Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Comparative analysis of Ensemble Machine Learning Techniques for Early Predicting the Risk of Chronic Kidney Disease

Chithra K
Associate Professor, Department of Computer Science, *Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, Vaishnava College Road, Shanthi Nagar, Chromepet, Chennai-600044, University of Madras, Tamil Nadu, India.*
E-mail: chithra.k@sdnbvc.edu.in
ORCID iD https://orcid.org/0000-0002-9421-1401

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction: Chronic Kidney Disease (CKD) is a condition marked by the gradual deterioration of kidney function. Timely detection and effective treatment can enhance the chances of a positive outcome.

Objectives: The proposed research focus on predicting CKD status by incorporating a naval framework for effective identification using ensemble machine learning for early prediction of Chronic Kidney Disease prediction.

Methods: The proposed research focus on predicting CKD status by incorporating data pre-processing, attribute selection using standard Lasso, Lasso with Cross-Validation, Multitask Lasso feature selection and combined features selected by above three feature selection methods with six different Boosting ensemble machine learning classifiers.This research has explored the potential of various Ensemble Machine Learning techniques such as Gradient Boosting Classifier, Histogram Gradient Boosting Classifier, adaBoost Classifier, XGBoosting Classifier, CatBoost Classifier and Light GBM Classifier for enabling early diagnosis of CKD using the dataset taken from kaggle.

Results: The performance is evaluated using confusion matrix. The efficiency of the methodologies is measured in terms of metrics. The overall result shows that the Gradient Boosting Classifier gives highest accuracy of 99% when compared to other five classifier used in this research work.

Conclusions: This study examines the performance of Boosting classifers in predicting chronic kidney disease (CKD) outcomes. The results show that Gradient Boosting achieved the highest accuracy 99% across all four feature selection categories. While Histogram Gradient Boosting and CatBoost showed better performance with Lasso CV, Multitask Lasso, and combined feature selections. Additionally, the AdaBoost classifier performed better with Multitask Lasso and combined features. Both XGBoost and Light GBM classifiers performed better when using combined feature selection. The combined feature set yielded 99% accuracy across all classifiers. Overall, the findings demonstrate that the Gradient Boosting classifier achieved the highest accuracy and sensitivity in identifying CKD patients, highlighting its potential for early detection in clinical settings compared to other classifiers.

Keywords: Chronic Kidney Disease, Ensemble Machine Learning, Lasso, Gradient Boosting, Histogram Gradient Boosting, adaBoost, XGBoosting, CatBoost, Light GBM.

INTRODUCTION

Chronic Kidney disease(CKD) is a loss of kidney function or the presence of kidney damage that leads to renal replacement therapy like dialysis or transplantation [1]. The mortality increase to 95% between 2000 and 202. The death due to kidney disease becomes ninth common cause globally [2]. More than two million people worldwide rely on dialysis or kidney transplantation that is approximately 14% of the world's population and the report says twelve people die every day while waiting for kidney transplantation [3].

LITERATURE REVIEW

In the review study, various studies have been conducted over the last five years to diagnose chronic renal disease.

In the article [4], the research enhances the multilayer perception(MLP) by integrating LIME for the better prediction of the kidney disease. In the article [5], cross validation in recursive feature elimination was used for feature selection. The three classifiers were applied on the dataset that proves Random forest gives better result. In the paper [6], nine Machine learning approaches are built for dataset. This study has compared these techniques and found that the LightGBM model gives better result. The article [7] proposed that the classifier Random Forest and AdaBoost works better with respect to accuracy, precision, Sensitivity than Gradient Boosting and Bagging.

EXPERIMENTAL DATASET

This article uses the Chronic Kidney Disease(CKD) dataset from Kaggle that contains information about patients, focusing on whether they have chronic kidney disease (CKD) or not. It contained 24 features and one target variable. The 'classification' variable has the value 'ckd' / 'notckd. The dataset consists of 11 numerical and 14 nominal features with 400 rows [8].

DATA PREPROCESSING

Data pre-processing involves converting unprocessed data into an easily interpreted format [9]. In data-mining process data preprocessing is used to increase the data quality [10]. Data cleaning entails filling in the blanks and eliminating erroneous, partial, and inaccurate data from the datasets. Various methods can be used to clean the data, such as substituting the attribute mean for missing values. When numerical form of data is used in classification the performs will be improved. Hence, categorical data including the classification in the dataset is transformed into integers using Label Encoding [11]. Instead than depending on conventional over-sampling approaches, the SMOTE is an over-sampling methodology that creates synthetic instances to enhance the minority class [12].

