International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Early Prediction of Chronic Kidney Disease using Machine Learning Techniques

K. K. Poovitha
Information Technology, Vivekananda College of Engineering for women ,
Elayampalayam, Tamil Nadu, India – 637205

S. Divya
Assistant Professor , Department of M.Tech Computer Science and Engineering, Erode Sengunthar Engineering College, Erode, Tamil Nadu, India – 637205

G. Gomathi
Assistant professor/Information Technology,
Vivekananda College of Engineering for women,
Elayampalayam, Tamil Nadu, India – 637205

M. Lakshmi Priya
Assistant Professor, Department of Information Technology, Gnanamani College of Technology,
Namakkal, Tamil Nadu, India.

***Abstract*- The proposed study uses a systematic strategy that includes multiple important elements to classify chronic kidney disease (CKD). First, the dataset with all the necessary characteristics loaded (age, sex, antivirals, steroids, weariness, malaise, anorexia). Data cleaning techniques are then applied to guarantee the dataset's dependability and integrity for analysis. Then, feature selection techniques are used to determine which attributes have the greatest bearing on the classification of CKD. After that, the dataset is subjected to a variety of classification techniques, such as Naive Bayes, Decision Tree, Kstar, Logistic Regression, and Support Vector Machine (SVM), to assess how well they perform in correctly categorizing cases of chronic kidney disease. The findings show that all methods have the same level of classification accuracy, with Logistic Regression performing the best at 86.452%.**

Keywords: Chronic Kidney Disease, Machine Learning. Naive Bayes, Decision tree, Support vector machine, Logistic regression

I. INTRODUCTION

Chronic kidney disease (CKD) is the incapacity of the kidneys to carry out their normal blood-filtering role and other functions. The gradual deterioration of kidney cells over an extended period of time is referred to as "chronic." This illness is classified as significant

renal failure, in which the body accumulates a large amount of fluid and the kidneys are unable to filter blood [1]. This causes the body's levels of calcium and potassium salts to rise dangerously. High concentrations of these salts cause the body to suffer from a number of other illnesses. The kidneys' main function is to filter excess water and waste from blood [2]. For the minerals and salts in our bodies to be balanced, this mechanism must operate effectively. To regulate blood pressure, trigger hormones, produce red blood cells, and other processes, the proper balance of salts are required. Women who have cystic ovaries and other bone problems are caused by high calcium concentrations [3]. Additionally, CKD may result in an unexpected sickness or medication allergies. We refer to this condition as acute kidney injury (AKI). Cardiac issues and heart attacks can result from elevated blood pressure.

Kidney transplants or permanent dialysis are frequently the results of CKD. Additionally, a family history of kidney illness increases the likelihood of CKD [4]. Research indicates that nearly one in three diabetics also have chronic kidney disease (CKD). The literature also offers proof that treating and diagnosing CKD early on can enhance a patient's quality of life. Machine learning prediction algorithms offer an early medicine technique that can be used intelligently to forecast the occurrence of CKD. The extensive literature review demonstrates how several machine learning methods are applied to predict chronic kidney disease [5]. This study proposes the optimal prediction model and attempts to predict CKD utilizing classifiers such as Support Vector Machine, Random Forest, and Decision Tree.

II. LITERATURE REVIEW

When a patient has end-stage renal disease (ESRD), their only options are still dialysis or a kidney transplant. Early detection of CKD and appropriate dietary management can slow down or even stop the disease's progression in a favourable situation. According to J. Aljaaf et al. [5], applying machine learning algorithms in conjunction with predictive analytics turns out to be a wise way to anticipate a disease's onset early on.

Data mining models employ group methods known as boosting to improve a model's prediction. Usually, AdaBoost and Logit Boost are used to compare how well categorization algorithms work. In order to identify

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

CKD, Arif-Ul-Islam et al. [6] examined the effectiveness of boosting algorithms and developed rules that show the relationships between the different CKD characteristics. The study employed a decision tree and the Ant-Miner machine learning algorithm to develop rules.

