INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

CHRONIC KIDNEY DISEASE PREDICTION USING MACHINELEARNING ALGORITHMS

V SRIKANTH
MCA, MTECH, MBA AND PGDBM

ABSTRACT:

One of the biggest issues with death rate in medical field today is chronic kidney disease (CKD), a slowly progressing illness that is sometimes detected too late. The prevalence of Chronic Kidney Disease (CKD) is 14% worldwide., causing gradual kidney function loss. About 2 million receive dialysis or transplants, representing only 10% of those in need. CKD arises from factors like diabetes and hypertension. Enhanced awareness and access to healthcare are vital to address this global health issue. Imesh Udara Ekanayake Et al., predicted CKD using SVM and KNN algorithms with 100% accuracy with a data set of 400 instances and 25 attributes. Therefore, we utilize ensemble machine learning methods to predict CKD using a dataset containing 201 records and 29 attributes. The ensemble models effectively capture the diverse patterns and interactions within the data, which provides a tool for early detection and personalized management of CKD. This highlights how effective ensemble learning may be in raising the standard of medical judgment when it comes to chronic renal disease. We evaluate the prediction accuracies of a number of methods, such as Random Forest, Decision Tree, SVM, and AdaBoost. Furthermore, we pinpoint crucial features in the dataset that are essential for CKD prediction..

Key words: Random Forest, SVM, Decision tree, Adaboost, chronic kidney disease(CKD).

INTRODUCTION:

Chronic kidney disease, also known as chronic kidney failure, causes a progressive decline in kidney function. The kidneys filter waste and excess fluid from the blood, which is then expelled from the body as urine. 10% of people worldwide have CKD, with 17.2% of adult Indians affected by the condition. As chronic kidney disease (CKD) worsens, the kidneys' ability to eliminate waste and excess fluid from the bloodstream diminishes. In the early stages, this deterioration occurs gradually and without noticeable symptoms in the initial phases. The inability of the kidneys to perform their regular functions, which include blood filtering, is known as chronic kidney disease (CKD). The progressive loss of kidney cells over a long period of time is called "chronic." Managing chronic kidney disease involves primarily slowing down its progression. However, halting the cause might not always prevent

ISSN: 2057-5688

further kidney damage. CKD typically develops slowly and may go unnoticed in its early stages, often asymptomatic. As the disease progresses, however, symptoms such as fatigue, swelling, changes in urination patterns, and high blood pressure may become apparent. Depending on the degree of kidney damage, CKD is usually divided into stages, with the end stage requiring either dialysis or a kidney transplant to maintain life-sustaining functions. If left untreated, CKD can advance to end-stage kidney failure, necessitating dialysis or a transplant. While symptoms like stomach discomfort, vomiting, muscle cramps, changes in urination patterns, and others typically appear in later stages, they might also emerge due to other conditions. Rapid disease prediction could save many lives globally before irreversible damage occurs. Detecting this growing condition can be aided by a training model using some algorithms on medical patient data. The challenge lies in achieving precise predictions promptly. The primary parameters that are included to predict CKD are red blood cell count, white blood cell count, red blood cells, pus cell, sodium, potassium, packed cell volume, age, and gender.

One of the primary concerns associated with CKD is the development of high blood pressure, or hypertension. This condition can exacerbate the progression of CKD, creating a detrimental cycle wherein impaired kidney function leads to elevated blood pressure and vice versa.CKD often disrupts the delicate balance of electrolytes in the body. Electrolytes, including sodium, potassium, calcium, and others, play pivotal roles in various physiological functions. The impaired regulation of these essential minerals can lead to a range of health issues, from muscle weakness to cardiac arrhythmias.As CKD progresses, individuals may face an increased likelihood of developing kidney stones, painful mineral and crystal formations that can obstruct the urinary tract and cause considerable discomfort. Moreover, metabolic acidosis may occur as the kidneys' ability to regulate the body's acid-base balance diminishes. This condition might result in a number of symptoms, such as exhaustion, confusion, as well as respiratory distress. With advanced stages of CKD, a condition known as uremia may arise. This occurs when waste products, normally filtered out by the kidneys, accumulate in the body. Symptoms of uremia can be severe and encompass nausea, vomiting, loss of appetite, and cognitive impairment. For individuals with advanced CKD, dialysis or kidney transplantation may be necessary to sustain life. While these treatments offer a lifeline, they are not without their own set of potential complications, underscoring the need for comprehensive, ongoing care.