FEATURE SELECTION

Feature selection methods facilitates the elimination of redundant or unnecessary features. Feature selection for a dataset d entails selecting a subset of features from the original feature set and optimizing the target function T, that is, maximizing the value of T [13]. This research concentrates on three feature selection techniques such as Lasso, Lasso CV and Multitask Lasso. This research concentrates on three Lasso feature selection techniques. They are as follows:

1. Standard Lasso (L1 Regularization):

A penalty proportion is added to the absolute values of the coefficient to zero. These features with zero coefficients are removed from the model that leads to automatic feature selection [14].

2. Lasso with Cross-Validation (Lasso CV):

This technique is used to identify the most relevant features by applying LASSO along with cross-validation to prevent overfitting by applying regression coefficients with a penalty that leads to zero value for less important feature that effectively removing irrelevant features [15] [16].

3. Multitask Lasso:

In multitask LASSO, regularization term is applied across all the tasks that helps to lean the relationship among them and leverage to enhance prediction accuracy [15].

The features selected by the above selection techniques are taken as input for the classifiers used in this article for the prediction of disease.

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

EMSEMBLE MACHINE LEARNING APPROACH

It is a supervised technique that combines the models to generate a more resilient a potent model that solves the same problem. The predictions of these models are then joint to improve overall performance. The Ensemble techniques that are used in this research are listed below:

1. **Bagging:** In bagging, many models uses different subsets to train the dataset. In order to construct each subset, Bootstrap sampling is used with data points chosen at random and replaced. The overall result is then generated by combining the predictions from the models using majority voting [16] [17].

2. **Boosting:** The stability and accuracy of machine learning classification are enhanced by reducing the bias in learning by transforming weak learners into strong learners [18] [19].

3. **Stacking:** In stacking, two layers of estimators are used. In first layer all the baseline models are executed are the results on the test dataset. In the second layer, the meta-classifier creates new predictions by using the baseline models' predictions as input [19].

The classifiers used in this research are as follows:

Gradient Boosting Classifier(GBoost): The mulitple base learner predictions are combined, usually decision trees, to create a model. In this method, every new model is trained to fix the mistakes (or residuals) caused by the models that came before it. Then the loss function is minimized by gradient descent the gradient descent and a new model is fitted to the residuals (errors) of the old model and the process is repeated with each new model. The final prediction is the sum of all weights of all models' predictions [20][21].

eXtreme gradient boosting(XGBoost): This method optimezes for speed and performance in structured/tabular datasets, especially for classification and regression tasks, This XGBoost creates an group of decision trees, each of which fixes the mistakes of the ones before it. These mistakes are calculated following the training of the first tree to produce new tree. Finally, model is created by combining the trees with weighted sums to reduces overfitting and guarantees improved generalization by using regularization and a particular objective function [22].

Adaptive Boosting (AdaBoost): In this classifie, several weak clssifier such as decision trees are combined. Each new clssifie are added iterratively that concetrates on mistakes generated by the previous classifiers. The weak classifier uses a weighted dataset for traiing. Higher weights ae assigned to the misclassified model instance to give priority in subsequent training cylce. Based on thee Error rate the weight of the weak classifier is modified. The weighted outputs of each weak classifier are combined to create the final prediction [23].

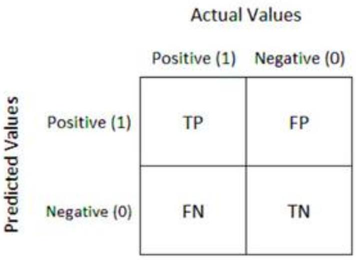
Categorical Boosting (CatBoost): This classifier works efffectively on categorical vaiables to perform beteer on dataset than other gradient boosting classifier. The predictions made by the preceding decision tree resuduals or gradients ae computed and this is reduced by the new predictions. The algorithm integrates categorical feactures effectively during this processes. The total weights of each tree gives the final prediction[24].

