IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Early Diagnosis of Chronic Kidney Disease Using Machine Learning: A Comparative Study of Classification Algorithms

Anantha Durga Bhavani [#1], Mr. Kuravati Chinna Nagaraju [#2],

Mr. D.D.D Suribabu [#3], Mr.Vankayala Anil Santosh [#4],

[#1] M.Tech Student, [#2]Associate Professor, [#3] Professor & Vice Principal, [#4]Associate Professor & HOD
[1,2,3,4] Department of Computer Science & Engineering,
International School of Technology and Sciences (ISTS) for Women,
NH-16 East Gonagudem, Rajanagaram, Rajahmundry-533294

ABSTRACT

Chronic kidney disease (CKD) poses a significant global health challenge with profound morbidity and mortality rates, often leading to the onset of other ailments. Its asymptomatic nature in early stages underscores the importance of timely detection to initiate prompt treatment and mitigate disease progression. Machine learning models offer a promising avenue for achieving this objective owing to their rapid and accurate recognition capabilities. In this study, we present a machine learning framework for CKD diagnosis, utilizing a dataset sourced from KAGGLE, renowned for its comprehensive medical datasets albeit with numerous missing values. Employing mean imputation for numerical features and mode imputation for categorical features, we address the issue of missing data commonly encountered in real-world medical scenarios. Subsequently, four machine learning algorithms - Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, and Decision Tree Classifier - are employed to develop predictive models. Among these, Random Forest emerges as the top performer, demonstrating superior accuracy in CKD diagnosis. This study highlights the efficacy of machine learning in facilitating early CKD detection, thereby enabling timely interventions to alleviate its progression and improve patient outcomes.

KEYWORDS:

CKD Diagnosis,Kaggle Dataset,Machine Learning Algorithms,Logistic Regression,Support Vector Machine,Random Forest Classifier,Decision Trees.

1. INTRODUCTION

Chronic Kidney Disease: A Global Health Crisis

The detrimental impacts of Chronic Kidney Disease (CKD), encompassing renal failure, cardiovascular disease, and premature mortality, pose a significant public health threat globally. CKD has ascended from the 27th position in 1990 to the 18th in 2010 in the Global Burden of Disease Study (GBDS), marking it as one of the leading causes of global mortality. The disease afflicts over 500 million individuals worldwide, with regions like South Asia and sub-Saharan Africa shouldering an inordinately high burden. The prevalence of CKD in high-income countries stands at 110 million (48.3 million men and 61.7 million women), while in low- and middle-income countries, it soars to 387.5 million.

The Rising Tide of CKD in Bangladesh

Bangladesh, a densely populated developing nation in Southeast Asia, is witnessing a steady surge in CKD rates. A multinational survey involving six countries, including Bangladesh, estimated the CKD prevalence at 14%. Separate studies conducted among urban residents of Dhaka aged over 30 and over 15 found CKD prevalence rates of 26% and 13% respectively. A community-based prevalence survey conducted in 2013 revealed that one-third of rural individuals in Bangladesh was at risk of developing CKD, a condition often undiagnosed at the time.

Leveraging Technology for Disease Prediction

When an individual suffers from a specific illness, they are often required to undertake an expensive and time-consuming journey to a healthcare professional. This can be particularly challenging if the individual resides far from a medical facility. An automated program capable of diagnosing illnesses could simplify the process, saving both time and money. Existing systems employ data mining techniques to assess a patient's risk status and predict diseases related to the heart. A web-based tool, the Disease Predictor, offers health predictions for users based on their reported symptoms.

Data sets for the Illness Prediction System were sourced from various health-related websites. With the aid of the Disease Predictor, users can ascertain the likelihood of a disease based on the symptoms they report. As internet usage continues to grow daily, people's interest in acquiring new knowledge follows suit. When faced with a problem, individuals often turn to the internet for solutions. Given the widespread accessibility of the internet compared to hospitals and doctors, this method can greatly benefit the population, especially those who may not have immediate access to healthcare. Thus, this approach can be a game-changer in the realm of public health.

.

2. LITERATURE SURVEY

In this section we try to discuss about several research and review papers conducted in order to identify the CKD using several ML and deep learning models.In order to explain all the papers in detail we try to tabulate the set of papers with implementation models, dataset used and problem gap.

