Interpretable Chronic Kidney Disease Risk Prediction from Clinical Data using Machine Learning

Vijay Simha Reddy Chennareddy[1], Santosh Tirunagari[1], Senthilkumar Mohan[2*], David Windridge[1], and Yashaswini Balla[3]

[1] Department of Computer Science, Middlesex University, London, United Kingdom
vc381@live.mdx.ac.uk & {s.tirunagari, d.windridge}@mdx.ac.uk
[2] School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India
senthilkumar.mohan@vit.ac.in
[3] Neurosciences Department, Alder Hey Children's NHS Foundation Trust, Liverpool, United Kingdom
yashaswiniballa@doctors.org.uk

Abstract. Chronic Kidney Disease (CKD) is a major cause of illness and death worldwide, with over 2 million cases diagnosed in the U.K. and potentially up to 1.8 million undiagnosed. However, there is a lack of longitudinal studies on CKD in India, resulting in limited data on its prevalence. CKD is often asymptomatic until 70% of the kidneys are severely damaged, and once this occurs, there is no cure. Patients may require dialysis or a kidney transplant to survive. Detecting the risk of CKD early is therefore crucial. In developing countries like India, many people cannot afford regular laboratory blood tests. This study aims to develop machine learning models to predict the likelihood of CKD using limited blood test results collected in India, including blood pressure, albumin, red and white blood cell count, blood urea, serum creatinine, HbA1Cs, and other biomarkers. Decision Trees and Logistic Regression classification algorithms were used, with hyperparameter tuning, achieving an F-score of 1. These promising results suggest that state-of-the-art results may be achievable with just six laboratory tests.

Keywords: CKD · Classfication · Feature Selection.

1 Introduction

Laboratory tests play a vital role in helping doctors and caregivers keep track of patients' health [1]. These tests can provide valuable insights into diseases like diabetes and chronic kidney disease (CKD). Typically, a laboratory appointment is taken to measure various parameters such as sugar levels, blood pressure, weight, and different cell counts. In India, millions of records from private labs are available, unlike in the UK, where records are not centralized[2]. It's essential for laboratory assistants to quickly identify any signs of disease progression from these records. However, reviewing each record manually is time-consuming and labour-intensive. Additionally, there are many other factors to consider when evaluating disease risk. For instance, serum creatinine levels less than 1.5 combined with abnormal protein

albumin are a reliable indicator of acute kidney injury [3]. Evaluating these test results, along with various other factors, for millions of records is a significant challenge. Fortunately, machine learning can help overcome this bottleneck and make it possible to analyze these records more efficiently.

1.1 Motivation

CKD is a significant cause of morbidity and mortality in recent times. Although 2 million people in the UK are diagnosed with CKD, a significant portion of the population, ranging from 1 to 1.8 million, is still undiagnosed, costing the NHS 1.4 billion [7]. Unfortunately, due to lack of research, there is not enough data available to validate how many patients are affected by CKD in India. The major problem with CKD is that it is asymptomatic, and no deterioration of health is observed until 70% of the kidneys are already severely damaged. Currently, medical science cannot cure patients who experience CKD at any stage. The only possible solution is either dialysis or a kidney transplant, and without either of these, patients will eventually succumb to the disease. Therefore, AI may help us evaluate lab records and various other health information related to the patients. The techniques of machine learning can be quite vital for early CKD detection, and in the past, ML methods have proved to be quite effective for CKD analysis. These methods include OneR, ZeroR, clustering, Naïve Bayes, decision trees, K-Nearest Neighbors, and Support Vector Machines (SVMs) [4]. However, machine learning algorithms, such as SVMs, Logistic Regression, Decision Trees, and Adaboost algorithms, have not been widely utilized with this dataset, as mentioned in previous research articles [4]. Although previous studies have attempted to create a machine learning model for classifying laboratory test results as indicating CKD or not, they have not identified which laboratory tests are critical in diagnosing CKD [5, 6].

