Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 10 Issue 02
2026

Research Article
https://doi.org/10.56979/1002/2026/1213

LightWeight Probabilistic Ensemble Model for Chronic Kidney Disease Detection

Ramya R[1*], Vakaimalar Elamaran[2], Muthulakshmi K[1], Anandh A[3], and Praveen Kumar Premkamal[1]

[1]Department of Artificial Intelligence and Data Science, Kamaraj College of Engineering and Technology, K.Vellakulam, 625 701, India.
[2]Department of Information Technology, Kamaraj College of Engineering and Technology, K.Vellakulam, 625 701, India.
[3]Department of Computer Science and Engineering, Kamaraj College of Engineering and Technology, K.Vellakulam, 625 701, India.
[*]Corresponding Author: Ramya R. Email: ramyaads@kamarajengg.edu.in

Received: December 08, 2025 **Accepted:** February 15, 2026

Abstract: CKD is the prolonged disease caused by gradual damage and reduction in the function of kidneys. The symptoms develop slowly and are non-specific to the disease. It is marked by high morbidity, mostly leading to the development of other serious health issues or mortality. Due to its asymptomatic nature, it is rarely diagnosed at early stages until it has progressed. Henceforth, early detection of CKD is crucial as there is more possibility of reducing the progression of the illness and improving treatment outcomes, once it is detected at an early stage. This work proposes a LightWeight Probabilistic Ensemble model that uses Logistic Regression and Naïve Bayes (LWPE-LRNB) with soft voting strategy to identify CKD with more accuracy and effectiveness. Several Machine Learning (ML) algorithms namely Decision Tree (DT), Logistic Regression (LR), KSTAR, Support Vector Machine (SVM) and Naïve Bayes (NB) are also compared with the proposed system. The metrics considered for evaluating the performance of ML models are Accuracy, Precision, Recall and F1-measure. Based on the performance values, it is inferred that our proposed system achieved an accuracy of 91.5%, a precision of 0.95, a recall of 0.92 and an F-measure of 0.93. Compared to other ML models, our lightweight probabilistic ensemble model implies the highest performance to ascertain CKD symptoms as the hybrid probabilistic model captures both the linear relationship and feature distribution. With its enhanced performance, it is concluded that our system would be the highly promising tool to the early detection of CKD which in turn supports timely diagnosis and treatment to improve patient outcomes.

Keywords: Clinical Predictive Model; Chronic Kidney Disease; Early Diagnosis; Feature Selection; Machine Learning

1. Introduction

CKD is a long-term disorder that is considered as the progressive loss in kidney performance over time. People suffering from CKD see a continuous decline in their kidney functionality over several months or years. At first, there are no symptoms, but later on, tiredness, limb swelling, nausea and disorientation could appear. Anemia, hypertension, fractures and cardiac issues are some of the other complications. Numerous conditions, such as polycystic kidney disease, diabetes and high blood pressure can make CKD worse. One risk factor of CKD is a history of long-lasting renal infection in the family. Normally, the diagnosis is made with a urine test to measure albumin levels and a blood test to determine the estimated glomerular filtration rate (eGFI). Moreover, to find the reason, an ultrasonography or biopsy needs to be performed.

ML algorithms employ sample data or training data to develop a model that enables one to make decisions autonomously, based on patterns and associations identified in the data. Subsequent sections will cover several key aspects such as Dataset Description, Framework, Proposed model, ML Algorithms used for comparison and Evaluation Metrics. These sections collectively contribute to a thorough exploration of developing and evaluating CKD prediction models using ML algorithms with various evaluation metrics. The research aims to deliver valuable perceptions into the predictive capabilities and interpretability of the models, ultimately contributing to advancements in healthcare decision-making and disease management.

