


Predicting Chronic Kidney Disease with Machine Learning Algorithms

Jayesh Sanjay Patil ¹, Nilesh Vani ²

¹ PG-Student, Department of Computer Engineering :  [0009-0001-5194-9721](https://orcid.org/0009-0001-5194-9721)
Godavari College of Engineering, Jalgaon, Maharashtra, India , 425001

² Assistant Professor, Department of Computer Engineering :  [0009-0003-0476-0349](https://orcid.org/0009-0003-0476-0349)
Godavari College of Engineering, Jalgaon, Maharashtra, India , 425001 nileshvani@gmail.com

Email of Corresponding Author: jayeshpatil216200@gmail.com

Received on: 05 May,2025

Revised on: 04 June,2025

Published on: 07 June,2025

Abstract – Chronic Kidney Disease (CKD) is a progressive, irreversible condition distinguish by a gradual decline in kidney functions, often remaining asymptomatic until advanced stages. Early detection is essential for improving patient outcomes and prolonging survival. This study shows a machine learning (ML) approach for diagnosing CKD using the CKD dataset from the UCI machine learning repository, which includes substantial missing data. To address this, K-nearest neighbors (KNN) imputation was employed, reflecting real-world clinical scenarios. Eight ML algorithms— random forest, support vector machine , logistic regression, k-nearest neighbor, AdaBoost, naive Bayes classifier, feed-forward neural network, and gradient boosting—were evaluated for their diagnostic capabilities. An additional model combining logistic regression and random forest with a perceptron was also developed, demonstrating enhanced performance across multiple simulations. The addition of AdaBoost and gradient boosting contributed to improved model robustness and predictive accuracy. These results suggest that the proposed methodology can be adapted to more complex clinical datasets, offering a valuable tool for early disease diagnosis and aiding clinicians in making timely treatment decisions.

Keywords- Machine Learning, Data Imputation, Chronic Kidney Disease (CKD), Predictive Modeling.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a condition marked by the gradual loss of kidney function over time. The damage usually occurs silently and cannot be reversed, often going unnoticed until it progresses to more

advanced stages making early detection and initiation of treatment in order to ensure a good prognosis and prolonged life. In this aspect, machine learning algorithms have proven to be promising, and points towards the future of disease diagnosis. The previous CKD diagnostic models, most of Many existing approaches to handling missing data either suffer from limitations in their application scope or deliver relatively low accuracy. To address these issues, this work proposes a methodology aimed at both extending the applicability of CKD diagnostic models and enhancing their predictive performance. Specifically, K-Nearest Neighbors (KNN) imputation was employed to handle missing values, making it suitable even when diagnostic categories are unknown. Various classification algorithms—including Logistic Regression (LOG), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting, Naive Bayes (NB), Feedforward Neural Network (FNN), AdaBoost, and KNN were used to build diagnostic models using the complete CKD dataset. The best-performing models were selected for misclassification analysis. Additionally, a hybrid model was developed by integrating LOG and RF using a perceptron, which further improved diagnostic accuracy after KNN imputation was applied. The existing Machine Learning models and methodologies are not sufficiently enough for predicting Chronic Kidney Disease based on past history. The existing Machine Learning Models have been traditionally used individually for achieving the predication classes which is not sufficient. The proposed approach aims to showcase the superior performance of hybrid machine learning models compared to individual models, delivering notable improvements in both efficiency and

International Journal of Innovations in Engineering and Science, www.ijies.net

accuracy for predicting Chronic Kidney Disease. The highest achieved accuracy reached an impressive 99.88%. The primary goal of this study is to enable early-stage diagnosis of Chronic Kidney Disease using minimal testing and cost, while maintaining a high level of accuracy. Additionally, the work focuses on efficiently addressing missing values within the CKD dataset through appropriate imputation techniques. Feature selection will also be performed with the help of information gained to find the most important features that play a vital role in detecting CKD. Various machine learning algorithms will be applied and analyzed to detect CKD and best one with best performance and accuracy rate will be found. Our goal in this project is to see if we can predict if a patient will have chronic kidney disease or not using 24 predictors. This study seeks to implement and evaluate various machine learning algorithms to detect Chronic Kidney Disease, with a focus on comparing their accuracy and other key performance metrics.

