Computational Journal of Mathematical and Statistical Sciences
x(x), 1–16
DOI:10.21608/cjmss.2025.374590.1150
https://cjmss.journals.ekb.eg/

Research article

Using Machine Learning Models with Elephant Herd Feature Selection Method for Diagnosing Chronic Kidney Disease

N.Senthil Selvan[1], **B.Elangovan** [2], **R.Bala Krishnan**[3], **G. Manikandan**[4,*]

[1] School of Computing, SASTRA Deemed To Be University, Thanjavur 613401, India; senthilselvan@cse.sastra.edu

[2] Department of CSE, Koneru Lakshmaiah Education Foundation, Guntur 522302, India; elan77@gmail.com

[3] Department of CSE, SASTRA Deemed to be University, Kumbakonam 612001, India; balakrishnan@src.sastra.edu.

[4] School of Computing, SASTRA Deemed To Be University, Thanjavur 613401, India; manikandan@it.sastra.edu.

* **Correspondence:** manikandan@it.sastra.edu

Abstract: The kidneys filter waste, toxins, and excess water from the bloodstream, promoting body balance. Chronic kidney disease (CKD) is defined as a progressive deterioration of kidney function. Chronic kidney disease is a serious health condition that affects people all over the world. Early detection of kidney disease is crucial due to the lack of visible symptoms. The timely diagnosis of chronic kidney disease (CKD) significantly impacts the patient's health development and allows for prompt treatment. The primary goal of this study is to detect the presence or absence of CKD in the human body by analyzing various features obtained from medical tests. So, to accurately detect and diagnose chronic kidney disease (CKD), we use machine learning techniques such as classifiers K-Nearest Neighbor (KNN), Naive Bayesian (NB), Logistic Regression (LR), and Decision Tree (DT). The study used the 400 instances and 25 features of the chronic kidney disease dataset obtained from Kaggle. The elephant herd optimization feature selection algorithm identifies eleven key features. Decision tree outperformed other classification algorithms, achieving an accuracy of 98.75%.

Keywords: Chronic Kidney Disease, KNN, Naive Bayes, Decision Tree, Machine Learning, Logistic Regression, Classification Model.

Mathematics Subject Classification: 62J12, 62H30, 91A27.

Received: 24 February 2025; Revised: 2 June 2025; Accepted: 5 June 2025 2025; Online: 17 June 2025 2025.

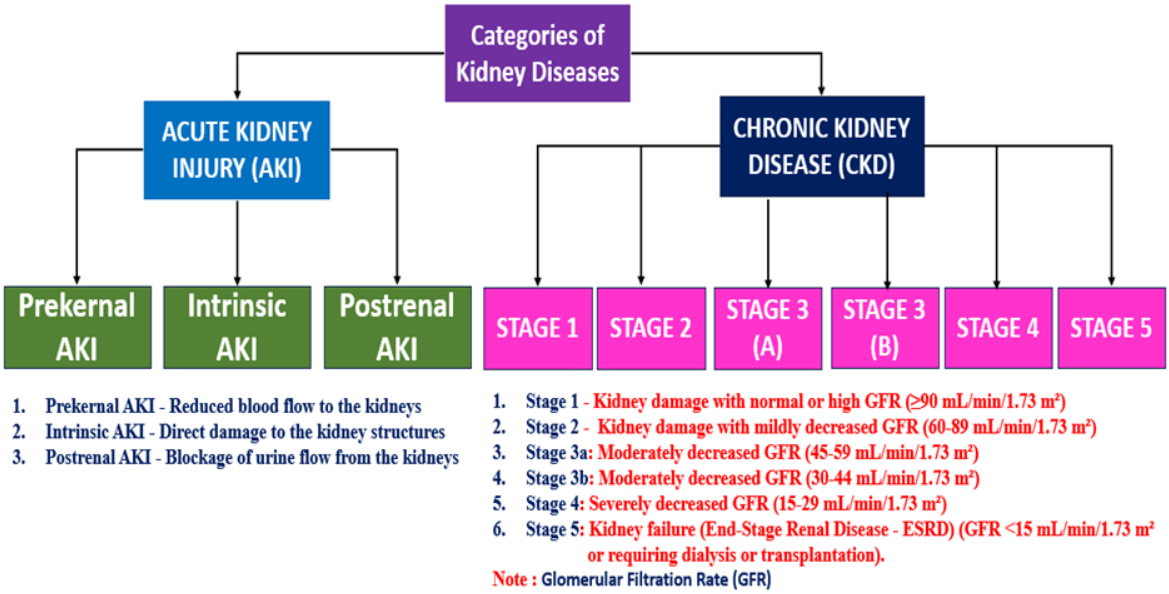

Figure 1. Broad Category of Kidney-related Illness for Humans

1. Introduction

In today's world, kidney problems are becoming more and more of a concern due to the severe damage they provide to public health. Millions of individuals worldwide are impacted by these illnesses, which compromise their physical health, put pressure on healthcare systems, and cause financial hardship. To increase awareness and encourage preventative measures, it is essential to comprehend the various forms of renal illnesses, their risk factors, and the increasing prevalence of these conditions. For several causes, kidney illnesses are becoming more common in today's culture. One main offender is the rising incidence of diseases like high blood pressure and diabetes, both of which can seriously harm the kidneys over time [1, 2].