Decisions are made by extracting hidden information from chronic illness datasets through datamining techniques. This necessitates the manipulation and storage of substantial volumes of semi structured, unstructured, and structured data. Big data has a critical part in this as well. G. Kaur et al. [7] used a variety of data mining techniques in a Hadoop environment to forecast chronic renal disease. In the study, classifiers like SVM (Support Vector Machine) and KNN (K-Nearest Neighbor) are employed.

When it comes to determining whether a patient receiving dialysis will survive or need a kidney transplant, blood levels of creatinine, salt, and urea are crucial factors. A straightforward K-means approach was employed by V. Ravindra et al. [8] to extract information regarding the relationship between a number of these CKD markers and patient survival. He came to the conclusion that the dialysis patients' survival period is predicted by the clustering process.

In order to forecast CKD A classification model for predicting the transitional interval of kidney disease stages 3 to 5 was created by R. Devika et al. [9] after analysing the accuracy, preciseness, and execution time of Naive Bayes, K-Nearest Neighbour (KNN), and Random Forest classifiers. Decision trees, KNN, Naive Bayes, and artificial neural networks were also used to elicit knowledge and create a classification model with the chosen set of attributes.

S. Vijayarani.at.al [10] used artificial neural networks (ANN) and support vector machines (SVM) to predict kidney disorders. The study evaluated the accuracy and execution time of the two aforementioned algorithms. Using feature selection algorithms, found a set of characteristics that effectively predict kidney disorders. Because of the fewer features, there are less expenses, more time savings, and less uncertainty. The missing data for each incomplete sample were processed using KNN imputation by selecting multiple complete samples with the most comparable measurements. Missing values are common in real-world medical scenarios because people may miss certain measures for a variety of reasons. A novel ensemble learning paradigm with three phases is presented in this paper for medical diagnosis with unbalanced data: data pre-processing, the training base classifier, and the final ensemble to address shortcomings in existing classification methods

III. RELATED WORK

The incidence, prevalence, and development of chronic kidney disease (CKD) have changed throughout time, particularly in nations with diverse social determinants of health. In most countries, diabetes and hypertension are the leading causes of CKD. According to the global recommendations, CKD is a disorder that results in decreasing kidney function over time, as seen by glomerular filtration rate (GFR) and kidney damage markers. People with CKD are likely to die at a young age. Doctors must diagnose various CKD-related diseases early on because early detection can prevent or even reverse renal damage. Early detection can lead to better therapy and care for patients. In many regional hospitals and clinics, there is a paucity of nephrologists or general practitioners who can diagnose the symptoms. This has resulted in patients having to wait longer for a diagnosis. As a result, this study argues that establishing an intelligent system to classify patients into 'CKD' or 'Non-CKD' categories will assist doctors in dealing with several patients and providing diagnoses more quickly.

IV. MATERIAL AND METHODS

The suggested solution uses machine learning techniques to create an effective and precise classification model for chronic kidney disease (CKD). The system is made up of multiple modules that handle feature selection, data loading, cleaning, and performance evaluation of categorization. The dataset containing crucial characteristics like age, sex, steroid use, and different CKD symptoms is loaded in the first module. Then, stringent methods are used in the data cleaning module to guarantee the dataset's integrity, including addressing missing values, resolving discrepancies, and standardizing data formats. The feature selection module finds the most pertinent attributes that have a substantial impact on the classification of chronic kidney disease (CKD) after data preparation. This is a critical step in lowering computational complexity and increasing model efficiency. Lastly, a number of machine learning methods, including Naive Bayes, Decision Tree, Kstar, Logistic Regression, and Support Vector Machine (SVM), are used to predict CKD cases in the classification performance evaluation module. The accuracy, precision, recall, F-measure, and other metrics are used to assess each algorithm's performance. The suggested system employs a systematic method to equip healthcare workers, including physicians, with a reliable tool for early diagnosis and precise classification of chronic kidney disease (CKD). This will enable prompt intervention and customized treatment plans for affected people.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

A. Loading the Dataset

This module deals with loading the dataset, which is the first stage in the analysis process. The dataset includes a number of characteristics, such as anorexia, fatigue, malaise, age, sex, steroid use, and antivirals. These characteristics are essential for comprehending and forecasting the classification of chronic kidney disease (CKD).