Studies revealed kidney dysfunction that is described by a GFR under 60 mL/min/1.73 m², albuminuria with the value of 30 mg or more per 24 hours are the indication of kidney impairment persisting like hematuria or physical abnormalities like polycystic or dysplastic kidneys [1-4]. Paper-based studies involving large populations indicates that generally given medications are Proton Pump Inhibitors (PPIs) and have been linked to cases of acute interstitial nephritis in individual reports with an increased danger of incident CKD. Clinics based on a Kidney Failure Risk Equation (KFRE) have score exceeding 10% on referrals to multidisciplinary CKD [5]. Whether using a Kidney Failure Risk Equation (KFRE) based strategy leads to better management of CKD is currently being assessed by an ongoing study [6]. Studies have indicated that limiting protein intake may be beneficial in slowing the progression of CKD or in decreasing the danger of end stage kidney disease (ESKD) [7]. Artificial intelligence is increasingly being leveraged in medical applications to enhance diagnostic accuracy, streamline patient care and optimize treatment plans, making healthcare more efficient and personalized [8-14]. Various classifiers of different diagnostic capabilities are examined by R. Biswas et al. and S. Akter et al. For evaluation, the confusion matrix which is also called as the contingency matrix is used in ML. When classifiers accurately and confidently identify a condition, a True Positive (TP) occurs; when they incorrectly identify a condition, a False Positive (FP) occurs [15- 16].

Chen et al., focus was on risk calculation for patients with CKD using various single model approaches [17]. Conventional risk prediction approaches, such as the regression-based model offered strong clinical interpretability but are limited in modelling complex nonlinear relationships among risk factors [18]. ML techniques, including SVM enhance classification performance through feature selection strategies [19]. Although these models are transparent and computationally efficient, their performance might not be sufficient when complex interactions are involved.

Ensemble methods have proven their effectiveness in improving discrimination for CKD classification. Boosting-based XGBoost has proven its effectiveness by refining weak classifiers. Higher AUC values are achieved with structured clinical data [20]. Based on the given dataset, it is stated that ensemble model leverages the strengths of multiple individual classifiers while effectively managing and compensating for the weakness exhibited by each classifier [21]. Using various ensemble models, the risk of CKD is predicted based on various factors such as gender, age, hemoglobin levels, smoking habit and urine protein levels [22 - 23]. However, ensemble methods are not transparent and are computationally complex.

Khan et al., used DT, LR and random forest for CKD prediction [24]. Seven classifier algorithms such as Chi-square Automatic Interaction Detector (CHAID), Artificial Neural Network (ANN), C5.0, Linear SVM with L1 and L2 regularization, Random Forest (RF) and LR were used to predict CKD [25].

The structured preprocessing steps of an end-to-end tabular clinical prediction pipelines usually involve missing data imputation, normalization, categorical encoding, feature selection, handling class imbalance issues and Hyperparameters tuning. Methodological rigor in the development of prediction pipelines has been shown to be as important as the selection of a prediction algorithm itself in determining performance in clinical prediction research [26 – 27]. In our work, we are specifically concerned with a streamlined tabular clinical prediction pipeline.

Though several studies used ML for CKD detection, they focused on limited existing models without using hybrid probabilistic model that capture both the linear relationship and feature distribution. This work also performs analysis across parametric and nonparametric approaches.

2. Materials and Methods

The CKD dataset, with all of its attributes, serves as the input. Pre-processing eliminates redundant data and unidentified attributes. Every trait and characteristic is chosen after eliminating redundant attributes. A LightWeight Probabilistic Ensemble framework that uses LR and NB (LWPE-LRNB) is used.

Our work integrates LR and NB to utilize the complementary benefits of both the algorithms. Our proposed work is termed as a LightWeight framework as it involves few numbers of trainable parameters typically in tens where LR requires n+1 weights and NB requires 4n+1, n represents the no. of features, as compared to millions of parameters in deep learning architectures. This results in less computational complexity, low memory usage and faster training. The proposed framework deviates from the traditional soft voting paradigm in its use of a structured probabilistic fusion of models. This maintains the low computational cost and intrinsic transparency of the model while providing additional robustness. This places the model in the interesting space between interpretable models and powerful ensemble systems. Algorithms like NB, DT, KSTAR, LR and SVM are used to compare classification performance with the proposed model. There will be a visual depiction of comparison of these algorithms. We used the classifiers to determine the changes in kidney's morphology, using a huge volume of training data and test data. The usage of ML in the diagnosis of chronic renal disease is examined in this work. Several classification works are being benefited by the effective use of ML algorithms, which is becoming prominent in the recognition of anomalies in a variety of physiological data.