Kidney diseases or Kidney-related illnesses are presented in Figure 1. It is a broad category of diseases that impair the kidneys' capacity to remove waste from your blood. Certain conditions, such as acute kidney injury (AKI), progress quickly as a result of urinary tract obstructions or dehydration. When treated quickly, AKI can occasionally be reversed, but in extreme circumstances, it can cause irreversible kidney damage [3]. On the other hand, diabetes and high blood pressure can damage the delicate kidney filtering units, causing chronic kidney disease (CKD), which advances slowly. The most severe level of CKD, stage 5, necessitates dialysis or a kidney transplant in order to survive. Any ailment that harms your kidneys and impairs their capacity to operate is referred to as kidney disease. Acute kidney injury (AKI) and chronic kidney disease (CKD) are the two primary forms. Acute kidney injury can be brought on by a number of conditions, such as acute dehydration, urinary tract obstructions, infections, or other fast-onset events that typically occur within hours or days [4]. When treated quickly, AKI can sometimes be reversed, but it can also occasionally result in irreversible kidney damage. On the other hand, chronic kidney disease progresses gradually over several years. Diabetes and high blood pressure are the most common causes, as they can harm the kidneys' small filters. Chronic kidney disease is a savvy plunder that frequently goes undetected until much later in life [5, 6]. Early

detection is essential for problems like renal failure to be successfully managed and avoided. This is where using machine learning to combat chronic kidney disease becomes revolutionary.

Machine learning has made it possible for computers to analyze massive datasets of medical records, including blood test results, blood pressure readings, urine results, and even patient demographics. Machine Learning (ML) algorithms that identify complex patterns in the data can forecast the risk of developing chronic kidney disease (CKD) with impressive Accuracy. This facilitates prompt response by medical practitioners, perhaps impeding the progression of an illness and improving patient results.

Moreover, kidney scans and other massive imaging datasets can be analyzed using machine learning. Machine learning (ML) algorithms can help physicians diagnose chronic kidney disease (CKD) and distinguish it from other illnesses by automatically detecting subtle anomalies, such as kidney size, shape, or density changes. This can lessen the necessity for invasive treatments like kidney biopsies and increase the Accuracy of diagnoses. ML may even be able to identify CKD earlier than conventional techniques in certain circumstances, enabling early intervention and maybe enhancing long-term patient outcomes.

By using machine learning algorithms to forecast the course of an illness, physicians can customize treatment regimens to meet the needs of each patient. An ML model may predict a patient with early-stage chronic kidney disease (CKD) to have a rapid loss in kidney function. With this information, the physician can recommend a more vigorous course of therapy, which may delay or even stop the disease's progression. Moreover, ML can be used to identify patients who have a high chance of acquiring osteoporosis or cardiac disease as a result of their chronic kidney illness. Patient outcomes for these problems can be significantly enhanced by early intervention. ML has the potential to completely change the way CKD is managed and improve the lives of millions of patients globally by identifying patients who are at high risk of complications and customizing treatment programs for them. One potent tool in the fight against chronic renal disease is machine learning. Machine learning (ML) has the potential to transform the management of chronic kidney disease (CKD) and enhance the lives of millions of patients globally by facilitating earlier detection, identifying latent risk factors, and aiding in diagnosis.

The study's main contributions are as follows:

- KNN imputer in the suggested strategy resolves the missing values in the data set.
- The elephant herd optimization algorithm is used to identify the significant features in the data set.
- To perform a comparative analysis, machine learning models like logistic regression (LR), decision tree (DT), Naive Bayes (NB), and support vector machine (SVM) have been used. The proposed model's effectiveness is compared using metrics like Accuracy, Precision, Recall, and F1-Score to analyze its capabilities thoroughly
- The proposed model enables medical experts to diagnose and treat patients, increasing their chances of quickly survival.

The remaining sections of the paper have the following structure: Section 2 explores ML applications for CKD health conditions. Section 3 outlines the proposed approach. Section 4 is devoted to confronting and hashing out the study's findings. Section 5 provides a conclusion and outlines possible future directions.

2. Related works

The technique has certain limitations that the authors Qin, J. et al. [7] suggested for highly accurate CKD diagnosis assistance. The concern arises from the study's dependence on a publicly accessible dataset because these datasets might not translate well to actual situations. The whole range of variance observed in real-world patients may not be represented in public databases, and they may be biased due to the methods used in data collection and labelling. Furthermore, there is a chance that the process of imputing missing values—KNN imputation—contains biases of its own. KNN imputation makes the potentially unfounded assumption that data points with comparable attributes would likewise have comparable missing values. This could distort how the model interprets the connections between the target variable and the features.

A recent study by the researchers De Almeida et al. [8] presents a viable method for machine learning-based early detection. It emphasizes how crucial early diagnosis results in better treatment results. Nonetheless, a useful criticism would be to take into account the particular machine-learning methods covered in the paper. The study by I. U. Ekanayake et al. [9] provides a valuable method for using machine learning to predict chronic kidney disease (CKD). A critical assessment, however, necessitates a detailed examination of the techniques applied to feature selection, missing value handling, and data preprocessing.

The possibility of a particular deep learning architecture for chronic kidney disease (CKD) diagnosis is examined in the paper by the authors Ma, F. et al. [10]. An innovative solution to this issue is provided by the authors' suggested neural network design, a heterogeneous modified artificial neural network. In this study, Nimmagadda, S. M. et al. [11] investigate the possibilities of machine learning in identifying chronic kidney disease (CKD). For example, defining the machine learning methods under investigation and comparing their performance might improve the research. El-Kenawy et al. [12] discussed predicting sleep disorders: leveraging sleep health and lifestyle data with dipper throated optimization algorithm for feature selection and logistic regression for classification.

An insightful comparison of different ensemble approaches for disease prediction is provided by the research conducted by P. Mahajan et al. [13]. Its strength resides in consistently assessing several ensemble methodologies over a wide range of disease datasets, going beyond the bounds of established research. This increases the scope of our knowledge on which ensemble techniques are generally more useful for tasks related to illness prediction. To address a specific nephrology challenge, the authors Takkavatakarn, K. et al. [14] employ machine learning algorithms to identify patients with stage 4 chronic kidney disease (CKD) who are at high risk of developing into end-stage kidney disease (ESKD) within three years. The researchers successfully forecasted the risk of ESKD after putting numerous machine learning models into practice, such as random forests and artificial neural networks.

