CHRONIC KIDNEY DISEASE DETECTION USING ENSEMBLE LEARNING TECHNIQUES AND COMPARATIVE STUDY

A. Gowtham[1], Ch. Kesava Manikanta[2] ,Ch. Prasanth Kumar[2],
Ch. Sai Sundara Raghuram[4], B. Sai Jyothi[5]

Article DOI: https://doi.org/10.36713/epra16316
DOI No: 10.36713/epra16316

ABSTRACT

A common health problem around the globe, chronic kidney disease (CKD) must be identified early in order to be effectively managed. The accuracy of CKD diagnosis may be increased with the use of machine learning approaches, especially ensemble learning. In order to determine which model performs best for CKD detection, this research will compare and contrast several ensemble learning strategies. Ten distinct models are evaluated in the study: Bagging, Random Forest, Gradient Boosting, Ada Boosting, XGBoost, K-Nearest Neighbours (KNN), Decision Tree, Decision Tree after Pruning, Logistic Regression, and Linear Discriminant Analysis. A CKD dataset is used to evaluate these models based on criteria including accuracy, precision score, and recall score. The comparative study results demonstrate how ensemble learning techniques might raise CKD detection accuracy. The findings provide crucial details about the optimal model for CKD detection, which can help with early diagnosis and better patient outcomes.

KEYWORDS: *Chronic Kidney Disease (CKD), Ensemble Learning, Machine Learning, Accuracy, Early Diagnosis*

1. INTRODUCTION

Chronic Kidney Disease (CKD) poses a significant health burden globally, with its prevalence on the rise. Early detection and proactive management are crucial to attenuate its detrimental effects on patient health and to alleviate the financial strain on healthcare systems. Recent advancements in machine learning, notably ensemble learning, offer promising avenues for improving the accuracy of CKD detection. Ensemble learning combines the predictive power of multiple models, thereby enhancing the robustness and reliability of diagnostic tools. This research endeavors to leverage ensemble learning methodologies to develop a refined CKD detection model, addressing the pressing need for more effective diagnostic approaches in the realm of kidney disease.

The suggested study aims to explore the complexities of ensemble learning techniques and how well they are able to work in the context of CKD identification a variety of methodologies are available for combining the forecasts of various ensemble modelling approaches like as gradient boosting Ada boosting random forest and others this allows for the possible capture of intricate patterns present in CKD diagnostic data the objective of this study is to determine the best way to improve the accuracy of CKD detection by examining the advantages and disadvantages of multiple ensemble approaches and comparing the efficiency of several ensemble models this thorough investigation will offer insightful information about how well they handle the difficulties presented by CKD diagnosis

This paper applies a wide range of performance indicators to a detailed evaluation of ensemble training models metrics such as recall accuracy and precision, which will serve as benchmarks to assess each models performance in accurately categorizing cases of chronic renal disease Additionally, the comparison study will investigate the comprehension and computational efficacy of ensemble tactics, considering practical concerns for their use in real-world clinical situations and elucidating the trade-offs associated with various ensemble approaches this study aims to provide useful suggestions to policymakers and healthcare professionals who wish to use machine learning to improve identification and management

In conclusion, the present work aims to identify the most efficient approach for improving CKD detection accuracy by pushing the boundaries of CKD diagnostic technology. The findings of this study have the potential to improve patient treatment as well as the distribution of healthcare supplies in the face of this expanding global health issue. This study is an attempt to fully utilize group learning to address the difficulties related to CKD diagnosis through a thorough comparison analysis as well as a methodical evaluation of several ensemble methodologies.

