Advanced Machine Learning Methods for Chronic Kidney Disease Identification

Ashok Lamichhane
Department of Computer Engineering
BRAC University
Dhaka, Bangladesh
ashok.lamichhane@g.bracu.ac.bd

Mukund Prasad Singh
Department of Computer Engineering
BRAC University
Dhaka, Bangladesh
mukund.prasad.singh@g.bracu.ac.bd

Sumaiya Haque
Department of Computer Engineering
BRAC University
Dhaka, Bangladesh
sumaiya.haque2@g.bracu.ac.bd

***Abstract*—In today's world medical diagnosing and proper treatment is a crucial part for everyone. Chronic kidney disease is a type of disease which can't filter the blood as it does usual and can lead to damage of kidneys if proper treatment is not applied on time. The common reasons for Chronic Kidney Disease (CKD) are family history of kidney diseases or failure, high blood pressure and types of diabetes. If this CKD reaches its worst condition it can cause kidney damage as well as lead to heart disease, bone diseases, anemia, high potassium and calcium etc. We used the dataset taken from one of the hospitals of India over two months for classifying CKD using five different machine learning models: Random Forest, Logistic Regression, K-Nearest Neighbours, Support Vector Machine & Kernel Naive Bayes. We used two types of precision metices: Accuracy and F1 Score of each model and got that high accuracy and high F1-score in the K-Nearest Neighbors Model. Remaining machine learning models are also performed well almost 2 % less than the best one.**

***Index Terms*—Chronic Kidney Disease, Random Forest, Support Vector Machine, K-Nearest Neighbours, Kernel Naive-Bayes, Logistic Regression.**

I. Introduction

Chronic kidney disease (CKD) is a global public health problem affecting approximately 10% of the world's population. It is a progressive and irreversible pathologic syndrome.The term "chronic" describes the slow degradation of the kidney cells over a long period. This disease is a major kidney failure where the kidney sans the blood filtering process and there is a heavy fluid buildup in the body. This leads to an alarming increase of potassium and calcium salts in the body. Hence, the prediction and diagnosis of CKD in its early stages is quite essential, it may be able to enable patients to receive timely treatment to ameliorate the progression of the disease.

Chronic kidney disease (CKD) is becoming a more significant issue in both industrialized and developing nations. People in emerging nations are leading unhealthy lifestyles that lead to diabetes and high blood pressure as a result of increased urbanization. For example, in Pakistan, CKD spreads quickly due to the ingestion of harmful and low-quality foods, self-medication, extreme use of drugs, polluted water, obesity, high blood pressure, hypertension, anemia, diabetes, and kidney stones [1]. A report from 1990 to 2013 indicated that the global yearly life loss caused by CKD increased by 90% and it is the 13th leading cause of death in the world [2]. In this study, Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Kernal Naive Bayes (KNV) have been used to detect CKD. Most of the previous research focused on two classes, which make treatment recommendations difficult because the type of treatment to be given is based on the severity of CKD.

II. Literature Survey

J. Qin. et al [3] used the Chronic Kidney Disease dataset from the University of California Irvine and used six machine learning algorithms: Logistic Regression, Random Forest, Support Vector Machine, K-nearest neighbor, Naive Bayes classifier, and feed forward Neural Network. Random forest achieved the best performance with a 99.75% diagnosis accuracy. An integrated model combining logistic regression and random forest using perceptron achieved an average accuracy of 99.83% after ten simulations. The methodology showed potential for more complicated clinical data for disease diagnosis. However, the study did not provide information on its generalizability to other diseases or conditions.

S. Revathy et al [4] discusses the use of machine learning algorithms and data mining methods to predict Chronic Kidney Disease (CKD) early. It uses large datasets and traditional data mining methods for data preparation and preprocessing. Three machine learning algorithms, Decision tree, Random Forest, and Support Vector machines, are used for early CKD prediction. The study also analyzes the performance of boosting algorithms like AdaBoost and LogitBoost. The proposed prediction framework shows promising results, with the Random Forest classifier model providing the highest accuracy. Although the paper does not explicitly mention limitations, the performance of the ML models can be improved.

Iftikhar, H. et al [5] focuses on predicting chronic kidney disease using various machine learning models, including logistic regression, random forest, decision tree, k-nearest

neighbor, and support vector machine. Diagnostic test reports were observed for hundreds of patients at the Medical Complex in Buner, Khyber Pakhtunkhwa, Pakistan. The authors used k-nearest neighbors and SVM classifiers for the CKD dataset. They compared performance measures, including accuracy, Brier score, sensitivity, Youdent, specificity, and F1 score. The support vector machine with Laplace kernel function outperformed all models, while the random forest was competitive. The SVM-LAP model outperformed other models in all three scenarios, while the RF model was competitive.

