IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Chronic Kidney Disease Prediction By Using Logistic Regression And Random Forest Model

[1]Hariom M. Bade, [2]Abhishek A. Bandwal, [3] Suraj P. Gore, [4]Yash P. Harne, [5]Prof. Dr. P. D. Bhamre

[1,2,3,4] Students, Department of Information Technology

[5] Head of Department Information Technology

K. K. Wagh Institute of Engineering Education and Research, Nashik, India

Affiliated under Savitribai Phule Pune University

Abstract: Chronic Kidney Disease (CKD) is a global health problem with a high morbidity and mortality rate. In such a disease, there are very few symptoms at an early stage; however, as time passes, the disease results in several damages. Early detection of CKD enables the patient to receive timely treatment and gives them a chance to minimize the damage. The proposed project aims to use a combination of Logistic Regression and Random Forest algorithms for detecting CKD. According to the literature survey, the combination of Logistic Regression Random Forest yields the best accuracy. Voting classifier is used to ensemble both the algorithms. Hence, machine learning methodology is employed, and the CKD dataset from the University of California Irvine (UCI) is used for training and detecting CKD in the early stage.

***Index Terms* - Chronic Kidney Disease, Logistic Regression, Random Forest, KNN**

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a prevalent and serious global health issue that affects a significant portion of the world's population. It is estimated that around 10% of people worldwide are affected by CKD, with the prevalence varying across different countries. In India, the percentage of CKD cases is approximately 13%, while it ranges from 10% to 15% in the United States. Moreover, recent studies have shown a prevalence of 14.7% in the Mexican adult general population [1]. CKD is characterized by a gradual deterioration in renal function, leading to the eventual loss of kidney function. Unfortunately, in its early stages, CKD often exhibits no obvious symptoms, which makes early detection and timely treatment challenging. Typically, CKD remains undetected until the kidneys have already lost around 25% of their function. The consequences of CKD are severe, with high morbidity and mortality rates. It is also strongly associated with the development of cardiovascular diseases. CKD is a progressive and irreversible condition [2]. Therefore, accurate prediction and early diagnosis of CKD are crucial to enable timely intervention and minimize the associated damages.

Machine learning techniques have gained significant attention in the medical field as promising tools for accurate and efficient disease diagnosis. Leveraging the power of computer algorithms, machine learning can analyze vast amounts of data to extract valuable insights and make predictions. Consequently, it holds great potential for diagnosing CKD effectively [3]. With the advancement of information technology, machine learning has become an asset in the medical domain. It has been successfully applied to monitor the human body's physiological status, analyze disease-related factors, and diagnose various ailments. In the context of CKD diagnosis, many previous studies have utilized the CKD dataset available from the UCI machine learning repository [4]. However, these studies often employ the K-Nearest Neighbors (KNN) algorithm for data imputation, primarily relying on mean imputation to handle missing values [5]. This approach assumes that missing measurements in patients' data can be substituted with average values based on diagnostic categories. However, in practice, patients mayhave missing measurements due to various reasons even before the diagnosis, rendering mean imputation inaccurate. Moreover, when dealing with missing values in categorical variables, mean imputation can introduce significant deviations from the actual values.

The aim of this project is to develop a machine learning system for predicting and detecting Chronic Kidney Disease (CKD) using Logistic Regression and Random Forest models. A voting classifier will be used to combine the predictions of these models for improved accuracy. The CKD dataset from the University of California Irvine (UCI) will be utilized for training and evaluation. The objective is to enhance early CKD detection, providing healthcare professionals with a valuable tool for intervention and better patient outcomes. This research aims to utilize machine learning to improve CKD diagnosis and management, contributing to global efforts against this prevalent disease.