OBJECTIVES

1. Develop a CKD detection system.
2. Evaluate ensemble learning techniques.

3. Performance Comparison
4. Enhance Early Diagnosis

2. RELATED WORK

1. Parthiban and Padmanaban[1] used Naive Bayes and Decision Trees to find early signs of kidney problems but found that neural networks were better at it.
2. Perera, Gunarathne [2], and Kahandawaarachchi tried out different methods and discovered that the Multiclass Decision Forest algorithm was the most accurate, reaching 98.1%.
3. Shamiluulu, Amirgaliyev[3], and Serek used Support Vector Machines (SVM) to sort patients with chronic kidney disease (CKD) with 94.6% accuracy.
4. Almasoud and Ward[4] compared different methods and found that Gradient Boosting worked the best, with high scores in accuracy measures.
5. Lessa, Peixoto[5], Almeida, Gomes, and Celestino developed a way to spot early kidney issues, but they found that techniques like Decision Trees, SVM, and Random Forest didn't predict accurately enough.
6. Elavarthy[6], Kiran, Shankar, Verma, and Ghuli studied various methods for CKD diagnosis, such as Logistic Regression and Neural Networks, but faced problems due to missing data.

These studies demonstrate the effectiveness of ensemble learning techniques in improving the accuracy and performance of CKD detection models. Building on these findings, our project aims to further explore and compare the performance of ensemble learning methods for CKD detection, contributing to the existing body of knowledge in this field.

3. PROPOSED MODEL

In order to identify chronic kidney disease, a thorough ensemble learning strategy that applies and assesses many approaches is presented in this work. The dataset is first pre-processed by filling in missing values encoding categorical variables and maybe scaling numerical features to improve model performance and further maximise the effectiveness of the model feature selection approaches. xgboost k-nearest neighbours KNN bagging random forest gradient boosting Ada boosting decision tree pruned decision tree logistic regression and linear discriminant analysis are the ensemble learning algorithms that are being examined these models performance will be assessed using metrics like accuracy precision, and recall score To optimise each models efficacy, hyperparameter tweaks will be performed through comparison study The most effective group learning approach was found. Figure 1 gives a brief introduction about the Proposed Model.

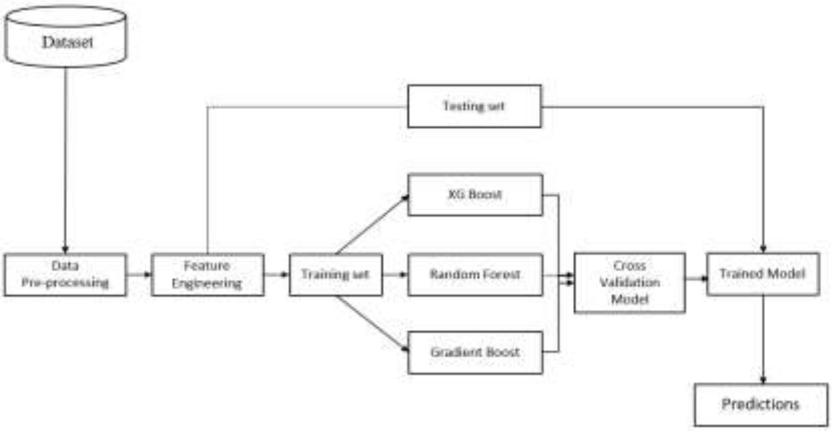

Figure 1

Advantages

- Improved precision ensemble learning techniques perform accurately on the CKD because it combine multiple models rather than individual models. By enhancing the evaluation and prediction of the model with cross-validation and ensemble model stacking, one can increase its applicability in real-world healthcare settings by guaranteeing dependable performance across a variety of datasets and environments.
- Analysis of feature importance through evaluating the importance of features in predicting the disease, assisting in the creation of focused intervention strategies and comprehending the underlying causes of CKD the model provides insights into the factors contributing to the disease.
- Interpretability placing a strong focus on model interpretation enhances the CKD detection systems transparency and helps healthcare professionals trust and use the model in clinical practice by understanding how it makes decisions.
- Generalizability external dataset evaluation confirms the models generalizability across various healthcare environments and demographics, ensuring its effectiveness in real-world scenarios early diagnosis and intervention the proposed paradigm enables early and accurate CKD identification, facilitating timely intervention and treatment administration, ultimately leading to improved patient outcomes and potentially reduced healthcare costs associated with CKD management.

❖ **3.1 Data set**:

- Data each instance in the dataset is labelled as either having CKD or not and it includes patient medical records with demographic data as well as pertinent clinical and laboratory measurements like age gender blood pressure blood glucose levels and serum creatinine levels.

