JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Analysis of performance of feature optimization techniques for the diagnosis of Chronic Kidney Disease using Machine Learning

Mr. Chayan Bhattacharjee
Department of Information Technology
Chikitsak Samuha's Patkar Varde College
Mumbai, India

Abstract: This review examines the paper "Analysis of the Performance of Feature Optimisation Techniques for the Diagnosis of Machine Learning-Based Chronic Kidney Disease." The research looks into using machine learning to improve the accuracy of chronic kidney disease (CKD) diagnosis. The study uses the Cleveland Kidney Disease dataset to embark on a comprehensive journey that includes data pre-processing, feature selection, model application, and evaluation. The Minimal Redundancy-Maximal-Relevance (MRMR) algorithm is an important tool for feature reduction, which results in a refined dataset. For CKD diagnosis, machine learning models such as linear regression, support vector machine (SVM), decision tree, k-nearest neighbours (KNN), and Linear Discriminant Analysis (LDA) are used. To assess model effectiveness, rigorous performance evaluation metrics are used. LDA stands out as the best performer, with a 99.5% accuracy rate. Further evaluations will involve comparing the new LDA-based model to existing ones and demonstrating accuracy improvements. The review emphasises the importance of feature optimisation and appropriate model selection in improving CKD diagnosis accuracy, paving the way for improved patient care and early intervention.

Keywords: *Machine learning, CKD, diagnosis, Feature optimization.*

I. Introduction

Chronic kidney disease (CKD) is a severe stage of kidney damage in which the kidneys gradually lose functionality and may eventually fail completely. High blood pressure, cardiovascular disease, diabetes, age, and a family history of kidney failure are all risk factors for CKD. Obesity, autoimmune diseases, infections, and kidney-related disorders are all secondary risk factors.

Treatment for CKD varies depending on the patient's condition and may include lifestyle changes, medication, dialysis, and kidney transplantation. In the United States alone, approximately 37 million people were estimated to have CKD in 2021, and globally, approximately 10% of the population suffers from CKD, resulting in approximately 2.4 million deaths each year.

II. Methods and Models

2.1 Dataset: The Cleveland Kidney Disease dataset, which contains 400 instances and 25 attributes, is central to the research paper. This dataset serves as the foundation for using machine learning techniques to diagnose chronic kidney disease (CKD). Each instance represents a patient's data, including various medical related attributes. The dataset's attributes, such as blood pressure, age, and other health indicators, provide critical insights into factors that may contribute to CKD diagnosis.
Descriptive statistics: such as calculating the minimum, maximum, average, standard deviation, variance, and mean of attributes, aid in understanding the distribution and characteristics of the data.
Visualisation of Data Quality: Graphs and heat maps are created to visually assess the data's quality. These visualisations aid in the identification of trends, patterns, and irregularities in the dataset.
Outlier Detection and Removal: Outliers, or data points that differ significantly from others, can skew analysis results. To avoid distortion, outliers are identified and removed.

2.2 Data Pre-processing

Handling Missing Values: Missing values, which are frequently denoted by blanks or NaN (Not a Number), can distort analysis results. To maintain data integrity, missing values are removed from the dataset in this study.