1.2 Hypotheses & Limitations

The hypothesis of this study suggests that not all lab results are required for diagnosing CKD, especially with the dataset used in this study. However, there are some limitations to this hypothesis, which are explained below.

This paper uses a dataset of laboratory test results to predict the risk of CKD rather than using Glomerular filtration rate (GFR) values, which are commonly used in hospitals to identify CKD stage. The dataset does include serum creatinine values, which could be used to calculate estimated GFR, but that is not the focus of this study.

It's important to note that the dataset used in this study only includes data from 400 patients, with 250 having CKD and 150 without. Also, all of the patients in the dataset are from India, so the classifiers used may be biased towards that population.

The study only uses machine learning classifiers like logistic regression and Decision Trees and not deep neural networks. This is because deep learning models require a large amount of data for training and are not feasible with this dataset. Additionally, deep learning models are not interpretable, and their predictions lack explainability.

1.3 Contributions

This section discusses the unique contributions of this study. Previous research on CKD mainly relied on estimated GFR and its progression, and the available datasets were also based on estimated GFR. However, this study uses a publicly available dataset with all laboratory tests labelled for CKD or not.

The hypothesis of this study is that not all 24 features (laboratory tests) in the dataset are necessary for diagnosing CKD. The contribution of this work is in investigating this hypothesis using machine learning models such as logistic regression and decision trees.

2 Data and Methods

This study utilized a dataset [4] that was collected at Apollo Hospitals in Tamil Nadu over the course of two months. This dataset, which contains information on 400 patients, has been made publicly available on the UCI machine learning repository for research purposes. Among the patients, 250 have been diagnosed with CKD, while the remaining 150 have not. The dataset includes a total of 24 variables, 11 of which contain numeric values and 13 of which are categorical.

Some of the numeric values recorded in the dataset include blood pressure, random blood glucose level, serum creatinine level, sodium, and potassium. For patients who suffer from hypertension or diabetes, their conditions are recorded as either 0 or 1, indicating no or yes respectively. On average, CKD patients in the dataset are around 60 years old.

In terms of other features recorded, there is a bi-modal distribution for some variables, meaning that there are only two unique values present. For instance, almost 250 patients have an albumin value of 0. The majority of the patients in the dataset have a good appetite, and around 90% do not have any coronary artery-related comorbidities. Since the dataset is focused on CKD, it was not surprising to find that serum creatinine values were less than 40 for 95% of the patients in the cohort. Additionally, it is interesting to note that 60% of the patients in the cohort were not diabetic.

In this study, two classification techniques were utilized, namely Decision Trees (DT) and Logistic Regression (LR), to predict the risk factor of CKD using patient data.

Decision Trees are a non-parametric supervised learning method that works by recursively splitting the data based on the features' values, resulting in a tree-like model. Each node in the tree represents a feature, and the branches represent the feature's possible values, leading to a final decision. DTs are robust to normalization and scaling of data, making them advantageous for this study. DTs can handle categorical and numerical data, and they are easy to interpret. However, they can be highly unstable and expensive to run, especially for complex datasets. DTs may overfit the training data, leading to low accuracy on new data.

On the other hand, Logistic Regression is a parametric supervised learning method that models the probability of a binary outcome. It works by finding a linear relationship between the input features and the log-odds of the output. Logistic

[4] https://archive.ics.uci.edu/ml/datasets/Risk+Factor+prediction+of+Chronic+Kidney+Disease

Regression requires the data to be normalized and scaled, which can be time-consuming in the pre-processing stage. LR has a lower risk of overfitting and is more computationally efficient than DTs. However, LR is less robust to outliers and non-linear relationships in the data.