Figure 1 presents an overview of the system design, offering insights into the architectural layout and the integration of various elements. Figure 2 illustrates the flow diagram of our system, providing a visual representation of the key components and their interconnections.

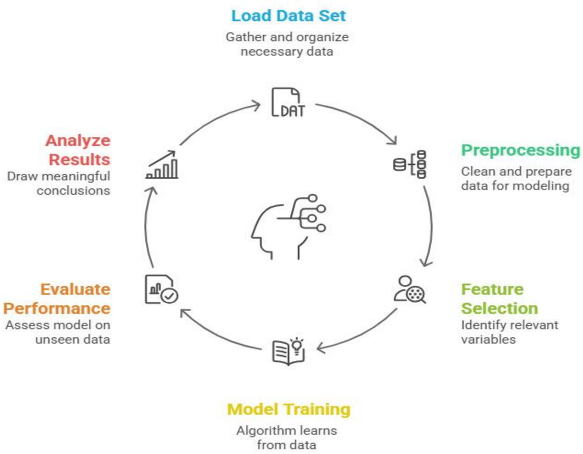

Figure 1. System Design of the proposed CKD prediction framework

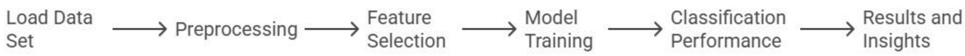

Figure 2. Flow Diagram of the proposed CKD prediction framework

2.1. Loading the Dataset

Here, we used CKD dataset. Our dataset consists of range of features or factors depicting each patient's CKD diagnosis results. The CKD data set, which is downloaded from the UCI Machine Learning Repository (Version 2), is utilized in this study. The data set is provided by the Apollo Hospitals Research Foundation in India. Each of the 400 cases in the data set includes one target variable and 24 predictor variables. This is a binary classification problem because the target variable has two possible outcomes: "ckd" (meaning the person has chronic kidney disease) and "notckd" (meaning the person does not have chronic kidney disease). In the data set, 250 cases (which are 62.5%) are labeled as "ckd" and 150 cases (which are 37.5%) are labeled as "notckd," making the data set slightly imbalanced. In every data analysis work involving ML, the initial step is to load the data. After loading the data set, data can be retrieved and used to carry out further analysis.

2.2. Pre-Processing

The next step is data analysis process. To perform statistical analysis of the data, the data must be cleaned, transformed and organized for better results. Missing value handling, duplicate value elimination, numerical feature normalization or scaling, categorical feature encoding is done in preprocessing. To make sure the data is accurate and the methods are reliable, a reproducible data preparation process is followed. Missing values in numeric data are filled in using the median, which helps to reduce the effect of extreme values and skewed data. Missing values in categorical data are replaced using the mode, which ensures that the most common category is kept. Before building models, any exact duplicate entries are removed to avoid any bias. Outliers are handled using the 1.5 times Interquartile Range (IQR) method, and instead of removing them, they are adjusted to the nearest limit. This is important because some of these outliers are meaningful in a clinical context. To prepare the data for ML models, all nominal variables are encoded using one-hot encoding. Finally, numerical variables are normalized using the Z-score method to make sure all variables are on the same scale. This step is very crucial and essential for enforcing the quality and integrity of the data to be used to train various ML models.

2.3. Feature Selection

Once the dataset is pre-processed, the most dominant and relevant features must be identified using the process of feature selection. The ultimate aim of feature selection is to achieve dimensionality reduction while preserving the most prominent features. These prominent features that are crucial in the prediction of CKD can be found out using various methods. Using a threshold of $|r| \geq 0.10$, the feature selection step applied point biserial correlation for the binomial labels and Pearson correlation for continuous features. This work removed one feature from each pair of highly correlated features when their correlations exceeded 0.85. This step helped to prevent multicollinearity. Notably, only the training data for each cross-validation fold is used for this step. Here, based on the correlation results, the features are selected from the data set.

2.4. Classifiers for Comparison

Various ML models can be utilized to examine patient data to predict CKD. In this work, KSTAR, SVM, DT, LR and NB are used to predict CKD effectively.