Imesh Udara Ekanayake Et al., predicted CKD using SVM and KNN algorithms with 100% accuracy with a data set of 400 instances and 25 attributes[1]. Therefore, we aim to utilize the Decision tree, Random forest, SVM, and AdaBoost algorithms which are ensemble learning algorithms with a data set comprising 201 instances and 29 attributes in order to enhance the best accuracy and also we identify

ISSN: 2057-5688

key attributes within the dataset that play a pivotal role in predicting CKD.

ISSN: 2057-5688

LITERATURE SURVEY

Imesh Udara Ekanayake researched a methodology that included attribute selection, collaborative filtering for missing values, and data preprocessing. The most accurate and least biased of the 11 machine learning techniques are provided by random forest and extra tree classifiers. The study stresses the importance of domain expertise in improving CKD prediction and stresses the practicality of data collection.[1]

JIONGMING QIN examined the application of KNN imputation to address missing values in a University of California dataset., Irvine to use machine learning for CKD diagnosis. Out of six machine learning approaches that were applied, random forest produced best accuracy, coming in at 99.75%. Additionally, an integrated model with an average simulation accuracy of 99.83% is proposed by the study, combining logistic regression and random forest together. The use of complex clinical data in illness detection using this paradigm appears promising [2].

Reshma S support using machine learning techniques, specifically the Support Vector Machine (SVM) classifier and Ant Colony Optimisation (ACO) method, to predict Chronic Kidney Disease (CKD). The main goal is to use as few information as possible to determine if a person has CKD while maximizing diagnostic accuracy and resource efficiency.[3]

Pronab Ghosh used a UCI machine learning dataset to test four different methods (SVM, AB, LDA, and GB). Gradient Boosting (GB) achieved an astounding 99.80% accuracy. The analysis offers a variety of performance measures to help choose the best algorithms for CKD prediction. In the medical field, it tackles the crucial problem of delayed diagnosis.[4]

Surya Krishnamurthy studied the healthcare challenges posed by Chronic Kidney Disease by developing machine-learning algorithm for predicting CKD in first 6 or 12 months. It employs Convolutional Neural Networks with high AUROC values to examine medication and comorbidity information from Taiwan's National Health Insurance Research Database. Diabetes, age, gout, and specific medications all emerged as significant predictors. To combat the healthcare impact of CKD, the algorithm serves as a valuable tool for policymakers, facilitating CKD trend prediction, proactive monitoring, early detection, efficient resource allocation, and patient-centric management.[5]

Md. Ariful Islam investigates different machine learning approaches for early CKD diagnosis,

ISSN: 2057-5688

employing predictive modeling to connect data factors and target class characteristics. It reduces the initial 25 variables to a 30% subset for CKD detection. XgBoost outperforms the other machine learning-based classifiers regarding F1-score, recall, accuracy, and precision. This research emphasizes the promise of recent machine learning advancements in improving prediction correctness in kidney disease as well as beyond.[6]

METHODOLOGY

The proposed system considered a data set comprising 201 records and 29 attributes, the Chronic Kidney Disease-related dataset (CKD). Data preprocessing was performed to handle missing values. The workflow includes removing outliers from the dataset. Employing statistical analysis to identify the most significant attributes. Applying Principal Component Analysis (PCA) to reduce the dimensionality by eliminating attributes with high inter-correlation. Evaluating classification algorithms - Random Forests, Decision trees, Support Vector Machines (SVM), and AdaBoost. Utilizing measures such the F1-score, recall, accuracy, and precision for comprehensive performance evaluation. Leveraging scikit-learn, a versatile machine learning library in Python, for its comprehensive implementation of SVM, Decision Trees, Random Forests, and AdaBoost.