II. OBJECTIVES

The objective of this project is to develop a machine learning-based system for the early detection of Chronic Kidney Disease (CKD). CKD is a global health issue with significant morbidity and mortality rates, and its symptoms are often absent in the early stages, making timely detection crucial for effective treatment and minimizing damage. To achieve this, a combination of Logistic Regression and Random Forest algorithms, with one more machine learning algorithm K-Nearest Neighbors (KNN) will be studied and evaluated. The project will utilize the CKD dataset from the University of California Irvine (UCI) for training and developing a voting classifier ensemble model using Logistic Regression and Random Forest. By leveraging machine learning techniques, this project aims to enhance the early detection of CKD, enabling healthcare professionals to intervene promptly and improve patient outcomes.

III. LITERATURE SURVEY

Chronic kidney disease (CKD) is a global health issue with a high morbidity and mortality rate. It often goes unnoticed in its early stages due to the absence of obvious symptoms. Timely detection of CKD is crucial for providing early treatment and minimizing disease progression. Machine learning models have emerged as effective tools for accurate and efficient CKD diagnosis, enabling clinicians to achieve this goal. The CKD dataset, sourced from the University of California Irvine (UCI) machine learning repository, contains a significant number of missing values. To address this, the incomplete dataset was appropriately imputed, and machine learning algorithms such as logistic regression and random forest were employed to build predictive models. Among these models, random forest demonstrated the highest diagnostic accuracy, achieving an impressive 99.75% accuracy rate [1].

Machine learning algorithms have revolutionized disease prediction by alleviating the burden on doctors and medical staff. According to the World Health Organization (WHO), approximately 85% of deaths related to heart disease are attributed to heart attacks and strokes. Numerous factors contribute to heart and diabetes problems, including age, gender, blood pressure, glucose levels, skin thickness, and insulin levels. These factors are easily measurable in primary care facilities, but accurate estimation of heart disease and diabetes data can greatly aid in predicting future cardiovascular issues. In this study, we utilize the Logistic Regression (LR) model and Support Vector Machine (SVM) model for disease prediction. The LR and SVM models have proven to be effective in predicting heart and diabetes diseases, offering accuracies of 85% and 78%, respectively. By leveraging these machine learning models, healthcare professionals can obtain accurate predictions and identify individuals who may be at a higher risk of developing heart disease or diabetes. This enables proactive interventions and targeted preventive measures, ultimately leading to improved patient outcomes and a reduction in the prevalence of these diseases [2].

In a comparative study focused on segregating data from patients with Chronic Kidney Disease (CKD), the researchers utilized the CKD dataset available from the UCI machine learning repository. The study aimed to compare the classification results of 400 CKD patients using different models. Specifically, the Random Forest algorithm was employed in the comparative model to analyze and sort patient data [3].

The utilization of data mining and machine learning techniques is widespread across various fields. Real-life datasets often exhibit an imbalance, with a scarcity of significant samples compared to less important ones, primarily due to the challenges in collecting representative positive examples. Conventional approaches aimed at reducing overall classification accuracy may not be suitable for imbalanced problems. Hence, in this study, we compare the performance of random forest and logistic regression models in predicting imbalanced datasets. To address the imbalanced nature of the data, we propose several enhancements based on cost-sensitive learning. These enhancements aim to provide more accurate predictions when dealing with imbalanced datasets. By evaluating and comparing the performance of these models, we aim to identify the most effective approach for handling imbalanced data. This research contributes to the development of improved strategies for dealing with imbalanced datasets and enhances the predictive capabilities of machine learning algorithms in real-world applications [4].

Collaborative healthcare data analytics is a systematic approach to analyzing healthcare data in order to identify opportunities for improvement in health system management. This approach involves three main components: pre-processing, attribute selection, and classification algorithms. The objective of this research is to develop a machine-based diagnostic approach using machine learning techniques for mining the risk factors associated with chronic kidney diseases. Over the years, several improved algorithms have been proposed to enhance the effectiveness of mining the risk factors of chronic kidney disease. However, these algorithms still have their limitations. In this study, we employed the Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN) algorithms to achieve an early diagnosis of CKD patients with high accuracy. By leveraging these machine learning algorithms, we aim to identify and predict CKD at an early stage, enabling timely interventions and improved patient outcomes. This research contributes to the advancement of diagnostic techniques in the field of chronic kidney disease and holds promise for improving healthcare decision-making and patient care [5].

