Vol. 3 No. 3 (2025)

Spectrum of Engineering Sciences

Online ISSN
3007-3138

Print ISSN
3007-312X

Enhancing Chronic Kidney Disease Prediction with a Hybrid Hard Voting Classification Approach

Kavita Tabbassum[1]
Information Technology Center, Sindh Agricultural University Tandojam, 70060, Pakistan. kavita@sau.edu.pk
Waqas Ahmed Gadahi[2]
Information Technology Center, Sindh Agricultural University Tandojam, 70060, Pakistan. winformation86@gmail.com
Saima Shaikh[3]
Information Technology Center, Sindh Agricultural University Tandojam, 70060, Pakistan. ss2kcs@gmail.com
Shahnawaz Farhan Khahro[4]
Energy Department, Government of Sindh, Pakistan.
shahnawazfarhan@gmail.com

Abstract

Chronic kidney disease (CKD) affects over 10% of the global population, equating to more than 800 million people. It is prevalent among older adults, women, minorities, and individuals with diabetes and hypertension. CKD also presents a significant health burden in low- and middle-income countries. The growing prevalence of CKD and its adverse impacts emphasize the urgent need for enhanced prevention and treatment strategies. Machine learning (ML), a prominent application of artificial intelligence, has made significant strides in healthcare research. The aim of this study is to design an ensemble method for more accurate prediction of CKD in patients. The proposed ensemble hard voting classifier integrates four machine learning algorithms—Support Vector Machine, Logistic Regression, K-Nearest Neighbors Classifier, and Random Forest Classifier. The model is trained on a dataset sourced from Kaggle, which includes data from 400 patients with 25 features.

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X
SPECTRUM OF ENGINEERING SCIENCES

Keywords: Chronic kidney disease, machine learning, ensemble method, hard voting classifier.

Introduction

Over 800 million people globally have chronic kidney disease. Those with diabetes, hypertension, old age, women, and people of dark color are at higher risk for chronic kidney disease. Low- and middle-income nations suffer most from chronic kidney disease. Chronic renal disease is a leading cause of death worldwide, with related fatalities increasing over the past two decades. More prevention and treatment needed for chronic renal disease due to high number of sufferers and negative effects. (Csaba et al., (2022)).

In China, around 119.5 million individuals, or 10.8%, are estimated to be impacted by this illness. CKD patients who lose kidney function may develop ESKD, which requires KRT. Prompt management can improve quality of life and reduce morbidity, mortality, and healthcare costs. A model predicting ESKD risk in early CKD stages is crucial for personalized treatment, improving prognosis and reducing financial burden. Factors used to predict ESKD include age, gender, lab findings (particularly albuminuria and eGFR).

New techniques for creating a prediction model, which previously relied on conventional statistics, become available with the introduction of the big data age. Artificial intelligence (AI) is a subset that includes machine learning (ML), which enables the computer to carry out a certain activity without explicit instructions (Dritsas et al., 2022.).

Hybrid Technique

To do the ensemble, the Hard voting method is implemented, which gives us a robust result as the one presented by (Jorge A. Morgan-Benita (2022) for prediction of CKD disease in patients.

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X
SPECTRUM OF ENGINEERING SCIENCES

Review Of Literature

The slow adversity of kidney work over time could be a characteristic of persistent kidney disease. Since most individuals do not appear any side effects, an illness goes unnoticed. A strategy based on supervised learning is depicted in this work to form successful models for foreseeing the probability of CKD event by utilizing probabilistic, outfit learning-based, and tree-based models. The author too looked at SVM, LR, SGD, ANN, and k-NN. The created comes about highlighted the Turn Woodland show, which exhausted the other models with an AUC of 100 percent, Accuracy, Review, F-Measure, and Exactness of 99. 2 percent. (Dritsaset al., 2022).

The reason of this study about was to see in case machine learning (ML) may precisely foresee the probability that patients with persistent kidney illness would create end-stage kidney disease (ESKD). The use of machine learning to assess the forecast of CKD based on promptly accessible information. Three ML models with high ratings for performance and affectability propose that they may well be utilized for quiet screenings. Three ML models—logistic regression, naiv Bayes, and binary forest—performed in terms of forecast and affectability to the KFRE. The KFRE had the most elevated levels of specificity, accuracy, and precision. (Qiong et al., 2022).