To tune the performance of the models, we need to adjust hyperparameters. For DT, hyperparameters like maximum depth, minimum samples split, and minimum samples leaf can be adjusted. For LR, hyperparameters like regularization strength, penalty type, and solver can be adjusted. Tuning hyperparameters can significantly affect the performance of the models. For example, increasing the maximum depth of the DT can lead to overfitting, while decreasing it can lead to underfitting. Similarly, increasing the regularization strength of LR can reduce overfitting, while decreasing it can lead to underfitting. Therefore, it is important to carefully choose and adjust hyperparameters to achieve the best performance for the models.

In summary, both DT and LR have their merits and demerits. DTs are advantageous for handling non-linear relationships and robust to normalization and scaling. However, they are unstable and can be expensive to run for complex datasets. On the other hand, LR is more computationally efficient and less prone to overfitting. However, it requires normalization and scaling and may not handle non-linear relationships well. DTs and LR are also known for their ability to handle imbalanced datasets, which is a common problem in medical datasets.

3 Experiments and Results

In this section, we will be discussing the experiments conducted and their results. Our hypothesis is that not all 24 laboratory tests are necessary when diagnosing CKD. We believe that fewer tests can potentially predict the risk of CKD. We also note that machine learning methods, such as logistic regression or decision trees, can automatically select features during the classifier learning process, making them interpretable.

We conducted three experiments in this section, which are as follows:

- **Baseline performance:** Since the dataset is imbalanced, it is crucial to estimate the baseline performance of the machine learning classification algorithms. The baseline performance acts as a benchmark that the classification algorithms should perform better than.
- **Hyperparameter Tuning:** In this experiment, we tuned the classification algorithms to find the optimal parameter settings that maximize their performance. We considered various parameters such as criterion for splitting, maximum depth of the tree, minimum samples per leaf node, and minimum samples in split for the Decision Trees classifier.
- **Feature Selection & Classification:** In this experiment, we selected the features that contributed the most to the classification performance, while discarding others. We then performed classification using only the selected features.

We believe that these experiments can help us determine which laboratory tests are essential in diagnosing CKD and how machine learning algorithms can aid in this process.

3.1 Baseline performance

In our dataset, out of the total 370 patient records (after removing missing values), 221 (59.73%) patients were diagnosed with CKD, while 149 (40.27%) patients did not have CKD. This indicates that there is an imbalance in the class distribution, with a higher number of CKD cases compared to non-CKD cases.

To estimate the baseline performance of our machine learning models, we need to consider this class imbalance. One simple approach is to use the "most frequent" representation, which means always predicting the most frequent class in the training set. In our case, this would be CKD. Thus, the baseline performance for our classification models would be 59.73%, which is the percentage of CKD cases in the dataset.

It is important to consider this imbalance when evaluating the performance of our models. A high accuracy rate alone does not necessarily indicate a good model, especially when dealing with imbalanced datasets. Therefore, we need to use appropriate evaluation metrics such as precision, recall, and F1-score, which take into account both true positive and false positive rates.

3.2 Hyperparameter Tuning

For the Logistic Regression classifier, we tested different values for the penalty parameter, C, and solver. The penalty parameter controls the regularization strength and the type of penalty used in the model. We tried L1 and L2 regularization penalties. The solver parameter specifies the algorithm to use in the optimization problem. We chose 'liblinear' solver as it is suitable for small datasets.

For the Decision Trees classifier, we tuned the criterion for splitting, maximum depth of the tree, minimum samples per leaf node, and minimum samples in split. The criterion parameter specifies the measure used to evaluate the quality of a split. We tried both 'gini' and 'entropy' criteria. The maximum depth parameter limits the depth of the tree. We considered depths of 4, 6, 8, and 12. The minimum samples per leaf node and minimum samples in split parameters determine the minimum number of samples required to be at a leaf node and in a split, respectively.

The best parameter values for Logistic Regression were a penalty of 'l1', C of 4.64, and solver of 'liblinear'. For Decision Trees, the best parameter values were a criterion of 'gini', maximum depth of 4, minimum samples per leaf of 1, and minimum samples in split of 2.

Overall, tuning the parameters of these classifiers helped us improve their performance in predicting the presence of CKD.