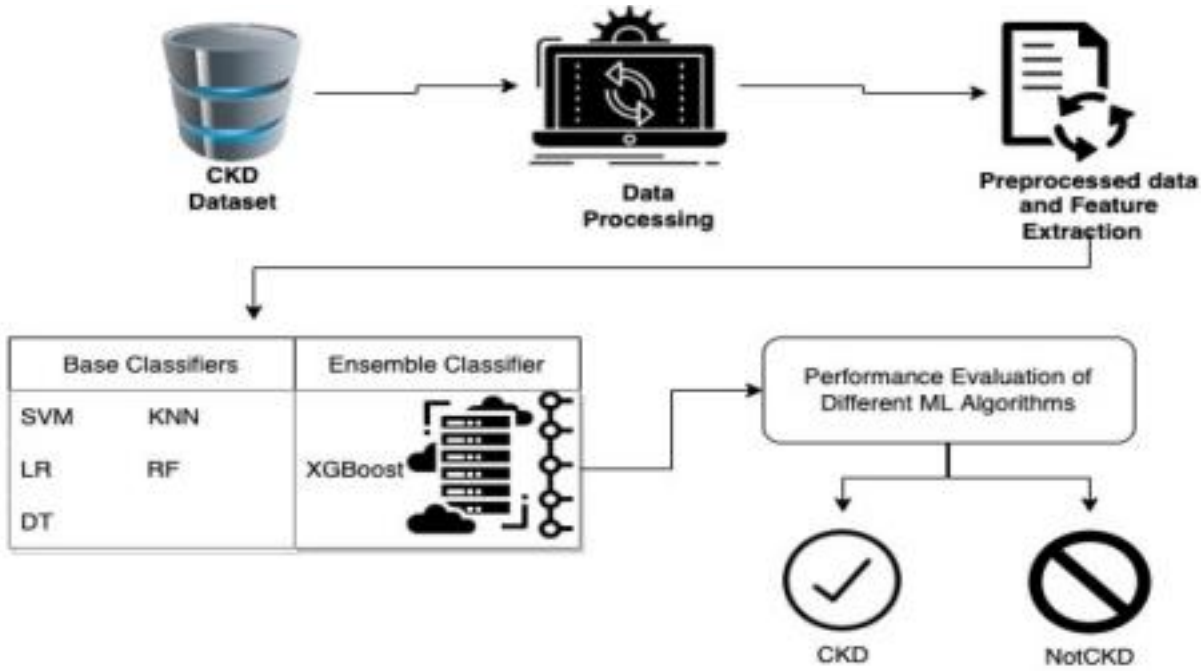

Fig 1: Structure of detection of CKD

2.3 Feature Selection

The MRMR algorithm: is concerned with selecting features that maximise relevance while minimising redundancy. To ensure a well-rounded set of features, it considers both the individual importance of attributes and their interactions.
Attribute Reduction: The MRMR algorithm identifies and retains the top performing features from the initial 25 attributes. The goal of this reduction is to simplify the dataset, increase model efficiency, and improve predictive accuracy.
Importance of Features: The features chosen have the strongest relationships with the target variable, CKD diagnosis. These characteristics play an important role in predicting the presence or absence of chronic kidney disease.

2.4 Machine Learning Models

Linear Regression: This model predicts outcomes based on the linear relationships between the input variables and the target variable.
Support Vector Machine (SVM): SVM determines the best hyperplane for separating different classes of data. It is useful for both classification and regression tasks.
Decision tree: To make decisions, decision trees split data based on attribute values. They are intuitive and are capable of capturing complex relationships.
K-Nearest Neighbours (KNN): KNN classifies data points in the feature space based on the majority class of their k nearest neighbours.
Linear Discriminant Analysis (LDA): LDA is a classification technique that focuses on determining the best linear combination of attributes for separating different classes.

2.5 LDA Performance
Accuracy: LDA achieves an impressive 99.5% accuracy. This means that the model's predictions are very close to the actual CKD diagnosis outcomes.
Decision boundary: that best distinguishes cases with chronic kidney disease from those without. The linear combination of attributes that maximises class separation yields this boundary.
Dimensionality Reduction: In addition to classification, LDA performs implicit dimensionality reduction. It maps data points onto a lower-dimensional subspace that separates classes optimally.
Clinical Importance: The high accuracy of LDA in diagnosing chronic kidney disease emphasises its clinical importance. Early and accurate diagnosis can lead to timely interventions and better patient outcomes.

III. Model Combination Comparison Analysis

Combination of LDA and SVM: achieves an accuracy of 98.5% using 10-fold cross validation and 98.75% without any folds.
Combination of LDA and Gradient Boosting: yields an accuracy of 98.125% with 10-fold cross-validation and a perfect 100% accuracy without any folds.
Combination of LDA and Linear SVM: achieves an accuracy of 98.25% with 10- fold cross-validation and 98.75% without any folds.
Combination of LDA and KNN: Accuracy of 97.875% with 10-fold cross validation and 98.75% without any folds.
Accuracy Improvements and Model Selection: According to the study, combining LDA with different machine learning models consistently results in improved accuracy in diagnosing chronic kidney disease. This observation emphasises the importance of using LDA as a foundational model because of its ability to improve diagnostic precision.

IV. Recommendations and future scope

4.1 Multi-Dataset Validation: Extend the research by validating the findings across multiple datasets from various sources. This will contribute to the generalizability of the identified optimal feature optimisation technique and strengthen the conclusions.

4.2 Hybrid Feature Optimization Approaches: Consider combining multiple feature optimisation techniques to capitalise on their respective strengths. Combining feature selection and feature reduction techniques, for example, may result in improved performance by combining the benefits of both approaches.

4.3 Integration of Clinical Expertise: Collaborate with medical professionals and domain experts to incorporate their perspectives into feature selection. A hybrid approach that combines clinical knowledge with feature optimisation algorithms may result in more relevant and interpretable feature subsets.