The study by Li, Y. et al. [15] offers unique machine-learning techniques to evaluate patient data and identify biomarkers that may be used to distinguish between distinct CKD-MBD subtypes. Although the study's machine learning models perform well on the training set, certain restrictions limit how broadly the findings may be applied. The study's comparatively small sample size raises questions regarding how reliable the results are. The study conducted by Zhu E. et al. [16] examines the relationship between calcific aortic valve disease (CAVD) and chronic kidney disease (CKD), with a particular emphasis on the immune system's function.

The enormous resources needed to develop reliable reference ranges (RIs) for laboratory tests are

a critical challenge in clinical diagnostics discussed in the article by Velev, J. et al.[17]. Conventional techniques can be costly and time-consuming, frequently resulting in the adoption of generic RIs that don't consider unique patient factors. By utilizing unsupervised machine learning on a sizable dataset of routine test results from the Puerto Rican population, the authors offer a revolutionary alternative. A potential 3D convolutional neural network (CNN) for autonomous kidney segmentation in patients with chronic kidney disease (CKD) is proposed [18]. Selection bias is a risk when relying solely on a single-center cohort because the patients enrolled may not be typical of the larger CKD community.

The research investigation by Van Nynatten et al. [19] offers a novel and potentially proper diagnostic technique for renal issues. This 21-protein panel provides a more accurate and effective way to detect kidney disease, which could significantly improve patient care. The absence of information on the patient sample is one major drawback. Because the number and makeup of the groups under study are not specified, it is challenging to evaluate how well the test generalizes to a larger population. Age, ethnicity, and underlying medical disorders are only a few factors that can affect how kidney disease manifests in different demographics and the expression of biomarkers. The black box dilemma, as described in the paper by Ghosh, S. K. et al.[20]. It represents a significant obstacle for machine learning applications in medicine. Even if the study finds important features like creatinine and has excellent Accuracy with XGBoost, a more thorough evaluation would look at the constraints of the selected interpretability approaches (SHAP, LIME) and investigate potential biases in the dataset. A more comprehensive examination would yield a more profound comprehension of the model's applicability and efficiency in real-world scenarios.

The possibility of machine learning for identifying people at high risk of osteoporosis using nationwide chronic disease data is investigated by Tu, J. B. et al. [21]. A dataset of 10,000 patients forms the basis of the analysis, which may not be sufficient to guarantee generalizability. Furthermore, there may be biases in the data that the authors are unaware of, which could distort the model's predictions. The potential of machine learning for the detection of chronic kidney disease (CKD) is examined in the study by the researchers Chittora, P. et al. [22]. Its usage of a single, publicly accessible dataset may be a significant limitation, even though it probably assesses several categorization algorithms and provides accurate reports. Because real-world data can differ significantly, this restricts the conclusions' generalizability.

The study by Halder, R. K. et al. [23] significantly adds to the field of CKD prediction. It focuses on data preparation methods and an intuitive web application. Concerns are raised, nonetheless, by the claimed 100% accuracy on several classifiers. Achieving such high Accuracy in medical applications is unusual. It may be a sign of overfitting, a phenomenon in which the model performs well on training data but may not generalize to new data. The dependability of the model would be increased by testing outside datasets more or employing methods to deal with overfitting. The potential of machine learning algorithms in detecting and forecasting chronic kidney disease (CKD) is investigated by the writers Rahman, M. M. et al. [24]. Research on the model that uses machine learning to diagnose chronic kidney disease is very promising. The authors of the study Rehman, A. et al.[25] introduced a novel approach for forecasting chronic diseases called HCDP-DELM. This approach makes use of Deep Extreme Learning Machines (DELM) to manage the heterogeneous (composed of multiple sources) and complex character of medical data. Another strength of the model is its incorporation of temporal characteristics, which enable it to take into account the temporal variations in medical data.

The application of a fuzzy logic system and adaptive backpropagation neural network (ABNN) for

the categorization of chronic kidney disease (CKD) was investigated in a recent study by KR Vineetha et al. [26]. Although there is potential for using these techniques in medical diagnostics, a rigorous evaluation is required. The fuzzy logic system's design and the selected ABNN architecture determine how effective the suggested approach will be. Details on the precise CKD diagnosis criteria employed and the performance comparison with current CKD categorization methods will also enhance the article's credibility.

An innovative method for relating single-cell RNA sequencing data to clinical outcomes in renal illness was presented by the authors Legouis, D. et al. [27]. Though encouraging, the analysis is dependent on combining a small number of new patient samples with pre-existing single-cell datasets. The K-CLIER framework's generalizability for other clinical features of kidney disease remains unexplored, as the validation concentrates on dimensionality reduction and its association with fibrosis. The study by Tang et al. [28] explores an excellent use of machine learning in critical care; nonetheless, a thorough analysis would necessitate evaluating the particular data, technique, and model performance. The way the authors handled issues like bias and data cleaning, the selection and rationale of machine learning models, and the models' relative performance versus other approaches are all possible topics for future study.

The possibility of a traditional Chinese medication for the treatment of chronic renal disease is examined in the research by Liu, Y. et al. [29]. Although the study identifies constituents of Shen Qi Wan and investigates its anti-inflammatory properties in mice, its generalizability is limited since it exclusively uses adenine-induced models and AQP1 mutant mice. Furthermore, more investigation is required into how Shen Qi Wan affects DEFB1 expression to validate the suggested AQP1-mediated pathway. The authors Reznichenko, A. et al.'s study [30] suggests an alternative strategy to the current indirect measure-based classification of chronic kidney disease (CKD): molecular analysis of renal tissue. This provides a more accurate knowledge of the root problems and may result in customized treatment plans. However, the study depends on kidney biopsy data, which can be intrusive, and the urine protein signature that was found needs to be validated before it can be used in a clinical setting.