The precise prediction of chronic renal disease appears to be one of the challenging research issues in medical science. The author evaluated and compared the performance of various machine learning (ML) techniques, such as Random Forest, Decision Tree, AdaBoost, Gradient Boost, K-Nearest Neighbors, XGBoost, GNB, and Extra Tree Classifier. For identifying chronic renal disease, each of these classification systems was examined and tested. In detecting CKD, the Gradient Boost classifier achieved an accuracy of 98%, while the recommended KNN, EXT method achieved an accuracy of 99%. The

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X

Gaussian Naive Bayes classifier was 96% accurate. Extreme Gradient Boosting, Decision Tree, and Random Forest were the three classifiers that achieved 95 percent accuracy. The accuracy with the AdaBoost classifier is as low as 94%. (Deepanita et al., (2022).

Materials And Methods

In order to increase accuracy, a hybrid technique has been proposed. These machine-learning techniques will be a part of this hybrid approach, 1. Support Vector Machine, 2. Logistic Regression, 3. K-Neighbors Classifier, 4. Random Forest Classifier). This study focused on analyzing a dataset comprised of 400 patient records and 25 features for the classification of the presence of chronic kidney disease. Every algorithm will generate its own set of classification results. After that, the hybrid approach with hard voting classification will be created using the same four algorithms. The dataset will be used to train the algorithms, with 70% of the data collected for training and 30% for testing. To determine how well the newly developed hybrid technique works in terms of Accuracy, Precision, Recall, and F1 score, the classification results of each algorithm as well as the results of the proposed algorithm have been compared.

Materials and Methods

Dataset Description

The dataset used for this study was sourced from Kaggle (insert specific dataset name or link). It contains data from 400 patients, with a total of 25 features related to demographic, medical, and clinical information. These features include variables such as age, gender, blood pressure, specific medical histories (e.g., diabetes, hypertension), and other relevant biomarkers. The dataset is labeled with the target variable, which indicates whether the patient has chronic kidney disease (CKD) or not.

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X
SPECTRUM OF ENGINEERING SCIENCES

Data Preprocessing

Prior to training the machine learning models, several preprocessing steps were carried out to ensure the quality and consistency of the data:

- **Handling Missing Data**: Missing values were either imputed using statistical techniques (such as mean or median imputation) or removed if the amount of missing data was significant.
- **Feature Scaling**: Continuous features were standardized to ensure that all variables contribute equally to the machine learning models.
- **Categorical Encoding**: Categorical features were encoded using one-hot encoding or label encoding, depending on the nature of the feature.

Machine Learning Algorithms

To predict chronic kidney disease, an ensemble hard voting classification approach was employed. This method combines the predictions of multiple machine learning models to improve accuracy and robustness. The following classifiers were used in the ensemble:

1. **Support Vector Machine (SVM)**: A powerful classifier known for its ability to handle high-dimensional data. The SVM model was trained with a radial basis function (RBF) kernel to capture non-linear relationships between features.
2. **Logistic Regression**: A simple yet effective classifier, logistic regression was used as a baseline model for comparison. It models the probability of CKD using a linear decision boundary.
3. **K-Nearest Neighbors (K-NN)**: A non-parametric classifier that predicts the class of a data point based on the majority class of its nearest neighbors. It is particularly useful for datasets with complex decision boundaries.
4. **Random Forest Classifier**: A versatile ensemble method based on decision trees, random forests combine multiple trees to increase prediction accuracy and reduce overfitting.

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X
SPECTRUM OF ENGINEERING SCIENCES

Ensemble Hard Voting Classifier

The ensemble method used in this study is the **Hard Voting Classifier**, which combines the predictions from the four individual classifiers. In hard voting, each classifier makes an independent prediction, and the final prediction is determined by majority voting (i.e., the class that receives the most votes is selected as the final prediction). This technique improves predictive performance by leveraging the strengths of each individual model.

Model Training and Evaluation

The models were trained using 70% of the dataset for training and 30% for testing. Cross-validation was performed to ensure the models generalize well to unseen data. The performance of the ensemble model was evaluated using various metrics, including:

- **Accuracy**: The proportion of correctly predicted instances.
- **Precision**: The proportion of true positive predictions among all positive predictions.
- **Recall**: The proportion of true positive predictions among all actual positive instances.
- **F1-Score**: The harmonic mean of precision and recall.
- **Confusion Matrix**: To visually assess the classification performance across different classes (CKD vs. non-CKD).

Software and Tools

The models were implemented using Python and the following libraries:

- **Scikit-learn**: For implementing machine learning models and preprocessing techniques.
- **Pandas**: For data manipulation and cleaning.
- **NumPy**: For numerical operations.
- **Matplotlib** and **seaborn**: For data visualization.

