SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, VOL.03 NO. 2 2024
DOI: 10.51984/SUCP.V3I2.3331

وقائع مؤتمرات جامعة سبها
Sebha University Conference Proceedings

Confrence Proceeeding homepage: http://www.sebhau.edu.ly/journal/CAS

Predicting Chronic Kidney Disease Using Filter and Wrapper Feature Selection Methods with Machine Learning Techniques

*Mohammed Shantal [a,c], Almahdi Alshareef [b,c], Omar Ahmid [a]

[a] Computer Science Department, College of Technology Science, Sebha, Libya
[b] Computer Science Department, Sebha University, Sebha, Libya
[c] Information Development Center, Sebha University - Sebha, Libya

Keywords:
Feature Selection
kidney prediction
Filter selection
wrapper selection
f_classif
chi^2
RFE
RFECV

ABSTRACT

Chronic kidney disease (CKD) is a condition characterized by the gradual loss of kidney function over months or years. Predicting this disease is a crucial issue in the medical field. Therefore, an automated tool utilizing Machine Learning (ML) techniques to assess a patient's kidney condition would be beneficial for doctors in predicting CKD and improving treatment. In the ML process, the preprocessing stage is a vital step that enhances data quality. Feature selection, a key preprocessing method, removes irrelevant or redundant features, thereby simplifying the model and reducing the number of features. This research explores the potential of various feature selection methods. The feature selection methods are categorized into filter methods (f_classif, chi^2) and wrapper methods (Recursive Feature Elimination with Cross-Validation RFECV) using Random Forest classifier and Support Vector Machine. The accuracy of classifiers was used to evaluate the performance of the full dataset compared to subsets created using feature selection (FS). The results show that the RFECV+SVM feature selection method outperforms others, yielding the best performance by improving accuracy in 5 out of 6 classifiers.

التنبؤ بمرض الكلى المزمن باستخدام طرق اختيار الميزات بالتصفية والتغليف مع تقنيات التعلم الآلي

*محمد شنتال[1,3] و المهدي الشريف[2,3]و عمر أحميد[1]

[1] قسم الحاسوب، كلية العلوم التقنية ،سبها، ليبيا
[2]قسم الحاسوب، كلية تقنية المعلومات، جامعة سبها ، ليبيا
[3] مركز التطوير المعلوماتي، جامعة سبها، ليبيا

الكلمات المفتاحية:
Feature Selection
اختيار الميزات
التنبؤ بمرض الكلى
wrapper selection
f_classif
chi^2
RFE
RFECV

الملخص

مرض الكلى المزمن (CKD) هو حالة تتميز بفقدان تدريجي لوظيفة الكلى على مدى شهور أو سنوات. يُعد التنبؤ بهذا المرض قضية حيوية في المجال الطبي. لذلك، فإن أداة آلية تستخدم تقنيات التعلم الآلي لتقييم حالة الكلى لدى المريض ستكون مفيدة للأطباء في التنبؤ بمرض الكلى المزمن وتحسين العلاج. في عملية التعلم الآلي، تعتبر مرحلة المعالجة المسبقة خطوة أساسية لتحسين جودة البيانات. يُعد اختيار الميزات إحدى طرق المعالجة المسبقة الرئيسية، حيث يزيل الميزات غير ذات الصلة أو الزائدة، وبالتالي يبسط النموذج ويقلل من عدد الميزات. يستكشف هذا البحث إمكانات طرق اختيار الميزات المختلفة. تصنف طرق اختيار الميزات إلى طرق التصفية (f_classif، chi^2) وطرق الالتفاف (إزالة الميزات التكرارية مع التحقق المتقاطع (RFECV) باستخدام مصنف الغابات العشوائية وSVC. تم استخدام دقة المصنفات لتقييم أداء مجموعة البيانات الكاملة مقارنة بالمجموعات الفرعية التي تم إنشاؤها باستخدام اختيار الميزات. تظهر النتائج أن طريقة اختيار الميزات RFECV+SVM تتفوق على غيرها، حيث تقدم أفضل أداء من خلال تحسين الدقة في 5 من أصل 6 مصنفات.

*Corresponding author:
E-mail addresses: moh.shantal@sebhau.edu.ly, (A. Alshareef) alm.alshareef@sebhau.edu.ly, (O. Ahmid) oma.ahmid@sebhau.edu.ly
Article History : Received 13 June 2024 - Received in revised form 07 September 2024 - Accepted 06 October 2024

1. Introduction

Kidney disease, also referred to as kidney failure, ranks among the significant global health concerns. Numerous medical professionals invest considerable time and effort in diagnosing this condition through frequent lab visits and consultations, aiming to ascertain its presence in patients [1]. Early identification and treatment frequently mitigate the worsening of CKD [2].