2.4.1. Description

KSTAR learning method is a prototype-based method that can handle imbalanced datasets, making it mainly suitable for predicting diseases where positive cases are few as compared to negative cases. This method works by generating a group of prototype models from training data that can be used to classify new test instances effectively based on the similarity with this group of prototypes.

SVM is a method that aims at finding the optimal hyperplane that is used to group various classes in feature space. This method is more effective in high dimensional feature spaces as it is common with clinical data involving various numerical attributes. By the use of appropriate kernel functions that helps to capture nonlinear relationships, SVM tries to maximize the margin between various classes so that it can achieve high accuracy in CKD prediction.

DTs are easy to interpret models that split the dataset into various subgroups based on the value of dominant features. This forms a tree like structure that aids in decision making. This model provides better insights regarding which of the features contributes the most to the hazard of developing the disease. Moreover, both the numerical and categorical data can be handled by decision tress making it suitable for clinical datasets.

LR is a commonly used method that is based on how likely the event is to occur such as the occurrence or non-occurrence of CKD. This is used in various applications because of its simplicity and interpretable nature. It estimates the relationship between diagnosis results of the patients and the possibility of CKD providing valuable insights that help in patient management.

NB is also a probability-based classifier that has its roots in Bayes theorem that assumes independence among predictive features. It performs well with high dimensional data because of its simplicity. As it can handle large datasets effectively and makes fast computation, it is a better choice for CKD prediction that allows assessments in real clinical environments.

2.4.2. Model Configuration

The KStar algorithm has a global blending parameter (b) of 20. It uses the default entropic distance measure. The SVM classifier uses a Radial Basis Function (RBF) kernel having a regularization parameter, C, set at 1.0. It also sets gamma to "scale". The DT classifier uses Gini impurity as its splitting criterion. It requires a minimum of two samples to split a node and has a maximum tree depth of 10. To ensure convergence, the LR model applies L2 regularization. It uses a liblinear solver and allows for a maximum of 1000 iterations. The Gaussian Naïve Bayes classifier is used for continuous features based on the assumption of normal distribution.

2.5. LightWeight Probabilistic Ensemble model

In this proposed system, a Lightweight Probabilistic Ensemble framework that uses LR and NB (LWPE-LRNB) is employed. Our work integrates LR and NB to utilize the complementary benefits of both the algorithms. LR is a discriminative model that uses posterior class probabilities to learn optimal decision boundary whereas NB is a generative model that uses feature distribution with an assumption of conditional independence.

Initially, both the LR and NB models are independently trained on the data set which is preprocessed to find a probabilistic prediction. This prediction indicates the likelihood of the CKD presence. Then the predictions are combined together using soft voting strategy by averaging out the probabilities generated by both the models. Thus, this approach captures both the global linear relationship feature from LR and probabilistic feature distribution from NB.

2.6. Classification Performance Metrics

The performance achieved by various ML models that categorizes the outcome as the existence or absence of CKD is measured using various metrics. The performance is generally assessed using several metrics such as accuracy, precision, recall, F1-score and confusion matrix.

By comparing the amount of correctly predicted instances to all of the occurrences in the dataset, the accuracy ratio calculates the correctness of the model. Precision evaluates the number of true positives among all instances that the model classified as positive, regardless of whether the actual label is positive or negative. Recall evaluates the capability of the model to identify the actual positive instances appropriately. F1-score poises both the precision and recall whereas the confusion matrix illustrates the results of prediction as true positive classification versus predicted classification.

3. Results

CKD dataset is divided into four data sets evenly so that the sample distribution is retained. The performance of the model is measured comprehensively. Using the selected ML models, all the subsets are used once in testing while the remaining subsets are being used as training dataset. The final performance is assessed using all the results obtained.

Figure 3 illustrates the CKD dataset sample that is used in our study. The dataset serves as the basis for our research, providing vital information for the evaluation and validation of our used algorithms. The attributes present in the dataset is characterized in Table 1. Figure 4 depicts the cleaned sample data after preprocessing. This important process helps to remove the inconsistent and missing information, improving the reliability of subsequent analysis. Figure 5 shows the attributes that are selected from the CKD dataset. The selection of relevant attributes is important to enhance the accuracy of our predictive models, inducing the overall system performance.

