Chronic Kidney Disease Prediction Using Machine Learning, Deep Learning and Explainable AI: A Review and a Hybrid Model Concept

Miss. Walzade Gayatri Vinayak[1], Prof. Late S.D.[2]

[2]Guide

[1,2]Dept. of Computer Engineering SND Collage of Engineering and Research Center, Babhulgaon, Yeola, Nashik, India

gwalzade515@gmail.com, gwambhure04@gmail.com

Abstract:

Chronic Kidney Disease (CKD) damages the kidneys slowly and is often missed in its early stages, so early detection matters. Normal diagnosis depends on clinical tests and expert analysis, which take time and resources. This paper reviews how researchers have used machine learning and deep learning to predict CKD from patient health parameters. Eight recent studies form the core of the survey: random forest and boosting models explained with SHAP and LIME, a neural network explained with LIME, XGBoost on hospital records, stage-wise prediction on hospital data, a comparison of classifiers with feature selection and a deep network, an ensemble study with bagging and random subspace, and a KNN and Naive Bayes application. The methods are grouped into traditional and learning-based families and compared on data, accuracy and interpretability. The comparison shows that accuracy is very high on the small public UCI dataset but more modest on several hospital datasets and on stage-wise prediction, and that few studies explain their results. On this basis the paper presents a project concept. Patient data pass through missing-value handling, normalisation and feature selection; machine learning and deep learning models are trained; a hybrid of two algorithms joins their strengths; models are judged with accuracy, precision, recall, F1-score and confusion matrix; and Explainable AI shows the health factors behind each prediction. The concept is defined together with its evaluation plan; implementation and experiments remain future work.

Keywords: Chronic Kidney Disease, Machine Learning, Deep Learning, Hybrid Model, Explainable AI, SHAP, LIME, Feature Selection.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a long-term condition in which the kidneys slowly lose the ability to clean waste from the blood. In the early years it often gives no clear signs, so many people learn about it only when the damage is already large [1], [4]. The reviewed studies put the share of affected adults at roughly 8% to 16%, and one of them mentions over 800 million affected people across the world [2], [4], [6]. CKD also raises the risk of heart disease, and its last stage needs dialysis or a kidney transplant, which are costly for families and for health systems [1], [9].

Today a doctor mainly relies on blood and urine tests, especially serum creatinine and the estimated glomerular filtration rate (eGFR), and uses imaging or a biopsy in difficult cases [10]. These steps work well, but they need laboratories and specialist time, and both are scarce in many regions [8]. Machine learning offers a helpful second opinion. Given a patient's test values, a trained model can estimate how likely it is that the person has CKD.

Many researchers have tried this idea on the public UCI CKD dataset [11] or on hospital records, and most report high accuracy [1]–[8]. Even so, three problems appear again and again. Most results come from small datasets. Many near-perfect scores come from the same 400-record UCI file. And only a few studies explain their predictions in a way a doctor can check [1]–[3]. In addition, none of the reviewed studies joins two different algorithms into one model and then explains that model.

This paper therefore has three aims. The first is to give a clear overview of the methods used for CKD diagnosis, from traditional clinical practice to modern learning-based systems. The second is to review eight recent studies [1]–[8] and compare what each one adds and what it leaves open. The third is to describe a project concept that answers the gaps found: patient data are cleaned and reduced to useful features, machine learning and deep learning models are built, a hybrid of two algorithms is formed, and Explainable AI (XAI) shows which health factors drive each prediction.

The rest of this paper is arranged as follows. Section II gives an overview of the field, Section III describes traditional approaches and Section IV advanced methods. Section V is the literature survey and Section VI compares the methods. Section VII presents the project concept, Sections VIII and IX discuss challenges and future directions, Section X sets out the expected results, and Section XI concludes.

II. OVERVIEW OF CKD PREDICTION

Predicting CKD brings together medicine, data processing, machine learning and explainable AI. Fig.No.1 arranges the main families of approach discussed in this paper.