Data mining (DM) involves uncovering concealed insights from vast datasets. It finds applications across diverse domains, including image, text, sequential, web, graph, and spatial mining. DM techniques serve various purposes such as fault diagnosis, anomaly detection, medical diagnosis, e-mail filtering, face recognition, and oil spill detection [3]. Several studies have explored the application of ML algorithms to CKD prediction, aiming to develop robust and efficient models for identifying patients at risk. Baidya et al. [4] proposed a method utilizing eight ML algorithms to rapidly detect CKD based on patients' health data, achieving promising results in terms of accuracy and performance. Similarly, [5] introduced an ML approach for diagnosing CKD, highlighting the effectiveness of Random Forest (RF) algorithms in achieving high diagnostic accuracy.

Khan et al. [6] conducted experimental analysis of various ML techniques to categorize CKD patients, showcasing the superior performance of Composite Hypercube on Iterated Random Projection (CHIRP) in terms of accuracy and error reduction. These studies underscore the potential of ML-based approaches in enhancing CKD diagnosis and prognosis.

Additionally, Ekanayake et al. [7] presented a workflow for predicting CKD status using clinical data, emphasizing the importance of data preprocessing and FS in developing accurate prediction models. Similarly, Tikariha et al. [8] analyzed various DM techniques for CKD prediction, with the *k*-nearest neighbor (*k*-NN) algorithm demonstrating superior accuracy compared to other methods.

Moreover, Sinha et al. [9] proposed a decision support system for forecasting CKD, evaluating the performance of SVM and *k*-NN classifiers. The study concluded that *k*-NN classifier outperformed SVM in terms of accuracy and execution time, further highlighting the potential of ML techniques in CKD prediction.

Finally, Rabby et al. [10] introduced a method for real-time kidney disease prediction, monitoring, and application, demonstrating the effectiveness of ML techniques such as Decision Tree (DT) Classifier and Gaussian Naive Bayes (NB) in achieving high accuracy and recall scores.

Almansour et al. [11] explored various machine learning classification algorithms using a dataset of 400 patients and 24 attributes relevant to CKD diagnosis. They primarily focused on Artificial Neural Networks (ANN) and Support Vector Machine (SVM). Initially, missing values were replaced with attribute means. Through extensive parameter tuning, the study optimized ANN and SVM models. Empirical results revealed that ANN surpassed SVM, achieving accuracies of 99.75% and 97.75%, respectively. This underscores the potential of these methodologies in medical diagnosis.

[12] focused on predicting whether patients have CKD using various ML classification algorithms. They developed multiple models using different algorithms to differentiate between CKD and non-CKD statuses. After comparing the outcomes, it was determined that the model using the Multiclass Decision Forest algorithm performed the best, achieving an accuracy of 99.1% on a reduced dataset with 14 attributes.

Polat, et al. [13] explored ML techniques for CKD diagnosis, focusing on FS methods to reduce dataset dimensions. Using SVM classification and two key FS approaches, wrapper and filter methods, they achieved the highest accuracy rate (98.5%) with the filtered subset evaluator and Best First search engine. Shrivas [14] developed a robust ensemble model for CKD diagnosis, employing RF, CART, and SVM classifiers. They used various ranking-based Feature Selection Techniques (FST) and a proposed Union Based FST, finding that the ensemble model with the proposed FST outperformed existing methods and individual classifiers. Atallah, et al. [15] introduced an intelligent prediction method for kidney transplantation outcomes using DM techniques. Their method combines three feature selectors—gain ratio, NB, and genetic algorithms—and modifies the *k-NN* algorithm. The results demonstrated superior performance compared to existing methods

In this study, the impact of Filter and Wrapper Feature Selection methods on CKD prediction has been extensively examined. The study aims to synthesize and critically evaluate existing feature selection methods applied in ML techniques for CKD prediction, with a focus on their effectiveness and the reported performance metrics. By providing a comprehensive overview of current methodologies, this study seeks to identify the most effective feature selection methods that enhance CKD prediction accuracy.