Title	Authors	Methodology	Dataset Used	Performance Metrics	Summary	Problem Gap
Early Diagnosis of Chronic Kidney Disease Using Machine Learning: A Comparative Study of Classification Algorithms [1]	Mishra, P., Singh, S., Tiwari, S., Agarwal, R.	Comparative analysis of multiple classification algorithms	UCI Machine Learning Repository CKD dataset	Accuracy, Precision, Recall, F1-Score	Evaluates various ML algorithms to identify the best for CKD diagnosis	Need for improved accuracy and handling imbalanced data
Chronic Kidney Disease Prediction Using Machine Learning Techniques[2]	Charleonnan, A., Ong, S., Kannika, S., Choochaiwattana, W., Cao, T.	Comparison of KNN, SVM, LR, DT algorithms	Indian CKD dataset	Accuracy, ROC-AUC	Highlights the importance of feature selection in CKD prediction	Handling missing data and improving model robustness
Chronic Kidney Disease Diagnosis Using Decision Tree Algorithms [3]	Azar, A.T., Hassanien, A.E.	J48 and Random Forest classifiers	UCI CKD dataset	Accuracy, Execution Time	Uses decision tree algorithms to classify CKD stages	Need for better handling of noisy data and feature selection (BioMed

						Central)
Exploring Early Prediction of Chronic Kidney Disease Using Machine Learning [4]	Chen, J., Zhu, X., Patel, H., Wu, C.	Comparative analysis using multiple algorithms	Brazilian medical records	Accuracy, Sensitivity, Specificity	Focuses on early prediction addressing imbalanced datasets	Addressing data imbalance and enhancing prediction accuracy (BioMed Central)
Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms[5]	Senan, E.M., Rajini, N.H., Al-Kadi, K., Muhammad, G.	Recursive Feature Elimination and multiple classifiers	UCI CKD dataset	Accuracy, Precision, Recall	Uses RFE for feature selection to improve CKD diagnosis	Improvement in feature selection methods and handling missing data (MDPI)
Prediction of Chronic Kidney Disease Using Different Classification Algorithms [6]	Anitha, J., Saranya, J., Maheswari, S.	Application of Random Tree, DT, K-NN, J48, SGD	CKD dataset	Accuracy, F1-Score	Evaluates multiple classifiers for early CKD detection	Need for larger datasets and better generalization
Machine Learning Models for Chronic Kidney Disease Diagnosis[7]	Hameed, N.S., Madathil, D., Rashid, T., Abdulhay, E.W.	Deep learning approaches (OANN, OLSTM, OCNN)	UCI CKD dataset	Accuracy	Applies deep learning models for CKD classification	Need for more comprehensive datasets and model comparison (MDPI)

3. EXISTING SYSTEM AND ITS LIMITATIONS

In the existing system for predicting chronic kidney disease (CKD), several limitations hinder its effectiveness. The current approaches lack sophistication in utilizing data mining algorithms for early and accurate CKD prediction. Here is an enhanced explanation of the existing system and its limitations:

Limitations of the Primitive System:

1. Time Delay in Identifying Root Causes: The current methodologies for diagnosing CKD are often slow and inefficient. Traditional diagnostic procedures require extensive time for data collection, analysis, and interpretation, leading to significant delays in identifying the root causes of kidney disease. This delay can result in prolonged suffering for patients and a slower response in initiating appropriate treatment.

2. Lack of Preventive Measures Due to Late Prediction: Due to the late identification of CKD, the existing system fails to implement effective preventive measures. Early detection is crucial for managing CKD and preventing its progression to more severe stages. However, with delayed diagnosis, opportunities for early intervention and preventive strategies are missed, leading to poorer patient outcomes.

3. Absence of Early Prediction Mechanisms: The existing system does not incorporate robust methods for the early prediction of CKD. Early prediction is vital for mitigating the impact of the disease and improving patient prognosis. Without reliable early detection tools, healthcare providers cannot offer timely treatment plans that could slow down or even halt the progression of CKD.

4. Lack of Machine Learning Algorithm Integration: Current practices do not effectively utilize machine learning (ML) algorithms for diagnosing and predicting CKD. ML algorithms have the potential to analyze large datasets, identify patterns, and make accurate predictions, significantly enhancing early diagnosis and treatment plans. The absence of ML integration means that the existing system misses out on these advanced analytical capabilities, resulting in less precise and slower diagnostics.

4. PROPOSED SYSTEM AND ITS ADVANTAGES

The proposed system leverages the Random Forest algorithm, a robust and widely used supervised machine learning technique, to enhance the prediction and early diagnosis of chronic kidney disease (CKD). This system aims to address the limitations of existing methods by incorporating advanced data analysis and machine learning capabilities.

