When BUN-CR is high, it indicates a prerenal etiology; when BUN-CR is less than 10, it indicates an intrinsic renal cause. The value of creatinine levels lies between 0.7 and 1.3 mg/dL. Figures 3 and 4 illustrate how the EGFR and BUN-CR tests are conducted in considering these variables in order to limit the exact Sc concentration that each patient possesses.

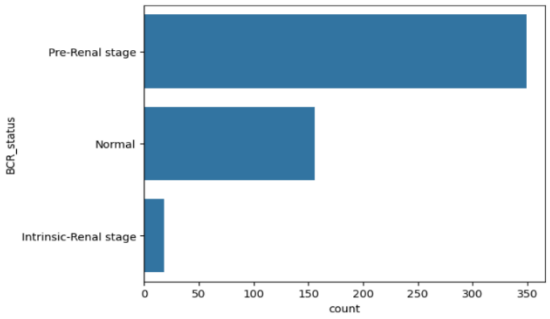

Figure 3 The status of BCR to detect CKD stages

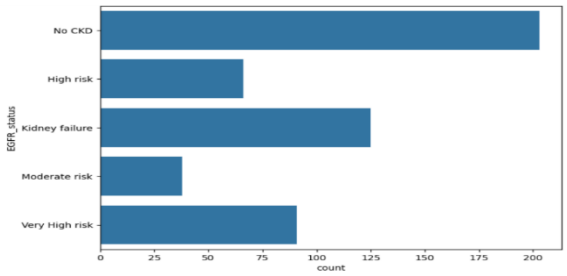

Figure 4 The status of EGFR to detect CKD stages

Then preprocessing the data involves identifying any missing values within the dataset, which can be achieved using the null(). sum() function in the Pandas library. This function helps count the occurrences of missing values. In order to manage scaling the all-variable unit as unique, the data is pre-processed using RobustScaler and label encoder following missing imputation. This dataset includes categorical elements like Male and Female in its Gender column. Since the data consists of string labels, these labels do not have an optimal order, and ML algorithms have presumed incorrectly that the labels have a hierarchy. Label encoding is one method for solving this issue; in this method, each label is given a number value, such as Male and Female being mapped to 0 and 1. But since our model will start to favor the Female parameter more when 1>0, this could add bias into it, even though in the ideal scenario, the two labels would have equal weight in the dataset. We are going to apply the One Hot Encoding method to address this problem. The categorical parameters in One Hot Encoding will set up distinct columns for the labels of Male and Female. Therefore, the value in the Male column will be 1 and the value in the Female column will be 0, and vice versa, wherever there is a Male. Figure.5 illustrates the overall proposed diagram.

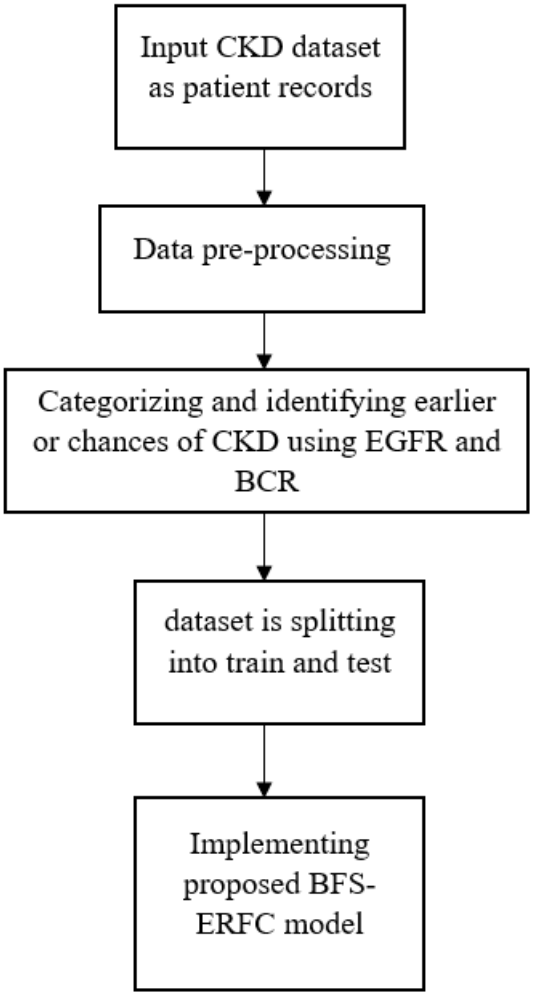

Figure 5 Proposed flow diagram of proposed BFS-ERFC model

Boruta Feature Selection (BFS) technique with Ensemble Based Random Forest Classifier (ERFC)