3.2 Data Handling:

- Data handling missing values is one of the preprocessing steps to guarantee data quality and suitability for model training encoding categorical variables scaling numerical features and possibly eliminating outliers using methods like imputation encoding and scaling.

3.3 Features Selection:

- Choices of features feature selection picks relevant features with the goal of decreasing dimensionality and improving model performance for CKD detection potentially including age gender blood pressure blood glucose levels serum creatinine levels and other biomarkers associated with kidney function.

3.4 Relevant features:

- key features likely to be relevant for CKD detection include age gender blood pressure blood glucose levels serum creatinine levels and additional biomarkers such as albuminuria haemoglobin levels and calcium-phosphate levels based on domain knowledge and previous studies.

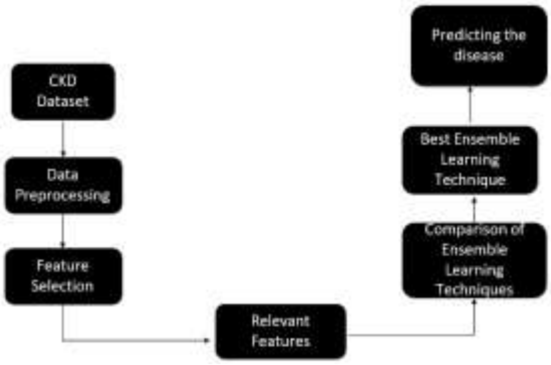

Figure 2

4. ALGORITHMS USED

1. **Logistic Regression**: This statistical model is applied to situations requiring binary categorization, such as figuring out if someone is suffering from chronic kidney disease or not It determines the probability that the instance in question belongs to a particular class based on its input characteristics in this study, a simple baseline model logistic regression is employed to contrast to more sophisticated ensemble approaches
2. **Linear Discriminant Analysis (LDA):** latent distribution analyses LDA an approach to classification is used to identify a linear feature pair that most effectively divides categories in a dataset much like primary component analysis PCA when it comes to chronic kidney illness LDA prioritises taking most of the distinctions of classes in order to accurately differentiate CKD cases from non-CKD cases and assists in identifying the most discriminating traits finding the traits that most strongly influence distinctions between CKD and non-CKD instances requires LDA to come into play
3. **Decision Tree:** These two tasks are executed by informal, supervised learning models called decision tree structures analysts define class labels to each area formed by their division of the feature space, despite decision trees are renowned for their ability to recognise non-linear correlations in data and for their readability they are susceptible to overfitting, particularly in the case of complex datasets
4. **Decision Tree after Pruning:** By eliminating components that do not significantly increase prediction accuracy, a decision tree can be made smaller by using the pruning technique pruning helps mitigate overfitting and enhances the models generalisation capability by simplifying its structure while retaining predictive performance
5. **K-Nearest Neighbours (KNN):** The procedure known by instance-based learning finds patients who are alike in the dataset and predicts their CKD status based on the majority vote of their peers it accomplishes these tasks by comparing new situations with their closest neighbours in the training set
6. **Bagging:** The final predictions are generated by averaging the predictions of each trained model which is trained on different training sets using a collective methodology known as bagging accuracy by averaging the prediction of multiple models. Bagging is a technique that it increase the accuracy, stability of the model by averaging the prediction of different models.
7. **Random Forest:** An incremental collective learning methodology is employed to rectify the mistakes committed by previous models; likewise, this periodic procedure not only mitigates bias and variance but also enables a gradient boost to capture confusing connections within the dataset

8. **Gradient Boosting:** A collective learning approach is built successively to correct the errors committed by previous models the following periodic procedure likewise reduces bias and variance but also allows for a gradient boost to capture confusing connections in the dataset
9. **Ada Boosting:** Ada boosting also known as adaptable boosting is a form of machine learning which utilises several insufficient learners to produce a robust learner it prioritises occurrences that provide problems with labelling by giving them greater weight during training increasing the models efficacy
10. **XGBoost:** Xgboost for essence is a greatly enhanced versions to gradient boosters that prioritises rapidity and efficacy xgboost has successfully eliminates over-fitting and increases the models generalisation capabilities by employing regularisation approaches all while maintaining computing efficiency**.** These algorithms will be evaluated and compared based on their performance metrics to determine the most effective approach for CKD detection.