In their research, Sasikaladevi N. et al. [31] explored a possible method for CKD identification. Although the idea of a digital twin for early diagnosis is intriguing, more information about the precise algorithms employed and how they were validated would be helpful for a thorough assessment. It would also be beneficial to comprehend how CT-radiography integration fits into the larger system and how it handles certain drawbacks like radiation exposure.

3. Proposed Method

The system leverages the Elephant Herd Optimization feature selection algorithm to extract key features from the Chronic kidney dataset from Kaggle. This dataset is widely used in many research papers and is a standard point of comparison in evaluating similar systems. Figure 2 illustrates the proposed system's workflow.

3.1. Data Preprocessing

The Chronic Kidney Disease dataset sourced from Kaggle comprises 400 entries and 25 attributes, encompassing 11 numeric and 14 nominal attributes. To handle null values in a data set, comprehensive data preprocessing is necessary prior to analysis. Several steps are involved in data preprocessing,

including data transformation, cleaning, and normalization. The initial step involves handling the dataset's missing values. The underlying distribution is kept intact when using the median imputation approach to fill in the null values for numerical attributes. The null values in a single attribute are filled using the KNN-based imputation technique. The KNN-based imputation technique has the benefit of preserving the intrinsic structure of the data. The label encoding technique converts nominal properties to numerical equivalents for analysis purposes. This conversion enables seamless integration of categorical data into machine learning algorithms. Standardizing the range of features ensures fair comparisons and prevents certain traits from influencing model training. The features are scaled within a predetermined range using the min-max normalization technique. The following pseudocode

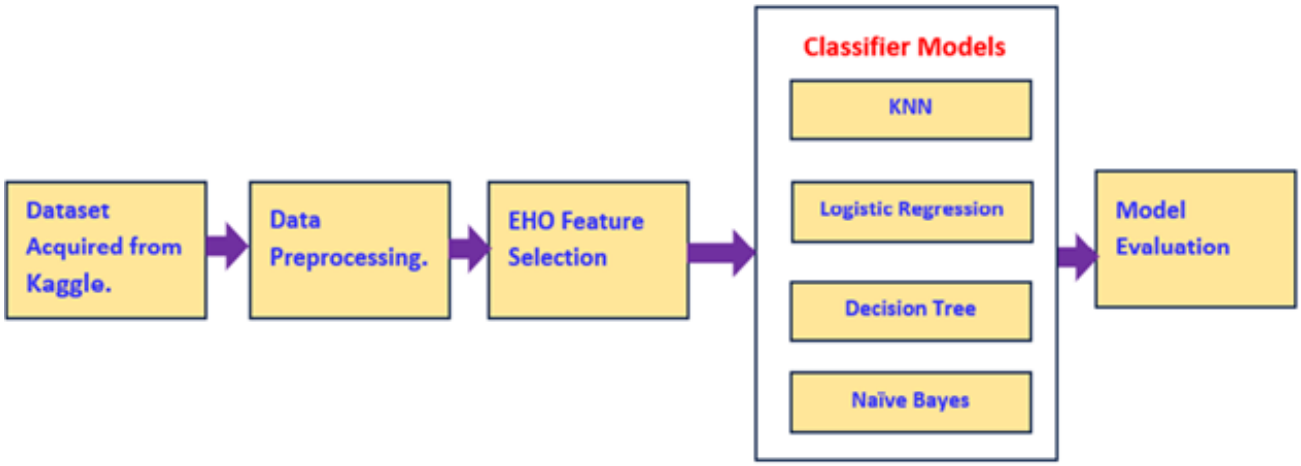

Figure 2. Proposed System for CKD

encapsulates the comprehensive steps within the proposed system.

1. Load the CKD dataset. Replace null values using the KNN-based imputation technique.
2. Apply label encoding to convert nominal properties to numerical equivalents.
3. Scale the features using the min-max normalization technique.
4. Create a population of "elephants" representing candidate feature subsets.
5. Use a fitness function to evaluate the performance of each elephant's selected feature subset.
6. Update each elephant's position using the best-found position so far and the global best position.
7. Repeat the evaluation and update steps until a stopping criterion is met.
8. The best elephant will indicate the optimal subset of features.
9. Train the models: k-NN, Logistic Regression, Decision Tree, and Naive Bayes using the feature subset selected by the Elephant Herd algorithm.
10. Evaluate the model using metrics like Accuracy, Precision, Recall, F_1-score, and AUC (Area Under Curve) on the testing set.
11. Use the trained model to predict whether new/unseen patient data has Chronic Kidney Disease (CKD) or not.

3.2. *Elephant Herding Optimization as Feature Selection*

Feature selection is an essential phase of the modelling process. It removes redundant characteristics and focuses on the most relevant traits that contribute to prediction objectives. While retaining predictive Accuracy, feature selection lowers dimensionality, enhances the model's interpretability, and boosts computational efficiency. This study uses Elephant Herd Optimization (EHO) as a feature selection technique. EHO, a nature-inspired optimization algorithm, mimics elephant herd behaviour to find optimal solutions. In EHO, virtual elephants navigate a solution space by representing different dataset features. Elephants use collective knowledge and personal experience to adjust their location with each iteration. This adjustment probabilistically moves towards regions with higher fitness values based on feature subset quality. A machine learning model is trained and evaluated using specific features to assess fitness. A performance metric, such as Accuracy or F1-score, is then calculated. Virtual elephants exhibit herding behaviour by focusing on promising feature subsets with higher fitness values, similar to how real elephants gather around favourable resources. The algorithm excels at navigating complex search areas, making it ideal for feature selection tasks. As the iterations go on, the population converges towards subsets of features associated with the best possible performance for the given task. Finally, the program determines which feature subset is most promising based on the fitness evaluations of the virtual elephants. Elephant Herd optimization (EHO) identifies the top 11 important features influencing model predictions, improving both Accuracy and interpretability. By selecting the most important features, it reduces noise, prevents overfitting, and ensures the model generalizes well to unseen data. The selected subset represents the optimal combination of characteristics that, when considered collectively, yield the best outcomes for the specified prediction task. The feature that has been chosen could be age, sg, al, su, pcc, bg, hemo, wc, rc, htm and pe. A brief comparison of various feature selection algorithms is shown in Table 1