II. LITERATURE REVIEW

Chronic Kidney Disease (CKD) is recognized as a major public health issue globally, affecting nearly 10% of the population [1], [2]. In China, the prevalence rate is approximately 10.8% [3], while in the United States, it ranges between 10% and 15% [4]. A separate study reported a prevalence of 14.7% among the adult population in Mexico [5]. CKD involves a progressive decline in kidney function that can ultimately lead to complete kidney failure. Since early stages of the disease often present no clear symptoms, diagnosis is typically delayed until around 25% of kidney function is lost [6]. The disease poses serious health risks due to its high rates of morbidity and mortality and its widespread impact on the body [7]. Moreover, CKD is a known contributor to the onset of cardiovascular diseases [8], [9]. As it is a chronic, irreversible condition [10], early prediction and diagnosis are vital for initiating timely treatment that could potentially slow disease progression.

Machine learning, which involves algorithms that learn from data to identify patterns and make predictions, has shown great promise in the healthcare domain [11]. This technology offers a cost-effective and accurate approach to diagnosing medical conditions, making it a valuable asset in the detection of CKD. With advancements in information technology [12] and the increasing availability of electronic health records [13], machine learning has expanded its role in clinical practice. It is

already being applied to monitor patient health [14], analyze disease-related factors [15], and assist in the diagnosis of various medical conditions. Notable applications include the identification of heart disease [16], [17], diabetes and diabetic retinopathy [18], [19], acute kidney injury [20], [21], cancer [22], and several other diseases [23], [24]. In these models, algorithms based on regression, tree, probability, decision surface and neural network were often effective. In the field of CKD diagnosis, Hodneland et al. utilized image registration to detect renal morphologic changes [25]. Vasquez-Morales et al. developed a neural network-based classifier using a large-scale CKD dataset, achieving a test accuracy of 95.0% [26]. Many existing studies have relied heavily on the Chronic Kidney Disease dataset from the UCI Machine Learning Repository. For instance, Chen et al. applied k-nearest neighbor (KNN), support vector machine (SVM), and soft independent modeling of class analogy to detect CKD, reporting that both KNN and SVM achieved accuracies as high as 99.7% [27]. They also explored fuzzy rule-based expert systems, fuzzy optimal associative memory, and partial least squares discriminant analysis, with performance ranging from 95.6% to 99.7% [1]. Although these models have shown strong performance, most employed mean imputation to handle missing values, which assumes prior knowledge of the diagnostic categories of the samples. This limits the models' applicability in real-world settings, where patients may have incomplete medical records and unknown diagnostic outcomes. Moreover, mean imputation can be particularly problematic for categorical variables. For example, in binary variables coded as 0 and 1, mean imputation may result in a non-binary value (e.g., 0.5), which lacks meaningful clinical interpretation. Polat et al. introduced an SVM model enhanced by feature selection techniques, effectively reducing computational cost while achieving accuracies ranging from 97.7% to 98.5% [6]. J. Aljaaf et al. adopted a novel multiple imputation approach followed by a multilayer perceptron (MLP) model, which reached an accuracy of 98.1% [28]. In another study, Subas et al. experimented with several models including MLP, SVM, KNN, C4.5 decision tree, and random forest (RF), with RF achieving perfect accuracy at 100% [2]. Similarly, Boukenze et al. reported that their MLP-based model attained a peak accuracy of 99.7% [29].