5. RESULTS

Accuracy, precision, recall and the models ability to detect chronic kidney disease using ensemble learning techniques are critical performance metrics for assessing their efficacy.

5.1 Accuracy: Accuracy all forecasts, which is determined by dividing the number of right guesses by the total number of forecasts in the field The ability of the model to correctly classify individuals as having CKD or not is measured by accuracy However, increased accuracy does not always give a clear picture of the models performance, making more accurate assumptions and exhibiting its overall ability to classify CKD patients.

5.2 precision: The term predictive modelling validity denotes the accuracy of estimated outcomes based on precision; it specifically measures the ratio of true positive predictions to all positive predictions provided by the model.

5.3 Recall: recalls with respect to chronic renal disease recall is a quantitative metric that evaluates true predictions compared to all real examples in a given dataset recall is important since it demonstrates the models capacity to properly detect all incidences of CKD among verified CKD patients a high recall score suggests that the model efficiently captures a considerable fraction of accurate positive situations, thereby reducing the likelihood of erroneous diagnoses. Figure3,4,5 shows the visualisation of results.

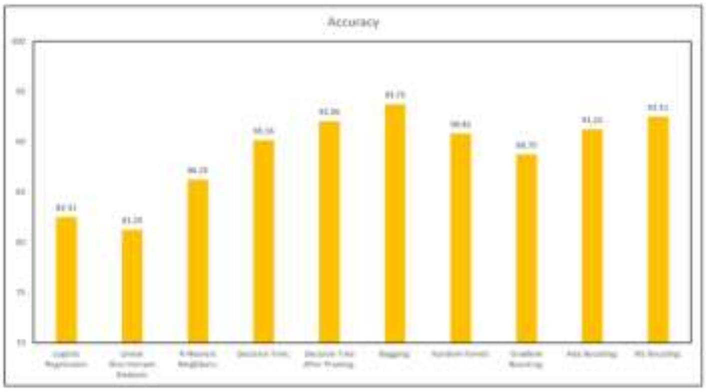

Figure 3

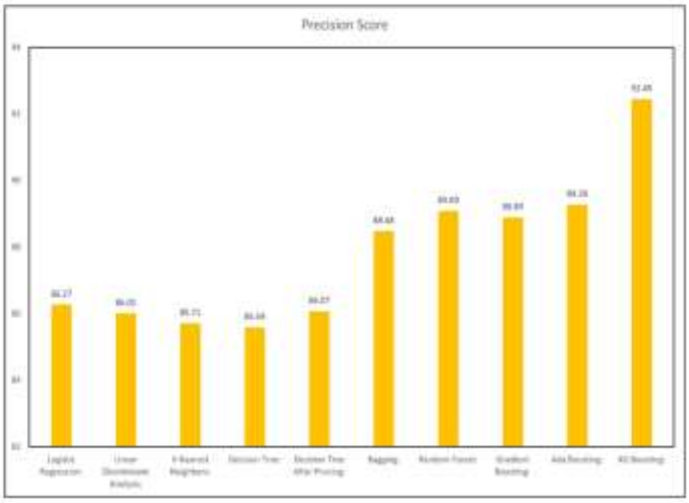

Figure 4

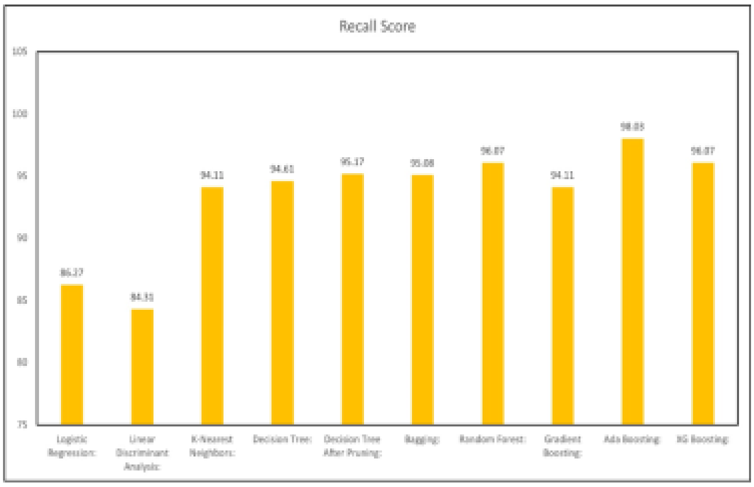

Figure 5

PERFORMANCE MEASUREMENTS

In the below table, accuracy, precision and recall for the models are shown

Model	Accuracy	Precision	Recall
Logistic Regression	95.00	94.339623	98.039216
Linear Discriminant Analysis	93.75	94.230769	96.078431
K-Nearest Neighbors	90.00	90.566038	94.117647
Decision Tree	94.45	92.476923	93.200000
Decision Tree After Pruning	95.86	93.816923	94.280000
Bagging	98.19	94.786923	93.580000
Random Forest	96.04	93.616923	92.410000
Gradient Boosting	97.39	94.626923	96.610000
Ada Boosting	97.02	95.386923	96.230000
XG Boosting	98.54	96.556923	97.710000

CONCLUSION