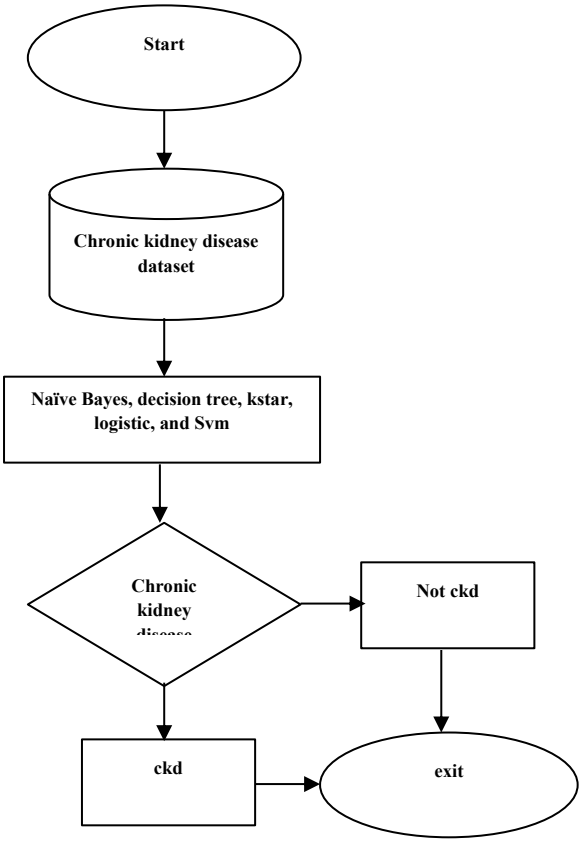


Figure 1. Process flow

B. Data Cleaning

To guarantee the dataset's integrity and dependability for analysis, this module thoroughly cleans it. A sample extract from the dataset is given, which displays the characteristics and matching values for two people. In the cleaning process, missing data must be handled, inconsistencies must be fixed, and data format uniformity must be guaranteed.

C. Feature Selection

Choosing the most pertinent qualities that make a major contribution to the classification problem is a crucial step in the building of a machine learning model. This module presents a list of qualities and the indices that go with them. These characteristics were chosen because of their possible influence on the categorization of CKD.

Id	Index	Attribute Name
1	0	Age
2	1	Sex
3	5	Malaise
4	10	Spiders
5	11	Ascites
6	12	Varices
7	13	Bilirubin
8	16	Albumin
9	17	Protime
10	18	Histology