Debal, D.A., Sitote, T.M. [6] uses data from St. Paulo's Hospital, Ethiopia's second-largest public hospital, to study chronic kidney disease records. The dataset contains 1718 instances with 19 features, with oversampling techniques used to balance minority and majority class values. The study uses binary and multi-class prediction models, including K-nearest neighbors, support vector machine, logistic regression, and decision tree. SVM has the highest classification accuracy of 98.3% and sensitivity of 0.99. Recursive feature elimination (RF) with cross validation has better performance than SVM and DT. The highest accuracy was achieved with RF, SVM, and XGBoost, with 99.8% for binary class and 82.56% for five-class datasets. DT produced the lowest performance compared to RF.

Arif, M.S.; Mukheimer, A.; Asif, D [7] proposes a novel ML model for predicting CKD, incorporating various preprocessing steps, feature selection, and ML algorithms, achieving outstanding performance. The model's effectiveness in enhancing early CKD detection highlights the potential of ML techniques in improving clinical support systems. The study utilized the UCI CKD dataset for developing a robust ML model for early CKD detection. The model achieved exceptional performance with 100% accuracy, precision, recall, and F1 score on the UCI CKD dataset. Advantages include the model's reliability and potential for clinical application, demonstrating effectiveness in enhancing early CKD detection. Although disadvantages were not explicitly mentioned in the provided sources, models accuracy can be improved to certain conditions.

Bai, Q. et al [8] utilized a dataset of 748 subjects with CKD, focusing on predicting the risk of ESKD using ML models. ML algorithms such as logistic regression, naive Bayes, random forest, decision tree, and K-nearest neighbors were employed, with performance metrics including accuracy, precision, recall, specificity, F1 score, and AUC. Results showed that simpler models like logistic regression performed comparably to complex ML algorithms due to the small sample size and limited predictor variables, indicating the importance of traditional regression models in disease risk prediction. Limitations included the small dataset size, lack of urine variables, and the need for external validation to enhance model predictability and generalizability.

III. Data Preprocessing & Model Description

Data preprocessing is a crucial step in the data analysis pipeline that involves cleaning, transforming, and organizing raw data into a format suitable for analysis and modeling. It includes tasks such as handling missing values, removing duplicates, standardizing data formats, encoding categorical variables, and scaling numerical features. Data preprocessing is important because it ensures the quality and reliability of the data used for analysis, which directly impacts the accuracy and effectiveness of machine learning models. By preparing the data properly, data preprocessing helps improve the model's performance, reducing bias, and enhancing the interpretability of the results. It also aids in identifying patterns, relationships, and insights that can drive informed decision-making and actionable outcomes.

There were several issues in the dataset that we used for this project. Initially, the presence of '?' characters in specific columns ('rc', 'wc', 'pcv') posed a problem as they needed to be removed for accurate analysis. The code successfully removed these characters and extracted numeric values from these columns, converting them to float data type. Moreover, missing numeric data in various columns ('age', 'bp', 'sg', 'al', 'su', 'bgr', 'bu', 'sc', 'sod', 'pot', 'hemo', 'rc', 'wc', 'pcv') were imputed with the mean values to ensure data completeness. Additionally, categorical variables like 'htn', 'dm', 'cad', 'pe', 'ane' were encoded into numerical values for better model interpretation. These data pre-processing steps were essential to handle missing values, standardize the features, and prepare the dataset for further analysis and model training.