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X

Hyperparameter Tuning

To improve the performance of the individual classifiers, hyperparameter tuning was carried out using grid search and cross-validation. The optimal set of hyperparameters was determined for each model to achieve the best possible prediction accuracy.

Kidney Disease Dataset

The data set was collected in India over a two-month period. It has 400 rows and 25 features, such as sugar, red blood cells, and edoema in the feet. Identifying whether a patient has chronic renal disease or not is the goal. A property called "classification" serves as the basis for categorization, and its values are "ckd" (chronic kidney disease) or "notckd." I cleaned up the dataset by mapping the text to numbers and making a few minor modifications. After cleaning, I performed some exploratory data analysis, separated the dataset into training and testing groups, and then used the models on each group. It has been noted that the classification findings are initially not very satisfying. Therefore, I used the lambda function to replace the rows with Nan values with mode for each column rather than deleting them. Following that, I once again split the dataset into training and testing sets and ran models on them. This time, the findings are more favorable, and we can see that random forest and logistic regression perform best, with accuracy ratings of 1.0 and 0 misclassifications, respectively. Printing confusion matrices, classification reports, and accuracy are used to measure classification performance.

Data Pre-Processing

Information Pre-processing may be a pivotal step that changes information into a usable and effective arrange that can be boosted to a machine-learning calculation. Information normalization is the primary strategy utilized for information pre-processing. This strategy is utilized to perform direct information changes. It is moreover known as Min-max

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

normalization, since the attributes' entire values drop between [0,1]. Name encoding is the following pre-processing strategy utilized. This strategy is utilized on the subordinate variable, which is whether the person has kidney disease. As a result, all of the string values within the yield variable are supplanted with and 1.

Algorithm

1: Procedure choose (Dataset)
2: Select Features and Target of Dataset
3: Procedure Split_data (Dataset)
Training_data, Testing_data=split (attributes,label)
Return Training_data, Testing_dataVoting="hard"
M1 = SVM (Training_data, Testing_labels, Testing_data)
M2 = LogisticRegression(Training_data, Testing_labels, Testing_data)
M3 = K_Neighbors(Training_data, Testing_labels, Testing_data)
M4=RandomForest(Training_data, Testing_labels, Testing_data)
S_V_CLASSIFIER=concatenate (M1, M2, M3, M4)
S_V_CLASSIFIER.FIT (Training_data, Testing_labels, Testing_data)
Predictions = S_V_CLASSIFIER.predict(Testing_data)

Implementation of Ensemble Classifier

This figure represents the code written in python programming language that implements the Ensemble classifier with hard voting.

```
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier
import numpy as np


df = pd.read_csv("C://programs//kidney_disease.csv")
feature_columns = ['id','age','bp','sg','al','su','rbc','pc','pcc','ba','bgr','bu','sc','sod','pot

from sklearn.model_selection import train_test_split
X = df[feature_columns]
y = df.classification

df.fillna(df.median(numeric_only=True).round(1), inplace=True)

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.70,random_state=42)
clf1 = SVC(probability=True,random_state=42)
clf2 = LogisticRegression(random_state=1)
clf3=KNeighborsClassifier(n_neighbors=1,metric='minkowski',p=2)
clf4 = RandomForestClassifier(random_state=42)

print ("Result")

from mlxtend.classifier import EnsembleVoteClassifier

eclf = EnsembleVoteClassifier(clfs=[clf1,clf2,clf3,clf4],voting = 'hard',weights=[1,1,1,1],)
```

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X
SPECTRUM OF ENGINEERING SCIENCES

Results

The aims of this research is to design an ensemble method for better prediction of chronic kidney disease in patients. The proposed ensemble hard voting classifier uses the ensemble of four-machine learning algorithms (Support Vector Machine, Logistic Regression, K-Neighbors Classifier and Random Forest Classifier) and have been trained with dataset that has been taken from kaggle.com and contains 400 patients with 25 features.

The resulting accuracy scores were as follows: in an effort to rectify any potential imbalances in the dataset, K5-cross-validation (Stratified) techniques were employed. Support Vector Machine- 0.62, Logistic Regression- 0.95, Random forest -0.99, K-Nearest Neighbors- 0.89, and Ensemble_With_Hard_Voting- 0. 97.