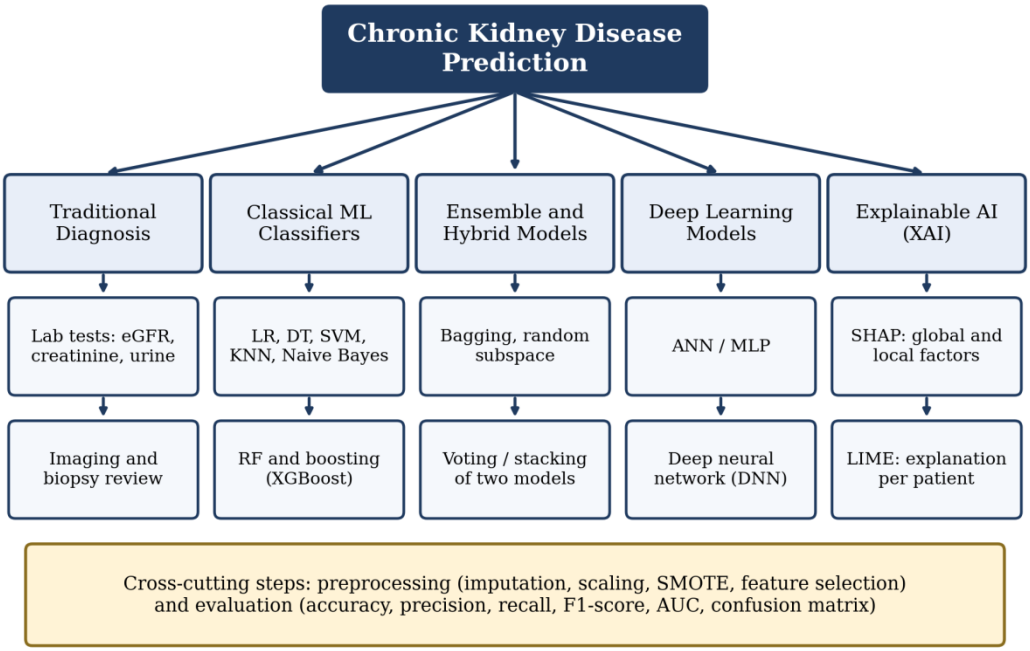


Fig.No.1: Classification of CKD Prediction Approaches

What CKD Is and Why Early Detection Matters:

CKD is defined by reduced kidney function, shown by an eGFR below 60 mL/min per 1.73 m² for more than three months, or by other signs of kidney damage [6], [10]. It is grouped into five stages. Early stages have few symptoms, but treatment at that point can slow the damage, which is why a simple screening tool is valuable [4], [8].

Patient Data and the UCI Dataset:

Models learn from records of past patients. The most used source is the UCI CKD dataset, which holds 400 records with 24 health parameters and one CKD or not-CKD label [11]. The parameters include age, blood

pressure, specific gravity, albumin, sugar, red blood cells, blood glucose, blood urea, serum creatinine, sodium, potassium, haemoglobin and signs such as hypertension, diabetes and anaemia [8]. Other studies use hospital records instead: 398 patients in Pakistan [1], 491 patients in a clinical cohort [3], 1,718 records from an Ethiopian hospital [4] and 224 subjects in China [6]. Because the sources differ in size and content, accuracy figures from different papers cannot be compared directly.

Data Preprocessing:

Medical data are rarely clean. The UCI file in particular has many missing values, so studies fill the gaps with the mean, median, most frequent value, KNN imputation or the next available value [2], [3], [7]. Values are then brought to a common scale, for example by min–max scaling or standardisation [2], [7]. When one class is much smaller than the other, the Synthetic Minority Over-sampling Technique (SMOTE) [15] or other oversampling is used to balance the classes [3], [4], [5].

Feature Selection:

Not every test is equally useful. Feature selection keeps the most informative parameters and drops the rest, which cuts training time and can reduce overfitting. The reviewed studies used mutual information, ANOVA, recursive feature elimination, LASSO, correlation-based and wrapper methods [1], [2], [4], [5]. In [1] a reduced set of ten features performed almost as well as the full set.

Classical Machine Learning Classifiers:

Logistic regression, decision trees, Naive Bayes, K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) [13] are simple and fast. Random Forest (RF) [12] builds many decision trees and takes a vote, while XGBoost [14] adds trees one after another so that each corrects the mistakes of the last. Tree-based models handle mixed numeric and yes-or-no data well, which suits clinical records.

Deep Learning Models:

An Artificial Neural Network (ANN) or Multilayer Perceptron (MLP) passes the inputs through layers of connected units and learns non-linear patterns [18]. A Deep Neural Network (DNN) simply has more layers. In [2] an MLP with two hidden layers was used for CKD, and in [5] a DNN gave the best accuracy among the models tested.

Ensemble and Hybrid Models:

An ensemble combines several models so that their errors cancel out. Bagging and random subspace methods were applied to KNN, Naive Bayes and decision trees in [7]. A hybrid model goes one step further and joins two suitable algorithms, for example by averaging their predicted probabilities or by stacking one on top of the other [19], so that each covers the weak points of the other.

Explainable AI:

A prediction that cannot be explained is hard for a doctor to trust. SHAP [16] gives each feature a score for how much it pushed one prediction up or down, and it can also be summed to show global importance. LIME [17] fits a small, simple model around one patient to show which values mattered most for that case. These tools were used in [1] and [3], and LIME was used in [2].

Evaluation Measures:

Models are judged with accuracy, precision, recall (sensitivity), specificity, F1-score, AUC and the confusion matrix. For a disease screen, recall deserves special care, because a missed patient costs more than a false alarm. A prediction should be treated as support for the doctor and not as a final diagnosis.