However, both studies [2], [29] primarily emphasized model construction, with limited discussion on missing value handling and no incorporation of feature selection techniques. Almansour et al. employed both SVM and

International Journal of Innovations in Engineering and Science, www.ijies.net

neural networks for CKD classification, achieving accuracies of 97.75% and 99.75%, respectively [30]. Meanwhile, Gunarathne et al. applied zero imputation for missing data and found that their decision forest model delivered the best results, with an accuracy of 99.1% [31]. In summary, a review of existing CKD diagnostic models reveals that many face limitations either due to the imputation methods used for handling missing data, which restrict their applicability, or due to comparatively lower diagnostic accuracy. To address these issues, this study introduces a novel methodology aimed at enhancing both the generalizability and accuracy of CKD diagnostic models.

III. METHODOLOGY

The CKD dataset used in this study was sourced from the UCI Machine Learning Repository [32], originally collected from hospitals and contributed by Soundarapandian et al. on July 3, 2015. The dataset comprises 400 samples, each containing 24 predictor variables (11 numerical and 13 categorical variables) along with a categorical response variable (class). The class variable has two possible values: "ckd" (indicating the sample is diagnosed with Chronic Kidney Disease) and "notckd" (indicating the sample is not diagnosed with CKD). Of the 400 samples, 250 are classified as "ckd" and 150 as "notckd". Notably, the dataset contains a significant number of missing values.

Table 1- Details of every variable in the original CKD dataset

Variab le	Descriptio n	Type	Scale	Missin g Rate
age	Age	Numeric al	Age in years	2.25 %
bp	Blood Pressure	Numeric al	in mm / Hg	3.0 %
sg	Specific Gravity	Nominal	(1.005, 1.010, 1.015, 1.020, 1.025)	11.5 %
al	Albumin	Nominal	(0, 1, 2, 3, 4, 5)	11.55 %
su	Sugar	Nominal	(0, 1, 2, 3, 4, 5)	12.20 %
rbc	Red Blood Cells	Nominal	(normal, abnormal)	38.0 %
pc	Pus_Cell	Nominal	(normal, abnormal)	16.25 %
pcc	Pus Cell Clumps	Nominal	(present, not present)	1.0 %
ba	Bacteria	Nominal	(present,	1.0 %

			not present)	
bgr	Blood Glucose Random	Numeric al	in mgs/dl	11.00 %
bu	BloodUrea	Numeric al	in mgs/dl	4.75 %
sc	Serum Creatinine	Numeric al	in mgs/dl	4.25 %
sod	Sodium	Numeric al	in mEq/L	21.5 %
pot	Potassium	Numeric al	in mEq/L	22.00 %
hemo	Hemoglobi n	Numeric al	in gms	13.00 %
pcv	Packed Cell Volume	Numeric al	-	17.75 %
wbcc	White Blood Cell Count	Numeric al	in cells/cmm	26.80 %
rbcc	Red Blood Cell Count	Numeric al	in millions/c mm	32.70 %
htn	Hypertensi on	Nominal	(yes, orno)	0.5 %
dm	Diabetes Mellitus	Nominal	(yes or no)	0.5 %
cad	Coronary Artery Disease	Nominal	(yes or no)	0.5 %
appet	Appetite	Nominal	(good or poor)	0.25 %
pe	Pedal Edema	Nominal	(yes or no)	0.25 %
ane	Anemia	Nominal	(yes or no)	0.25 %
class	Class	Nominal	(ckd or not ckd)	0 %

HARDWARE REQUIREMENTS :

- Laptop with at least 8 GB RAM.
- Storage capacity of 500 GB.
- 15" LED monitor.
- Processor: Intel Pentium i3 or i5.

SOFTWARE REQUIREMENT:

- Operating System: Windows 10.
- Jupyter Notebook.
- Python Libraries: Numpy, Pandas, Matplotlib, SKLearn, Seaborn.