Table 1. Attributes in CKD dataset

Feature	Type	Description
htn	Qualitative	Hypertension
dm	Qualitative	Diabetes mellitus
cad	Qualitative	Coronary artery disease
appet	Qualitative	Appetite
pe	Qualitative	Pedal edema
ane	Qualitative	Anemia
rbc	Qualitative	Red blood cells (normal/abnormal)
pc	Qualitative	Pus cells
pcc	Qualitative	Pus cell clumps
ba	Qualitative	Bacteria
bgr	Quantitative	Blood glucose random
bu	Quantitative	Blood urea
sc	Quantitative	Serum creatinine

sod	Quantitative	Sodium
pot	Quantitative	Potassium
hemo	Quantitative	Hemoglobin
pcv	Quantitative	Packed cell volume
wc	Quantitative	White blood cell count
rc	Quantitative	Red blood cell count
age	Quantitative	Patient age
bp	Quantitative	Blood pressure
sg	Quantitative	Specific gravity
al	Quantitative	Albumin
su	Quantitative	Sugar
class	Target	ckd / notckd

```
29 @data
30 48,80,1.020,1,0,?,normal,notpresent,notpresent,121,36,1.2,?,?,15.4,44,7800,5.2,yes,yes,no,good,no,no,ckd
31 7,50,1.020,4,0,?,normal,notpresent,notpresent,?,18,0.8,?,?,11.3,38,6000,?,no,no,no,good,no,no,ckd
32 62,80,1.010,2,3,normal,normal,notpresent,notpresent,423,53,1.8,?,?,9.6,31,7500,?,no,yes,no,poor,no,yes,ckd
33 48,70,1.005,4,0,normal,abnormal,present,notpresent,117,56,3.8,111,2.5,11.2,32,6700,3.9,yes,no,no,poor,yes,yes,ckd
34 51,80,1.010,2,0,normal,normal,notpresent,notpresent,106,26,1.4,?,?,11.6,35,7300,4.6,no,no,no,good,no,no,ckd
```

Figure 3. Sample data of CKD Dataset

age	bp	sg	al	su	rbc	pc	pcc	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	class
48	80	1.02	1	0	normal	notpresent	notpresent	121	36	1.2	140	4.4	15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
7	50	1.02	4	0	normal	notpresent	notpresent	110	18	0.8	138	3.8	11.3	38	6000	4.8	no	no	no	good	no	no	ckd
62	80	1.01	2	3	normal	normal	notpresent	423	53	1.8	135	4.0	9.6	31	7500	4.6	no	yes	no	poor	no	yes	ckd
48	70	1.005	4	0	normal	abnormal	present	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
51	80	1.01	2	0	normal	normal	notpresent	106	26	1.4	141	4.2	11.6	35	7300	4.6	no	no	no	good	no	no	ckd

Figure 4. After applying Preprocessing workflow - Data Cleaning to the sample data of CKD dataset

sg
al
hemo
pcv
sc
bgr
bu
sod
rc
bp
age

Figure 5. Attributes selected after Preprocessing workflow – Feature Selection

A reproducible evaluation framework has been implemented using stratified four-fold cross-validation with a fixed random seed of 42 that is chosen for consistency. During each and every iteration, one fold is chosen for testing whereas remaining folds are used for training purpose. The pre-processing and feature-selection steps are applied on the training data first and then applied to the test fold to avoid data leakage. This step is repeated for all the folds so that every fold served once as a test set. The aggregated results obtained across all folds generate a more reliable estimate of model generalization which in turn improves the overall performance.

Table 2 shows the performance comparison of various algorithms such as NB, DT, KStar, LR and SVM with the proposed system (LWPE-LRNB). The metrics included for evaluation are accuracy, precision, recall and F1 score. Our LWPE-LRNB system has been implemented and the resultant confusion matrix is tabulated in Table 3. The results show that the model effectively identifies true CKD cases and minimizes false positives which make it suitable for clinical decision-support where the accuracy of precision and recall are indispensable. The occurrences of errors have been noticed primarily for patients with near-threshold laboratory values, where early-stage CKD look like normal renal variation. Figure 6 depicts the visual representation of accuracy that is achieved among all the algorithms. This visual analysis provides valuable insights about the capabilities of each algorithm.