III. TRADITIONAL APPROACHES FOR CKD DIAGNOSIS

Before prediction models appeared, CKD was found and tracked by doctors using tests and examinations. These methods are still the base of care and cannot be replaced, but they have limits when many patients must be screened.

1. Clinical History and Physical Examination

The doctor asks about symptoms, diabetes, high blood pressure and family history, and checks for swelling, tiredness and changes in urine. Early CKD often shows none of these, so this step alone misses many cases [1], [8].

2. Blood and Urine Tests

Serum creatinine and urea show how well the kidneys clear waste, and eGFR is worked out from creatinine. Urine tests look for albumin and other signs of damage [6], [10]. These tests are the standard way to confirm CKD, but eGFR has known limits and does not tell the whole story of kidney health [1].

3. Imaging Studies

Ultrasound and other scans show the size and shape of the kidneys and can reveal stones, blockage or scarring. They need equipment and a trained reader, and a normal scan does not rule out early CKD [8].

4. Kidney Biopsy

A small piece of kidney tissue is examined under a microscope to find the cause and extent of damage. It is accurate for difficult cases but invasive, costly and not suitable for routine screening [1], [8].

5. Expert Review and Staging

A nephrologist combines test results, scans and history to decide the stage and plan treatment. This judgement is valuable, yet it depends on a small pool of specialists [4].

Limitations of Traditional Approaches

Although traditional methods remain part of good care, they share several limitations:

- **Late detection,** because early CKD gives few signs and many patients reach the clinic at an advanced stage [1], [8].
- **Costly and slow,** since repeated tests, scans and specialist visits are needed.
- **Dependent on resources,** as laboratories and nephrologists are scarce in many regions [8].
- **Varying judgement,** because different doctors may read the same results differently.
- **Limited use of combined data,** since it is hard for a person to weigh dozens of test values together.
- **Poor scalability,** as screening a large population is beyond the capacity of specialists.
- **Invasive steps,** such as biopsy, which are unsuitable for routine screening.

IV. ADVANCED METHODS FOR CKD PREDICTION

Learning-based methods replace fixed rules with models that learn the difference between CKD and non-CKD patients from past data. The families most relevant to CKD are described below.

Classical Machine Learning Classifiers:

Simple classifiers are the usual baseline. In [3], logistic regression, random forest, decision tree, Naive Bayes and XGBoost scored 87.98%, 92.67%, 89.00%, 82.49% and 93.29% accuracy. In [1], RF reached about 98% on hospital records. In [6], XGBoost gave an accuracy of 0.93 and an AUC of 0.99 against SVM and logistic regression. In [5], a linear SVM reached 98.86% when SMOTE was applied to the full feature set.

Ensemble Learning:

Ensembles often beat single models. In [7], random subspace and bagging improved KNN, Naive Bayes and decision trees on the UCI data, and random subspace with KNN reached 100% accuracy. Boosting models such as XGBoost, CatBoost and AdaBoost were also strong in [1].

Deep Learning Models:

Neural networks can learn complex links between test values. The MLP in [2] reached 100% accuracy on the UCI test split, and the DNN in [5] reached 99.6%. Deep models need more data and are harder to explain, so they are best used together with an explanation method.

Stage Prediction:

Most work only answers yes or no. The study in [4] also predicted the five stages of CKD. Accuracy was 99.8% for yes or no but only 82.56% across five stages, which shows that stage prediction is much harder.

Hybrid Models:

A hybrid joins two algorithms to use their different strengths, for example a tree-based model that is strong on tabular data and a neural network that captures non-linear patterns. None of the reviewed studies does this directly, so it forms one of the gaps addressed in Section VII.

Explainable AI for CKD:

SHAP and LIME turn a model's output into readable factors. In [1] the important features were eGFR, creatinine and urea; in [3] they were creatinine, HbA1c and age; and in [6] carotid intima-media thickness ranked first. These results match medical knowledge, which helps doctors trust the model.

Table No.1: Comparison Between Traditional and Learning-Based Approaches

Sr. No.	Feature	Traditional Approach	Learning-Based / Proposed Approach
1	Decision basis	Doctor's reading of tests and history	Patterns learned from patient data
2	Speed	Depends on lab time and specialist time	Fast once the model is trained
3	Consistency	Can vary between doctors	Same input gives the same output
4	Early detection	Early CKD gives few signs [1]	Can flag risk from routine test values
5	Accuracy	Depends on test and expert	93% to 100% reported for yes-or-no prediction [1]–[8]
6	Data needed	Individual tests per patient	A cleaned set of past patient records
7	Explanation	Doctor's reasoning	SHAP / LIME factor scores [1]–[3]
8	Cost	Repeated tests and visits	Low per patient after set-up
9	Scalability	Limited by specialists	High once trained
10	Availability	Limited where specialists are few [8]	Can run as a simple computer or web application [8]

V. LITERATURE SURVEY

Eight recent studies were examined in detail. Each is summarised below in terms of its data, method, results and the problems it leaves open.