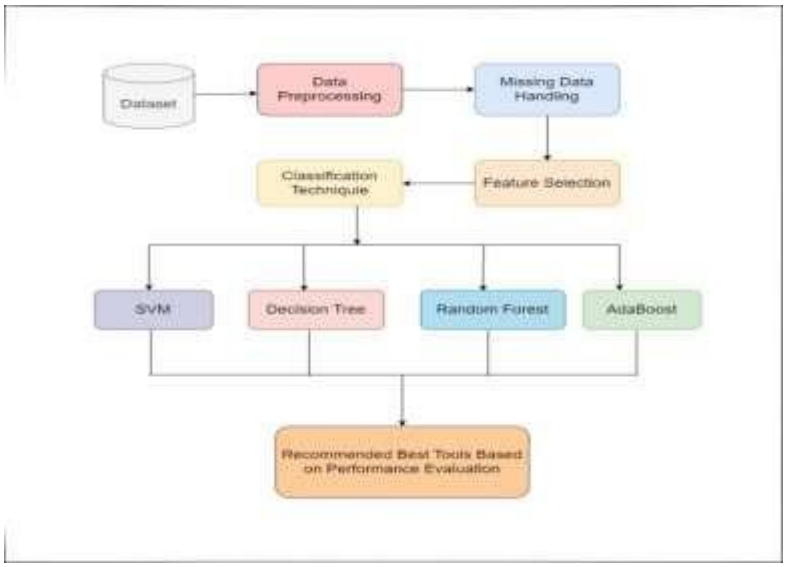


Fig 1:Proposed Methodology

ISSN: 2057-5688

Data Preprocessing:

The proposed system methodology is done by Data preprocessing i.e handling missing values, Model training and Model selection Data preprocessing was performed in two steps to handle missing values.Initially, missing values within the dataset was thoroughly identified, and those features were excluded from the analysis, ('pe,' 'stage,' 'ane,' 'grf,' 'age,' and 'affected,'). This can be done by removing these columns from the dataset. Managing missing values in the remaining data is the second phase in the preparation process of data.. After excluding the characteristics lacking values, the dataset is left with the remaining attributes with no missing values. Principal Component Analysis (PCA) is one method for addressing and handling missing values during the data preprocessing stage. PCA assists in lowering the dataset's dimensionality while keeping crucial information. Any missing values that remain after PCA can be filled in by imputation methods or by just removing the relevant rows. It's crucial to remember that eliminating entire columns following PCA should only be done sparingly because they can include crucial data for further investigation. Dropping columns should only be done after carefully evaluating the particular dataset, the type of missing values.

Attributes	MissingValues	Percentage
bp(Diastolic)	1	0.49505
bp limit	1	0.49505
stage	1	0.49505
grf	1	0.49505
ane	1	0.49505
pe	1	0.49505
appet	1	0.49505
cad	1	0.49505
dm	1	0.49505
htn	1	0.49505
wbcc	1	0.49505
rbcc	1	0.49505
pcv	1	0.49505