2. Recent work

Elhoseny, et al. [16] developed a CKD diagnosis system using density-based FS and Ant Colony Optimization (ACO). Their system, employing wrapper methods for FS, showed higher classification accuracy with fewer features when tested with a benchmark CKD dataset. [17, 18] both emphasized the importance of FS in medical DM for CKD diagnosis. M & Balakrishnan proposed an Improved Teacher Learner Based Optimization (ITLBO) algorithm, achieving a 36% feature reduction and improved classification accuracy. Dey et al. used a hybrid FS approach combining Chi-squared test (Chi^2), Mutual Information (MI), and Pearson correlation matrix, achieving 98% accuracy with the Extra Trees classifier. [19, 20] both focused on developing ML-based diagnosis systems for early CKD detection. Senan et al. used Recursive Feature Elimination (RFE) for FS and evaluated four classification algorithms, with the RF algorithm achieving 100% accuracy. Singh et al. used RFE and a deep neural network, which outperformed other classifiers, achieving 100% accuracy. [21] evaluated a bagging ensemble technique combined with FS on a CKD dataset. Using the RF algorithm for FS and ensemble aggregation of NB, KNN, and DT classifiers, they achieved 100% accuracy in CKD diagnosis. [22, 23] both utilized CBFS for FS in CKD prediction. Hassan [22] combined CBFS with Principal component analysis (PCA) and found the ensemble learning algorithm achieved the highest performance. [23] found that CBFS and Chi^2 improved model performance, with Sequential Minimal Optimization and Multi-layer Perceptron (MLP) achieving the highest accuracy. [24] examined CKD detection using RF and Logistic Regression(LR) classifiers, with FS through correlation analysis. RF achieved the highest accuracy and F1 score, and a predictive app was developed for CKD presence determination. [25] proposed a method combining information-gain-based FS and a cost-sensitive AdaBoost classifier for efficient CKD detection, achieving superior performance with 99.8% accuracy. [26] diagnosed CKD using a hybrid ML model with Pearson correlation for FS. The hybrid model, combining Gaussian NB, gradient boosting, and decision tree classifiers, achieved 100% accuracy. [27] focused on CKD diagnosis using ML algorithms. Mehta et al. emphasized feature extraction via PCA and feature selection using Lasso regularization, with Naive Bayes showing the highest accuracy and sensitivity. Hema et al. investigated the impact of Exhaustive Feature Selection (EFS) on various classifiers, enhancing KNN accuracy from 77% to 83%, promoting early CKD diagnosis and healthy lifestyle adoption. Mamatha and Terdal [28] proposes a deep learning system with a CNN architecture and Grasshopper Optimization Algorithm (GOA) for feature selection, enhancing early CKD detection accuracy, achieving improved predictive performance and interpretability in identifying key CKD-related features. [29] study aims to develop a forecasting model for early CKD detection using machine learning classifiers like GB, XGBoost, DT, RF, and KNN, emphasizing the impact of Exhaustive Feature Selection (EFS) on predictive accuracy. Experiments on standard and real-time datasets showed improved performance, measured by Accuracy, Precision, Recall, and F1-score, validating the proposed approach.

Yashwante, et al. [30] evaluates the impact of feature extraction methods (LDA, PCA, ICA) and meta-heuristic feature selection techniques (PSO, ACO, ABC) on CKD prediction using classifiers like ANN, RF, MLP, and KNN. Results show that meta-heuristic optimization improves model performance by around 19% compared to feature extraction methods, addressing overfitting and underfitting issues, with evaluations based on accuracy and AUC-ROC scores.

3. Materials and methods

In this study, feature selection methods were applied to assess their impact on the accuracy of DKD classification. CKD caused by diabetes is also known as diabetic kidney disease (DKD) [31]. Figure 1 illustrates the block diagram of the proposed research. Initially, all categorical data in the dataset were converted to numerical values using label encoding; for example, 'not-CKD' and 'CKD' were encoded as 0 and 1, respectively. To clean the dataset, the following preprocessing steps were then applied: missing data imputation was performed using *k*-NN imputation to address the dataset's missing values [32]. Additionally, Min-Max normalization was employed to standardize the dataset, ensuring equal contribution from each feature. Four feature selection methods were compared to the full dataset. These methods, sourced from the scikit-learn library, include kBest using f_classif, which selects the top k features based on the ANOVA F-value; kBest using chi, which selects the top k features based on the Chi-squared test; Recursive Feature Elimination with Cross-Validation (RFECV) using RF, which iteratively eliminates features and evaluates performance using a Random Forest classifier; and RFECV using SVM, which performs the same process using a Support Vector Machine classifier. kBest using f_classif and kBest using chi are filter methods, while RFECV using RF and RFECV using SVM are wrapper methods. Six machine learning classification techniques were employed, and their outcomes were compared to identify the best-performing technique for predicting DKD.