1. Explainable ML Framework on Electronic Health Records [1]

Rizwan et al. studied 398 patients (241 with CKD and 157 without) from a kidney centre in Pakistan. They tested logistic regression, decision tree, KNN, gradient boosting, CatBoost, AdaBoost, XGBoost and random forest. Features were ranked by mutual information, decision tree and random forest, and models were run on all features and on the top ten, with stratified 5-fold and 10-fold cross-validation. Random forest reached about 98% accuracy. Statistical tests showed no significant difference among the best models, which suggests that the strong clinical features, mainly eGFR, creatinine and urea, matter more than the choice of algorithm. SHAP and LIME confirmed these features. The authors note that the data came from one centre, the sample was small, and medication, lifestyle and other illnesses were missing.

2. Explainable MLP Model with LIME [2]

Arif et al. used the UCI dataset of 400 records. Missing values were filled with KNN imputation for numbers and the most frequent value for categories, and features were scaled with min–max scaling. Twelve features were chosen with mutual information. An MLP with two hidden layers was trained on 75% of the data and tested on 25%. It was compared with Naive Bayes, logistic regression, random forest, decision tree and others, and it scored 100% on the test split. LIME was then used to show why single patients were classified as CKD or normal. A perfect score on a small public dataset needs careful reading, and the authors themselves call for larger and more varied data.

3. XAI-Based Prediction with XGBoost, SHAP and LIME [3]

Ghosh and Khandoker used clinical data from 491 patients (56 with CKD and 435 without) with 22 features. Missing values were filled with the median and SMOTE balanced the classes. Data were split 7:3 and checked with 10-fold stratified cross-validation. Five models were compared, and XGBoost was best with 93.29% accuracy and an AUC of 0.9689. SHAP and LIME showed creatinine, HbA1c and age as the most important features. The authors suggest adding richer healthcare data and deep learning in future work.

4. Binary and Stage-wise Prediction on Hospital Data [4]

Debal and Sitote built a dataset of 1,718 records with 19 features from patients at a hospital in Ethiopia. The data cover five classes (normal and four CKD stages), and for the yes-or-no task the smaller class was oversampled. ANOVA and recursive feature elimination with cross-validation were used for feature selection, and RF, SVM and decision tree were tested with 10-fold cross-validation. Random forest with eight selected features reached 99.8% for yes or no, while the best result for five stages was 82.56% with XGBoost. The authors suggest trying deep learning and a mobile application. The study does not explain individual predictions.

5. A Machine Learning Perspective with Feature Selection [5]

Chittora et al. used the UCI dataset and tested seven classifiers: ANN, C5.0, CHAID, logistic regression, linear SVM with L1 and L2 penalties, and random tree. Each was run on six feature settings, including full features, correlation-based, wrapper and LASSO selection, with and without SMOTE. The linear SVM with L2 penalty scored the highest at 98.86% with SMOTE on full features, and a deep neural network later reached 99.6%. Accuracy, precision, recall, F-measure, AUC and the GINI coefficient were compared. The work stays on one small public dataset and gives no explanation of predictions.

6. Predictive Analysis with XGBoost, SVM and Logistic Regression [6]

You collected data from a hospital in China for 69 CKD patients and 155 healthy people and compared XGBoost, SVM and logistic regression. XGBoost was best, with accuracy 0.93, specificity 0.89, sensitivity 0.94, F1-score 0.91 and AUC 0.99. Carotid intima-media thickness came out as the most important indicator. The author admits that the data were few, that the key features were not checked for medical practicality, and that neural networks were left out.

7. Ensemble Approach with Bagging and Random Subspace [7]

Jongbo et al. used the UCI dataset with 25 attributes. Missing values were filled with the next available value, features were standardised, and the data were split 70:30. Bagging and random subspace were applied to KNN, Naive Bayes and decision tree. Ensembles beat the single learners, random subspace did better than bagging in most cases, and random subspace with KNN reached 100% accuracy. Results are limited to a small public dataset and carry no explanation of the output.

8. Early Prediction with KNN and Naive Bayes [8]

Deepika et al. built a working application on the UCI dataset with KNN and Naive Bayes. The application has four user roles (admin, receptionist, doctor and patient), lets the doctor check whether a patient has CKD and find the stage, and stores treatment details. KNN gave 97% accuracy and Naive Bayes 91%. The study is practical, but it uses only simple classifiers and offers no explanation of how a result was reached.