Figure 7 depicts the visual comparison of precision, recall and F1 score of all the algorithms that are used in our study. Thus, the detailed analysis of all these metrics provides a better understanding regarding all the algorithms.

Table 2. Performance comparison of ML models for CKD classification

Algorithm	Precision	Recall	F-Measure	Accuracy
NB	0.849	0.845	0.847	84.516
DT	0.802	0.819	0.806	81.935
KStar	0.831	0.839	0.833	83.871
LR	0.86	0.865	0.862	86.452
SVM	0.807	0.826	0.808	82.581
LWPE-LRNB (Proposed work)	0.95	0.92	0.93	91.5

Table 3. Confusion Matrix obtained for LWPE-LRNB (Proposed work)

Actual / Predicted	CKD	notCKD
CKD	229	21
notCKD	13	137

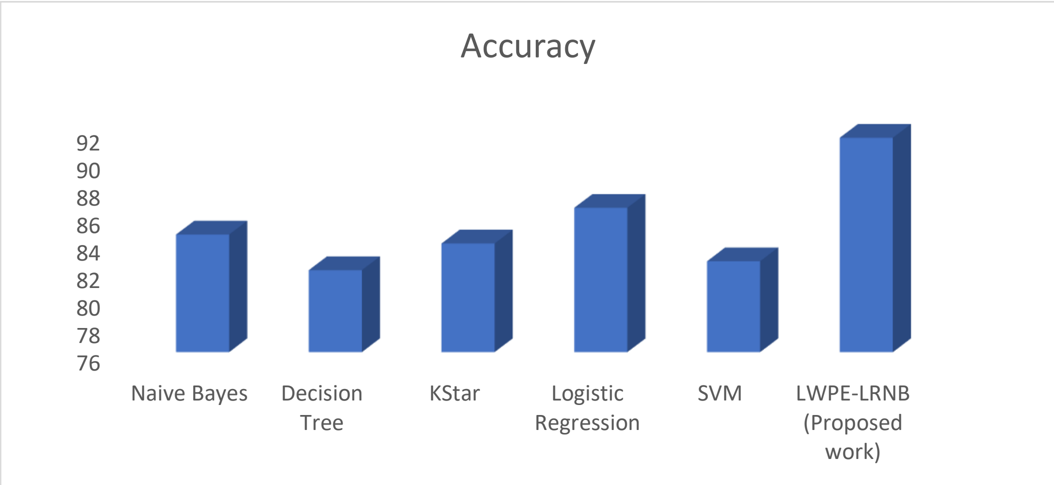

Figure 6. Accuracy comparison of evaluated classifiers for CKD detection

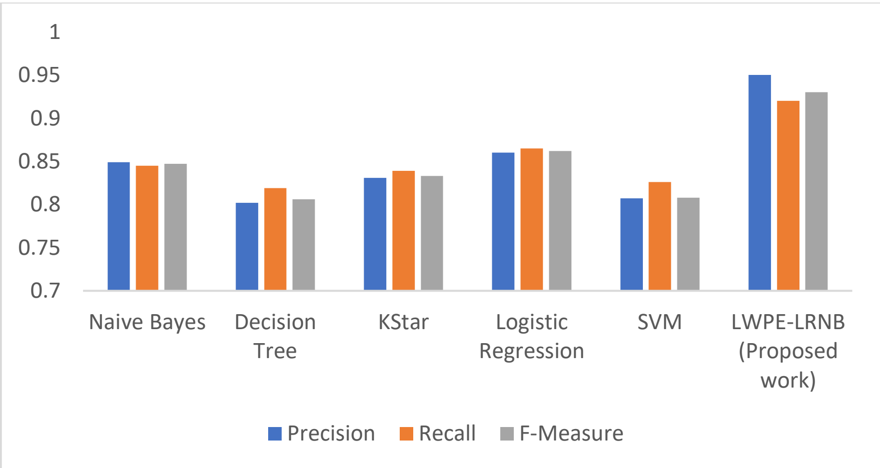

Figure 7. Graphical comparison of precision, recall and F-measure for various evaluated classifiers for CKD detection