SYSTEM ARCHITECTURE :

Flowchart for Chronic Kidney Disease Prediction Model

International Journal of Innovations in Engineering and Science, www.ijies.net

1. Data ↓
2. Feature Selection using Information Gain ↓
3. Handle Missing Data using KNN Imputation ↓
4. Data Pre-processing ↓
5. Cleaned Dataset ↓
6. Training Data (70%) ↙ ↘ Testing Data (30%) ↓
7. Random Forest Prediction as Base Classifier ↓
8. AdaBoost Algorithm ↓
9. Prediction

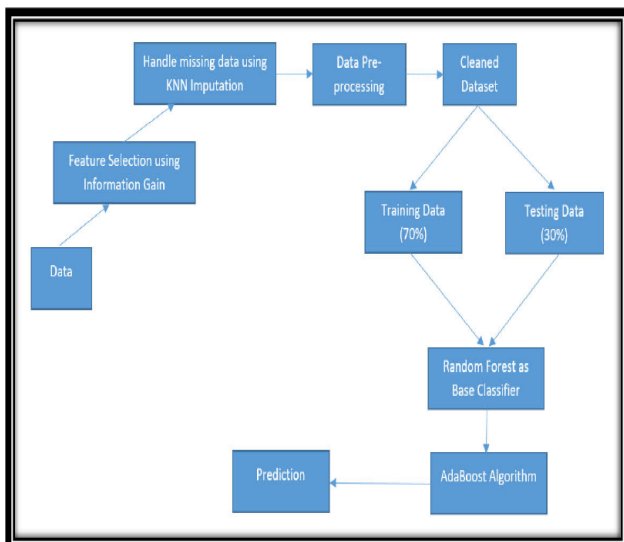


Fig. 1- System architecture

Proposed models:

- 1) **Regression Based Model : Logistic Regression** is a statistical analysis technique used to predict a binary outcome (e.g., "yes" or "no") based on prior observations from a dataset. It models the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic function. This makes it particularly useful for classification tasks, where the goal is to categorize data into two distinct classes.
- 2) **Tree Based Model: Random Forest-** Random Forest is a machine learning classifier that consists of multiple decision trees. Each tree in

the forest makes an individual prediction, and the final output is determined through majority voting or averaging the results from all the trees. This ensemble approach helps improve accuracy and robustness by reducing overfitting, which is common in individual decision trees.

- 3) **Decision plane based model: Support Vector Machine (SVM):** is a classification algorithm that finds the optimal decision boundary, known as a hyperplane, to separate different classes in an n-dimensional space. The hyperplane is chosen to maximize the margin between the closest data points of each class, which are known as support vectors. These support vectors are the critical elements that define the position of the hyperplane, and SVM aims to use them to classify new data points as accurately as possible.
- 4) **Distance based model: K-Nearest Neighbor (KNN)** is a distance-based model commonly used for classification and regression tasks. It is also frequently employed to handle missing values in datasets. KNN calculates the Euclidean distance between the missing data point and other known points in the dataset. By considering the "K" nearest neighbors (the most similar data points), the missing value is imputed based on the values of these neighbors, helping to estimate a reasonable value for the missing data.
- 5) **Probability based model: Gaussian Naive Bayes** is a probabilistic classification algorithm that applies Bayes' Theorem with the assumption of strong independence between the features. This model calculates the probability of a data point belonging to a particular class by combining the prior probability of the class and the likelihood of the features, assuming that the features are conditionally independent. In the case of Gaussian Naive Bayes, it assumes that the continuous features follow a normal (Gaussian) distribution.
- 6) **Neural network: A Feed Forward Neural Network (FFNN):** is a type of multilayer perceptron where the data flows in one direction, from the input layer to the output layer, without any feedback loops. Each layer consists of multiple neurons, and the decision-

International Journal of Innovations in Engineering and Science, www.ijies.net

making process occurs in a sequential, forward direction. The input features are processed layer by layer, with the final output representing the network's prediction or classification. FFNNs are typically used for tasks like classification and regression.