Table 1. Comparison of Various Works Using Feature Selections

References	Feature Selection Algorithm	No. of Selected Features
[7]	Variance Threshold	14
[9]	Correlation	16
[23]	Chi-square	18
[24]	Recursive Feature Elimination	12
[24]	Boruta based on Recursive Feature Elimination	20
Proposed	Elephant Herd Optimization Algorithm	11

3.3. *K – Nearest Neighbour*

A flexible machine learning technique called K-Nearest Neighbours (KNN) is applied to regression and classification problems. Because it is non-parametric, that is, it doesn't make any explicit assumptions about the distribution of the underlying data and it is straightforward. Rather, KNN works on the similarity principle, which is to find the k closest neighbours in the feature space of a given data point given a predefined distance metric (usually Euclidean distance). Next, the majority class or average value of the new data point's closest neighbours is used to establish its class or value. This approach, which makes KNN particularly easy to understand, is based on the idea that comparable instances usually belong to the same class or have similar target values. Despite its simplicity, KNN can be very helpful, especially when dealing with small to medium-sized datasets or when there are nonlinear

decision limits. Its performance, however, may vary depending on the chosen distance metric and the number of neighbors (k).

3.4. Logistic Regression

Logistic regression is a linear classification method used in a variety of disciplines. Logistic regression is simple and easy to interpret. Logistic regression attempts to estimate the likelihood that an instance belongs to a particular class in binary classification tasks. The algorithm applies the logistic function to a linear combination of characteristic values, resulting in a probability score between 0 and 1. Logistic regression is effective for identifying complex correlations between data and target variables. However, its effectiveness in nonlinear datasets may be limited due to linearity assumptions and the need for independent features. Logistic regression is a popular classification method due to its ability to balance predictability and simplicity. It is the preferred choice when model transparency is crucial due to its interpretability and robustness.

3.5. Naïve Bayes

Under the presumption of feature independence, the Naive Bayes (NB) probabilistic classifier predicts class probabilities using the Bayes theorem. This algorithm is well-known for its efficiency, which makes it particularly effective at tasks such as spam filtering and text classification. Naive Bayes frequently outperforms more complex models due to its quick and simple handling of high dimensional input. Large datasets with a high number of features can benefit from NB because it reduces computational load by assuming feature independence. Its reliance on probability distributions simplifies interpretation and model assessment. The assumption of feature independence may not always be accurate, leading to poor performance for datasets with significant feature dependencies.

3.6. Decision Tree

Decision Tree (DT) is a flexible supervised learning approach that organizes data into hierarchical trees. Internal nodes represent characteristics, and leaf nodes represent class labels or values. To achieve optimal prognostic Accuracy, decision trees use recursive feature partitioning based on attribute values. The splitting process effectively creates boundaries around decisions by minimizing impurity measurements such as entropy or Gini impurity. Because of its interpretability and capacity to handle both numerical and categorical data, DT is a widely used option across various fields. But when working with narrow decision boundaries in large datasets, DTs are especially prone to overfitting. Despite its limitations, this method is still an effective tool for classification and regression tasks.

4. Results and Discussion

This study employs a chronic kidney disease dataset obtained from Kaggle. The dataset includes 400 observations, some of which may have incomplete or missing values. There are records for 250 CKD patients and 150 non-CKD people among these observations. This distribution shows that 37.5% of cases are unrelated to CKD and 62.5% are related to CKD cases. The people in the dataset are between the ages of 2 and 90. There are 24 features in the CKD dataset: 11 numerical features and

13 nominal features. Furthermore, the 25th feature is a classification indicator for CKD. Table 2 and Table 3 displays the performance metrics of various classification algorithms with and without EHO algorithm. Table 4 and Figure 7 shows the Accuracy of the proposed system in comparison with the existing works. This work was implemented in Google Colab. This work uses the following software versions: Python 3.11.11, NumPy 1.26.4, Pandas 2.2.2, Scikit-learn 1.6.1, and PyTorch 2.5.1+cu124, and the random seed is set to 42 to ensure that training and testing subsets are consistently selected.

We used 80/20 train-test split and 10-fold stratified cross-validation to evaluate the models' performance and generalisability. The train-test split strategy randomly picked 80% of the dataset for training and 20% for testing to retain class distribution. We also performed 10-fold stratified cross-validation, which divided the dataset into 10 equal-sized folds while preserving the class distribution. One-fold was tested and nine were trained per cycle. Average performance measures (accuracy, precision, recall, F1-score, and AUC) were provided after 10 repetitions. Dual validation ensures reliable and unbiased model prediction evaluation.

Table 2. Performance comparison of algorithms without Feature Selection

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
KNN	72.50	56.82	89.29	69.44
Naïve Bayesian	92.75	96.55	99.98	99.25
Logistic Regression	93.50	96.43	96.43	96.43
Decision Tree	94.75	96.55	100.00	98.25

Table 3. Performance comparison of algorithms with EHO Feature Selection

Elephant Herd Optimization with 11 Features				
Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
KNN	93.75	87.10	96.43	91.53
Naïve Bayesian	98.00	96.55	99.98	98.75
Logistic Regression	97.50	96.43	96.43	96.43
Decision Tree	98.75	100.00	96.43	98.18

Table 4. Comparison of Related Works

References	**Method**	**Accuracy (%)**
[9]	Gaussian Naïve Bayes	93.33
[10]	Heterogeneous Modified Artificial Neural Network	89.90
[22]	Random Tree	94.00
[24]	Naïve Bayes	97.33
[26]	Adaptive Backpropagation Neural Network	92.54
[32]	J48	85.50
Proposed	**Decision Tree**	**98.75**