The RF classification technique included in the RF package is wrapped in the boruta algorithm. The RF algorithm provides a numerical approximation of the feature relevance, can typically be used without the need for parameter modification, and is quite fast. It is an ensemble method in which several unbiased weak classifier decision trees are voted on to do classification. These trees are separately constructed using various training set bagging samples. The significance measure of an attribute is the reduction in classification accuracy caused by the random permutation of its values among objects. It is determined separately for every tree in the forest that uses a particular characteristic to categorize data. Next, the accuracy loss's average and Standard Deviation (SD) are calculated. As an alternative, the important metric can be the Z score, which is designed by dividing the average loss by its SD. Unfortunately, because the Z score's distribution is not N (0,1), it is not directly correlated with the statistical significance of the feature importance returned by the RF technique. However, as the Z score takes into account variations in the average accuracy loss among the forest's trees, we utilize it as the significance metric in Boruta.

Since Z score cannot be used to determine importance directly, we must utilize an outside reference to determine whether an attribute's significance—that is, whether its importance can be distinguished from chance fluctuations—is required. In order to do this, we have included random attributes by design to the information system. Then establish a comparable "shadow" attribute for every attribute, the values of which are derived by shuffle the original attribute among objects. Next, we classify this extended system using all of its attributes and determine how essential each attribute. Only random fluctuations can cause a shadow attribute's relevance to be nonzero. In order to determine which qualities are actually essential, the set of importance of shadow attributes is consulted. The RF classifier's stochasticity causes variations in the significance measure itself. Furthermore, it is sensitive to the existence of unimportant attributes—including shadow ones—in the information system. Furthermore, it is reliant on how specific shadow properties are realized. Consequently, in order to get statistically meaningful findings, we must repeat the reshuffle process. To put it briefly, Boruta is predicated on the same theory that behind the RF classifier: one can lessen the misleading impact of random fluctuations and correlations by introducing randomization into the system and gathering data from the ensemble of randomized samples. In this case, the additional unpredictability will help us which characteristics are truly crucial.

The steps of the Boruta algorithm are as follows:

1. Makes a duplicate of the features from the training set and combines it with the original features.
2. To eliminate any link of any sort between these synthetic features and the target variable y, random permutations are created on them. In essence, these synthetic features are randomized permutations of the actual feature that they are derived from.
3. Each time a synthetic characteristic is used, it is randomly generated.
4. Calculates the z-score for each original and synthetic feature at each iteration. If a feature's importance exceeds the total significance of all synthetic features, it is deemed relevant.
5. It performs a statistical test to every one of the original attributes and retains track of the outcome. The greatest significance of synthetic characteristics equals the significance of a feature, according to the null hypothesis. The original and synthetic features are tested for equivalence using a statistical test. When a feature's relevance is noticeably greater or lower than one of the synthetic features, the null hypothesis is rejected.
6. Eliminates features from the synthetic and original datasets that are deemed irrelevant.
7. After n iterations, repeat all the procedures until all features are eliminated or deemed significant.

Boruta's implementation in Python

Boruta uses its own library to function in Python. The load_CKD() dataset from Sklearn.datasets will be used to evaluate Boruta on a regression issue. When the script is executed, the terminal will display the graphic that illustrates how Boruta is constructing its inferences in figure.6

```
Iteration:      9 / 10
Confirmed:      14
Tentative:      5
Rejected:       13


BorutaPy finished running.

Iteration:      10 / 10
Confirmed:      14
Tentative:      3
Rejected:       13
```

Figure 6 The final result of Boruta's ten iterations on the Sklearn CKD dataset

```
------Support and Ranking for each feature------
Passes the test:  age  - Ranking:  1
Doesn't pass the test:  Gender  - Ranking:  7
Doesn't pass the test:  bp  - Ranking:  2
Passes the test:  sg  - Ranking:  1
Passes the test:  al  - Ranking:  1
Doesn't pass the test:  su  - Ranking:  7
Doesn't pass the test:  rbc  - Ranking:  3
Doesn't pass the test:  pc  - Ranking:  4
Doesn't pass the test:  pcc  - Ranking:  10
Doesn't pass the test:  ba  - Ranking:  15
Passes the test:  bgr  - Ranking:  1
Passes the test:  bu  - Ranking:  1
```