4. Discussion

Precision signifies the accuracy of positive predictions. Recall indicates the percentage of actual positives that are correctly recognized. Accuracy measures how correct the proposed model is. F1-score balances recall and precision. These metrics provide thorough evaluation of the used algorithms. As shown in Table 2, NB achieved precision of 0.849, recall of 0.845, F-measure of 0.847 and accuracy of 84.516%. DT achieved 0.802, 0.819, 0.806, and 81.935% for precision, recall, F-measure and accuracy respectively. KStar achieved 83.871% accuracy, 0.831 precision, 0.839 recall, and 0.833 F-measure. LR showed the results as 86.452% for accuracy, 0.86 for precision, 0.865 for recall and 0.862 for F1-score. SVM achieved an accuracy rate of 82.581%, with a precision score of 0.807, recall of 0.826, and F1-score of 0.808. Our proposed LWPE-LRNB method achieved a precision of 0.95, recall of 0.92, F-measure of 0.93 and 91.5% accuracy. Our proposed LWPE-LRNB method performed better over all other models. These results ensure the feasibility of the proposed work. Based on the misclassification analysis, LWPE-LRNB is chosen as the model that performed best with an accuracy of about 91.5%.

5. Conclusions

The usage of ML techniques to detect CKD is experientially observed in this work. ML techniques provide more advantages such as cost effectiveness, scalability, efficiency, accuracy and personalization over other conventional diagnostic approaches. Meticulous selection and design of input and output parameters of the model is a critical aspect in developing a ML model for CKD. It is interpreted that the proposed LWPE-LRNB model consistently outperforms well, which is illustrated using significant performance metrics including accuracy, precision, recall and F1 score as compared to other models. It gains accuracy of 91.528% which proves the efficacy of the proposed model in identifying CKD at early stages which is significant for timely intervention and better disease management. Such models could be integrated into clinical practice which enhances early detection efforts and enable more effective treatment strategies, eventually improving patient outcomes and slowing disease progression.

Funding: This research received no external funding.
Acknowledgments: The authors would like to acknowledge Kaggle (https://www.kaggle.com) for providing the publicly available dataset used in this study.
Conflicts of Interest: The authors declare no conflict of interest.