Further findings of the study reveal that the resulting accuracy scores with direct fitting data on algorithms were as follows:

MACHINE LEARNING ALGORITHM	ACCURACY:	F1 SCORE:	RECALL:	PRECISION:
SUPPORT VECTOR MACHINE	0.633333333	0.775510204	1	0.633333

MACHINE LEARNING ALGORITHM	ACCURACY:	F1 SCORE:	RECALL:	PRECISION:
LOGISTIC REGRESSION	1	1	1	1

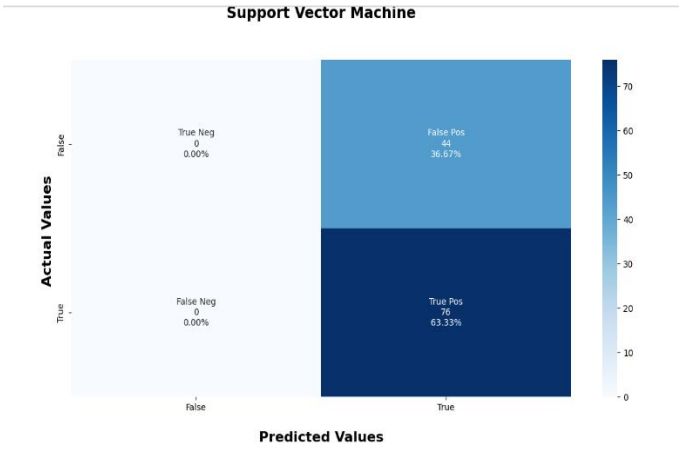
Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X

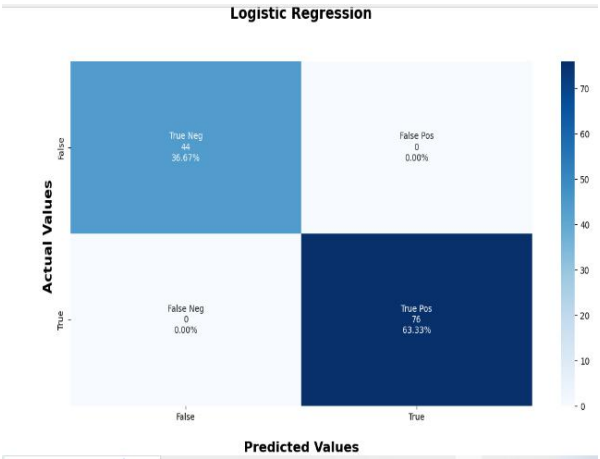
MACHINE LEARNING ALGORITHM	ACCURACY:	F1 SCORE:	RECALL:	PRECISION:
KNEIGHBORS	0.941666667	0.953020134	0.934210526	0.972602744

MACHINE LEARNING ALGORITHM	ACCURACY:	F1 SCORE:	RECALL:	PRECISION:
RANDOM FOREST	1	1	1	1

MACHINE LEARNING ALGORITHM	ACCURACY:	F1 SCORE:	RECALL:	PRECISION:
ENSEMBLE_WITH-HARD_VOTING	1	1	1	1

The confusion matrix of Support Vector Machine, Logistic Regression Kneighbors, Random Forest, and confusion matrix of proposed ensemble hybrid technique:



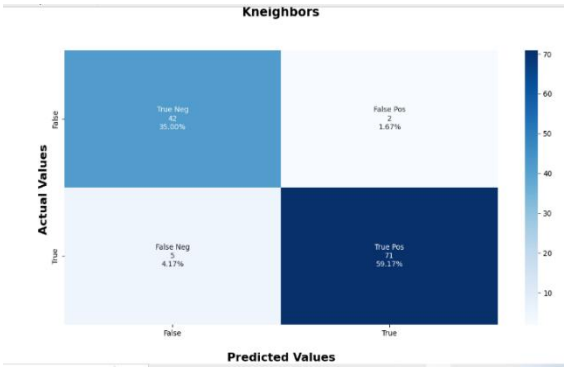


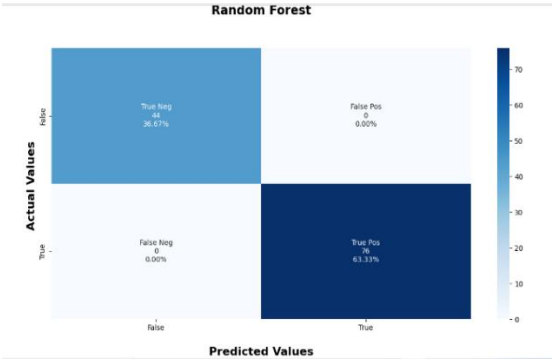
Vol. 3 No. 3 (2025)