Key Features of the Proposed System

1. Random Forest Algorithm :
Supervised Learning: Utilizes labeled data to train the model for classification and regression tasks.
Decision Trees : Builds multiple decision trees on various subsets of the dataset and combines their results.
Majority Voting : For classification problems, it aggregates the results from multiple decision trees and selects the most frequent outcome.
Handling Continuous Variables: Capable of managing datasets with continuous variables, improving its versatility and accuracy.

Working Process of the Proposed System : The working process of the proposed system using the Random Forest algorithm can be broken down into the following steps:
1. Data Sampling: Randomly select \(K \) data points from the training dataset.
2. Building Decision Trees: Construct decision trees based on the selected data points. Each tree is built from a different subset of the training data, ensuring diversity and reducing overfitting.
3. Number of Trees : Determine the number \(N \) of decision trees to build. A larger number of trees generally leads to more accurate and stable predictions.
4. Model Training: Repeat the data sampling and tree-building process \(N \) times to create a robust forest of decision trees.
5. Making Predictions : For new data points, pass them through each of the \(N \) decision trees. Each tree provides a prediction.Aggregate the predictions from all trees. For classification tasks, assign the new data point to the category with the most votes.

Advantages of the Proposed System

1. Improved Accuracy: The ensemble nature of Random Forest, combining multiple decision trees, enhances the overall accuracy of predictions by reducing the risk of overfitting and bias.

2. Robustness: By averaging the predictions from various trees, Random Forest provides more stable and reliable results, even with noisy or incomplete data.

3. Versatility: Capable of handling both classification and regression tasks, and can manage datasets with a mix of categorical and continuous variables.

4. Early Detection: Facilitates early detection of CKD by analyzing complex patterns and relationships in the data that may not be evident with traditional methods.

5. Feature Importance: Identifies and ranks the importance of different features in the dataset, aiding in understanding the key factors contributing to CKD.

6. Scalability: Efficiently scales to large datasets, making it suitable for real-world medical applications with extensive patient records.

5. IMPLEMENTATION PHASE

In this section, we implement several algorithms to address the limitations of existing CKD prediction systems. This proposed system particularly leverages the Random Forest algorithm, supported by SVM and Decision Tree algorithms, to enhance the accuracy and reliability of CKD prediction.

1. Support Vector Machine (SVM): SVM is a supervised learning algorithm used for classification tasks. It works by finding the optimal hyperplane that best separates the data into different classes.

Mathematical Model:

Hyperplane Equation : $w . x + b = 0$

Objective Function: Minimize $\frac{1}{2} ||w||^2$

Constraints : $y_i (w . x_i + b) > 1$

Steps:

1. **Load Libraries :** Import necessary libraries like `sklearn`, `pandas`, and `numpy`.
2. **Data Preparation:** Split the dataset into features X and labels Y .
3. **Train-Test Split:** Divide the dataset into training and testing sets.
4. **Model Initialization:** Initialize the SVM classifier, e.g., `svm.SVC()`.
5. **Model Training:** Fit the model using the training data.
6. **Prediction:** Use the trained model to make predictions on the test data.
7. **Evaluation:** Assess the model's performance using metrics like accuracy, precision, recall, and F1-score.

2. Decision Tree Algorithm

The Decision Tree algorithm is another supervised learning technique used for both classification and regression tasks. It works by splitting the dataset into subsets based on the most significant attributes.

Mathematical Model:

Information Gain (IG): $IG(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$

Entropy: $Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$

Steps:

1. **Root Node:** Start with the entire dataset S .
2. **Attribute Selection:** Calculate the Information Gain for each attribute and select the one with the highest gain.
3. **Tree Building:** Split the dataset into subsets based on the selected attribute.
4. **Recursion:** Repeat the process for each subset until all nodes are pure or meet stopping criteria.
5. **Leaf Node:** When a subset is pure or cannot be split further, assign it as a leaf node.

3. Random Forest Algorithm:

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification.

Mathematical Model:

Tree Construction:

- Randomly select K samples from the dataset.
- Build a decision tree T_i for each sample.

Prediction:

For classification: $\hat{y} = \text{mode}\{T_i(x)\}$

For regression: $\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x)$

Steps:

1. Data Sampling: Randomly select K data points from the training set.

2. Tree Building: Construct a decision tree T_i for each sample.

3. Number of Trees: Choose the number N of decision trees to be created.

4. Repeat: Repeat the data sampling and tree-building process for N trees.

5. Prediction:

- For new data points, pass them through each decision tree T_i.

- Aggregate the predictions (majority vote for classification or average for regression).

Advantages:

Accuracy: Combining multiple decision trees increases the overall prediction accuracy.