Light gradient boosting machine (Light GBM): This classifier is used for handling big and high dimensional dataset in effective manner. This uses decision tree model and used histogram data to find best splits after continuous d=features discretizing in to bins. The trees are built by cultivating the most promising leaf. The boosting iteration is used for model training in which each new tree fixes the mistakes which was made by previous tree. The prediction from all the tree are combined and the weights by each tree’s accuracy yields the final prediction [25].

Histogram Gradient Boosting (HGBoosting) Classifier: Histogram-based algorithmsare used to increase computational efficiency and decision tree us is fitted to the dataseet at the beginning. The discretizing continuous variables are used to create histograms. The residuals are calculated for every sample. Ther iterative approach is used to minimize the loss function inorder to enhce the overall model [26].

PERFORMANCE METRICS

The performance of the proposed methodology is tested by creating the confusion matrix that used for calculating the metrics [27] [28][29]. The confusion matrix used in evaluation is shown in the figure 1.



	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Fig. 1. Confusion Matrix.

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+TP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

PROPOSED METHODOLOGY

This study uses the Chronic Kidney Disease dataset from Kaggle to predict the presence of the disease. The dataset undergoes preprocessing to handle missing values by imputing them with the mean, mode, or median, and categorical variables handled by label encoding, which assigns a integer value to category. The dataset contains 250 patients (62.5%) with disease and 150 patients (37.5%) without it, resulting in an imbalanced distribution. To address this, a Random Over Sampler is applied such that the minority class are replicates balance the dataset.

Following preprocessing, feature selection is carried out using three different methods: Lasso, Lasso CV, and multitasking Lasso. The chosen features are then passed into various classification algorithms. Six classifiers are evaluated to predict kidney disease: GBoost, HGBoosting, AdaBoost, LightGBM, XGBoost, and CatBoost. The performance of these models is measured using metrics. The overall methodology is illustrated in figure 2.

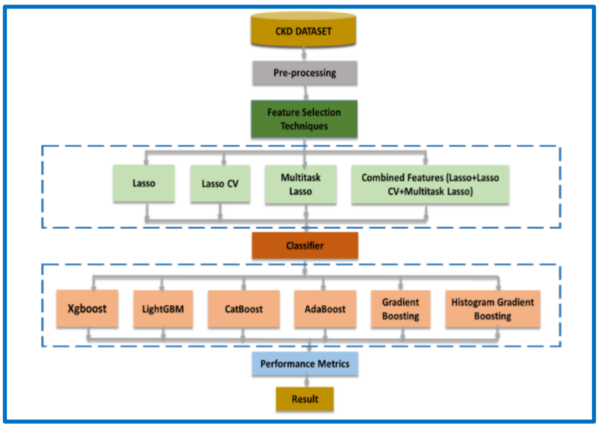


Fig. 2. Proposed Methodology.

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

RESULTS

This study explores the working of machine learning methods using input features selected by using above mentioned feature selection techniques, as well as a combined set of features derived from all three methods.

Using the Lasso (L1 regression), 9 features were selected, The Lasso CV technique selected 14 features, Multitask Lasso method identified 12 features and the union of the features selected by all three methods resulted in a combined 15 features. These four different feature sets were used as inputs to six different classifiers, and their performance in disease detection was evaluated using the confusion matrix.

The results show that the Gradient Boosting classifier performs best with Lasso feature selection. For Lasso CV feature selection, GBoost, HGBoosting, and CatBoost classifiers yield the best performance. With Multitask Lasso feature selection, the Gradient Boosting, AdaBoost, and CatBoost classifiers achieve the best results. When using the combined set of features, all six classifiers perform equally well, each achieving 99% accuracy. The results, presented in Tables 1 through 4, demonstrate that the combined feature selection consistently delivers superior performance across all classifiers. Figure 3 illustrates the accuracy obtained by various classifiers using different feature selection techniques.