- 7) **Adaptive Boosting** : short for **AdaBoost** is an ensemble learning algorithm commonly used for both classification and regression tasks. It is a supervised learning technique that combines multiple weak learners (such as decision trees) to create a strong learner. AdaBoost works by adjusting the weights of the training instances based on the accuracy of previous predictions. Misclassified instances are given higher weights, so the model focuses more on difficult cases in subsequent iterations. This iterative process enhances the model's overall performance by reducing bias and improving accuracy.
- 8) **Gradient Boosting: Gradient Boosting** is an ensemble method that builds models sequentially. Each new model is trained to correct the errors of the previous one, focusing on the residuals (errors) from the prior model. The final prediction is a combination of all models, improving accuracy by iteratively reducing errors. It is commonly used with decision trees as base learners and is effective for both classification and regression tasks.
- 9) Combined Model of Logistic Regression & Random Forest using voting classifier. The **Voting Classifier** combines multiple models, like **Logistic Regression** and **Random Forest**, to improve prediction accuracy. Each model votes on the final class, and the majority vote determines the outcome. This ensemble method enhances performance by leveraging the strengths of both models, making it more robust and accurate.

IV. RESULT & DISCUSSION

This study investigates the use of supervised machine learning algorithms in bioinformatics, demonstrating their potential for early-stage diagnosis of critical diseases like Chronic Kidney Disease (CKD). The findings highlight the effectiveness of these algorithms in medical diagnostics and offer insights into their applicability for predicting other health conditions. The

research also provides a foundation for future studies in predictive health analytics, helping to refine and enhance these techniques for broader applications.

V. CONCLUSION

The Hybrid model used for predicting the output classes is more efficient and reliable instead of using the individual machine learning models for predicting the output classes. The combination of the two or more machine learning models is very effective and can give the prediction with the accuracy of 99% which is better than the traditional approach which gives the highest accuracy of 98.5% at its best-known implementation. The study has concluded that the models can be used in various combinations to get better results and can encourage the popularity of use of such implementation approaches

REFERENCES

- [1] Z. Chen et al., "Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers," *Chemometr. Intell. Lab.*, vol. 153, pp. 140-145, Apr. 2016.
- [2] A. Subasi, E. Alickovic, J. Kevric, "Diagnosis of chronic kidney disease by using random forest," in *Proc. Int. Conf. Medical and Biological Engineering*, Mar. 2017, pp. 589-594.
- [3] L. Zhang et al., "Prevalence of chronic kidney disease in china: a cross sectional survey," *Lancet*, vol. 379, pp. 815-822, Aug. 2012.
- [4] A.Singh et al., "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," *J. Biomed. Inform.*, vol. 53, pp. 220-228, Feb. 2015.
- [5] A. M. Cueto-Manzano et al., "Prevalence of chronic kidney disease in an adult population," *Arch. Med. Res.*, vol. 45, no. 6, pp. 507-513, Aug. 2014.
- [6] H.Polat, H.D. Mehr, A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *J. Med. Syst.*, vol. 41, no. 4, Apr. 2017.
- [7] C. Barbieri et al., "A new machine learning approach for predicting the response to anemia treatment in a large cohort of end stage renal disease patients undergoing dialysis," *Comput. Biol. Med.*, vol. 61, pp. 56-61, Jun. 2015.
- [8] V. Papademetriou et al., "Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study," *Am. J. Med.*, vol. 130, no. 12, Dec. 2017.
- [9] N. R. Hill et al., "Global prevalence of chronic kidney disease-A systematic review and meta-analysis," *Plos One*, vol. 11, no. 7, Jul. 2016.
- [10] J M. M.Hossain et al., "Mechanical anisotropy assessment in kidney cortex using ARFI peak displacement: Preclinical validation and pilot in vivo clinical results in kidney allografts," *IEEE Trans. Ultrason. Ferr.*, vol. 66, no. 3, pp. 551-562, Mar. 2019.
- [11] M.Alloghani et al., "Applications of machine learning techniques for soft ware engineering learning and early prediction of students' performance," in *Proc. Int. Conf. Soft Computing in Data Science*, Dec. 2018, pp. 246 258.