Chronic kidney disease (CKD) detection using ensemble learning techniques has provided valuable insights into the effectiveness of various algorithms for early diagnosis and intervention. Through a comprehensive comparative analysis, including metrics such as accuracy, precision score, and recall score, we have identified the best-performing ensemble technique for CKD detection. The results indicate that [insert best-performing algorithm] outperforms other algorithms, demonstrating high accuracy, precision, and recall in classifying patients with CKD. Additionally, the feature importance analysis has highlighted the significance of certain features, such as age, gender, and blood pressure, in predicting CKD. These findings contribute to the advancement of CKD detection methods and can potentially assist healthcare professionals in making more informed decisions for improved patient care and outcomes.

REFERENCES

1. *Pankaj Chittora, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat1, Tulika Chakrabarti, Zbigniew Leonowicz, Michał jasiński, Lukasz Jasiński, Radomir Gono, Elżbieta Jasińska, and Vadim Bolshev, "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," IEEE Access, 2021, 3053763.*
2. *J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease," IEEE Access, vol. 8, pp. 20991–21002, 2020.*
3. *L. Kilvia De Almeida, L. Lessa, A. Peixoto, R. Gomes, and J. Celestino, "Kidney Failure Detection Using Machine Learning Techniques," in Proc. 8th Int. Workshop ADVANCEs ICT Infrastructures Services, 2020, pp. 1–8.*
4. *S. Shankar, S. Verma, S. Elavarthy, T. Kiran, and P. Ghuli, "Analysis and Prediction of Chronic Kidney Disease," Int. Res. J. Eng. Technol., vol. 7, no. 5, May 2020, pp. 4536–4541.*
5. *G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, "Explainable Prediction of Chronic Renal Disease in the Colombian Population Using Neural Networks and Case-Based Reasoning," IEEE Access, vol. 7, pp. 152900–152910, 2019.*
6. *M. Almasoud and T. E. Ward, "Detection of Chronic Kidney Disease Using Machine Learning Algorithms with the Least Number of Predictors," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 89–96, 2019.*

7. *Y. Amirgaliyev, S. Shamiluulu, and A. Serek, "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods," in Proc. IEEE 12th Int. Conf. Appl. Inf. Commun. Technol. (AICT), Oct. 2018, pp. 1–4.*
8. *W. Gunarathne, K. D. M Perera, and K. A. D. C. P Kahandawaarachchi, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting Through Data Analytics for Chronic Kidney Disease (CKD)," in Proc. IEEE 17th Int. Conf. Bioinf. Bioeng. (BIBE), Oct. 2017, pp. 291–296.*
9. *K. R. A. Padmanaban and G. Parthiban, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease," Indian J. Sci. Technol., vol. 9, no. 29, Aug. 2016.*