Figure 7 Boruta result report

Figure 7 indicates that the features that are most important in constructing our predictive model are age, sg, al, bgr, and bu. Additionally, we use feat_selector.transform(np.array (X)) to filter our dataset and select only the features that are significant to Boruta; this function returns a Numpy array.

```
------Selected Features------

[[48 1.02 1.0 ... 1 71.07749408488438 30.0]
 [62 1.01 2.0 ... 1 39.45729007476052 29.444444444444443]
 [48 1.005 4.0 ... 1 13.295099194436789 14.73684210526316]
 ...
 [12 1.02 0.0 ... 0 153.1029652547277 43.333333333333336]
 [17 1.025 0.0 ... 0 110.1629433601936 50.0]
 [58 1.025 0.0 ... 0 73.60561791734244 16.363636363636363]]
```

Figure 8 attributes chosen by Boruta

Figure 8 illustrates how to feed a chosen set of X features into a RF Regressor model.

Pseudocode:

Step 1: Import the Python's scikit-learn library.

Step 2: Initialize Borutapy

Feat_selector =Borutapy

Verbose =2

Estimator = random forest model

N_estimator = auto

Number of iterations =10

Step 3: train Boruta

Step 4: Initialize x and y must be numpy arrays

Step 5: Print the support and ranking for each feature

```
for i in range(len(feat_selector.support_)):
    if feat_selector.support_[i]:
print ("Passes the test: ", X.columns[i],
     " - Ranking: ", feat_selector.ranking_[i])
else:
print("Doesn't pass the test: ",
X.columns[i], " - Ranking: ", feat_selector.ranking_[i])
```

Step 6: Filter the selected features

```
X_filtered = feat_selector.transform(np.array (X))
```

Step 7: Initialize logistic regression object and fit object

Step 8: Predict labels for test set and assess accuracy

Results and Discussions

Following data cleaning process, there are 523 occurrences in the dataset of patient records related to CKD; 80% of the dataset is used as the training set, and the remaining 20% of dataset is used as the testing set. The best classification according to CKD status, such as "No CKD," "High risk," "Kidney failure," "Moderate risk," and "very high risk," is determined using the suggested BFS with ERFC classifier models. The programming language Python is used to carry out the experiment. Below is a calculation and identification of the Confusion Matrix (CM) parameter for the suggested BFS with ERFC classifier models and the status of each classifier's testing.

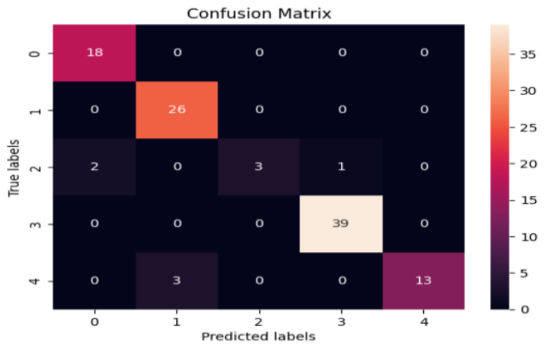

Figure 9 CM for the suggested model's multi-class classification

Figure 9 shows how this kind of categorization can be used to identify five different types of classification difficulties. The CM parameter plays important roles in classification, but they are not as common as binary classification. The suggested BFS with ERFC's the values in the arrays denoting 0 represents No CKD, 1 represents High risk, 2 represents Kidney failure, 3 represents Moderate risk and 4 represents Very High risk respectively. Multi-label binary quality can be evaluated using the Micro F1-score. According to equation 4, micro F1-score measurements that are computed worldwide which have been determined to be equal.

$$Precision = Recall = Micro\ F1 - Score = Accuracy \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

While the five classes are accumulated, the micro accuracy and micro recall can be computed using CM parameters. Therefore, according to equation 4, the suggested BFS with ERFC model's accuracy is said to be 0.92 based on the Micro F1-score. Therefore, the suggested model's accuracy is 0.92 (92.38%). Similarly, Table 1 and Figure 10 calculate and display different classifier techniques.