Chronic kidney disease (CKD) is a significant global health concern characterized by a gradual decline in kidney function. Detecting CKD at an early stage is crucial as it often remains asymptomatic until advanced stages. Therefore, the primary objective of this study is to determine the presence of chronic kidney disease by applying various classification algorithms to patient medical records. The focus is on identifying the most suitable classification algorithm for CKD diagnosis. This research investigates several classification algorithms, including Support Vector Machine, Random Forest, XGBoost, Logistic Regression, Neural Networks, and Naive Bayes Classifier. The aim is to evaluate their performance in accurately diagnosing CKD based on the patient's medical records. Through experimentation and analysis, the results indicate that Random Forest and XGBoost exhibit superior performance compared to other classification algorithms in terms of diagnostic accuracy. By identifying the most effective algorithm for CKD diagnosis, this research contributes to improving the accuracy and efficiency of diagnosing CKD, leading to earlier interventions and enhanced patient outcomes [6].

Chronic Kidney Disease (CKD) is a prevalent kidney condition that has a significant impact on a large number of individuals. It gives rise to various complications such as weakened bones, anemia, nerve damage, and high blood pressure, and in severe cases, it can result in complete kidney failure. CKD contributes to millions of deaths worldwide each year. Diagnosing CKD is challenging due to the absence of prominent symptoms. Therefore, this paper proposes the use of a Multi-Layer Perceptron Classifier, which employs a fully connected Deep Neural Network, to predict whether a patient is affected by CKD. The model is trained on a dataset consisting of approximately 400 patients and takes into account several symptoms including blood pressure, age, sugar level, red blood cell count, and others. Through extensive experimentation, the results demonstrate that the proposed model achieves a testing accuracy of 92.5% in classifying patients for CKD diagnosis. By leveraging the capabilities of Deep Neural Networks, this approach offers a promising solution for accurate CKD classification, aiding in early detection and timely interventions. Ultimately, the proposed model has the potential to improve healthcare outcomes and reduce the burden of CKD on individuals and healthcare systems [7].

Chronic kidney disease (CKD) is a progressive condition characterized by a gradual loss of kidney function. Early detection plays a vital role in determining appropriate treatment strategies for CKD. Timely and accurate diagnosis can help prevent further deterioration of a patient's health. In this research paper, the authors employ various machine learning algorithms including Random Forest (RF) Classifier, Logistic Regression (LR), K-Nearest Neighbor (K-NN), and Support Vector Machine (SVM) to predict CKD. The dataset used for prediction consists of 400 instances with 25 attributes, collected from the UCI Repository. Experimental results demonstrate that the K-NN algorithm achieves an accuracy of 94%, Logistic Regression achieves 98% accuracy, SVM achieves 93.75% accuracy, and the RF classifier achieves the maximum accuracy of 100%. These findings suggest that machine learning models, particularly RF, LR, K-NN, and SVM, can effectively contribute to the early and accurate diagnosis of CKD. This research highlights the potential of these algorithms in improving CKD detection, leading to better patient outcomes and informed treatment decisions [8].

An imbalanced dataset is a dataset that reflects an unequal distribution of classes within a dataset. It is difficult to deal with unbalanced datasets in classification problems, and many classification algorithms do not perform well in unbalanced datasets. In this paper, logistic regression analysis is presented with Python on imbalanced datasets [9].

In this paper, a performance comparison between random forest and logistic regression algorithms was made by using real banking marketing data. In addition, these algorithms were run on WEKA, Google Colab, and MATLAB platforms to compare performance on different platforms. At the end of the study, the most successful result was obtained with 94.8 percent accuracy, 93.9 percent sensitivity, 94.8 percent recall, 94.4 percent fl-score, and 98.7 percent [10].