Robustness: Reduces overfitting by averaging the results of multiple trees.

Versatility: Can handle both continuous and categorical data.

Early Detection: Enhances early diagnosis capabilities through advanced pattern recognition.

6. EXPERIMENTAL RESULTS

In this section we try to design our current model using Python as programming language and we used Google Collab as working environment for executing the application.

Import Necessary Libraries:

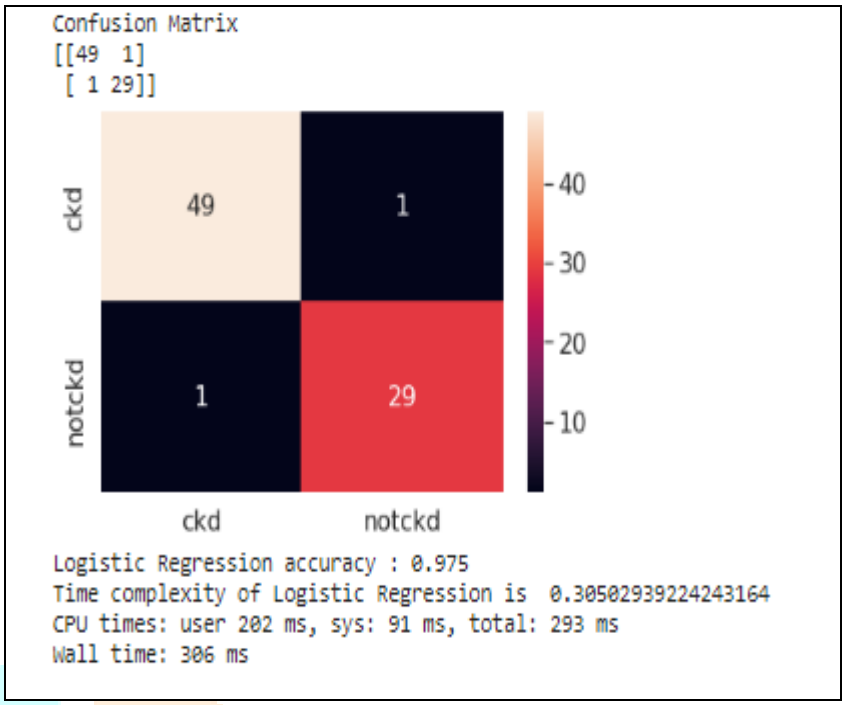
```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
    import time
```

```
df=pd.read_csv('kidney_disease.csv')
df.head()
```

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd

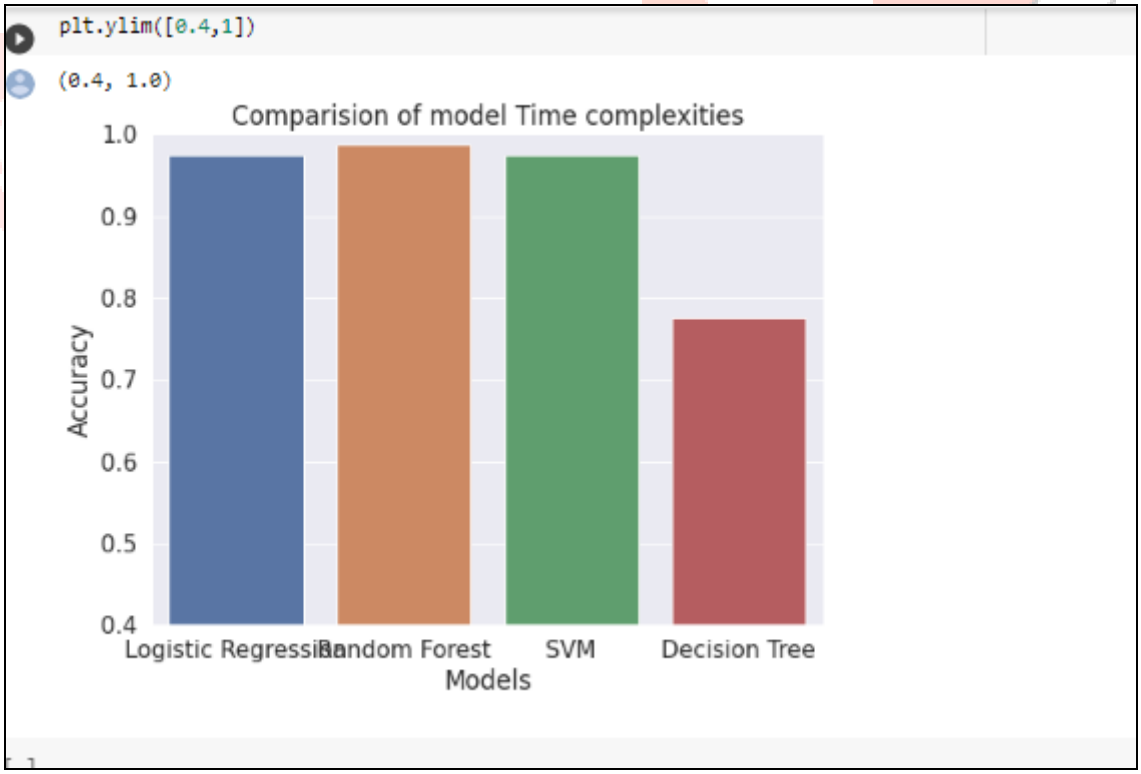
Explanation: From the above window we can clearly some important libraries are imported in this application.