References

1. Coresh, J.; et al. Prevalence of chronic kidney disease in the United States. JAMA 2007, 298, 2038–2047. https://doi.org/10.1001/jama.298.17.2038
2. Hsu, C.-Y.; Vittinghoff, E.; Lin, F.; Shlipak, M.G. The incidence of end-stage renal disease is increasing faster than the prevalence of chronic renal insufficiency. Ann. Intern. Med. 2004, 141, 95–101. https://doi.org/10.7326/0003-4819-141-2-200407200-00007
3. Plantinga, L.C.; et al. Patient awareness of chronic kidney disease. Arch. Intern. Med. 2008, 168, 2268–2275. https://doi.org/10.1001/archinte.168.20.2268
4. Jha, V.; et al. Chronic kidney disease: Global dimension and perspectives. Lancet 2013, 382, 260–272. https://doi.org/10.1016/S0140-6736(13)60687-X
5. Klahr, S.; et al. The effects of dietary protein restriction and blood-pressure control on the progression of chronic renal disease. N. Engl. J. Med. 1994, 330, 877–884. https://doi.org/10.1056/NEJM199403313301301
6. Hingwala, J.; et al. Risk-based triage for nephrology referrals using the kidney failure risk equation. Can. J. Kidney Health Dis. 2017, 4, Article ID 2054358117722782, 1–8. https://doi.org/10.1177/2054358117722782
7. Smekal, M.D.; et al. Patient and provider experience and perspectives of a risk-based approach to multidisciplinary chronic kidney disease care: A mixed methods study. BMC Nephrol. 2019, 20, Article 110, 1–11. https://doi.org/10.1186/s12882-019-1269-2
8. Hameed, E.K.; Al-Ameri, L.T. Artificial intelligence: The gateway to better medical diagnosis. Al-Kindy Coll. Med. J. 2024, 20, 1–3. https://doi.org/10.47723/w1ymg293
9. Wajeeh, Z.A.; Hamandi, S.J.; Al-Obaidi, W.S.; Kervan, U. Artificial intelligence-based prediction of intra-aortic balloon pump need in high-risk CABG surgery: Enhancing clinical decision support. Al-Kindy Coll. Med. J. 2024, 20, 145–151. https://doi.org/10.47723/sf00fq21
10. Nayak, V.; Vora, N. A machine learning-based career recommendation. J. Trends Comput. Sci. Smart Technol. 2024, 6, 374–390.
11. Ramya, R.; Anandh, A.; Premkamal, P.K.; Elamaran, V. Deep learning-based algorithm for prediction of heart disease using electrocardiogram. INFOCOMP J. Comput. Sci. 2024, 23, 1–9.
12. Mrudhhula, V.S.; Thirumahal, R. A comprehensive survey of arrhythmia classification techniques. J. Trends Comput. Sci. Smart Technol. 2024, 6, 357–373.
13. Mushtaq, M.; Mehmood, Z.; Butt, A.A.; Shabir, M.Y.; Khan, M.N.U. Causal and explainable machine learning framework for heart disease prediction using XGBoost and SHAP. J. Comput. Biomed. Inform. 2025, 10(1).
14. Akeel, M.; Butt, K.K.; Javed, K.; Tariq, M.; Yousaf, M. Email spam detection using machine learning with optimized feature engineering and classification techniques. J. Comput. Biomed. Inform. 2025, 10(1).
15. Biswas, R.; et al. A robust deep learning based prediction system of heart disease using a combination of five datasets. Proc. Int. Conf. Comput. Theory Appl. (ICCTA) 2021, 223–228. https://doi.org/10.1109/ICCTA54562.2021.9916601
16. Akter, S.; Shamrat, F.M.J.M.; Chakraborty, S.; Karim, A.; Azam, S. COVID-19 detection using deep learning algorithm on chest X-ray images. Biology 2021, 10, Article 1174. https://doi.org/10.3390/biology10111174
17. Chen, Z.; Zhang, X.; Zhang, Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. Int. Urol. Nephrol. 2016, 48, 2069–2075. https://doi.org/10.1007/s11255-016-1346-4
18. Kshirsagar, A. V.; Bang, H.; Bomback, A. S.; Vupputuri, S.; Shoham, D. A.; Kern, L. M.; Klemmer, P. J.; Mazumdar, M. A simple algorithm to predict incident kidney disease. Arch. Intern. Med. 2008, 168, 2466–2473. https://doi.org/10.1001/archinte.168.22.2466
19. Polat, H.; Danaei Mehr, H.; Cetin, A. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. J. Med. Syst. 2017, 41, 55. https://doi.org/10.1007/s10916-017-0703-x
20. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD '16) 2016, 785–794. https://doi.org/10.1145/2939672.2939785
21. Afrin, S.; et al. Supervised machine learning based liver disease prediction approach with LASSO feature selection. Bull. Electr. Eng. Inform. 2021, 10, 3369–3376. https://doi.org/10.11591/eei.v10i6.3242
22. Wang, W.; Chakraborty, G.; Chakraborty, B. Predicting the risk of chronic kidney disease (CKD) using machine learning algorithm. Appl. Sci. 2020, 11, Article 202. https://doi.org/10.3390/app11010202
23. Almasoud, M.; Ward, T.E. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int. J. Adv. Comput. Sci. Appl. 2019, 10, 179–184. https://doi.org/10.14569/IJACSA.2019.0100813

24. Khan, P.; Reddy, M.; Samatha, K.; Chowdary, R.; Rao, P. Predictive analytics of chronic kidney disease by using machine learning. Cardiff Metrop. Univ. Rep. 2019, Article ID 14617014, 1–6. https://doi.org/10.25401/cardiffmet.14617014
25. Chittora, P.; et al. Prediction of chronic kidney disease – A machine learning perspective. IEEE Access 2021, 9, 17312–17334. https://doi.org/10.1109/ACCESS.2021.3053763
26. Collins, G. S.; Reitsma, J. B.; Altman, D. G.; Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BJS, 2015, 102(3). https://doi.org/10.1136/bmj.g7594
27. Steyerberg, E. W. Clinical prediction models: A practical approach to development, validation, and updating; Springer: New York, NY, USA, 2009. https://doi.org/10.1007/978-0-387-77244-8