Table 1. Attribute selection

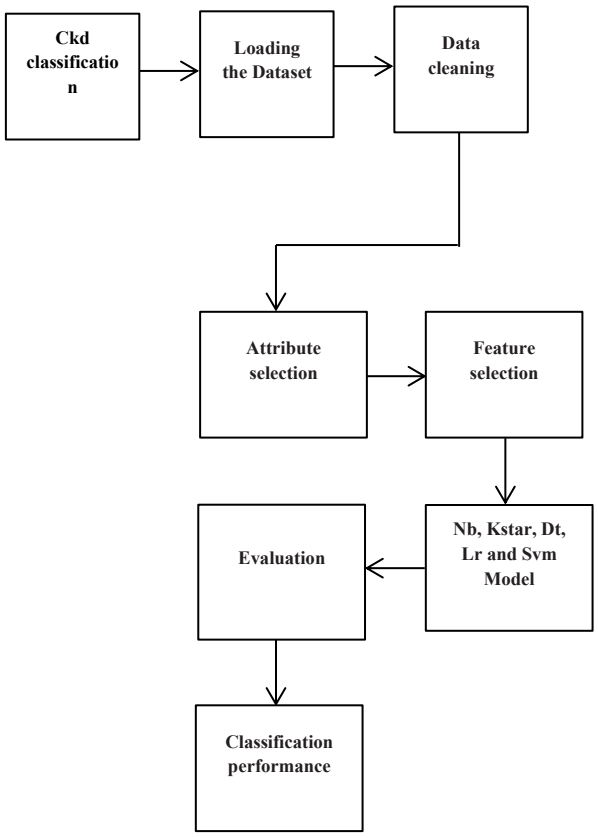


Figure 2. Proposed work flow

D. Classification Performance

The last module assesses how well different classification algorithms predict chronic kidney disease. Naive Bayes, Decision Tree, Kstar, Logistic Regression, and Support Vector Machine (SVM) are among the methods evaluated. Metrics for accuracy, precision, recall, and F-measure are given for every algorithm. The best algorithm for CKD classification can be chosen with the use of these measures,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

which measure how well each algorithm performs in correctly classifying CKD cases.

V. ALGORITHM DETAILS

A. Naive Bayes

Naive Bayes is a classification approach based on Bayes' Theorem that calculates the likelihood for each attribute. It chooses the outcome with the highest probability. This classifier assumes that the features are independent, and that the presence of one feature in a class does not imply the presence of another. Even if the traits are reliant on one another, they all contribute to the likelihood independently. The Naive Bayes approach is especially useful for large datasets. It is widely known to produce extraordinary outcomes. The Bayes theorem is based on conditional probability. Conditional probability refers to the likelihood of an event occurring if another event has previously occurred. The equation to compute conditional probability is as follows:

P (Hyp Evi) =P (Evi Hyp) *P(Hyp)/P(Evi)

Where, P (Hyp) is the possibility of hypothesis Hyp being true. P (Evi) is the possibility of the evidence (unrelated to the hypothesis). P (Evi Hyp) is the possibility of the evidence when the hypothesis is true.

B. J48 decision tree

It is a predictive method for analyzing the target value from a dataset using numerous attributes. It uses the training data to identify the attribute that distinguishes numerous occurrences. These events are further classified in order to get the most information possible. This technique is repeated for smaller selections until all instances are correctly placed in their respective classes.

C. Logistic Regression

Logistic Regression is a system in Machine Learning; it is most likely a quantifiable research strategy utilized to predict information value based on previous perceptions of an informative index. The logistic capacity is also known as sigmoid capacity; it produces an S-shaped bend; it can accept any true esteemed numbers or anything similar; and it is mapped to values 0 or 1.

1 / (1 + e^-value)

D. Svm

Support vector machines are classified as supervised learning, and they are often described by a separating hyperplane. SVM is a popular and helpful classification method for efficient computations and large datasets. SVM is a clever approach of preventing overfitting because they use a large number of features without requiring excessive processing. In SVM, support vectors are defined, and from these, we determine a margin with the shortest length while simultaneously maintaining the restriction such that features have good confidence in the plane. The margins can be (1) functional margin. 2) Geometric margin.

E. K Nearest Neighbour

Occurrence-based learning is one manner for comprehending assignments that approximate discrete or actual esteemed objective capacities, and KNN is one such grouping. In KNN, we store preparatory examples, and when a test model is presented, we find the nearest neighbor. At the forecast time

F. Algorithm

Input: Chronic Kidney Disease Dataset

Output: Hıgh Accuracy Classification Framework

Step1: Input data

Step2: Preprocess the data

Step 2.1: Convert Categorical values to numerical values

Step 2.2: Replace numerical missing values by Mean

Step2.3: Replace Categorical missing values by Mode

Step2.4: Attribute selection

Step3: Construct Classifier Models

Step3.1: Construct Naïve Bayes Model

Step3.2: Construct Decision Tree Model

Step 3.3: Construct SVM model

Step 3.4: Construct LR model

Step 3.5: Construct KSTAR model

Step 4: Check the accuracy of the constructed models

using Precision, Recall, F-measure and Accuracy

Step 5: Decide the best Classification model for CKD.