Table No.2: Summary of the Surveyed Studies

Ref.	Year	Method	Data	Reported result	Open gap
[1]	2026	RF, XGBoost and others with SHAP, LIME	398 patients, Pakistan	RF about 98% accuracy	Single centre; small sample
[2]	2024	MLP with LIME	UCI, 400 records	100% accuracy	Small public data; perfect score
[3]	2024	LR, RF, DT, NB, XGBoost with SHAP, LIME	491 patients, 22 features	XGBoost 93.29%, AUC 0.9689	One cohort; no deep learning
[4]	2022	RF, SVM, DT with ANOVA and RFECV	1,718 records, Ethiopia	99.8% binary; 82.56% five-stage	No explanation; no deep learning
[5]	2021	Seven classifiers, six feature settings, DNN	UCI, 400 records	LSVM 98.86%; DNN 99.6%	No explanation; small data
[6]	2021	XGBoost, SVM, LR	224 subjects, China	Accuracy 0.93, AUC 0.99	Few data; no neural network
[7]	2020	Bagging and random subspace on KNN, NB, DT	UCI, 400 records	100% with random subspace KNN	Small data; no explanation
[8]	2020	KNN and Naive Bayes application	UCI, 400 records	KNN 97%; NB 91%	Simple models; no explanation

VI. COMPARATIVE ANALYSIS OF PREDICTION METHODS

Table No.3 sets the main method families beside one another. The entries are qualitative and follow what the reviewed papers report; they are not results of new experiments.

Table No.3: Comparison of CKD Prediction Methods

Method	Basis	Evidence reported	Interpretability	Cost
Clinical tests and expert review	eGFR, creatinine, urine, history	Standard of care [10]	High (doctor's reasoning)	High per patient
LR, DT, Naive Bayes	Simple statistical rules	82.49% to 89.00% [3]	High	Low
KNN / SVM	Distance and margin	KNN 97% [8]; LSVM 98.86% [5]	Low to medium	Low
Random forest	Many trees voting	About 98% [1]; 99.8% binary [4]	Medium (feature importance)	Low to medium
XGBoost / boosting	Trees added in sequence	93.29% [3]; 0.93 [6]	Medium with SHAP	Medium

Method	Basis	Evidence reported	Interpretability	Cost
Bagging / random subspace	Many base learners	100% on UCI [7]	Low	Medium
ANN / MLP / DNN	Layered neural units	100% [2]; 99.6% [5]	Low without XAI	Medium to high
MLP + LIME	Neural net plus local explanation	100% [2]	High per patient	Medium
Proposed hybrid + XAI	Two algorithms + SHAP / LIME	To be measured	High	Medium

Three patterns emerge. First, studies on the UCI dataset report 97% to 100% accuracy [2], [5], [7], [8], while three of the four hospital-based studies report 93% to 98% [1], [3], [6]. The fourth [4] reaches 99.8% for yes-or-no prediction but only 82.56% for five stages. High accuracy on one small file therefore does not separate methods; testing on more varied data and on harder tasks does. Second, tree-based models, especially RF and XGBoost, win in most hospital-based studies [1], [3], [4], [6], and [1] suggests that strong clinical features matter more than the algorithm. Third, only three studies explain their predictions [1]–[3], and none of the eight joins two different algorithms into one hybrid model with an explanation layer.

VII. PROPOSED PROJECT CONCEPT

The gaps above lead to a hybrid CKD prediction system with built-in explanation. The concept follows seven stages, from data collection to the healthcare professional who reads the result. Fig.No.2 shows the flow.