IV. METHODOLOGY

The methodology used in this project involves preprocessing a dataset of 400 patients using the K-nearest neighbors (KNN) imputation algorithm to fill in missing values. The preprocessed dataset is then used for model building using two main algorithms: Random Forest and Logistic Regression. The Random Forest algorithm extracts the most relevant features and improves the model's efficiency, while the Logistic Regression algorithm classifies the data into respective classes. The predictions from both models are combined using a voting classifier as an ensemble method to determine the final classification. The ensemble model takes user input data and detects whether the patient has CKD or not.

Architecture: As shown in Fig. CKD dataset consisting of 400 patients is provided to the system for model development. The first step in CKD detection involves preprocessing the dataset. In the figure, the K-Nearest Neighbors (KNN) imputation algorithm is applied to fill in the missing values, a process known as data cleaning. After preprocessing the dataset, two main algorithms are used for model building. The Random Forest algorithm is employed to extract the most relevant and useful features and group them together, enhancing the efficiency of the model. Additionally, the Logistic Regression algorithm is utilized to classify the data into respective classes. These two models' results are combined using a voting classifier as an ensemble method. The voting classifier combines the predictions from both models to determine the final classification. The ensemble model takes input data from the user and detects whether the patient has CKD or not.

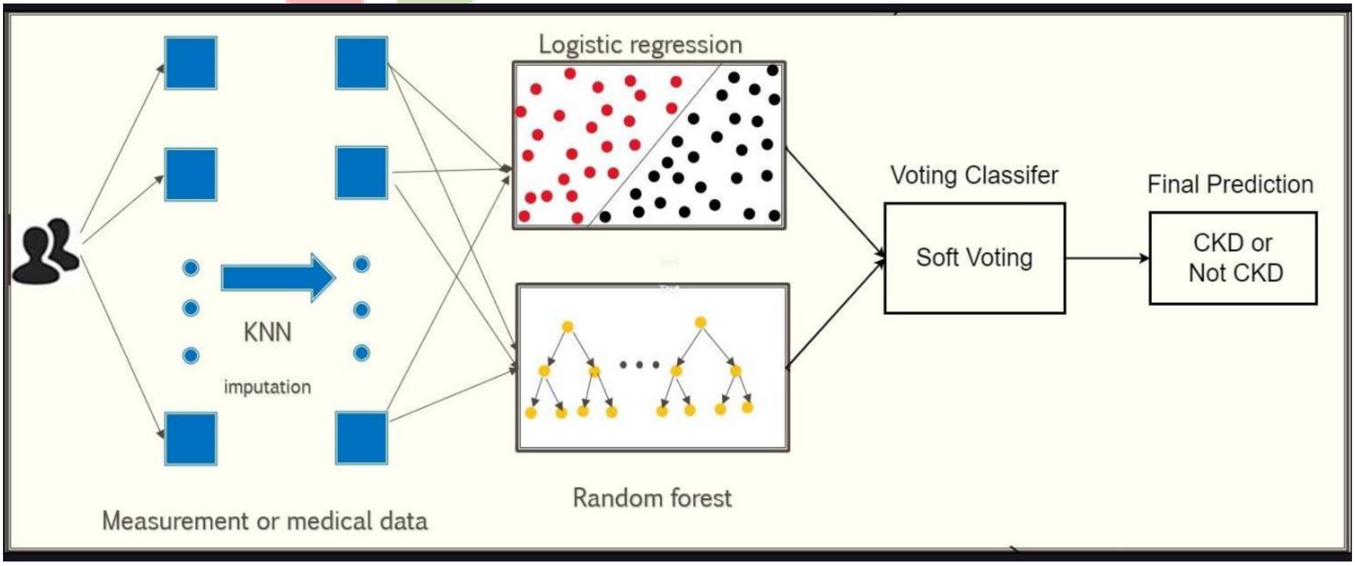

Fig: Architecture.