Table 1 The accuracy of different classifier models using the suggested BFS and ERFC model

Classifier Model	**Accuracy**
RWOM with NN	0.88
RWOM with LR	0.45
Ensemble based Random Forest	0.904
Proposed BFS-ERFC model	0.92

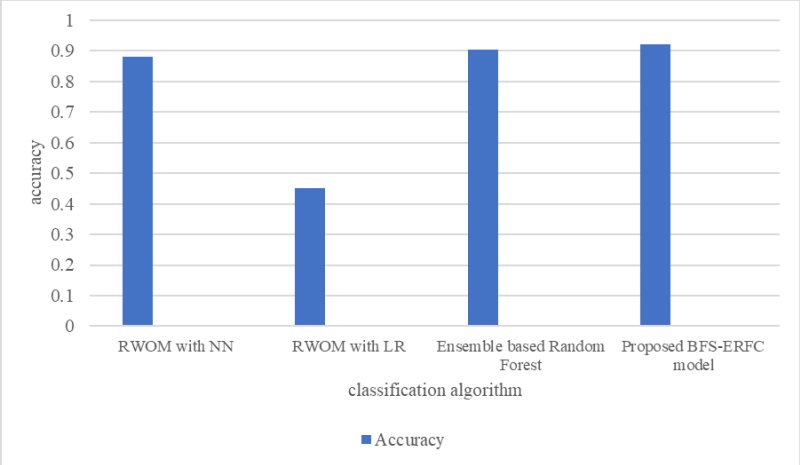

Figure 10 CKD status accuracy score with a classification method

The macro average F1-score for each category in the suggested BFS-ERFC model is determined by evaluating the CM parameters, such as precision, recall, and F1-Score, using the corresponding ML approaches. In the same manner, the weighted score is determined by adding the weights assigned to each category to the total number of samples. Table 2 displays the performance metrics for weighted and macro scores to all ML techniques.

Table 2 Performance metrics for weighted and macro scores for various classifier

ML Technique	Macro Avg precision score	Macro Avg Recall score	Macro Avg F1-score value	Weighted Avg Precision Score	Weighted Avg Recall Score	Weighted Avg F1-Score value
RWOM with NN	0.84	0.82	0.82	0.91	0.89	0.89
RWOM with LR	0.37	0.32	0.33	0.54	0.46	0.48
Ensemble based Random Forest	0.97	0.82	0.87	0.97	0.90	0.93
Proposed BFS-ERFC model	0.97	0.84	0.89	0.97	0.92	0.94

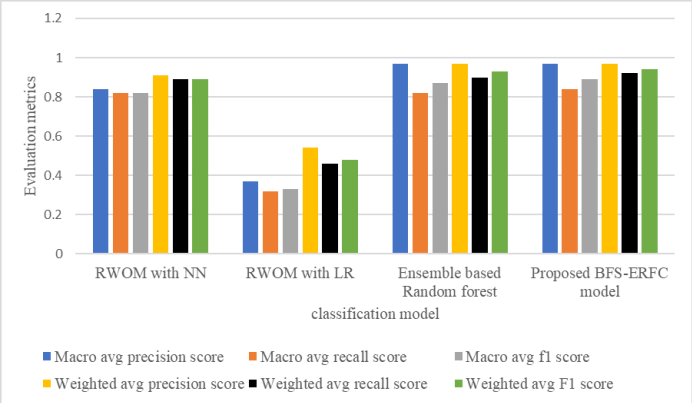

Figure 11 Performance metrics for weighted and macro scores for classify CKD status

The macro avg and weighted avg for the suggested model using different ML approaches are shown in Figure 11. Compared to previous ML techniques, the suggested BFS-ERFC model performs better in terms of macro and weighted avg precision, recall, and F1-Score.

Conclusion

The main objective of this research is to analyse the precise diagnosis of CKD and to determine the likelihood of CKD that may be justified by utilizing EGFR and BCR to identify and evaluate the degree of kidney disease before beginning therapy. Determining the risk of CKD with normal, pre-Renal, or intrinsic renal stages is necessary to define the stage of CKD, comprehend its severity, and identify patients who are in an early stage of the disease using eGFR and BCR. The proposed BFS-ERFC model has been introduced in this research to increase classification accuracy and facilitate efficient CKD diagnosis. As a result, the suggested Boruta algorithm has the advantage of being an easy-to-use and effective method that they should integrate into their pipeline. The study has improved the accuracy of CKD status classification by identifying and detecting the accuracy of these BFS-ERFC models using a kidney disease dataset. Based on BCR and EGFR scores, the suggested Boruta algorithm with Random Forest Classifier has been able to define the status of CKD with a high degree

of accuracy—92.38%. Additionally, help define the degree of CKD severity so that decisions about the appropriateness of treatment can be made.

Reference

[1] K. B. Naidu, B. R. Prasad, S. M. Hassen, et al., "Analysis of Hadoop log file in an environment for dynamic detection of threats using machine learning," Elsevier Measurement: Sensors, vol. 24, pp. 1–5, 2022.