Algorithm	**Class**	**Precision**	**Recall**	**F1-Score**	**Support**
Decision Trees	CKD	0.98	1	0.99	45
	non-CKD	1	0.99	0.99	66
Logistic Regression	CKD	0.93	0.96	0.95	45
	non-CKD	0.97	0.95	0.96	66
Weighted Avg F1-Score				0.97	111

Table 1. Comparison of classification performance of Decision Trees and Logistic Regression.

Table 1 summarizes the performance of Decision Trees and Logistic Regression models in classifying patients as having CKD or not. The results indicate that both models achieved good overall performance, with weighted average F1-score of 0.99 for Decision Trees and 0.96 for Logistic Regression.

Decision Trees achieved perfect recall for CKD patients, meaning that all patients with CKD were correctly identified by the model. The model also achieved a high precision score of 0.98, indicating that out of all patients classified as having CKD, 98% of them were correctly classified. Similarly, Logistic Regression achieved good recall and precision scores for CKD patients, with a recall score of 0.96 and a precision score of 0.93.

For Non-CKD patients, Decision Trees achieved a perfect recall score of 0.99 and a high precision score of 1, indicating that all Non-CKD patients were correctly classified by the model. On the other hand, Logistic Regression achieved a recall score of 0.95 and a precision score of 0.97 for Non-CKD patients, indicating that 95% of Non-CKD patients were correctly classified, out of all patients classified as Non-CKD.

Overall, both Decision Trees and Logistic Regression models performed well in classifying patients as having CKD or not. Decision Trees achieved slightly better overall performance, with perfect recall and high precision scores for both CKD and Non-CKD patients. However, Logistic Regression model achieved good performance as well and is faster to run, making it a good alternative for classification problems with categorical values.

3.3 Feature Selection & Classification

The Table 2 shown in this section displays the comparison of feature weights between two different classifiers, Decision Trees and Logistic Regression, for the classification of CKD and non-CKD patients. The table contains 24 features with their respective weights assigned by each classifier, and the higher the weight, the more significant the feature is in the classification.

The Decision Tree classifier assigned higher weights to six features, namely, albumin, hemoglobin, packedCellVolume, redBloodCellCount, sodium, and specificGravity. These six features were considered for classification, while the other 18 features were discarded for this experiment. Logistic Regression is more suited for mixed data types, such as data containing continuous and categorical features, and provided weight coefficients for all 24 features. However, the interpretation of Logistic Regression weights is more complex than that of Decision Trees as it is multiplicative.

The results show that only a few laboratory tests are needed to accurately predict the risk of CKD, with an F-score of 1 obtained using only the six features identified by the Decision Trees. Therefore, it can be concluded that the remaining features are unnecessary for CKD risk prediction based on this dataset.

This section shows which features were important for the classification between CKD and non-CKD. The greater the weight, the more important the feature. Table 2 shows the features that are sorted in the highest of their weights on the Decision Tree classifier.

LR is better with mixed data types, i.e., data containing continuous features and categorical features. The beauty of the LR is its simplicity of providing the

Feature	Decision Trees	Logistic Regression
albumin	0.032937429	78.99741
anemia	0	44.33313
appetite	0	10.61923
bacteria	0	-49.6613
bloodGlucoseRandom	0	45.03704
bloodPressure	0	-8.01845
bloodUrea	0	-2.34529
coronaryArteryDisease	0	3.083723
diabetesMellitus	0	28.2038
hemoglobin	0.708762166	-143.146
hypertension	0	73.90377
packedCellVolume	0.003614422	6.913668
pedalEdema	0	54.75435
potassium	0	-14.2495
pusCell	0	-5.05058
pusCellClumps	0	-31.1891
redBloodCellCount	0.015749964	-9.59441
redBloodCells	0	10.28026
serumCreatinine	0	102.3379
sodium	0.015974963	22.06235
specificGravity	0.222961056	12.74159
sugar	0	11.48266
whiteBloodCellCount	0	-68.9975
age	0	0.791261

Table 2. Comparison of feature weights in Decision Trees and Logistic Regression (sorted by Feature).

weight coefficients for the variables in model interpretation. Therefore, it is convenient to check what variables influence the prediction result. However, a drawback of LR model is that it is quite difficult to interpret as the weight representation is multiplicative. On the other hand, DTs are efficient in showing the weight representation.