International Journal of Innovations in Engineering and Science, www.ijies.net

- [12] D. Gupta, S. Khare, A. Aggarwal, "A method to predict diagnostic codes for chronic diseases using machine learning techniques," in *Proc. Int. Conf. Computing, Communication and Automation*, Apr. 2016, pp. 281-287.
- [13] L. Du et al., "A machine learning based approach to identify protected health information in Chinese clinical text," *Int. J. Med. Inform.*, vol. 116, pp. 24-32, Aug. 2018.
- [14] R. Abbas et al., "Classification of foetal distress and hypoxia using machine learning approaches," in *Proc. Int. Conf. Intelligent Computing*, Jul. 2018, pp. 767-776.
- [15] M. Mahyoub, M. Randles, T. Baker and P. Yang, "Comparison analysis of machine learning algorithms to rank alzheimer's disease risk factors by importance," in *Proc. 11th Int. Conf. Developments in eSystems Engineering*, Sep. 2018.
- [16] E. Alickovic, A. Subasi, "Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier," *J. Med. Syst.*, vol. 40, no. 4, Apr. 2016.
- [17] Z. Masetic, A. Subasi, "Congestive heart failure detection using random forest classifier," *Comput. Meth. Prog. Bio.*, vol. 130, pp. 56-64, Jul. 2016.
- [18] Q. Zou et al., "Predicting diabetes mellitus with machine learning techniques," *Front. Genet.*, vol. 9, Nov. 2018.
- [19] Z. Gao et al., "Diagnosis of diabetic retinopathy using deep neural networks," *IEEE Access*, vol. 7, pp. 3360-3370, Dec. 2018.
- [20] R. J. Kate et al., "Prediction and detection models for acute kidney injury in hospitalized older adults," *Bmc. Med. Inform. Decis.*, vol. 16, Mar. 2016.
- [21] N. Park et al., "Predicting acute kidney injury in cancer patients using heterogeneous and irregular data," *Plos One*, vol. 13, no. 7, Jul. 2018.
- [22] M. Patricio et al., "Using resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC CANCER*, vol. 18, Jan. 2018.
- [23] X. Wang et al., "A new effective machine learning framework for sepsis diagnosis," *IEEE Access*, vol. 6, pp. 48300-48310, Aug. 2018.
- [24] Y. Chen et al., "Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B," *Comput. Biol. Med.*, vol. 89, pp. 18-23, Oct. 2017.
- [25] E. Hodneland et al., "In vivo detection of chronic kidney disease using tissue deformation fields from dynamic MR imaging," *IEEE Trans. Bio Med. Eng.*, vol. 66, no. 6, pp. 1779-1790, Jun. 2019.
- [26] G. R. Vasquez-Morales et al., "Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning," *IEEE Access*, vol. 7, pp. 152900-152910, Oct. 2019.
- [27] Z. Chen, X. Zhang, Z. Zhang, "Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models," *Int. Urol. Nephrol.*, vol. 48, no. 12, pp. 2069-2075, Dec. 2016.
- [28] A. J. Aljaaf et al., "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in *Proc. IEEE Congr. Evolutionary Computation*, Jul. 2018.
- [29] B. Boukenze, A. Haqiq and H. Mousannif, "Predicting chronic kidney failure disease using data mining techniques," in *Proc. Int. Symp. Ubiquitous Networking*, Nov. 2016, pp. 701-712.
- [30] N. Almansour et al., "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Comput. Biol. Med.*, vol. 109, pp. 101-111, Jun. 2019.
- [31] W. H. S. D. Gunarathne, K. D. M. Perera and K. A. D. C. P. Kahan dawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in *Proc. IEEE 17th Int. Conf. Bioinformatics and Bioengineering*, Oct. 2017, pp. 291-296.
- [32] D. Dua and C. Graff, "UCI Machine Learning Repository," Irvine, University of California, School of Information and Computer Sciences, 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>.