Step 1: Data preprocessing.

In this step, the dataset underwent several preprocessing steps as part of the data preparation process before training the machine learning models. Initially, the 'id' column was removed since it held no relevance to the classification task at hand. To enable compatibility with machine learning algorithms, categorical variables were encoded into numerical representations. The K-Nearest Neighbors imputation method was employed to handle missing values in the dataset. This technique allowed for the estimation of missing values based on the values of their nearest neighbors. Additionally, the input features were standardized using the StandardScaler technique. This process transformed the features to have a mean of zero and a standard deviation of one, ensuring uniformity and facilitating model training. These preprocessing steps were crucial in improving the dataset's quality, making it suitable for subsequent analysis and model training.

Algorithm:

- Data Cleaning: The unnecessary 'id' column was removed from the CKD dataset as it was not relevant for the classification task.
- Data Encoding: Categorical variables were transformed into numerical representations to facilitate the utilization of machine learning algorithms. For instance, 'rbc' and 'pc' were converted to binary values, where 'abnormal' was assigned a value of 1 and 'normal' was assigned a value of 0. Similarly, 'pcc' and 'ba' were encoded as 1 for 'present' and 0 for 'not present'. The variables 'htn', 'dm', 'cad', 'pe', and 'ane' were encoded as 1 for 'yes' and 0 for 'no'. 'appet' was encoded as 0 for 'good' and 1 for 'poor'.
- Target Variable Encoding: The target variable 'classification' was encoded as 1 to represent 'ckd' (chronic kidney disease) and 0 to represent 'notckd' (not chronic kidney disease). Additional variations of the 'ckd' and 'notckd' labels were also accounted for during the encoding process.
- Handling Missing Values: Missing values in the dataset, denoted by '\t?', were replaced with NaN. Subsequently, K-Nearest Neighbors (KNN) imputation was applied to estimate and fill in the missing values in the 'pcv', 'wc', and 'rc' columns based on the values of their neighboring instances.
- Feature Scaling: The input features were standardized using the StandardScaler, which transforms each feature to have zero mean and unit variance. This standardization ensures that all features are on a similar scale, preventing those with larger values from dominating the learning process and facilitating fair comparisons between different features.

Step 2: Configuring and creating an instance of logistic regression algorithm.

In this step, the configuration of logistic regression is chosen to optimize the performance of the model by taking into account the specific characteristics of the dataset and potential class imbalance. Create instance of logistic regression algorithm and store as ‘log_reg’.

Logistic regression predicts the probability of an individual having CKD based on their health attributes. The logistic regression model applies a logistic or sigmoid function to transform a linear equation into a range of 0 to 1, representing probabilities.

$$P(y = 1|X) = 1 \,/\, (1 + e\hat{}(-z))$$

In above formula, the logistic function is defined, where P(y=1|X) is the probability of CKD being present given the input features X, and z is the linear combination of the input features and their corresponding coefficients.

During training, the model adjusts the coefficients using a maximum likelihood estimation algorithm, which finds the values that maximize the likelihood of observing the training data. Once trained, the logistic regression model can predict the probability of CKD for new individuals by applying the learned coefficients to their health attributes.

To classify individuals, a threshold (e.g., 0.5) is set, and if the predicted probability exceeds the threshold, they are classified as having CKD; otherwise, they are classified as not having CKD.

Algorithm:

- Algorithm configured as,
- Penalty='elasticnet': Utilizes a regularization penalty combining L1 (Lasso) and L2 (Ridge) regularization, facilitating feature selection and coefficient shrinkage.
- random_state=620: Sets the random seed for result reproducibility, ensuring consistent results if the same seed is used.
- solver='saga': Specifies the optimization algorithm ('saga') for handling both L1 and L2 regularization penalties.
- class_weight= {0.0:0.85, 1.0:0.15}: Assigns class weights to handle imbalanced data, giving higher importance to the minority class (‘NOT CKD’) during training.
- l1_ratio=0.7: Determines the mixing parameter between L1 and L2 regularization in the 'elasticnet' penalty, with 70% weight on L1 regularization and 30% weight on L2 regularization.