ISSN: 2057-5688

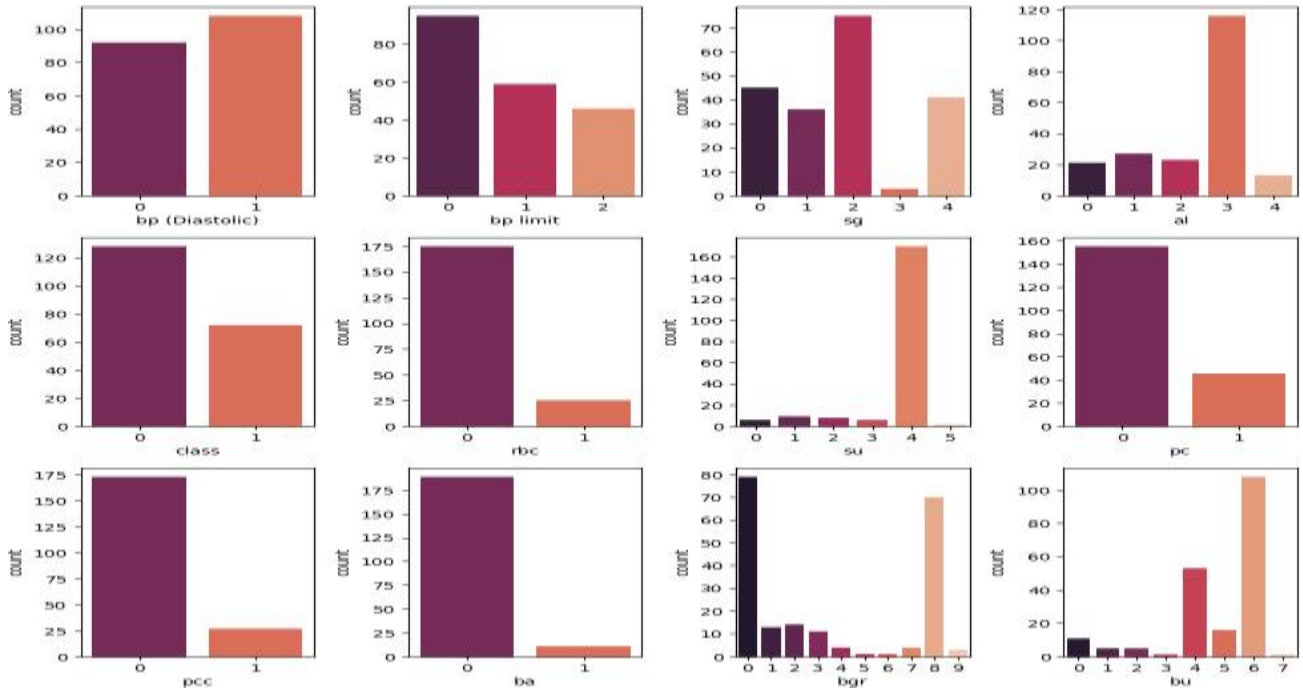
hemo	1	0.49505
pot	1	0.49505
sc	1	0.49505
sod	1	0.49505
bu	1	0.49505
bgr	1	0.49505
ba	1	0.49505
pcc	1	0.49505
pc	1	0.49505
su	1	0.49505
rbc	1	0.49505
class	1	0.49505
al	1	0.49505
sg	1	0.49505
affected	0	0.00000
age	0	0.00000

Table:1 Missing Values in CKD dataset

Exploratory Data Analysis:

A crucial stage of data analysis is exploratory data analysis (EDA), which is systematically going over datasets to find patterns, trends, and insights. EDA guides further analytical decisions by providing a basic understanding of the data distribution, linkages, and potential outliers through statistical and visual methodologies. It is essential for feature engineering, model selection, and overall interpretation of results in many fields,This visualization allows healthcare professionals to quickly identify patterns in key indicators like creatinine levels,aiding in the early detection and management of CKD progression.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688



The categorical count plot, provides a graphical depiction illustrating the distribution categorical variables in the data. Each subplot displays the count of occurrences for a specific categorical attribute. It aids in exploring the frequency of different categories, identifying imbalances, and understanding the overall distribution of categorical data. This visualization is crucial in Exploratory Data Analysis (EDA) to gain insights into the composition and prevalence of categorical features, facilitating a comprehensive understanding of the dataset's categorical characteristics.

Data Analysis: Correlations are analyzed using heat maps, guiding the removal of features with weak correlations. This optimizes thedataset for meaningful insights and enhances the overall model performance.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