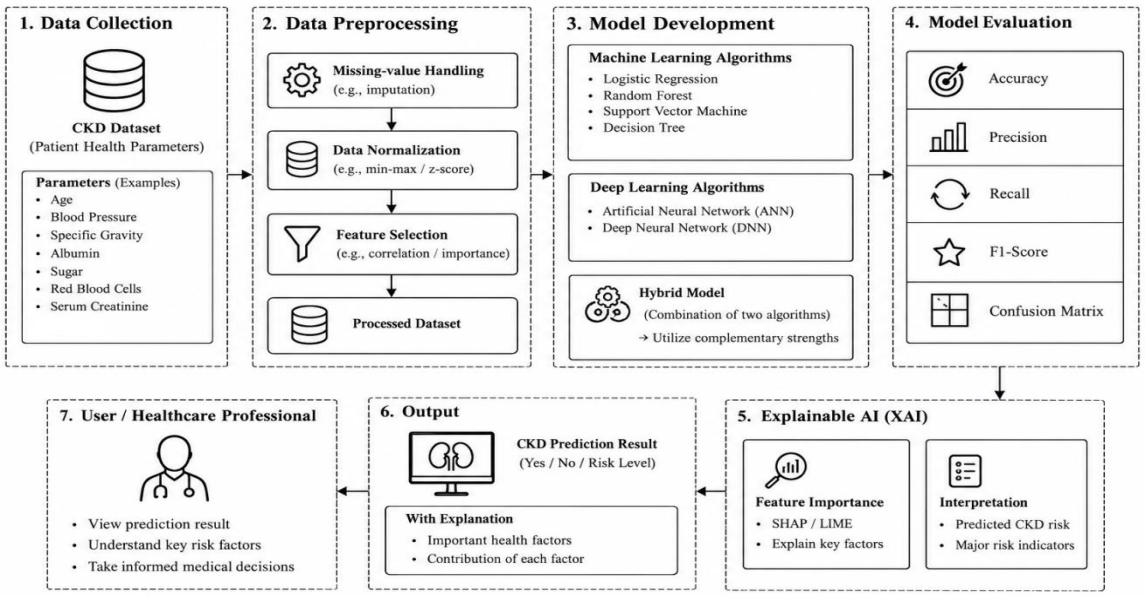


Fig.No.2: Architecture of the Proposed Explainable Hybrid CKD Prediction System

Data Collection:

The system starts from a CKD dataset of patient health parameters. Typical parameters are age, blood pressure, specific gravity, albumin, sugar, red blood cells and serum creatinine. The UCI CKD dataset [11] is the starting point, because it is public and widely used, and hospital data can be added later if permission is given.

Data Preprocessing:

Three steps prepare the data. Missing values are handled by imputation, using the median for numbers and the most frequent value for categories, or KNN imputation as in [2]. Numeric values are then normalised. With min–max scaling, each value x is converted to x′ as in (1):

$x' = (x - x_min) / (x_max - x_min)$ (1)

Here x_min and x_max are the smallest and largest values of that parameter in the training data, so every parameter lies between 0 and 1. Z-score scaling will be tried as an alternative. Finally, feature selection keeps the most useful parameters, using correlation and importance scores, as in [1], [2], [4]. The result is the processed dataset. If SMOTE is needed, it will be applied only to the training part so that no synthetic record leaks into the test data, a weakness reported for earlier work in [2].

Model Development:

Two groups of single models will be built. The machine learning group has logistic regression, random forest, support vector machine and decision tree. The deep learning group has an ANN and a deeper DNN. Their results serve as the baseline.

On top of these, a hybrid model combines two suitable algorithms to use their complementary strengths. The first plan is to join random forest, which is strong and steady on small mixed tables [12], with a neural network, which captures non-linear links between test values [18]. Each model outputs the probability of CKD, and the hybrid averages them as in (2):

$P(CKD \mid x) = w \cdot P_RF(x) + (1 - w) \cdot P_ANN(x)$ (2)

The weight w lies between 0 and 1 and is tuned on validation data. Other pairs, such as random forest with XGBoost, will also be tried so that the best pair is chosen from evidence and not by assumption.

Model Evaluation:

All models will be tested with stratified 10-fold cross-validation and scored on accuracy, precision, recall, F1-score and the confusion matrix, with AUC added as a further check. Four groups will be compared: single machine learning models, single deep learning models, the hybrid on all features and the hybrid on selected features. Recall will be watched closely, since a missed CKD patient is the costlier mistake.

Explainable AI:

After the best model is chosen, SHAP [16] and LIME [17] will be applied. Both only need the model's probability output, so they work on the hybrid as well. They give feature importance and an interpretation for each patient: the predicted CKD risk and the main risk indicators behind it.

Output and User:

The output page shows the result (CKD yes or no, with a risk level) together with the important health factors and the contribution of each. The doctor can view the result, understand the key risk factors and take an informed decision. The system supports the doctor and does not replace the doctor's judgement.

Implementation Setup:

The prototype is planned with widely available open-source tools, summarised in Table No.4.

Table No.4: Planned System Requirements

Category	Planned configuration
Software	Python 3.10 with pandas, scikit-learn, XGBoost, TensorFlow or PyTorch, SHAP and LIME, on Windows 11 or Ubuntu
Hardware	Core i5/i7-class CPU, 16 GB RAM, 256 GB or more disk; a GPU is helpful but not required for this data size

Category	Planned configuration
Dataset	UCI CKD dataset (400 records, 25 attributes), with hospital records to be added if available

VIII. CHALLENGES AND LIMITATIONS

Despite steady progress, several challenges remain. The first is data size and variety. Many studies rest on a few hundred records from one place: 398 patients at one centre in [1], 224 subjects in [6] and the 400-record UCI file in [2], [5], [7], [8]. A model trained on such data may not work on patients from another hospital or region, so outside testing is needed before any real use.

The second is over-optimistic results. Scores of 100% on a small public dataset [2], [7] are a warning and not a proof. Choices such as applying SMOTE before the train–test split can leak information and inflate accuracy, a weakness the authors of [2] point out in earlier studies. Careful splitting, cross-validation and testing on a second dataset are the safeguards.