Step 3: Configuring and creating an instance of random forest algorithm.

The third step involves the configuration of random forest algorithm is chosen to enhances the model's accuracy and robustness.
Create instance of random forest algorithm and store as ‘rf’.

To utilize Random Forest for CKD detection, the dataset is first divided into training and test sets. Then, bootstrap sampling is employed to create random subsets of the training data, allowing the construction of multiple decision trees. At each tree node, a random subset of features is considered, and the tree is split based on a criterion such as entropy or Gini impurity. This process is repeated for each subset, resulting in an ensemble of decision trees.

During the prediction phase, new instances from the test set are passed through each tree in the Random Forest. Each tree independently predicts the class, either CKD or not CKD, and the class with the majority votes is chosen as the final prediction.

Algorithm:

- Algorithm configured as,
- n_estimators=125: This parameter specifies the number of decision trees (estimators) that will be included in the random forest. By setting it to 125, the random forest classifier will consist of 125 decision trees.
- criterion='entropy': The criterion parameter determines the metric used for measuring the quality of a split during the construction of each decision tree. In this case, the 'entropy' criterion is chosen, which calculates the information gain to evaluate the purity of tree nodes.
- random_state=552: This parameter sets the random seed to ensure reproducibility of results. By using a specific value (552), the random forest classifier will produce the same outcomes if the code is run multiple times with the same seed.
- class_weight= {0.0:0.87, 1.0:0.1}: Assigns class weights to handle imbalanced data, giving higher importance to the minority class (‘NOT CKD’) during training.
- max_samples=75: This parameter determines the maximum number of samples randomly selected for training each decision tree in the random forest. By setting it to 75, each tree will be trained on a maximum of 75 samples.
- ccp_alpha=0.12: The ccp_alpha parameter controls the complexity of decision trees through post-pruning. It sets the complexity parameter used for pruning the trees based on the Cost-Complexity Pruning (CCP) criterion. A higher value (0.12) leads to more aggressive pruning, balancing the tree's complexity and generalization ability.

Step 4: Creating an ensemble model using voting classifier.

In the final step, the implemented soft voting classifier combines the predictions of logistic regression and random forest classifiers for chronic kidney disease (CKD) prediction. Soft voting involves aggregating the predicted probabilities from each classifier and selecting the class with the highest average probability.

In soft voting, the predicted probabilities from both classifiers are averaged. This averaging process allows for combining the strengths of logistic regression and random forest, leveraging their diverse predictive capabilities. The class label with the highest average probability is chosen as the final prediction for each instance.

The combined classifier captures both linear and nonlinear relationships, improving the model's robustness. The integrated model achieved a satisfactory accuracy.

Algorithm:

- A list of models is created to be included in the voting ensemble, without specifying their names.
- The voting classifier object is instantiated using the model list, with the voting method set to "soft".
- Soft voting is employed, which combines the predicted probabilities from multiple models.
- The ensemble model aggregates the predicted probabilities and selects the class with the highest average probability as the final prediction.
- By leveraging the strengths of diverse models, the ensemble model offers improved performance in detecting chronic kidney disease.

V. RESULTS AND DISCUSSION

The experimentation is done on CKD dataset.

DATASET:

CKD: The Chronic Kidney Disease (CKD) dataset was collected from a hospital and donated to the UCI Machine Learning Repository by Soundarapandian et al. on July 3, 2015. The dataset contains 400 samples, each with 24 predictive variables or features (11 numerical variables and 13 nominal variables) and a categorical response variable (class). Each class has two values: CKD (sample with CKD) and not CKD (sample without CKD). In the 400 samples, 250 samples belong to the category of CKD, and 150 samples belong to the category of not CKD. It is worth mentioning that there is a large number of missing values in the data.