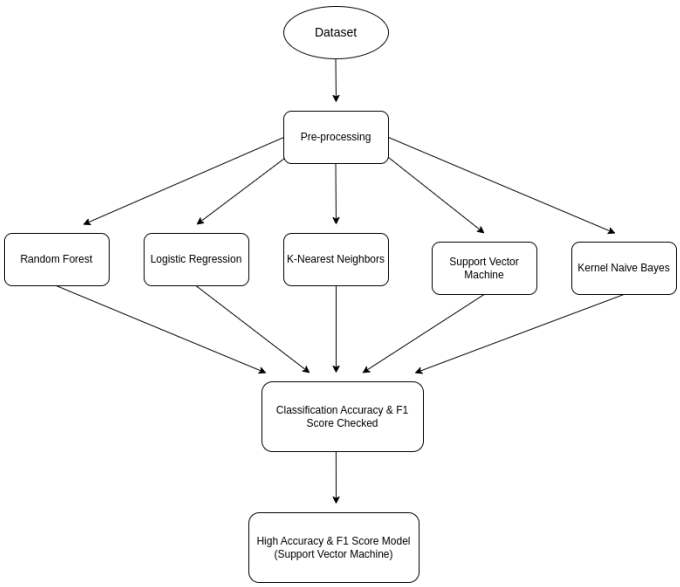

Fig. 1: Flow Chart of CKD Prediction Using Machine Learning Models

The following machine learning models have been obtained by using the corresponding subset of features or predictors on

the complete CKD data sets for diagnosing CKD.

A. **Random Forest**

Random forest algorithm constructs multiple decision trees to act as an ensemble of classification and regression processes. It generates a large number of decision trees by randomly sampling training samples and predictors. Each decision tree is trained to find a boundary that maximizes the difference between CKD and not CKD. Based on those multiple decision trees, it determines the final decision in disease diagnosis.

B. **Logistic Regression**

Logistic regression is a statistical method utilized for modeling the likelihood of a binary outcome. According to the effectiveness of the classification (CKD or not CKD) of the target variable, we choose logistic regression which obtains the wait of each predictor and a bias. If the sum of the effects of all predictors exceeds a threshold, the category of the sample will be classified as CKD or not CKD.

C. **K-Nearest Neighbors**

K-Nearest Neighbors (KNN) is a simple and intuitive algorithm used for both classification and regression tasks in machine learning. It is an algorithm that predicts a new data point which is determined by the majority of class or the mean of K nearest neighbors values. It finds the nearest training samples by calculating the distances between the test sample and the training samples and then determines the diagnostic category by voting.

D. **Support Vector Machine**

A linear model for classification and regression is a Support Vector Machine (SVM) which can be used to solve both linear and non-linear problems. The algorithm classifies data using a hyperplane. It divides different kinds of samples by establishing a decision Classification will be performed by finding the right hyper-plane that can differentiate the two classes efficiently. surface in a multidimensional space that comprises the predictors of the samples.

E. **Kernel Naive Bayes**

Kernel Naive Bayes applies a kernel function to transform the input features into a higher-dimensional space before applying the standard Naive Bayes algorithm. This allows it to handle non-linear decision boundaries effectively. Naive Bayes classifier calculates the conditional probabilities of the sample under the interval by the number of CKD and not CKD samples in each different measurement interval.

IV. Results & Analysis

Different types of machine learning models have been implemented to the original CKD dataset of India taken in 2 months from a hospital. The dataset has a total 400 unique instances having 25 clear features. We split the dataset in such a way that 80% of the data is used for training and 20% is for tests. After that we chose the best model with two performance metrics (f1 score & Accuracy).

A. *Random Forest*

Successfully generated confusion matrix from the Random Forest model with the help of 80 instances in which labels are 0: No CKD & 1 CKD. Among them 76 instances are classified properly and 4 are misclassified. The accuracy of this model is 95% whereas F1 score is 95%.

Table I: Confusion Matrix - Random Forest

	NO CKD	CKD
NO CKD	26	2
CKD	2	50

B. *Logistic Regression*

Successfully generated confusion matrix from the Logistic Regression model with the help of 80 instances in which labels are 0: No CKD & 1 CKD. Among them 77 instances are classified properly and 3 are misclassified. The accuracy of this model is 96.2% whereas F1 score is 96.2%.

Table II: Logistic Regression

	NO CKD	CKD
NO CKD	27	1
CKD	2	50

C. *K-Nearest Neighbors*

Successfully generated confusion matrix from the K-Nearest Neighbors model with the help of 80 instances in which labels are 0: No CKD & 1 CKD. Among them 78 instances are classified properly and 2 are misclassified. The accuracy of this model is 96% whereas F1 score is 96%.

Table III: Confusion Matrix - K-Nearest Neighbors

	NO CKD	CKD
NO CKD	28	0
CKD	2	50

D. *Support Vector Machine*

Successfully generated confusion matrix from the Support Vector Machine model with the help of 80 instances in which labels are 0: No CKD & 1 CKD. Among them 77 instances are classified properly and 3 are misclassified. The accuracy of this model is 96.3% whereas F1 score is 96.3%.