According to Table 2, the Decision Tree classifier assigned weights to only 6 out of 24 features, namely: 1) hemoglobin, 2) specificGravity, 3) albumin, 4) sodium, 5) redBloodCellCount, and 6) packedCellVolume. Hence, these 6 features were considered for classification, while the other 18 features, including bacteria, serumCreatinine, potassium, whiteBloodCellCount, hypertension, diabetesMellitus, coronaryArteryDisease, appetite, pedalEdema, and anemia, were discarded for this experiment. Both the Decision Trees and Logistic regression classifiers were trained using these 6 features, and an **F-score of 1** was obtained, validating our hypothesis that only a few laboratory tests were required to accurately predict the risk of CKD.

Thus, based on this dataset, only 6 laboratory tests are needed to detect the risk of CKD, and the remaining features are unnecessary.

4 Conclusion

This research highlights the potential of machine learning models in analyzing and detecting the risk of CKD from routinely collected laboratory tests. By automatically learning the interactions between different lab tests, it may be possible to accurately predict CKD risk with just a few tests, which could significantly reduce the cost and time associated with obtaining multiple tests. The study applied machine learning models on a dataset of 400 patients collected from Apollo hospitals in Tamil Nadu, India, and found that only six laboratory tests were required for an accurate prediction of CKD risk.

The findings of this study have important implications for healthcare professionals, as it could help them quickly identify patients at risk of CKD and provide them with appropriate interventions and treatments to prevent the progression of the disease. Moreover, the models developed in this study can be used to analyze large amounts of patient data and identify those at high risk of CKD in a hospital setting. However, more research is needed to validate the results of this study on a larger and more diverse population. Additionally, future studies should investigate the potential of other machine learning models and hyperparameters to improve the accuracy and generalization of the prediction models.

References

1. Faulkner, S.L. and Trotter, S.P. (2017). Data Saturation. In The International Encyclopedia of Communication Research Methods (eds J. Matthes, C.S. Davis and R.F. Potter). https://doi.org/10.1002/9781118901731.iecrm0060
2. Dalrymple LS, Katz R, Kestenbaum B, Shlipak MG, Sarnak MJ, Stehman-Breen C, Seliger S, Siscovick D, Newman AB, Fried L. Chronic kidney disease and the risk of end-stage renal disease versus death. J Gen Intern Med. 2011 Apr;26(4):379-85.
3. Rule AD, Larson TS, Bergstralh EJ, Slezak JM, Jacobsen SJ, Cosio FG. Using serum creatinine to estimate glomerular filtration rate: accuracy in good health and in chronic kidney disease. Ann Intern Med. 2004 Dec 21;141(12):929-37. doi: 10.7326/0003-4819-141-12-200412210-00009. PMID: 15611490.
4. Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. Data Mining: Practical Machine Learning Tools and Techniques (3rd. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
5. Reddy M, Cho J. (2016). Detecting Chronic Kidney Disease Using Machine Learning. Qatar Foundation Annual Research Conference Proceedings 2016: ICTSP1534 http://dx.doi.org/10.5339/qfarc.2016. ICTSP1534.
6. Bhattacharya, M., Jurkovitz, C., & Shatkay, H. (2017, November). Assessing chronic kidney disease from office visit records using hierarchical meta-classification of an imbalanced dataset. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 663-670). IEEE.
7. Tirunagari, S., Bull, S.C., Vehtari, A., Farmer, C., De Lusignan, S. and Poh, N., 2016, December. Automatic detection of acute kidney injury episodes from primary care data. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1-6). IEEE.