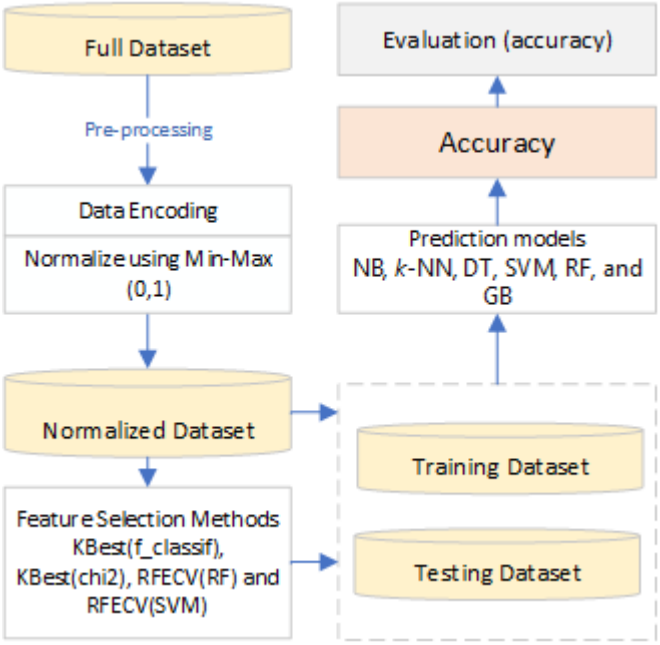

Fig. 1: The methodology

3.1 Datasets

In this study, the Chronic Kidney Disease Dataset from the UCI Machine Learning Repository was utilized [33]. The dataset obtained from 400 patients was sourced from the University of California, Irvine Machine Learning Repository. This dataset includes 24 features, with 11 being numeric and 13 categorical, along with the binary class labels which are 'CKD' and 'not-CKD' for classification purposes [34]. The dataset consists of 250 instances of the 'CKD' class and 150 instances of the 'not-CKD' class. The detailed characteristics of the UCI CKD dataset are presented in Table 1.

Table 1: Dataset description

	Features	Specification	Value
1	age	AGE (IN YEARS)	0 – 90
2	bp	BLOOD PRESSURE	0 – 180
3	sg	SPECIFIC GRAVITY	0 – 1.025
4	al	ALBUMIN	0 – 5
5	su	SUGAR	-
6	rbc	RED BLOOD CELLS	NORMAL, ABNORMAL
7	pc	PUS CELL	NORMAL, ABNORMAL
8	pcc	PUS CELL CLUMPS	PRESENT, NOTPRESENT
9	ba	BACTERIA	PRESENT, NOTPRESENT
10	bgr	BLOOD GLUCOSE RANDOM	0 – 490
11	bu	BLOOD UREA	0 – 391
12	sc	SERUM CREATININE	0 – 76
13	sod	SODIUM	0 – 163
14	pot	POTASSIUM	0 – 47
15	hemo	HAEMOGLOBIN	0 – 17.8
16	pcv	PACKED CELL VOLUME	0 – 54
17	wbcc	WHITE BLOOD CELL COUNT	0 – 26,400
18	rbcc	RED BLOOD CELL COUNT	0 – 8
19	htn	HYPERTENSION	NO, YES
20	dm	DIABETES MELLITUS	NO, YES
21	cad	CORONARY ARTERY DISEASE	NO, YES
22	appet	APPETITE	good/poor
23	pe	PEDAL EDEMA	NO, YES
24	ane	ANAEMIA	NO, YES
25	class	CLASS	CKD / not-CKD

3.2 Pre-processing

Real-world data is often inconsistent, which can affect model performance. Preprocessing the data before feeding it into classifiers is a vital part of developing a machine learning model [35]. In this study, the dataset contains both numerical and categorical data. Label encoding has been applied to convert all categorical features into numerical values, including the dataset labels. Additionally, the dataset contains missing values that need to be handled appropriately. The *k*-NN imputation method has been employed to address the issue of missing data. Finally, normalization is important to scale numerical features before fitting them to any models, as scaling is mandatory for some techniques such as nearest neighbors, SVMs, and deep learning. For normalization, the Min-Max 0-1 scaling method has been applied.

3.2.1. Feature Selection

In this study, three FS methods have been employed to reduce the dimensionality and choose the relevant features. First is "f_classif" which is a statistical technique commonly used in machine learning to select the most relevant features for a predictive model. It evaluates the relationship between each feature and the target variable using analysis of variance (ANOVA). Features with higher ANOVA F-values and lower p-values are considered more important and are selected for the model. This technique helps to improve model efficiency by reducing the number of features while retaining the most informative ones for accurate predictions.

Second is Chi^2: The feature selection method chi^2 is a statistical technique used in machine learning to select the most relevant features for a predictive model, particularly in classification tasks. It measures the dependency between each feature and the target variable using the chi-square statistic. Features with higher chi-square values indicate stronger associations with the target variable and are considered more important for the model. By selecting features with significant chi-square values, this technique helps improve model efficiency and accuracy by focusing on the most informative attributes for classification.

The third method is the feature selection method "RFECV" stands for Recursive Feature Elimination with Cross-Validation. It's a technique used to automatically select the most important features from a dataset while also optimizing the model's performance through cross-validation.

RFECV works by recursively removing features from the dataset and evaluating the model's performance using cross-validation at each step. It ranks the features based on their importance and eliminates the least important ones until the optimal subset of features is identified.