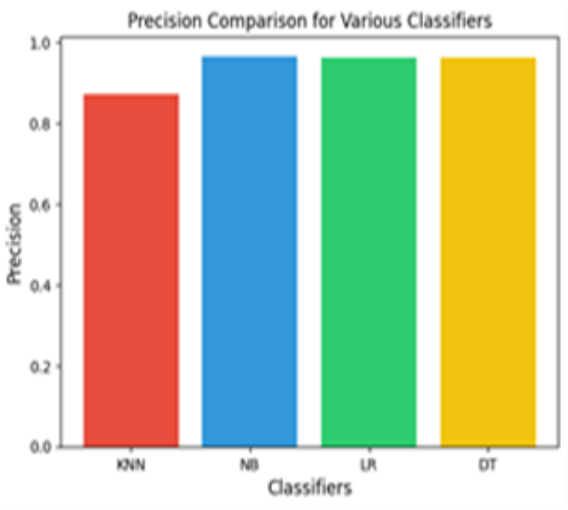

Figure 3. Accuracy Comparison

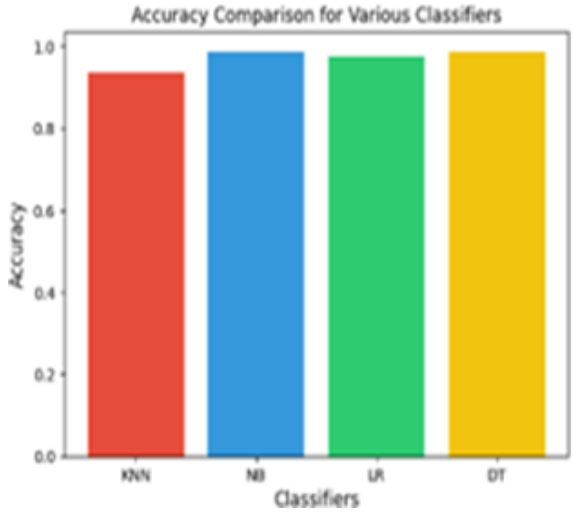

Figure 4. Precision Comparison

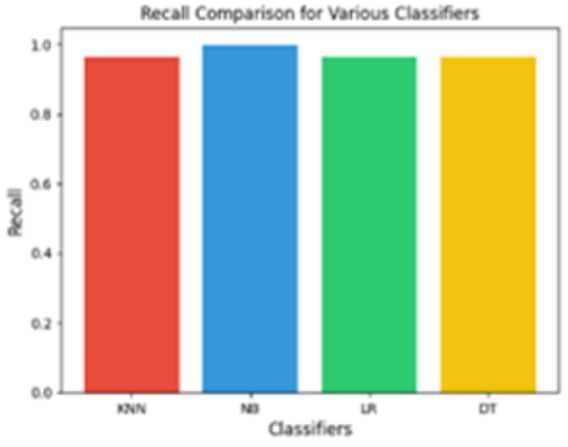

Figure 5. Recall Comparison

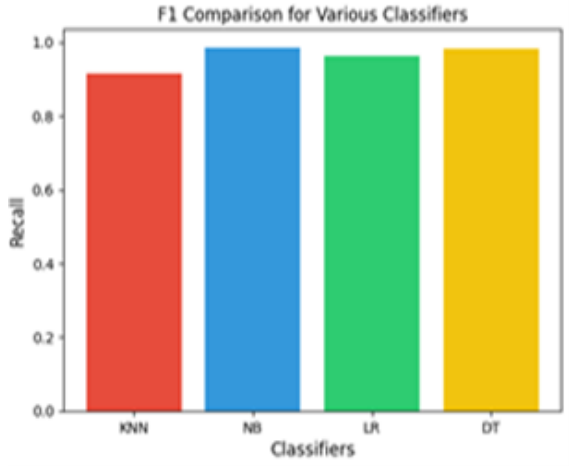

Figure 6. F1-Score Comparison

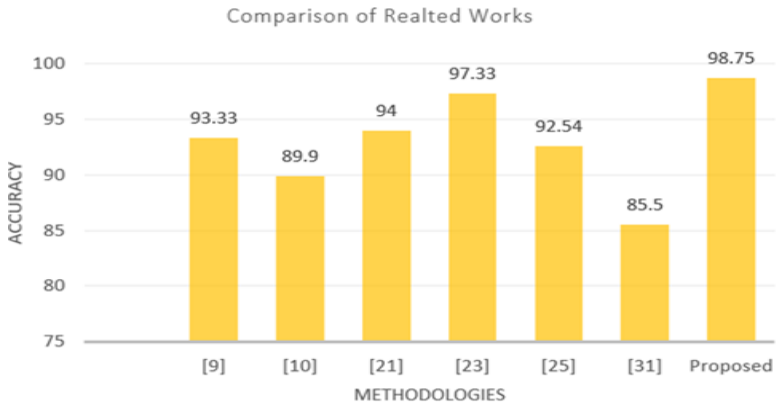

Figure 7. Performance Comparison of Related Works

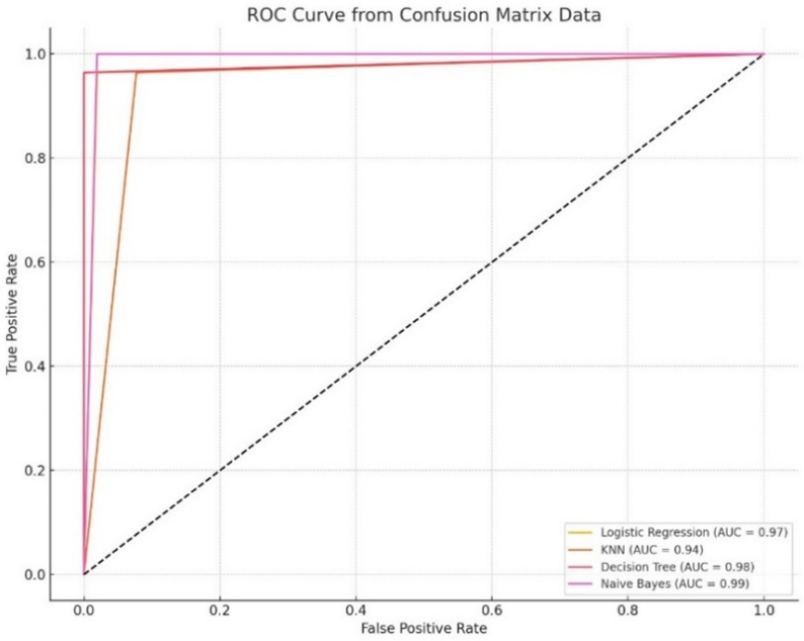

Figure 8. ROC Curve

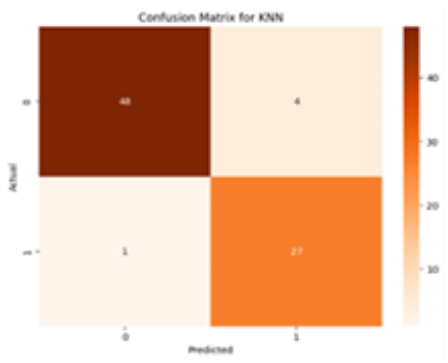

Figure 9. Confusion Matrix for KNN

Figure 10. Confusion Matrix for Decision Tree

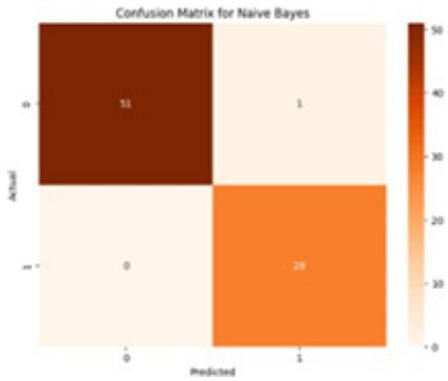

Figure 11. Confusion Matrix for Naive Bayesian

Figure 12. Confusion Matrix for Logistic regression

Figure 3 demonstrates that among the different algorithms, Decision Tree has the optimum Accuracy. Moreover, Figure 4 and Figure 5 further confirm the NB also achieved the highest precision and recall, respectively. From Figure 6, it is clear that NB has the best F1 score. The ROC Curve is shown in Figure 8. Figures 9, 10, 11, 12 show the obtained confusion matrix with feature selection.

5. Conclusion

This study aims to diagnose chronic kidney disease (CKD) early by analysing four different machine learning models. The CKD dataset was preprocessed, and dominant features were identified through elephant herd optimization. These dominating features are then used to train and validate the suggested models. Algorithm performance was evaluated primarily based on Accuracy, precision, recall, and F1 scores. Using the elephant-shredding optimization technique improved decision tree performance. The model Decision Tree achieved the highest Accuracy of 98.75%, outperforming the other models. The small number of available data samples may limit the model's generalization performance. The proposed models cannot accurately diagnose the severity of CKD due to the limited number of data categories (ckd and notckd). In the future, incorporating deep learning models, such as neural networks, can improve predictive performance and facilitate the identification of more intricate patterns in medical data, thereby expanding this approach. Additionally, the integration of other explainability techniques, such as SHAP or Counterfactual Explanations, in conjunction with LIME would offer a more thorough comprehension of model decisions. Our model can be tested with other datasets from various sources to ensure its