Methods	Technique	Accuracy
Polat et al	Filter feature selection + SVM	98.5 (using 10-fold)
Ghosh et al.	Feature selection + GB	99.80 (using no fold)
Chittora et al.	SMOTE + Linear SVM	98.86 (using no fold)
Deepika et al.	KNN	97 (using no fold)
Drall et al.	CD feature selection + KNN	100 (using no fold)
Proposed method	LDA feature reduction + Max voting Ensemble classification	99.5% (using 10-fold) 100% (using no fold)
Performance of LDA with different classifier	LDA + SVM	98.5 (using 10-fold) 98.75 (using no fold)
	LDA + GB	98.125(using 10-fold) 100 (using no fold)
	LDA + Linear SVM	98.25 (using 10-fold) 98.75 (using no fold)
	LDA + KNN	97.875 (using 10-fold) 98.75 (using no fold)

Fig 2: Comparison of different Machine Learning Based CKD diagnosis System

4.4 Application to Other Medical Diagnoses: Extend the research framework beyond CKD to other medical diagnoses. Analyse the performance of feature optimisation techniques in various disease contexts and investigate their utility in improving diagnostic accuracy across multiple healthcare domains.

V. Results

Missing values were handled using median substitution during rigorous pre-processing, and binary nominal values were converted to numerical values for consistency. The study concentrated on three main feature optimisation strategies: feature importance, feature selection, and feature reduction. Each strategy included two widely used techniques, for a total of six approaches considered.

The study evaluated each optimised feature set using an ensemble classifier and the max voting technique. Logistic Regression, Random Forest, Support Vector Machine, K-nearest Neighbours, and Xtreme Gradient Boosting were the five influential classification models in the ensemble.

The findings demonstrated the importance of feature optimisation on the performance of the diagnostic system. Experiment results showed that certain techniques within each optimisation approach significantly improved the accuracy of the machine learning-based CKD diagnostic system. The ensemble model amplified this effect, demonstrating superior predictive abilities over individual models. The findings highlighted the importance of selecting appropriate feature optimisation techniques to advance CKD diagnostics, as well as the potential of an ensemble approach for improving diagnostic accuracy.

6. Conclusion

The paper "Analysis of the Performance of Feature Optimisation Techniques for the Diagnosis of Machine Learning-Based Chronic Kidney Disease" makes an important contribution to the field of medical diagnostics. The study demonstrates the power of machine learning in improving chronic kidney disease diagnosis accuracy by navigating through data pre-processing, feature selection, model application, and evaluation. Notably, the Minimal Redundancy-Maximal-Relevance (MRMR) algorithm helps with feature reduction, streamlining the dataset for better results. A comparison of various machine learning models demonstrates the effectiveness of Linear Discriminant Analysis (LDA), which emerges as the clear winner with a remarkable accuracy of 99.5%. The findings highlight how meticulous feature optimisation, combined with astute model selection, can result in significant improvements in diagnostic accuracy. This research bridges the gap between healthcare and technology by providing insights that could significantly impact patient care, treatment strategies, and intervention approaches. The implications of this study go beyond academia, with the potential to revolutionise how chronic kidney disease is detected and managed in clinical

settings. This research paves the way for further exploration, validation, and application of these methodologies in broader medical contexts as the field evolves.

References

[1] M. Rashed-Al-Mahfuz, A. Haque, A. Azad, S. A. Alyami, J. M. W. Quinn and M. A. Moni, "Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 9, pp. 1-11, 2021,https://journalofbigdata.springeropen.com/ articles/10.1186/s40537-022-00657-5

[2] Deepika, B., Rao, V. K. R., Rampure, D. N., Prajwal, P., & Gowda, D. G. (2020). "Early prediction of chronic kidney disease by using machine learning techniques." American Journal of Computer Science and Engineering Survey, 8(2), 7

[3] Khalid, Hira, Khan, Ajab, Zahid Khan, Muhammad, Mehmood, Gulzar, Shuaib Qureshi, Muhammad, Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease, Computational Intelligence and Neuroscience, 2023, 9266889, 14 pages, 2023

[4] Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data." BMC Bioinformatics, 18(1), 1–14.

[5] Yan, K., & Zhang, D. (2015). "Feature selection and analysis on correlated gas sensor data with recursive feature elimination." Sensors and Actuators B (Chemical), 212, 353–363.

[6] P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in IEEE Access, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763