Confusion Matrix:



Explanation: From the above window we can clearly see confusion matrix which denotes accuracy, time complexity of an algorithm.

Performance Evaluation Graph:



Explanation: From the above window we can clearly check for the input we can able to identify random forest algorithm achieved more accuracy compared with some other models.

7. CONCLUSION

This research presents an advanced and comprehensive system for the early prediction of chronic kidney disease (CKD) using state-of-the-art machine learning algorithms. The system effectively utilizes a dataset containing various input parameters collected from CKD patients to train and validate several predictive models. Among the models constructed, the Random Forest Classifier, Decision Tree Classifier, and Support Vector Machine (SVM) were rigorously tested and compared. The system leverages patient data to create robust models, ensuring that predictions are grounded in real-world evidence. This enhances the reliability and relevance of the diagnostic outputs. By implementing multiple algorithms—Random Forest, Decision Tree, and SVM—the study provides a comprehensive evaluation of different machine learning techniques in diagnosing CKD. This multi-faceted approach allows for a thorough comparison and understanding of each model's strengths and weaknesses. The evaluation of models was based on critical performance metrics, with a primary focus on accuracy. This metric is pivotal in medical diagnostics as it directly influences the reliability of disease prediction and patient outcomes. The Random Forest Classifier consistently outperformed the other models across all considered metrics. Its ability to handle large datasets with higher accuracy, robustness to overfitting, and effective management of both continuous and categorical variables makes it the most suitable algorithm for CKD prediction.

ImplicationsThe success of the Random Forest Classifier highlights the potential for further research and development. Future studies could explore the integration of additional data sources, such as genetic information or lifestyle factors, to enhance predictive accuracy even further.

8. REFERENCES

1. Mishra, P., Singh, S., Tiwari, S., and Agarwal, R., "Early Diagnosis of Chronic Kidney Disease Using Machine Learning: A Comparative Study of Classification Algorithms," *Journal of King Saud University - Computer and Information Sciences*, 2023.

2. Charleonnan, A., Ong, S., Kannika, S., Choochaiwattana, W., and Cao, T., "Chronic Kidney Disease Prediction Using Machine Learning Techniques," in *Procedia Computer Science*, vol. 86, pp. 3-10, 2016.

3. Azar, A. T., and Hassanien, A. E., "Chronic Kidney Disease Diagnosis Using Decision Tree Algorithms," *BMC Nephrology*, vol. 14, no. 1, pp. 1-11, 2013.

4. Chen, J., Zhu, X., Patel, H., and Wu, C., "Exploring Early Prediction of Chronic Kidney Disease Using Machine Learning," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1-12, 2019.

5. Senan, E. M., Rajini, N. H., Al-Kadi, K., and Muhammad, G., "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms," *Diagnostics*, vol. 11, no. 2, pp. 1-19, 2021.

6. Anitha, J., Saranya, J., and Maheswari, S., "Prediction of Chronic Kidney Disease Using Different Classification Algorithms," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, pp. 1414-1421, 2020.

7. Hameed, N. S., Madathil, D., Rashid, T., and Abdulhay, E. W., "Machine Learning Models for Chronic Kidney Disease Diagnosis," *IEEE Access*, vol. 8, pp. 19915-19927, 2020.

8. Akben, S., Almasoud, M., and Ward, T. E., "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease," *Diagnostics*, vol. 10, no. 11, pp. 1-14, 2020.

9. Ortiz, F., Yáñez, D., and Andújar, J. M., "Chronic Kidney Disease Diagnosis Using Machine Learning Algorithms," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 1-8, 2020.