International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 6, Nov- Dec 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

PREDICTIVE MODELLING FOR CHRONIC KIDNEY DISEASE USING DEEP LEARNING

Kaviya Shree M
Computer Science and Engineering
Paavai Engineering College
Namakkal ,India
mkaviyashree364@gmail.com

Nandhini Priya P
Computer Science and Engineering
Paavai Engineering College
Namakkal ,India
Nandhusedhu195@gmail.com

Pavithra K
Computer Science and Engineering
Paavai Engineering College
Namakkal , India
pavithrakrishnappa02@gmail.com

Abstract:

Chronic Kidney Disease (CKD) has become one of the most serious global health problems, affecting millions of people and increasing the burden on healthcare systems. Early identification and prevention of CKD can significantly improve patient outcomes and reduce mortality rates. This study focuses on a comparative investigation of multiple machine learning and deep learning algorithms for CKD risk prediction. Eleven models were analyzed, including traditional approaches such as Naïve Bayes, K-nearest Neighbours, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, AdaBoost, and XG Boost, along with advanced neural models such as Artificial Neural Network(ANN), Simple Recurrent Neural Network(RNN), and Long Short-Term Memory(LSTM).The experiments were performing using a CKD dataset from the UCI Repository, evaluated under three dataset conditions-unbalanced, balanced with SMOTENC, and reduced by feature selection. Each model was tested for accuracy, precision, recall, F1-score and computational efficiency to determine its effectiveness for clinical applications. The findings revealed that while most algorithms achieved comparable accuracy levels, ensemble-based methods like Random Forest, AdaBoost, and XG Boost offered a better balance between speed and performance. Deep learning models did not demonstrate notable improvements due to the datasets limited size. Overall, this research emphasizes the potential of optimized machine learning models for reliable and efficient CKD risk prediction.

Keywords- ***Chronic Kidney Disease (CKD), Machine Learning Algorithms, Deep Learning, Risk Prediction, Clinical Data Analysis, Healthcare Informatics.***

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a long-term and progressive condition in which the kidneys gradually lose their ability to filter waste products and maintain fluid balance in the body. This disorder often develops silently over time and becomes clinically evident only in the advanced stages, by which point treatment options are limited. According to global health statistics, CKD affects nearly one in ten adults worldwide and is one of the major contributions to morbidity and mortality. The disease is commonly associated with risk factors such as diabetes, hypertension, obesity, and cardiovascular disorders. Early detection of CKD is therefore critical, as timely medical intervention can delay or even prevent kidney failure, significantly improving the patient's quality of life.

II. METHODS

A. Data Preparation and Preprocessing

The initial phase of this project involves collecting and preparing data to ensure the accuracy and efficiency of the prediction models. The dataset used for this research was obtained from the UCI Machine Learning Repository, a reliable and publicly accessible database widely used for academic research.

B. Model Development and Evaluation

After Preprocessing, a set of machine learning and deep learning algorithms were implemented to predict CKD and compare their performances. Eight traditional machine learning models – Naïve Bayes (NB), K- Nearest Neighbours (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), AdaBoost, and

XG Boost- and three deep learning models – Artificial Neural Network (ANN), Recurrent Neural Network (RNN), and Long Short- Term Memory (LSTM) were selected for experimentation.

III. LITERATURE SURVEY

Deep learning has become a leading paradigm in predictive modelling for chronic kidney disease (CKD) due to its remarkable advancement in handling complex and high dimensional health data over the past few years.

Earlier studies in CKD prediction primarily used conventional machine learning methods such as random forests, support vector machines, and logistic regression, achieving substantial accuracy but sometimes suffering from limited generalization and feature handling.

Overall, the state-of-the-art literature indicates that deep learning is transforming CKD risk prediction and screening, enabling more reliable and personalized clinical decision support tools, while persistent challenges around data quality, bias, and clinical validation warrant ongoing research

A.Traditional machine-learning approaches

- Early and mid-period work focused on classical supervised classifiers – Logistic Regression, Decision Trees, Naïve Bayes, K-Nearest Neighbors (KNN), Support Vector Machines(SVM) because of their simplicity, Interpretability and low computational cost. Key Points from this line of work:
- Decision-tree based ensembles(RF, XG Boost) are popular because they handle mixed feature types and missing value well, output feature-important scores and often deliver state-of-the-art results on small tabular medical datasets.
- Simpler models such as Logistic Regression remain useful baselines because results are interpretable and stable when sample sizes are small.