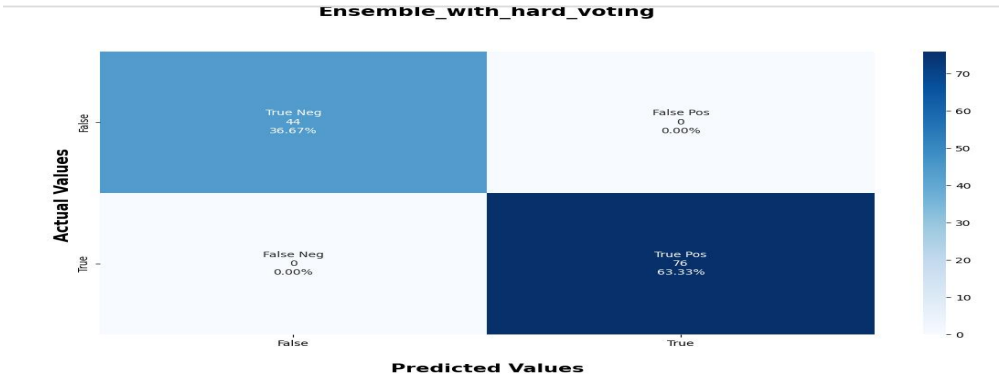
Spectrum of Engineering Sciences

Online ISSN
3007-3138

Print ISSN
3007-312X







Conclusion

After analysis of the results, it can be said that 0.62% accuracy is provided by Support Vector Machine, 0.95% accuracy by Logistic Regression, 0.89% accuracy by K-Nearest Neighbors, 0.99% accuracy by Random Forest and 0.97% accuracy by proposed method. It can be concluded that proposed model have represented highest accuracy 0.97% than all others except Random forest.

References

1. Csaba, P. (2022). Epidemiology of chronic kidney disease. Division of Nephrology, Department of Medicine, University of Tennessee Health Science Center, Memphis, Tennessee, USA.

2. Dritsas, E. (2022). Machine Learning Techniques for Chronic

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X

Kidney Disease Risk Prediction. Big Data Cogn. Comput. 2022, 6, 98. https://doi.org/10.3390/ bdcc6030098.

3. Jorge, A. (2022). Hard Voting Ensemble Approach for the Detection of Type-2 Diabetes, in Mexican Population with Non-Glucose Related Features; https://www.mdpi.com/journal/healthcare.

4. Qiong, B. (2022). Machine learning to predict end stage kidney disease in chronic kidney disease. Scientific Reports | (2022) 12:8377 | https://doi.org/10.1038/s41598-022-12316-z

5. Deepanita, B. (2022). A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method. Department of Computer Science & Engineering Daffodil International University Dhaka, Bangladesh deepanita15-2453@diu.edu.bd.

6. KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney International Supplements, 2013.

7. Jha, V., et al. "Chronic Kidney Disease: Global Dimension and Perspectives." *The Lancet*, vol. 382, no. 9888, 2013, pp. 260-272. DOI: 10.1016/S0140-6736(13)60687-X

8. Nguyen, D., & Ryu, K. "Machine Learning in Healthcare: A Review of Algorithms, Applications, and Trends." *Journal of Healthcare Engineering*, vol. 2020, 2020, pp. 1-14. DOI: 10.1155/2020/6945438

9. Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830.

10. Breiman, L. "Random Forests." *Machine Learning*, vol. 45, no. 1, 2001, pp. 5-32. DOI: 10.1023/A:1010933404324

11. Cortes, C., & Vapnik, V. "Support Vector Networks." *Machine Learning*, vol. 20, no. 3, 1995, pp. 273-297. DOI: 10.1007/BF00994018

12. Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN: 978-0-387-31073-2

13. Zhang, Z., et al. "K-Nearest Neighbor Algorithm in Machine

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

Learning." *Proceedings of the International Conference on Machine Learning*, 2008.

14. López, M., et al. "Application of Ensemble Methods for Predicting Chronic Kidney Disease." *International Journal of Data Science and Analytics*, vol. 3, 2017, pp. 105-115.

15. Chawla, N. V., & Davis, D. A. "Data Mining for Imbalanced Datasets: An Overview." *Data Mining and Knowledge Discovery Handbook*, 2009, pp. 853-867.

16. Raschka**, S.** *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*. Packt Publishing, 2019. ISBN: 978-1838552808