Table 1. Performance of Algorithms using Standard Lasso

Classifiers	Accuracy in %	Precision	Recall	F1 Score
GBoosting	**98**	**0.98**	**0.98**	**0.98**
HGBoosting	97	0.97	0.97	0.97
AdaBoost	96	0.96	0.96	0.96
XGBoost	97	0.97	0.97	0.97
CatBoost	97	0.97	0.97	0.97
Light GBM	96	0.96	0.96	0.96

Table 2. Performance of Algorithms using Lasso CV

Classifiers	Accuracy in %	Precision	Recall	F1 Score
GBoosting	**99**	**0.99**	**0.99**	**0.99**
HGBoosting	**99**	**0.99**	**0.99**	**0.99**
AdaBoost	97	0.97	0.97	0.97
XGBoost	97	0.97	0.97	0.97
CatBoost	**99**	**0.99**	**0.99**	**0.99**
Light GBM	98	0.98	0.98	0.98

Table 3. Performance of Algorithms using Multitask Lasso

Classifiers	Accuracy in %	Precision	Recall	F1 Score
GBoosting	**97**	**0.97**	**0.97**	**0.97**
HGBoosting	96	0.96	0.96	0.96
AdaBoosting	**97**	**0.97**	**0.97**	**0.97**

XGBoosting	96	0.96	0.97	0.97
CatBoost	**97**	**0.97**	**0.97**	**0.97**
Light GBM	96	0.96	0.96	0.96

Table 4. Performance of Algorithms using Combined Features

Classifiers	Accuracy in %	Precision	Recall	F1 Score
GBoosting	**99**	**0.99**	**0.98**	**0.99**
HGBoosting	**99**	**0.99**	**0.98**	**0.99**
AdaBoosting	**99**	**0.99**	**0.98**	**0.99**
XGBoosting	**99**	**0.99**	**0.98**	**0.99**
CatBoosting	**99**	**0.99**	**0.98**	**0.99**
Light GBM	**99**	**0.99**	**0.98**	**0.99**

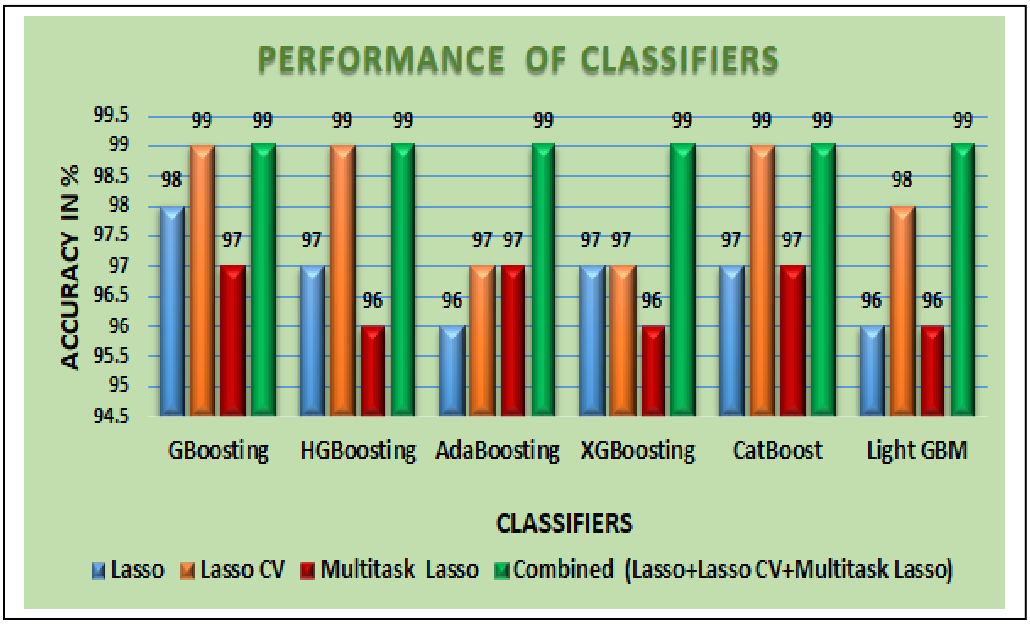


Fig. 3. Comparison of accuracy of Classifiers

DISCUSSION

This research investigates the impact of three different Lasso feature selections. The multiple feature selection techniques lead to the identification of different subsets of relevant features. The outputs of all the three methods are combined to produce a comprehensive feature set of 15. The evaluation of these feature sets is done using six classification techniques. The gradient boosting algorithm performs better in terms of accuracy when using Lasso feature selection. The GBoost, HGBoosting, and CatBoost achieved give optimal results for LassoCV feature selection. Multitask Lasso-selected features yielded the best performance with GBoost, AdaBoost, and CatBoost classifiers, reflecting the benefit of multitask learning in capturing subtle patterns when using boosting-based methods. When using a combined feature set, all the six classifiers achieve a high accuracy of 99%. In summary, a hybrid feature set combining outputs from multiple selection techniques can offer a universally strong foundation for classification tasks.