By iteratively selecting the best subset of features and evaluating the model's performance, RFECV helps to improve the efficiency of predictive models while reducing the risk of overfitting.

As the kBest method requires a specified number of features, the study used the number of features selected by RFECV (RF) as the k value for kBest (chi^2) and kBest (f_classif).

3.3 Classifiers

3.3.1. Naïve Bayes (NB):

Naïve Bayes (NB) is a straightforward probabilistic classifier rooted in Bayes theorem. This supervised learning algorithm employs

maximum likelihood estimation, generating a probability distribution by tallying the frequencies of dataset values. It operates under the assumption of attribute independence given the class variable and builds a class model from a finite set. Notably, Naïve Bayes offers the advantage of needing only a small amount of training data to compute the classification parameters [36]

3.3.2. k-Nearest Neighbors (k-NN):

k-Nearest Neighbors (*k*-NN) is a straightforward supervised algorithm applicable to both classification and regression tasks, although it is predominantly utilized for classification. It operates without a distinct training stage, incorporating all available data for training, earning its classification as a lazy learning algorithm. Furthermore, *k*-NN, being nonparametric, disregards underlying data characteristics. It retains the entire dataset as it lacks a specific model, thus necessitating no learning phase. During prediction, it evaluates *k* neighbors, requiring careful selection of *k*'s value. Distances between already labeled data points are computed, typically using the Euclidean method, to determine the nearest neighbor of new data [37].

3.3.3. Decision Tree (DT):

The DT Classifier Algorithm is employed in artificial intelligence for both categorization and prediction purposes. By utilizing a predefined set of values, one can trace various outcomes or choices within the decision tree structure. The decision tree comprises multiple steps that guide individual decision-making processes. Constructing a decision tree involves two main stages: Induction and Pruning [38]

3.3.4. Support Vector Machine (SVM)

SVM is a linear model utilized for both classification and regression tasks, capable of tackling linear and non-linear problems. It operates by classifying data points using a hyperplane. In SVM, each data point is represented as a point in an n-dimensional space (where n denotes the number of features), with each feature value corresponding to a particular coordinate. Classification is achieved by identifying the optimal hyperplane that effectively separates the two classes [39].

3.3.5. Random forest (RF)

RF finds utility in diverse fields like image classification, recommendation engines, and feature selection. It constructs decision trees using randomly sampled data. Then, it consolidates predictions from these trees and selects the best outcome through voting. A key advantage of this algorithm lies in its ability to provide a reliable measure of feature importance. Remarkably, it effortlessly computes the relative significance of each feature in prediction [40].

3.3.6. Gradient Boosting (GB):

Gradient Boosting is a machine learning technique for classification and regression that enhances model accuracy by sequentially adding weak learners, often starting with regression trees. This ensemble method minimizes the loss function, which measures the difference between expected and actual values, reducing bias and variance. Gradient Boosting is notable for its improved accuracy and the simplicity of its least squares regression setting, making it easier to understand and implement [40].

3.4 Performance Metrics

In our study, we employed the accuracy metric to evaluate the effectiveness of classifiers. The accuracy of the classifier reflects the rate of successful predictions, which is computed using the confusion matrix as outlined in Equation (1).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} * 100$$

All the code was developed using Python 3.7 in Jupyter Notebook (Anaconda 3). The SciKitLearn library was utilized to implement all the methods in this study to assess the impact of feature selection (FS) methods on performance and class outcome enhancement. Six classification techniques were employed: NB, *k*-NN, DT, SVM, RF, and GB. All experiments were conducted on numerical datasets.

4. Result and discussion

This paper presents the results in two sections: the first section focuses on analysing the selected features, highlighting the features chosen by different feature selection methods. The second section discusses the impact of these feature selection techniques on the performance of various classification methods, providing insights into how the selected features influence the accuracy and efficiency of the predictive models.

4.1 Analysis of Selected Features

The feature selection methods— kBest (chi^2), kBest (f_classif), RFECV(SVM), and RFECV(RF) produced varying sets of features significant for CKD prediction. The kBest (chi^2) method selected 16 features: sg, al, su, rbc, pc, pcc, ba, bgr, hemo, pcv, htn, dm, cad, appet, pe, and ane. The kBest (f_classif) method also identified 16 features: sg, al, su, rbc, pc, bgr, bu, sc, sod, hemo, pcv, rbcc, htn, dm, appet, and pe. In contrast, the RFECV(SVM) method narrowed its selection to 7 key features: sg, al, rbc, hemo, pcv, dm, and appet, emphasizing a smaller set of the most predictive attributes. The RFECV(RF) method, with 16 selected features, focused on a broader range, including bp, sg, al, rbc, pc, bgr, bu, sc, sod, hemo, pcv, rbcc, htn, dm, appet, and pe.