[2] U. Ekanayake and D. Herath, "chronic kidney disease prediction using machine learning methods," in Proc. 2020 Moratuwa Engineering Research Conference (MERCon), 2020, pp. 260–265.

[3] Mortazavi, B. J. et al. Analysis of machine learning techniques for heart failure readmissions. Circ. Cardiovasc. Qual. Outcomes 9, 629–640. https://doi.org/10.1161/CIRCOUTCOMES.116.003039 (2016).

[4] Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PLoS ONE 12, e0174944. https://doi.org/10.1371/journal.pone.0174944 (2017).

[5] Hsu, H.H., Hsieh, C.W., Lu, M.D.: Hybrid feature selection by combining filters and wrappers. Expert Syst. Appl. 38(7), 8144–8150 (2011)

[6] Shahinda Mohamed Mostafa Elkholy, Amira Rezk, and Ahmed Abo El Fetoh Saleh, "Early Prediction of Chronic Kidney Disease Using Deep Belief Network" Volume 9,2021, Digital Object Identifier 10.1109/ACCESS.2021.3114306.

[7] Poonia RC, et al. Intelligent Diagnostic Prediction and Classification Models for Detection of Kidney Disease. Healthcare. 2022; 10:2.

[8] Swain, D.; Mehta, U.; Bhatt, A.; Patel, H.; Patel, K.; Mehta, D.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A.; Manika, S. A Robust Chronic Kidney Disease Classifier Using Machine Learning. Electronics 2023, 12, 212.

[9] Ullah, Z.; Jamjoom, M. Early detection and diagnosis of chronic kidney disease based on selected predominant features. J. Healthc. Eng. 2023, 2023, 3553216.

[10] Farjana, A.; Liza, F.T.; Pandit, P.P.; Das, M.C.; Hasan, M.; Tabassum, F.; Hossen, M.H. Predicting Chronic Kidney Disease Using Machine Learning Algorithms. In Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference, Las Vegas, NV, USA, 8–11 March 2023; pp. 1267–1271.

[11] Islam, M.A.; Majumder, M.Z.H.; Hussein, M.A. Chronic kidney disease prediction based on machine learning algorithms. J. Pathol. Inform. 2023, 14, 100189.

[12] Hassan, M.M.; Hassan, M.M.; Mollick, S.; Khan, M.A.R.; Yasmin, F.; Bairagi, A.K.; Raihan, M.; Arif, S.A.; Rahman, A. A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records. Hum. -Centric Intell. Syst. 2023, 3, 92–104.

[13] Kaur, C.; Kumar, M.S.; Anjum, A.; Binda, M.B.; Mallu, M.R.; Al Ansari, M.S. Chronic Kidney Disease Prediction Using Machine Learning. J. Adv. Inf. Technol. 2023, 14, 384–391.

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

[15] E. M. Senan, M. H. Al-Adhaileh, and F. W. Alsaade, "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques," Journal of Healthcare Engineering, volume 2021, https://doi.org/10.1155/2021/1004767.

[16] M. M, Hossain, Reshma. A. Swarna, "Analysis of the performance of feature optimization techniques for the diagnosis of machine learningbased chronic kidney disease," Machine Learning with Applications, https://doi.org/10.1016/j.mlwa.2022.100330.

[17] Hosseinzadeh, M., Koohpayehzadeh, J., Bali, A. O., Asghari, P., Souri, A., Mazaherinezhad, A., and Rawassizadeh, R. (2021). A diagnostic prediction model for chronic kidney disease in internet of things platform. Multimedia Tools and Applications, 80(11),16933-16950.

[18] Krishnamurthy S., Ks K., Dovgan E., Lustrek M., Piletic B.G., Srinivasan K., Li Y.-C., Gradišek A., Syed-Abdul S. Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. Healthcare. 2021;9:546. doi: 10.3390/healthcare9050546.

[19] Singh V., Jain D. A Hybrid Parallel Classification Model for the Diagnosis of Chronic Kidney Disease. Int. J. Interact. Multimed. Artif. Intell. 2021 doi: 10.9781/ijimai.2021.10.008.

[20] Ifraz, G.M.; Rashid, M.H.; Tazin, T.; Bourouis, S.; Khan, M.M. Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods. Comput. Math. Methods Med. 2021, 2021, 6141470.

[21] Kaggle,“Chronic Kidney Disease Dataset,” https://www.kaggle .com/abhia1999/chronic-kidney-disease