Table 1: Details of each variable in original CKD data set.

Variable	Full Form	Class	Range/Unit	Missing Percentage
age	Age	Numerical	age in years (1-100)	2.25%
bp	Blood Pressure	Numerical	in mm/Hg	3%
sg	Specific Gravity	Nominal	(1.005,1.010,1.015,1.020, 1.025)	11.75%
al	Albumin	Nominal	(0,1,2.3.4.5)	11.5%
su	Sugar	Nominal	(0,1,2.3.4.5)	12.25%
rbc	Red Blood Cells	Nominal	(normal, abnormal)	38%
pc	Pus Cell	Nominal	(normal, abnormal)	16.25%
pcc	Pus Cell clumps	Nominal	(present, notpresent)	1%
ba	Bacteria	Nominal	(present, notpresent)	1%
bgr	Blood Glucose Random	Numerical	in mgs/dl	11%
bu	Blood Urea	Numerical	in mgsdl	4.75%
sc	Serum Creatinine	Numerical	in mgs/dl	4.25%
sod	Sodium	Numerical	in mEq/l	21.75%
pot	Potassium	Numerical	in mEq/l	22%
hemo	Hemoglobin	Numerical	in gms	13%
pcv	Packed Cell Volume	Numerical	-	17.75%
wbcc	White Blood Cell Count	Numerical	in cell/cumm	26.5%
rbcc	Red Blood Cell Count	Numerical	in millions/cmm	32.75%

htn	Hypertension	Nominal	(yes, no)	0.5%
dm	Diabetes Mellitus	Nominal	(yes, no)	0.5%
cad	Coronary Artery Disease	Nominal	(yes, no)	0.5%
appet	appet	Nominal	(good, poor)	0.25%
pe	Pedal Edema	Nominal	(yes,no)	0.25%
ane	Anemia	Nominal	(yes,no)	0.25%
class	Class	Nominal	(ckd, notckd)	0%

ANALYSIS:

PATIENT_1: The table presents a comparison of the effective accuracy of three machine learning algorithms in detecting chronic kidney disease (CKD) for Patient_1. The algorithms include Logistic Regression, Random Forest Classifier and a Voting Classifier Ensemble Model that combines both algorithms. Each row of the table represents a single algorithm and the columns provide information on the algorithm's name, accuracy, predicted result (CKD or no CKD) and the corresponding probability of the patient having or not having CKD.

The first algorithm logistic regression achieved an accuracy of 92.18% and predicted that Patient_1 has CKD with a high probability close to 1, while the probability of not having CKD was low, at 3.86 × 10−8.

The second algorithm random forest achieved an accuracy of 96.87% and predicted that Patient_1 has CKD with a probability of 0.53 and a probability of not having CKD of 0.47.

The third algorithm voting classifier ensemble model achieved an accuracy of 95.31% and predicted that Patient_1 has CKD with a probability of 0.77 and a probability of not having CKD of 0.23.

Algorithm	**Accuracy**	**Result**	**Probability of CKD**	**Probability of not CKD**
Logistic Regression	92.18%	CKD	99×10^{-1}	3.86×10^{-8}
Random Forest	96.87%	CKD	0.53	0.47
Ensemble	95.31%	CKD	0.77	0.23

PATIENT_2: The table presents a comparison of the effective accuracy of three machine learning algorithms in detecting chronic kidney disease (CKD) for Patient_2. The algorithms include Logistic Regression, Random Forest Classifier, and a Voting Classifier Ensemble Model that combines both algorithms. Each row of the table represents a single algorithm, and the columns provide information on the algorithm's name, accuracy, predicted result (CKD or no CKD), and the corresponding probability of the patient having or not having CKD.