robustness and applicability. The system can be made more adaptable to a wider range of healthcare applications by leveraging these advancements, which will enable it to manage a more diverse array of datasets. Ultimately, medical research and patient care will benefit from the continuous development of predictive analytics, which such innovations will facilitate.

Data Availability: The data can be accessed on this
https://www.kaggle.com/datasets/abhia1999/chronic-kidney-disease.

Funding Statement: The author(s) received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: R. Bala Krishnan; data collection: N. Senthil Selvan; analysis and interpretation of results: N. Senthil Selvan, B. Elangovan; draft manuscript preparation: G. Manikandan. All authors reviewed the results and approved the final version of the manuscript.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Webster, A. C., Nagler, E. V., Morton, R. L., Masson, P. (2017). Chronic kidney disease. The Lancet,389(10075), 1238–1252.
2. Zhong, J., Yang, H. C., Fogo, A. B. (2017).A perspective on chronic kidney disease progression. American Journal of Physiology-Renal Physiology, 312(3), F375–F384.

3. Gambaro, G., Croppi, E., Bushinsky, D., Jaeger, P., Cupisti, A., Ticinesi, A., Ferraro, P. M. (2017). The risk of chronic kidney disease associated with urolithiasis and its urological treatments: A review. The Journal of Urology, 198(2), 268–273.

4. Lin, M.-Y., Liu, M.-F., Hsu, L.-F., Tsai, P.-S. (2017). Effects of self-management on chronic kidney disease: A meta-analysis. International Journal of Nursing Studies, 74, 128–137.

5. Bouya, S., Balouchi, A., Rafiemanesh, H., Hesaraki, M. (2018). Prevalence of chronic kidney disease in iranian general population: A meta-analysis and systematic review. Therapeutic Apheresisand Dialysis, 22(6), 594–599.

6. Nicoll, R., Robertson, L., Gemmell, E., Sharma, P., Black, C., Marks, A. (2018). Models of care for chronic kidney disease: A systematic review. Nephrology, 23(5), 389–396.

7. Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., Chen, B. (2019). A machine learning methodology for diagnosing chronic kidney disease. IEEE Access, 8, 20991–21002.

8. De Almeida, K. L., Lessa, L., Peixoto, A. B., Gomes, R. L., Celestino, J. (2020). Kidney failure detection using machine learning techniques. 8th International Workshop on Advances in ICT Infrastructures and Services (ADVANCE 2020), 1–8.