V. RESULT ANALYSIS

The following table provides an overview of the analysis of the classification results. For every classification algorithm—Naive Bayes, Decision Tree, Kstar, Logistic Regression, and Support Vector Machine (SVM)—metrics such as precision, recall, F-measure, and accuracy are shown. With a precision of 0.860, the highest of these methods, Logistic Regression, shows a high percentage of accurately predicted instances of CKD among all cases categorized as positive. Furthermore, Logistic Regression shows the highest recall (0.865), indicating that it can correctly identify a sizable percentage of true positive cases of CKD. With an F-measure of 0.862, our

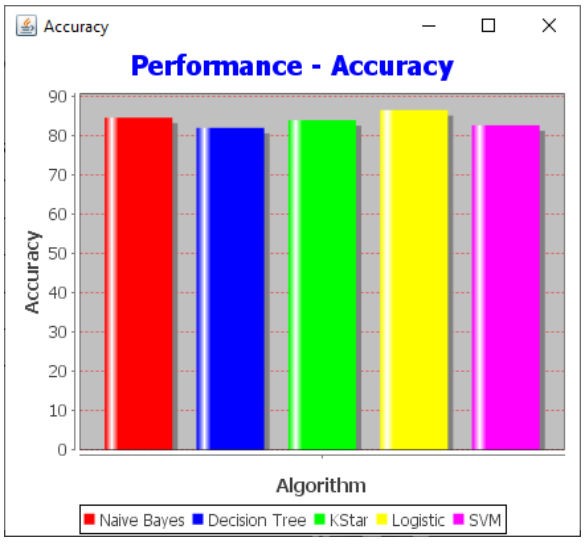
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

algorithm exhibits strong overall performance in CKD classification by striking a compromise between precision and recall. Additionally, Logistic Regression produces the best accuracy score of 86.452%, demonstrating its overall efficacy in accurately categorizing cases of CKD. This investigation shows that Logistic Regression is the best algorithm, with potential futures for precise CKD classification in clinical practice. Other algorithms, like Naive Bayes, Decision Tree, Kstar, and SVM, also show good performance metrics

.

Algorithm	Precision	Recall	F-measure	Accuracy
Naive Bayes	0.849	0.845	0.847	84.516
Decision Tree	0.802	0.819	0.806	81.935
Kstar	0.831	0.839	0.833	83.871
Logistic	0.86	0.865	0.862	86.452
SVM	0.807	0.826	0.808	82.581

Table 2. Comparison table



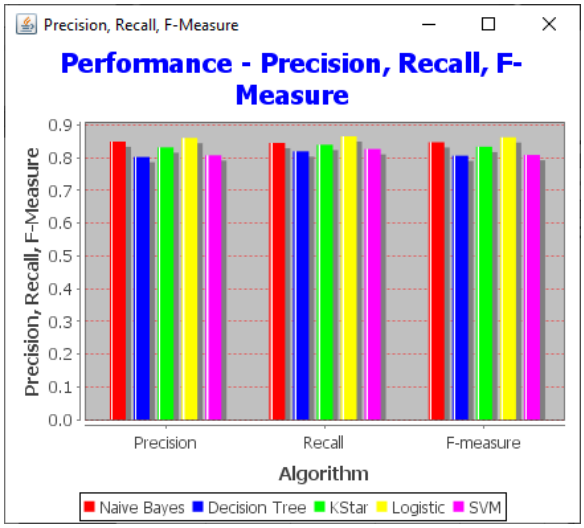


Figure 3. Comparison graph

VI. CONCLUSION

In summary, the system that has been built offers a thorough method for classifying chronic kidney disease (CKD) through the application of machine learning techniques. The system exhibits potential as a useful tool for early diagnosis and precise categorization of instances of chronic kidney disease (CKD) through careful data input, cleaning, feature selection, and performance evaluation. The evaluation of several classification algorithms demonstrates that Logistic Regression is the best approach for obtaining the best accuracy among the techniques examined. By utilizing these discoveries, medical professionals can improve patient care pathways and their diagnostic skills, which will ultimately improve outcomes and quality of life for CKD patients.