Among the selected features, seven were consistently chosen across all methods: sg, al, rbc, hemo, pcv, dm, and appet, underscoring their critical importance in CKD prediction models. Additionally, several features, such as pe and bgr, were frequently selected by three of the four methods, highlighting their relevance even though they were not universally chosen. These shared selections indicate that while each method has its unique focus, there is significant overlap on key features, emphasizing their importance in accurately predicting CKD and guiding effective feature selection strategies in machine learning models.

4.2 Analysis of performance

Table 2 shows the results of the full data compared with four feature selection methods. In the NB classifier, the best performance is achieved with RFECV(SVM) (96.08%), followed by RFECV(RF) (94.88%). Both feature selection methods significantly improve performance compared to using the full dataset (92.85%). f_classif also improves accuracy, while chi slightly decreases it. *k*-NN shows the best performance with RFECV(SVM) (98.15%), with RFECV(RF) and f_classif also improving accuracy compared to the full dataset. chi provides a slight improvement over the full dataset. The Decision Tree performs best with RFECV(SVM) (95.90%), and chi also shows a notable improvement. RFECV(RF) provides a small improvement, while f_classif shows moderate improvement over the full dataset. SVM performs best with RFECV(SVM) (98.30%), followed closely by RFECV(RF) (97.95%). Random Forest has the highest accuracy with the full dataset (98.45%). Feature selection methods f_classif and RFECV(RF) slightly reduce accuracy, while chi and RFECV(SVM) show a more significant decrease. Gradient Boosting performs best with RFECV(SVM) (98.55%), followed by the full dataset (98.13%). Other feature selection methods either slightly decrease accuracy (f_classif) or provide similar performance (chi^2, RFECV(RF)). In conclusion, RFECV(SVM) often yields the highest accuracy across most classifiers, showing significant improvement for NB, DT, SVM, and GB. RFECV(RF) also improves performance but is not as consistently strong as RFECV(SVM). f_classif and chi generally provide moderate improvements, but their impact varies by classifier.

Based on the accuracy of the full dataset across the classifiers, RF and Gradient Boosting show high accuracy even with the full dataset, with only slight variations due to feature selection methods. In addition, NB and DT benefit the most from feature selection, showing substantial accuracy improvements with the right methods.

Table 2: Accuracy of FSs compared across classifiers accuracy.

classifier	Full-dataset	f_classif	chi	rfecv(rf)	rfecv(svm)
NB	92.85%	93.98%	92.28%	94.88%	**96.08%**
***k*-NN**	97.13%	97.93%	97.60%	97.95%	**98.15%**
DT	94.20%	94.95%	95.45%	94.28%	**95.90%**
SVM	97.75%	97.70%	97.63%	97.95%	**98.30%**
RF	**98.45%**	98.38%	98.15%	98.33%	97.70%
GB	98.13%	97.93%	98.03%	98.03%	**98.55%**

5. Conclusion

This study demonstrates the effectiveness of feature selection methods in improving the accuracy of chronic kidney disease (CKD) prediction using ML techniques. By comparing filter methods (f_classif, chi^2) and wrapper methods (RFECV using RF and SVC), it was found that the RFECV+SVM method consistently yielded the highest accuracy across most classifiers, significantly enhancing the performance of NB, DT, SVM, and GB classifiers. Specifically, RFECV(SVM) improved accuracy in 5 out of 6 classifiers, indicating its robustness and effectiveness. While RFECV(RF) also improved performance, it

was not as consistently strong as RFECV+SVM. Filter methods f_classif and chi^2 provided moderate improvements, though their impact varied by classifier. Notably, RF and GB classifiers maintained high accuracy even with the full dataset, with only slight variations due to FS. Overall, the findings highlight the potential of feature selection to refine predictive models and enhance CKD diagnosis, particularly benefiting classifiers such as NB and DT.

6. Acknowledgments

With great pride and honor, the Information Development Center at Sebha University announces the completion of a significant joint scientific research project. This research is part of a series of studies conducted by the center with the aim of contributing to the development of local community institutions, particularly in the field of medicine using artificial intelligence technologies. Researchers from various disciplines contributed to this study, utilizing their skills and expertise in data analysis and algorithm development, resulting in valuable scientific outcomes. This research aims to improve medical performance and provide innovative solutions that aid in disease diagnosis and healthcare delivery more effectively.

We look forward to more future achievements in the field of scientific research and express our gratitude to everyone who contributed to this study, including the team members, administrators, and supporters.