The first algorithm logistic regression achieved an accuracy of 92.18% and predicted that Patient_2 has Not CKD with a high probability close to 1 , while the probability of having CKD was low, at 4.72 × 10−8.

The second algorithm random forest achieved an accuracy of 96.87% and predicted that Patient_2 has CKD with a probability of 0.48 and a probability of not having CKD of 0.52.

The third algorithm voting classifier ensemble model achieved an accuracy of 95.31% and predicted that Patient_2 has CKD with a probability of 0.25, and a probability of not having CKD of 0.75.

Algorithm	**Accuracy**	**Result**	**Probability of CKD**	**Probability of not CKD**
Logistic Regression	92.18%	Not CKD	4.72×10^{-11}	1
Random Forest	96.87%	Not CKD	0.48	0.52
Ensemble	95.31%	Not CKD	0.25	0.75

VI. CONCLUSION

In summary, Chronic Kidney Disease (CKD) poses a significant global health challenge. To address this issue, a system has been designed for early-stage detection of CKD using machine learning models. The system operates by collecting health data from patients' medical reports through a dedicated website. This data is then utilized to build a Machine Learning Model using Logistic Regression and Random Forest algorithms. The model effectively determines whether a patient is suffering from CKD or not. By applying this integrated machine learning methodology to the practical diagnosis of CKD, a desirable outcome is achieved. The project has been successfully completed, attaining an accuracy rate of over 95%.

REFERENCES

[1] Jiongming Qin, Lin Chen, Yuhua Liu, Chuanjun Liu, Changhao Feng, and Bin Chen, A Machine Learning Methodology for Diagnosing Chronic Kidney Disease, IEEE Access, vol. 8, pp. 13, Feb. 2020.

[2] N. Chumuang et al., “An Efficiency Random Forest Algorithm for Classification of Patients with Kidney Dysfunction,” 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), pp. 1-6, Nov 2020.

[3] H. Luo, X. Pan, Q. Wang, S. Ye and Y. Qian, “Logistic Regression and Random Forest for Effective Imbalanced Classification,” 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), pp. 916-917 July 2019.

[4] V. Shanmugarajeshwari and M. Ilayaraja, “Chronic Kidney Disease for Collaborative Healthcare Data Analytics using Random Forest Classification Algorithms,” 2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1- 14 April 2021.

[5] N. V. Ganapathi Raju, K. Prasanna Lakshmi, K. G. Praharshitha and C. Likhitha, “Prediction of chronic kidney disease (CKD) using Data Science,” 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 642-647, doi: 10.1109/ICCS45141.2019.9065309

[6] S. Vashisth, I. Dhall and S. Saraswat,” Chronic Kidney Disease (CKD) Diagnosis using Multi-Layer Perceptron Classifier,” 2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2020, pp. 346-350, doi: 10.1109/Confluence47617.2020.9058178.

[7] G. Nandhini and J. Aravinth, “Chronic kidney disease prediction using machine learning techniques,” 2021 International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT), 2021, pp. 227-232, doi: 10.1109/RTEICT52294.2021.9573971.

[8] H. Zhang, Z. Li, H. Shahriar, L. Tao, P. Bhattacharya and Y. Qian, “Improving Prediction Accuracy for Logistic Regression on Imbalanced Datasets,” 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, pp. 918-919, doi: 10.1109/COMPSAC.2019.00140.

[9] A. B. Varol Malkoolu and. Utku Malkooglu, “Comparative Performance Analysis of Random Forest and Logistic Regression Algorithms,” 2020 5th International Conference on Computer Science and Engineering (UBMK), 2020, pp. 25-30, doi: 10.1109/UBMK50275.2020.9219478.

[10]D. Sharathchandra and M. R. Ram, “ML Based Interactive Disease Prediction Model,” 2022 IEEE Delhi Section Conference (DELCON), 2022, pp. 1-5, doi: 10.1109/DELCON54057.2022.9752947.