Data quality is a third issue. The UCI file has many missing values, and the way they are filled changes the result [2], [7]. Classes are often unbalanced, with few CKD cases in some sets and few healthy cases in others [3], [4]. Important variables such as medication, lifestyle and other illnesses are missing from several datasets [1].

Fourth, explanations need care. SHAP and LIME show which inputs moved the prediction, but not why the body behaves that way, so they are guides and not medical proof. Some top features also need a medical check: the author of [6] notes that the important indicators were not tested for practical treatment use.

Fifth, harder tasks are still unsolved. Predicting the CKD stage fell to 82.56% accuracy in [4], far below the yes-or-no result. Finally, the ethical stakes are real. Medical records are private, a missed case can delay treatment, and a false alarm causes worry and cost. A prediction is a probability and not a diagnosis, so a doctor must confirm it.

IX. FUTURE DIRECTIONS

Future work can follow several lines. The first is larger and more varied data. Records from several hospitals and different populations would test whether models hold up, as the authors of [1] suggest. Adding medication, lifestyle and other-illness variables would give models more to learn from.

Second, deep learning deserves a fair trial next to tree-based models. The authors of [3] and [4] both name it as a next step, and the hybrid concept in this paper is one way to bring the two together. Stage-wise prediction also needs more work, since it is much harder than yes-or-no prediction [4].

Third, explanations can be made easier to use. The authors of [2] call for friendly interfaces for clinical staff, and [4] and [8] point to mobile and web tools. A short, plain-language note beside the SHAP or LIME chart would help doctors and patients alike.

Fourth, privacy-friendly training, such as learning across hospitals without moving patient records, could widen the data available while protecting patients. Finally, the concept in this paper can be turned into a web application, followed by a study with doctors to learn whether the explanations really help them trust or question the result.

X. EXPECTED RESULTS & DISCUSSION

Expected Results

The framework has been designed and the literature reviewed, but implementation and training are still to be done, so no new experimental figures are reported here. The finished system is expected to predict CKD with accuracy comparable to the better published results, to show whether the hybrid model gives a

measurable gain over its two parts, and to keep a high recall so that few patients are missed. The SHAP and LIME output is expected to point to medically sensible factors such as serum creatinine, albumin and haemoglobin, in line with [1], [2] and [3].

Discussion

The reviewed studies suggest that supervised classifiers are already strong on a single dataset [1], [4], [7], so the real test for the proposed concept is trust and steadiness on new data. If the random forest and the neural network make different kinds of mistakes, averaging them should give a more reliable result, which is the idea behind ensembles in [7]. The explanation layer adds a practical benefit that most reported systems lack. Against these hopes stand the limits in Section VIII, in particular small data and the fact that explanations are aids, not proof.

Comparative Performance Analysis

Table No.5 collects results as reported in the reviewed papers; the figures come from different datasets and splits and are therefore indicative only. The proposed concept is listed without a number because it has not yet been evaluated.

Table No.5: Reported Results of Existing Methods

Method	Reported result	Data / setting	Remark
Random forest [1]	About 98% accuracy	398 patients, Pakistan	Top-10 features kept performance
MLP + LIME [2]	100% accuracy	UCI, 75/25 split	Perfect score on small data
XGBoost [3]	93.29% accuracy, AUC 0.9689	491 patients	Best of five models
RF with RFECV [4]	99.8% (binary)	1,718 records, Ethiopia	Eight features selected
XGBoost [4]	82.56% (five stages)	1,718 records, Ethiopia	Stage prediction is harder
Linear SVM, DNN [5]	98.86%; 99.6%	UCI	SMOTE with full features
XGBoost [6]	0.93 accuracy, AUC 0.99	224 subjects, China	cIMT most important
Random subspace + KNN [7]	100% accuracy	UCI, 70/30 split	Ensemble beat single models
KNN / Naive Bayes [8]	97% / 91%	UCI	Working web application
Proposed hybrid concept	To be measured	UCI, hospital data if available	Accuracy, recall, F1, AUC, explanation