REFERENCES

[1] https://www.ncbi.nlm.nih.gov/books/NBK535404/

[2] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-ofdeath#:~:text=In%20contrast%2C%20kidney%20diseases%20have,95%25%20between%202000%20and%202021.

[3] https://www.kidney.org/news/newsroom/factsheets/KidneyDiseaseBasics.

[4] M. S. Arif, A. U. Rehman and D. Asif , "Explainable Machine Learning Model for Chronic Kidney Disease Prediction," Algorithms, vol. 17, no. 10, 2024, doi: https://doi.org/10.3390/a17100443.

[5] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," J Big Data, vol. 9, no. 109, 2022, doi: https://doi.org/10.1186/s40537-022-00657-5.

[6] A. Farjana et al., "Predicting Chronic Kidney Disease Using Machine Learning Algorithms," IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 1267-1271, 2023 doi: 10.1109/CCWC57344.2023.10099221.

[7] Nikhila, "Chronic Kidney Disease Prediction using Machine Learning Ensemble Algorithm," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 476-480, 2021, doi: 10.1109/ICCCIS51004.2021.9397144.

[8] https://www.kaggle.com/datasets/mansoordaku/ckdisease/data

[9] K. Chitha, "Comparative Analysis of Classification Techniques using Feature Selection Methods for Stroke Prediction," Indian Journal of Natural Sciences, vol.13, no.76, pp.52917-52926, 2023.

[10] https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide

[11] S. A. Alasadi and W. Bhaya, "Review of data preprocessing techniques in data mining," Journal of Engineering and Applied Sciences, vol. 12, no. 16, pp. 4102–4107, 2017.

[12] V. Nitesh, Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol.16, pp.321–357, 2002.

[13] Mustafa BUYUKKECECI and Mehmet Cudi OKUR,"A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning", Journal of Science, vol. 36, no. 4, pp.1506-1520, 2023.

[14] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 58, no. 1, pp.267-288, 1996.

[15] A. Argyriou, T. Evgeniou and M. Pontil, "Multitask learning for classification: A convex formulation," European Conference on Machine Learning (ECML), pp.10-21, 2007.

[16] Mohammed, Nouralden Mohammed Jadalla. University of the Witwatersrand, Johannesburg (South Africa) ProQuest Dissertations & Theses, 2023. 31787387.

[17] https://www.geeksforgeeks.org/ml-bagging-classifier/

[18] Leo Breiman,"Bias, variance, and arcing classifiers," (PDF). Technical Report, 2015.

[19] Zhou Zhi-Hua,"Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC. pp. 23, 2012.

[20] https://www.geeksforgeeks.org/boosting-in-machine-learning-boosting-and-adaboost/

[21] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, 2001.

[22] T.Chen, and C. Guestrin, XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[23] Y. Freund, and R.E, Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. Proceedings of the Second European Conference on Computational Learning Theory, pp. 23-37, 1997.

[24] L. Prokhorenkova, G. Gusev, A. Vorobev, A, A.V. Dorogush, and M. Frolov, "CatBoost: Unbiased Boosting with Categorical Features," Proceedings of the 32nd International Conference on Neural Information Processing Systems, vol. 31, pp. 6638–6648, 2018.

[25] G. Ke, Q. Meng, T. Finley, T. Wang , W. Chen, and W. Ma, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, Proceedings of the 31st International Conference on Neural Information Processing Systems, vol. 30, 2017.

[26] Scikit-learn contributors, “HistGradientBoostingClassifier and HistGradientBoostingRegressor,” In scikit-learn: Machine Learning in Python, 2020.

[27] D.M. Powers, “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation”, Journal of Machine Learning Technologies, vol.2, no.1, pp. 37-63, 2011.

[28] T.awcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.

[29] Giulia Solda and Rosanna Asselta, ”Applying artificial intelligence to uncover the genetic landscape of coagulation factors,” jth Journal of Thrombosis and haemostasis, vol.23, no.2, 2025.