VII. FUTURE WORK

Future research may take multiple approaches to further explore various avenues for enhancing the created system's capabilities. First, the predictive ability and robustness of the model could be improved by adding more pertinent features or investigating more advanced feature engineering approaches. Furthermore, comprehensive validation studies on larger and more varied datasets from other demographic and geographic regions may aid in extending the applicability and efficacy of the model to a wider range of populations.

VIII. REFERENCES

[1] "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," M. P. N. M. Wickramasinghe, D. M. Perera, and K. A. D. C. P. Kahandawaarachchi, 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, 2020, pp. 300-303.

[2] The article "Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease" was published in the 2022 2nd International Conference on Informatics and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Computational Sciences (ICICoS), Semarang, Indonesia, pp. 1-4. Wibawa, Wibawa, Malik, and Bahtiar.

[3] "Derivation of action guidelines for persistent kidney disease through Naïve Bayes classifier application," U. N. Dulhare and M. Ayesha, 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2022, pp. 1-5.

[4] "Chronic Kidney Disease Survival Prediction with Artificial Neural Networks," H. Zhang, C. Hung, W. C. Chu, P. Chiu, and C. Y. Tang, 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2021, pp. 1351-1356

[5] J. Aljaaf et al., "Machine Learning-Supported Predictive Analytics for Early Prediction of Chronic Kidney Disease," 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, 2022, pp. 1-9.

[6] Arif-Ul-Islam and S. H. Ripon, "Using Boosting Classifiers, Ant-Miner, and J48 Decision Tree for Rule Induction and Prediction of Chronic Kidney Disease," Cox's Bazar, Bangladesh, 2020 International Conference on Electrical, Computer, and Communication Engineering (ECCE), pp. 1-6,

[7] Using data mining algorithms in Hadoop, G. Kaur and A. Sharma "Predict chronic kidney disease," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2022, pp. 973-979.

[8] "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique," N. Tazin, S. A. Sabab, and M. T. Chowdhury, Proceedings of the 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Dhaka, 2020, pp. 1-6.

[9] "Discovery of significant parameters in kidney dialysis data sets by K-means algorithm," International Conference on Circuits, Communication, Control and Computing, Bangalore, 2021, pp. 452-454, V. Ravindra, N. Sriraam, and M. Geetha.

[10] "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN, and Random Forest," 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 679–684. R. Devika, S. V. Avilala, and V. Subramaniyaswamy.

[11] Second Asian Conference on Defence Technology (ACDT), Chiang Mai, 2022 pp. 145–150; P. Panwong and N. Iam-On, "Predicting transitional interval of kidney disease stages 3 to 5 using data mining method."

[12] S. Dhayanand and S. Vijayarani, "ANN ALGORITHMS AND SVM FOR KIDNEY DISEASE PREDICTION", International Journal of Computing and Business Research (IJCBR), vol. 6, no. 2, 2022.

[13] Misir R, Samanta RK, Mitra M. A Decreased Features List for Predicting Chronic Kidney Disease. 2017; 14; J Pathol Inform. "Chronic Kidney Disease DataSet: UCI Machine Learning Repository", Archive.ics.uci.edu, 2020.

[14] The Chronic Kidney Disease dataset is accessible at http://archive.ics.uci.edu/ml/datasets.

[15] Neural Computation based general disease prediction model, B. Bharati and S. Prince Mary (2021), International Journal of Recent Technology and Engineering, vol. 8(2), pp. 5646–5649.