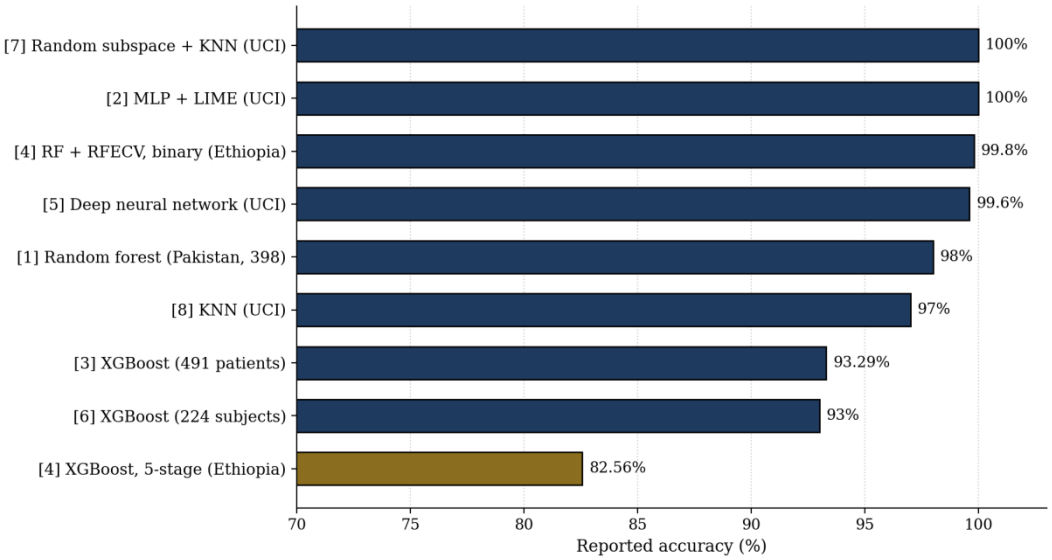


Fig.No.3: Accuracy Reported by Reviewed Studies (different datasets; indicative comparison only)

XI. CONCLUSION

Chronic Kidney Disease is common, grows quietly, and is costly to treat in its late stages. Traditional diagnosis relies on tests and specialists, which are slow and not available everywhere. This paper reviewed eight recent studies and the wider families of prediction method. Tree-based classifiers, neural networks and ensembles all reach high accuracy, especially on the UCI dataset, and SHAP and LIME have started to make the results readable for doctors. What is still missing is a system that joins two complementary algorithms, is tested on more varied data and explains every prediction.

The hybrid concept proposed here, with careful preprocessing, feature selection, machine learning and deep learning baselines, a hybrid of two algorithms and an XAI layer, was designed to fill that gap, and its evaluation plan was set out so the value of each part can be measured. Implementation and experiments are the next step, followed by tests on data from more hospitals and a simple web tool that returns a result, a risk level and the key health factors behind it.

REFERENCES:

[1] M. Rizwan, R. Naseem, M. A. Khan, A. Ahmad, M. Fozan and N. Muhammad, “Explainable AI machine learning framework for chronic kidney disease prediction utilizing electronic health records,” BMC Medical Informatics and Decision Making, vol. 26, art. no. 244, 2026, doi: 10.1186/s12911-026-03602-1.

[2] M. S. Arif, A. U. Rehman and D. Asif, “Explainable machine learning model for chronic kidney disease prediction,” Algorithms, vol. 17, no. 10, art. no. 443, 2024, doi: 10.3390/a17100443.

[3] S. K. Ghosh and A. H. Khandoker, “Investigation on explainable machine learning models to predict chronic kidney diseases,” Scientific Reports, vol. 14, art. no. 3687, 2024, doi: 10.1038/s41598-024-54375-4.

[4] D. A. Debal and T. M. Sitote, “Chronic kidney disease prediction using machine learning techniques,” Journal of Big Data, vol. 9, art. no. 109, 2022, doi: 10.1186/s40537-022-00657-5.

[5] P. Chittora et al., “Prediction of chronic kidney disease – A machine learning perspective,” IEEE Access, vol. 9, pp. 17312–17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[6] H. You, “Predictive analysis of chronic kidney disease based on machine learning,” Engineering Applications of Science Letters, vol. 4, no. 1, pp. 62–68, 2021, doi: 10.30538/psrp-easl2021.0062.

[7] O. A. Jongbo, A. O. Adetunmbi, R. B. Ogunrinde and B. Badeji-Ajisafe, "Development of an ensemble approach to chronic kidney disease diagnosis," Scientific African, vol. 8, art. no. e00456, 2020, doi: 10.1016/j.sciaf.2020.e00456.

[8] B. Deepika, K. R. Vasudeva Rao, D. N. Rampure, P. Prajwal and G. Devanand Gowda, "Early prediction of chronic kidney disease by using machine learning techniques," American Journal of Computer Science and Engineering Survey, vol. 8, no. 2, art. no. 7, 2020, doi: 10.36648/computer-science-engineering-survey.08.02.07.

[9] GBD Chronic Kidney Disease Collaboration, "Global, regional, and national burden of chronic kidney disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017," The Lancet, vol. 395, no. 10225, pp. 709–733, 2020.

[10] KDIGO CKD Work Group, "KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease," Kidney International Supplements, vol. 3, no. 1, pp. 1–150, 2013.

[11] L. Rubini, P. Soundarapandian and P. Eswaran, "Chronic kidney disease," UCI Machine Learning Repository, 2015.

[12] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[13] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.

[17] M. T. Ribeiro, S. Singh and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[18] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[19] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.