Table IV: Confusion Matrix - Support Vector Machine

	NO CKD	CKD
NO CKD	27	1
CKD	2	50

E. *Kernel Naive Bayes*

Successfully generated confusion matrix from the Kernel Naive Bayes model with the help of 80 instances in which labels are 0: No CKD & 1 CKD. Among them 76 instances are classified properly and 4 are misclassified. The accuracy of this model is 95% whereas F1 score is 95.1%.

Table V: Confusion Matrix - Kernel Naive Bayes

	NO CKD	CKD
NO CKD	28	0
CKD	4	48

Best Model Selection

Accuracy and F1 score of all five machine learning models are listed in the table and evaluating all Support Vector Machines have high accuracy as well as high F1 score so SVM is the best model among others for CKD classification.

Table VI: Accuracy & F1 Score of Different Classifier Models

S.No.	Classifer	Accuracy (%)	F1-Score (%)
1	Random Forest	95	95
2	Logistic Regression	96.2	96.2
3	K-Nearest Neighbors	96	96
4	Support Vector Machine	**96.3**	**96.3**
5	Kernel Naive Bayes	95	95.1

SVM performs well even in high-dimensional spaces, making it an excellent choice when dealing with data with many features whereas some ML models can't perform very well such as K-Nearest Neighbors. In the case of our dataset of CKD, 14 feature variables support the SVM very well. It is less prone to overfitting in comparison to other machine learning models, especially in high dimensional space. SVM is effective even when the number of dimensions exceeds the number of samples, unlike other models. It provides multiple options for tuning hyperparameters, such as the choice of kernel and kernel-specific parameters, which can lead to better performance. SVM works based on global optimization for finding the optimal decision boundary, unlike other local optimization ML algorithms. Those reasons are why it works better in the CKD dataset than other ML models.

V. Conclusion

The proposed CKD diagnostic methodology is used to predict CKD at an early stage and it is also feasible in terms of data imputation and sample diagnosis. The dataset shows input parameters collected from the CKD patients and the models are trained and validated for the given input parameters. However, the accessible data samples are relatively small, consisting of just 400 samples, during the model-building procedure because of the conditions' limitations. As a result, the model's ability to generalize may be limited. Random Forest, Logistic Regression, K Nearest Neighbors, Support Vector Machine, and Kernel Naive Bayes models are constructed to carry out the diagnosis of CKD. The performance of the models are evaluated based on the accuracy of the prediction. The result of this research showed that the Support Vector Machine model is much better at predicting CKD compared to the other proposed models. The comparison can also be based on the time complexity, and feature selection as the improvisation of the research. We believe that this model will be more and more perfect with the increase in size and quality of the data.

References

1 Mubarik, S., Malik, S. S., Mubarak, R., Gilani, M., and Masood, N., "Hypertension associated risk factors in pakistan: A multifactorial case-control study," *J Pak Med Assoc*, vol. 69, no. 8, pp. 1070–1073, 2019.

2 Yan, M.-T., Chao, C.-T., and Lin, S.-H., "Chronic kidney disease: Strategies to retard progression," *International Journal of Molecular Sciences*, vol. 22, no. 18, 2021. [Online]. Available: https://www.mdpi.com/1422-0067/22/18/10084

3 Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., and Chen, B., "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20 991–21 002, 2020.

4 Ramesh, R., "Chronic kidney disease prediction using machine learning models," *International Journal of Engineering and Advanced Technology*, vol. 9, p. 6364, 05 2020.

5 Iftikhar, H., Khan, M., Khan, Z., Khan, F., Alshanbari, H. M., and Ahmad, Z., "A comparative analysis of machine learning models: A case study in predicting chronic kidney disease," *Sustainability*, vol. 15, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2071-1050/15/3/2754

6 Debal, D. A. and Sitote, T. M., "Chronic kidney disease prediction using machine learning techniques," *Journal of Big Data*, vol. 9, no. 1, p. 109, Nov 2022. [Online]. Available: https://doi.org/10.1186/s40537-022-00657-5

7 Arif, M. S., Mukheimer, A., and Asif, D., "Enhancing the early detection of chronic kidney disease: A robust machine learning model," *Big Data and Cognitive Computing*, vol. 7, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2504-2289/7/3/144

8 Bai, Q., Su, C., Tang, W., and Li, Y., "Machine learning to predict end stage kidney disease in chronic kidney disease," *Scientific Reports*, vol. 12, no. 1, p. 8377, May 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-12316-z