B.Handling missing data and class imbalance

Real clinical datasets tend to have missing entries and class imbalance (fewer negative/positive cases depending on collection).Two trends dominate preprocessing:

Imputation – KNN imputation and other model-based imputers are commonly used to avoid discarding records: imputation choice affects downstream model variance and bias.

Resampling / Oversampling – SMOTE and its categoricalaware variations (Eg: SMOTENC) are widely used to address imbalance; recent reviews show SMOTE variants often help but can introduce synthetic-sample artifacts and may be less effective on very small datasets or when minority samples are very sparse. Systematic reviews of medical datasets suggest oversampling methods (SMOTE, ADASYIN and variants) can improve classifier performance but their effectiveness depends on sample size and the ratio of classes.

Recent large reviews also emphasize that there is no universal "best" oversampling strategy – effectiveness depends on datasets structure (feature types, noise) and classifier choice.

C. Evaluation Practices and reproducility issues

A recurring methodological concern is inconsistent evaluation. Reported accuracies differ because studies vary in:

- How they impute missing data,
- Whether they balance data (and by what method),
- The splits and cross-validation used (simple train/test vs. k-fold, whether folds respect patient grouping),and Hyperparameter tuning procedures.

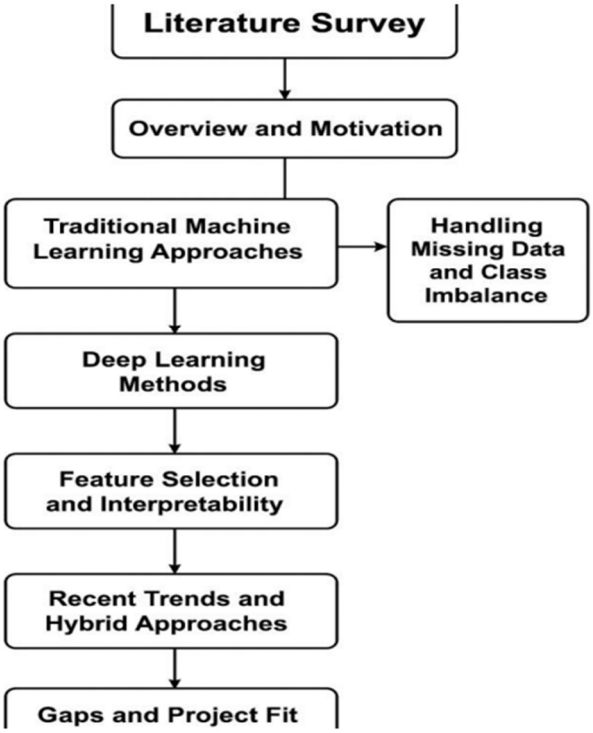
Systematic reviews call for standardized pipelines (transparent preprocessing + cross-validation + statistical testing) to enable fair comparisons. Many high-performing claims fall when strict cross-validation and held-out test sets are enforced.

D.Recent trends and Hybrid approaches

Recent literature (2023-2025) shows several directing gaining tractions:

- **Ensemble and hybrid methods** (stacking different learns, ensembles with feature selection) to combine strengths of different algorithms.
- **Improved oversampling** (variants and hybrid methods, including some learned/Generative approaches) to reduce synthetic-sample artifacts. Large methodological studies suggest careful selection of oversampling method depending on downstream classifier.
- **Explainability and clinically-oriented validation** – emphasis on SHAP/LIME, and testing whether selected feature align with known medical risk factors, to build clinician trust.

FLOWCHART



IV. FEATURES

Features are the fundamental building blocks of any machine learning model. In medical prediction systems such as Chronic Kidney Disease (CKD) diagnosis, features represent patient-specific physiological and clinical parameters that influence the likelihood of developing the disease.

The accuracy, interpretability, and reliability of a predictive model depend largely on the quality and relevance of these features. In this project, the features are drawn from the UCI Machine Learning Repository's CKD dataset, which includes a blend of numerical and categorical attributes.

These features describe biochemical measurements, medical, symptoms, and patient history that together contribute to disease prediction.