7. Abbreviations and Acronyms

Abbreviation	Meaning
Naïve Bayes	NB
***k*-nearest neighbor**	*k*-NN
Decision Trees	DT
Support Vector Machine	SVM
Random Forest	RF
GB	GB
Chronic kidney disease	CKD
Diabetic Kidney Disease	DKD
Feature Selection	FS
Feature Selection Techniques	FST
Data Mining	DM
Machine Learning	ML

8. References

[1]- R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," *Computational Intelligence and Neuroscience*, vol. 2022, p. 3820360, 2022/04/15 2022, doi: 10.1155/2022/3820360.

[2]- M. A. R. Rahat *et al.*, "Comparing Machine Learning Techniques for Detecting Chronic Kidney Disease in Early Stage," *Journal of Computer Science and Technology Studies*, vol. 6, no. 1, pp. 20-32, 2024.

[3]- S. Vijayarani, S. Dhayanand, and M. Phil, "Kidney disease prediction using SVM and ANN algorithms," *International Journal of Computing and Business Research (IJCBR)*, vol. 6, no. 2, pp. 1-12, 2015.

[4]- D. Baidya, U. Umaima, M. N. Islam, F. M. J. M. Shamrat, A. Pramanik, and M. S. Rahman, "A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 28-30 April 2022 2022, pp. 1305-1310, doi: 10.1109/ICOEI53556.2022.9776876.

[5]- J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease," *IEEE Access*, vol. 8, pp. 20991-21002, 2020, doi: 10.1109/ACCESS.2019.2963053.

[6]- B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy," *IEEE Access*, vol. 8, pp. 55012-55022, 2020, doi: 10.1109/ACCESS.2020.2981689.

[7]- I. U. Ekanayake and D. Herath, "Chronic Kidney Disease Prediction Using Machine Learning Methods," in *2020 Moratuwa Engineering Research Conference (MERCon)*, 28-30 July 2020 2020, pp. 260-265, doi: 10.1109/MERCon50084.2020.9185249.

[8]- P. Tikariha and P. Richhariya, "Comparative Study of Chronic Kidney Disease Prediction Using Different Classification Techniques," 2018, pp. 195-203.

[9]- P. S. Parul Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM," *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)* vol. Volume 04, no. Issue 12 (December 2015), 2015, doi: http://dx.doi.org/10.17577/IJERTV4IS120622.

[10]- A. K. M. S. A. Rabby, R. Mamata, M. A. Laboni, Ohidujjaman, and S. Abujar, "Machine Learning Applied to Kidney Disease Prediction: Comparison Study," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 6-8 July 2019 2019, pp. 1-7, doi: 10.1109/ICCCNT45670.2019.8944799.

[11]- N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," (in eng), *Comput Biol Med*, vol. 109, pp. 101-111, Jun 2019, doi: 10.1016/j.compbiomed.2019.04.017.

[12]- W. H. S. D. Gunarathne, K. D. M. Perera, and K. A. D. C. P. Kahandawaarachchi, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)," in *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, 23-25 Oct. 2017 2017, pp. 291-296, doi: 10.1109/BIBE.2017.00-39.

[13]- H. Polat, H. Danaei Mehr, and A. Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods," *Journal of Medical Systems*, vol. 41, no. 4, p. 55, 2017/02/27 2017, doi: 10.1007/s10916-017-0703-x.

[14]- A. Shrivas, S. K. Sahu, and H. Hota, "Classification of chronic kidney disease with proposed union based feature selection technique," in *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, 2018, pp. 26-27.

[15]- D. M. Atallah, M. Badawy, and A. El-Sayed, "Intelligent feature selection with modified K-nearest neighbor for kidney transplantation prediction," *SN Applied Sciences*, vol. 1, no. 10, p. 1297, 2019/09/27 2019, doi: 10.1007/s42452-019-1329-z.

[16]- M. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease," *Scientific Reports*, vol. 9, no. 1, p. 9583, 2019/07/03 2019, doi: 10.1038/s41598-019-46074-2.

[17]- M. M and S. Balakrishnan, "Feature Selection Using Improved Teaching Learning Based Algorithm on Chronic Kidney Disease Dataset," *Procedia Computer Science*, vol. 171, pp. 1660-1669, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.procs.2020.04.178.

[18]- S. K. Dey, K. M. M. Uddin, H. M. H. Babu, M. M. Rahman, A. Howlader, and K. A. Uddin, "Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease," *Intelligent Systems with Applications*, vol. 16, p. 200144, 2022.

[19]- E. M. Senan *et al.*, "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques," *Journal of Healthcare Engineering*, vol. 2021, p. 1004767, 2021/06/09 2021, doi: 10.1155/2021/1004767.

[20]- V. Singh, V. K. Asari, and R. Rajasekaran, "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease," *Diagnostics*, vol. 12, no. 1, 2022, doi: 10.3390/diagnostics12010116.