9. Ekanayake, I. U., Herath, D. (2020). Chronic kidney disease prediction using machine learning methods. 2020 Moratuwa Engineering Research Conference (MERCon), 260–265.

10. Ma, F., Sun, T., Liu, L., Jing, H. (2020). Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. Future Generation Computer Systems, 111, 17–26.

11. Nimmagadda, S. M., Agasthi, S. S., Shai, A., Khandavalli, D. K. R., Vatti, J. R. (2023). Kidney failure detection and predictive analytics for ckd using machine learning procedures. Archives of Computational Methods in Engineering, 30(4), 2341–2354.

12. El-Kenawy, E. S. M., Ibrahim, A., Abdelhamid, A. A., Khodadadi, N., Abualigah, L., & Eid, M. M. (2024). Predicting sleep disorders: leveraging sleep health and lifestyle data with dipper throated optimization algorithm for feature selection and logistic regression for classification. Computational Journal of Mathematical and Statistical Sciences, 3(2), 341-358.

13. Mahajan, P., Uddin, S., Hajati, F., Moni, M. A., Gide, E. (2024). A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets. Health and Technology, 14(3), 597–613.

14. Takkavatakarn, K., Oh, W., Cheng, E., Nadkarni, G. N., Chan, L. (2023). Machine learning models to predict end-stage kidney disease in chronic kidney disease stage 4. BMC Nephrology, 24(1), 376.

15. Li, Y., Lou, Y., Liu, M., Chen, S., Tan, P., Li, X., Shao, X. (2024). Machine learning based biomarker discovery for chronic kidney disease–mineral and bone disorder (ckd-mbd). BMC Medical In- formatics and Decision Making, 24(1), 36.

16. Zhu, E., Shu, X., Xu, Z., Peng, Y., Xiang, Y., Liu, Y., Zheng, Z. (2023). Screening of immune related secretory proteins linking chronic kidney disease with calcific aortic valve disease based on comprehensive bioinformatics analysis and machine learning. Journal of Translational Medicine, 21(1), 359.

17. Velev, J., LeBien, J., Roche-Lima, A. (2023). Unsupervised machine learning method for indirect estimation of reference intervals for chronic kidney disease in the puerto rican population. Scientific Reports, 13(1), 17198.

18. Inoue, K., Hara, Y., Nagawa, K., Koyama, M., Shimizu, H., Matsuura, K., Kozawa, E. (2023). The utility of automatic segmentation of kidney mri in chronic kidney disease using a 3d convolutional neural network. Scientific Reports, 13(1), 17361.

19. Van Nynatten, L. R., Miller, M. R., Patel, M. A., Daley, M., Filler, G., Badrnya, S., Fraser, D. D.(2023). A novel multiplex biomarker panel for profiling human acute and chronic kidney disease. Scientific Reports, 13(1), 21210.

20. Ghosh, S. K., Khandoker, A. H. (2024). Investigation on explainable machine learning models to predict chronic kidney diseases. Scientific Reports, 14(1), 3687.

21. Tu, J. B., Liao, W. J., Liu, W. C., Gao, X. H. (2024). Using machine learning techniques to predict the risk of osteoporosis based on nationwide chronic disease data. Scientific Reports, 14(1), 5245.

22. Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., Bolshev,V. (2021). Prediction of chronic kidney disease-a machine learning perspective. IEEE Access,9, 17312–17334.

23. Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., Saha, S., Hossen, R., Akter, M. F. (2024). Mlckdp: Machine learning-based chronic kidney disease prediction with smart web application. Journal of Pathology Informatics, 15, 100371.

24. Rahman, M. M., Al-Amin, M., Hossain, J. (2024). Machine learning models for chronic kidney disease diagnosis and prediction. Biomedical Signal Processing and Control, 87, 105368.

25. Rehman, A., Xing, H., Hussain, M., Gulzar, N., Khan, M. A., Hussain, A., Mahmood, S. (2024). Hcdp-delm: Heterogeneous chronic disease prediction with temporal perspective enabled deep extreme learning machine. Knowledge-Based Systems, 284, 111316.

26. Vineetha, K. R., Maharajan, M. S., Bhagyashree, K., Sivakumar, N. (2024). Classification of adaptive back propagation neural network along with fuzzy logic in chronic kidney disease. e-Prime Advances in Electrical Engineering, Electronics and Energy, 7, 100463.

27. Legouis, D., Rinaldi, A., Malpetti, D., Arnoux, G., Verissimo, T., Faivre, A., Cipp'a, P. E. (2024). A transfer learning framework to elucidate the clinical relevance of altered proximal tubule cell states in kidney disease. iScience, 27(3).

28. Tang, J., Huang, J., He, X., Zou, S., Gong, L., Yuan, Q., Peng, Z. (2024). The prediction of in-hospital mortality in elderly patients with sepsis associated acute kidney injury utilizing machine learning models. Heliyon, 10(4), 109271.

29. Liu, Y., Hong, X., Liu, L., Li, X., Huang, S., Luo, Q., Li, C. (2024). Shen qi wan ameliorates nephritis in chronic kidney disease via aqp1 and defb1 regulation. Biomedicine & Pharmacotherapy, 170, 116027.

30. Reznichenko, A., Nair, V., Eddy, S., Fermin, D., Tomilo, M., Slidel, T., Kretzler, M. (2024). Unbiased kidney-centric molecular categorization of chronic kidney disease as a step towards precision medicine. Kidney International, 105(6), 1263–1278.

31. Sasikaladevi, N., Revathi, A. (2024). Digital twin of renal system with ct radiography for the early diagnosis of chronic kidney diseases. Biomedical Signal Processing and Control, 88, 105632.

32. Ilyas, H., Ali, S., Ponum, M., Hasan, O., Mahmood, M. T., Iftikhar, M., Malik, M. H. (2021). Chronic kidney disease diagnosis using decision tree algorithms. BMC Nephrology, 22(1), 273.