A.Demorgraphic and Physiological Features

Demographic features provide general background information about the patient, which can influence CKD risk. **1) Age:** Age is one of the most important risk factors for CKD. As individuals grow older, the glomerular filtration rate (GFR) tends to decline naturally, making kidneys more vulnerable to damage.

2)Blood Pressure(bp): Elevated blood pressure is directly linked to kidney damage because it increases the stress on blood vessels in the kidneys. High blood vessels in the kidneys. High blood pressure can both cause and worsen CKD.

3)Body Mass Indicators (if available): Parameters such as body weight or BMI, when present, can also indicate metabolic conditions that contribute to CKD development.

B.Biochemical Features

Biochemical indicators measure substances in the blood in the blood and urine that reflect kidney function. These are the most significant function.

1) **Blood Urea (bu):** Urea is a waste product filtered by the kidneys. High levels in the blood suggest reduced kidney function.

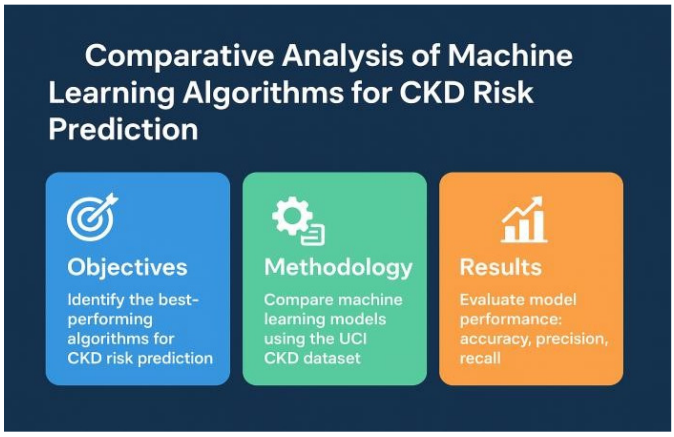
2) **Serum Creatinine (sc):** One of the most critical markers of kidney performance. Elevated serum creatinine levels indicate impaired filtration efficiency.

3) **Hemoglobin (hemo):** Low Hemoglobin is common in CKD patients due to reduced erythropoietin production by the kidneys.

4) **Packed Cell Volume (pcv):** Reflects the percentage of red blood cells in the blood. A low PCV level correlates with anemia, often associated with CKD.

C. Tables and Figure a) Key Features for CKD Prediction

Feature Category	Feature Name	Type	Clinical Significance
Demographic& Physiological	Age, Blood Pressure	Numerical	Older age and high BP increase
Biochemical Indicators	Serum Creatinine, Blood Urea	Numerical	Elevated levels indicate poor kidney filtration
Urinary& Microscopic Tests	Albumin, Red Blood Cells	Categorical	Presence of protein
Medical History	Diabetes, Hypertension	Categorical	Major comorbidities that cause



V. ACKNOWLEDGMENT

I would like to express my heartfelt gratitude to my project guide [Associate Professor Mrs. K. Sudha Devi] for their valuable guidance, support, and encouragement throughout the completion of this project titled "Comparative Analysis of Machine Learning Algorithms for CKD Risk Prediction". I am also thankful to the faculty and management of [Paavai Institution] for providing the necessary facilities and resources to carry out this work successfully.

Finally, I extend my sincere thanks to my friends for their continuous motivation and support during the course of this project.

REFERENCES

[1] S.Z.Fadem,Introduction to Kidney Disease: A Complete Guide for Patients,Springer,2022.

[2] M. Zisser and D.Aran,"Transformer-based time-to- event prediction for chronic kidney disease deterioration",J.of the Amer.Medi.Info.Asso,vol,31,no.4,pp. 980-990,2024. [3] A. L. Ammirati,"Chronic kidney disease,"Revista da Associacao Medica Brasileira, vol. 66, no. 1, pp.s03-s09, 2020.

[4] J. Catlett, "On changing continuous attributes into ordered discrete attributes", In Machine LearningEWSL-91European Working Session on Learning
Porto,pp. 164-178, 1991.

[5] John F. et al., "Aplliactions of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management" Annals of the New York Academy of Sciences, vol. 1020,pp. 239-262, 2004.

[6] A. J. Alijaaf et al., "Early prediction of chronic kidney disease using machine learning supported by predictive analytics", IEEE Congress on Evolutionary Computation (CEC), pp, 1-9, 2018.