[21]- O. A. Jongbo, T. A. Olowookere, and A. O. Adetunmbi, "Performance Evaluation of an Ensemble Method for Diagnosis of Chronic Kidney Disease with Feature Selection Technique," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*, 8-9 Nov. 2020 2020, pp. 959-965, doi: 10.1109/DASA51403.2020.9317190.

[22]- M. M. Hassan, T. Ahamad, and S. Das, "An Ensemble Learning Approach for Chronic Kidney Disease Prediction Using Different Machine Learning Algorithms with Correlation Based Feature Selection," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, 17-19 Dec. 2022 2022, pp. 242-247, doi: 10.1109/ICCIT57492.2022.10055471.

[23]- A. S. Turiac and M. Zdrodowska, "Data mining approach in

diagnosis and treatment of chronic kidney disease," *acta mechanica et automatica*, vol. 16, no. 3, pp. 180-188, 2022.

[24]- K. P. Babu and S. Noorullah, "Recognition of Chronic Kidney Disease Using Machine Learning," *Journal of Algebraic Statistics*, vol. 13, no. 1, pp. 910-917, 2022.

[25]- S. A. Ebiaredoh-Mienye, T. G. Swart, E. Esenogho, and I. D. Mienye, "A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease," *Bioengineering*, vol. 9, no. 8, 2022, doi: 10.3390/bioengineering9080350.

[26]- H. Khalid, A. Khan, M. Zahid Khan, G. Mehmood, and M. Shuaib Qureshi, "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease," *Computational Intelligence and Neuroscience*, vol. 2023, no. 1, p. 9266889, 2023/01/01 2023, doi: https://doi.org/10.1155/2023/9266889.

[27]- V. Mehta *et al.*, "Machine Learning based Exploratory Data Analysis (EDA) and Diagnosis of Chronic Kidney Disease (CKD)," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 10, 03/22 2024, doi: 10.4108/eetpht.10.5512.

[28]- B. Mamatha and S. P. Terdal, "A Review On Early Detection Of Chronic Kidney Disease," *Journal of Scientific Research and Technology*, pp. 35-43, 2024.

[29]- K. Hema, K. Meena, and R. Pandian, "Analyze the impact of feature selection techniques in the early prediction of CKD," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 66-77, 2024/01/01/ 2024, doi: https://doi.org/10.1016/j.ijcce.2023.12.002.

[30]- P. Yashwante, Y. Patil, K. Nadar, and A. Khade, "Comparative Analysis of Meta-heuristic Feature Selection and Feature Extraction Approaches for Enhanced Chronic Kidney Disease Prediction," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, 14-16 March 2024 2024, vol. 2, pp. 1-6, doi: 10.1109/IATMSI60426.2024.10502980.

[31]- A. Allen *et al.*, "Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus," *BMJ Open Diabetes Research and Care*, vol. 10, no. 1, p. e002560, 2022.

[32]- M. Shantal, Z. Othman, and A. A. Bakar, "Impact of Missing Data on Correlation Coefficient Values: Deletion and Imputation Methods for Data Preparation," *Malaysian Journal of Fundamental and Applied Sciences*, vol. 19, no. 6, pp. 1052-1067, 2023.

[33]- D. a. G. Dua, Casey. "UCI Machine Learning Repository." http://archive.ics.uci.edu/ml (accessed.

[34]- M. M. Hassan *et al.*, "A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records," *Human-Centric Intelligent Systems*, vol. 3, no. 2, pp. 92-104, 2023.

[35]- M. Shantal, Z. Othman, and A. Abu Bakar, "Missing data imputation using correlation coefficient and min-max normalization weighting," *Intelligent Data Analysis*, vol. Preprint, pp. 1-15, 2024, doi: 10.3233/IDA-230140.

[36]- P. Tikariha and P. Richhariya, "Comparative study of chronic kidney disease prediction using different classification techniques," in *Proceedings of International Conference on Recent Advancement on Computer and Communication: ICRAC 2017*, 2018: Springer, pp. 195-203.

[37]- P. Chittora *et al.*, "Prediction of chronic kidney disease-a machine learning perspective," *IEEE access*, vol. 9, pp. 17312-17334, 2021.

[38]- D. Bhattacharyya, B. P. Doppala, and N. Thirupathi Rao, "Prediction and forecasting of persistent kidney problems using machine learning algorithms," *Int J Current Res Rev*, vol. 12, no. 20, pp. 134-139, 2020.

[39]- S. Revathy, B. Bharathi, P. Jeyanthi, and M. Ramesh, "Chronic kidney disease prediction using machine learning models," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, pp. 6364-6367, 2019.

[40]- D. Baidya, U. Umaima, M. N. Islam, F. J. M. Shamrat, A. Pramanik, and M. S. Rahman, "A deep prediction of chronic kidney disease by employing machine learning method," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2022: IEEE, pp. 1305-1310.