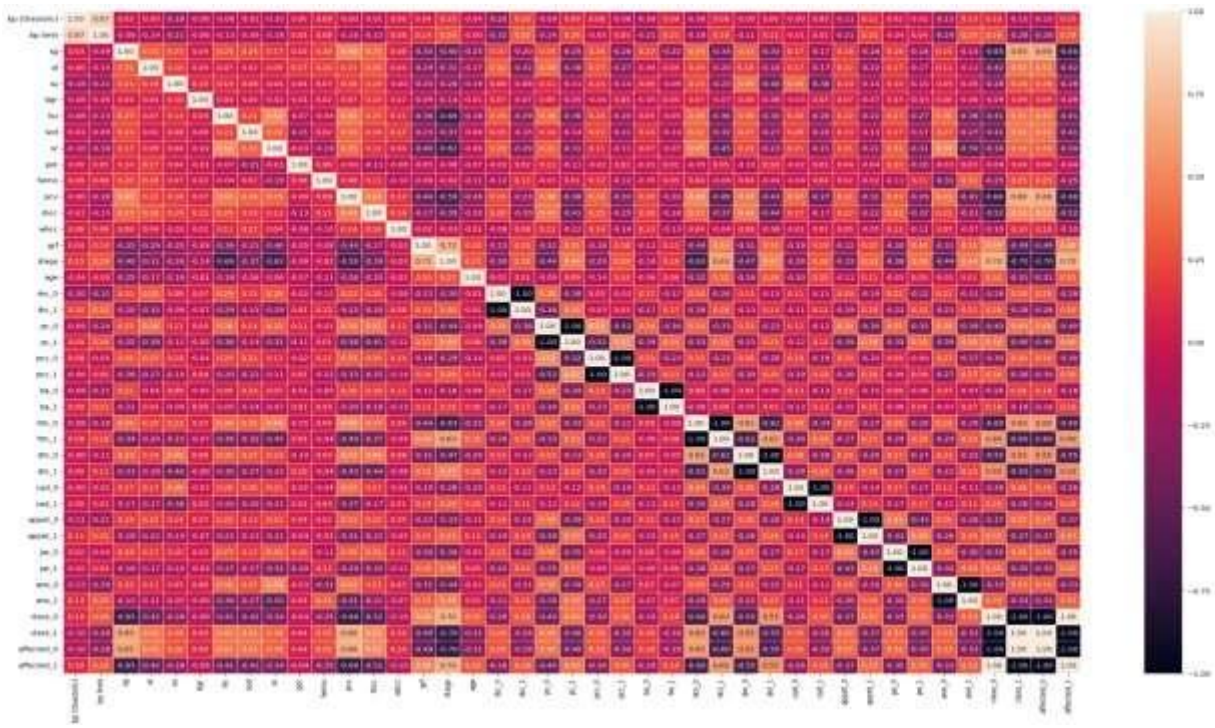
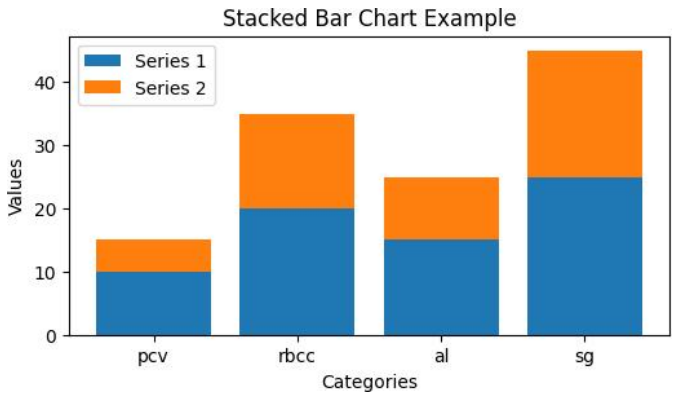


Fig 2: Correlation analysis using heatmap

ISSN: 2057-5688

A heat map can be employed to visually represent patient data, such as laboratory values over time. This visualization allows healthcare professionals to quickly identify patterns in key indicators like creatinine levels, aiding in the early detection and management of CKD progression.In Figure 2, a correlation analysis was performed to assess the relationships between different features



A stacked bar chart, in Exploratory Data Analysis (EDA) to visually compare the composition of different categories across two series. It
illustrates the contribution of each category to the total value while highlighting the relative proportions of the individual series. This graph aids in identifying patterns, trends, or disparities in the distribution of values within each category, offering a quick and informative overview of the dataset's structure.The columns denotes different categories, while the rows represents numerical values. This visualization allows for a quick comparison of the total values for each category and an assessment of the contribution of each series to the overall distribution.

Data Splitting:

A 70-30 split was used to carefully divide the dataset into training and testing halves, which is an essential stage in assessing machine learning algorithms. In this partitioning, 70% of the data was allocated to the training set, providing a foundation for algorithms to discover patterns and relationships in the dataset. The remaining 30% constituted the testing set, serving as an as a separate standard by which to evaluate the model's capacity to generalize to fresh, untested data. This careful division secures a thorough evaluation of the model's predictive performance and gauges its effectiveness in real-world scenarios.

ISSN: 2057-5688

Model Building:

Model building involved training the training dataset is used by a number of machine learning methodologies, including Random Forests, Decision trees, SVM, AdaBoost to generate prediction models. Models were then evaluated on the testing dataset to assess their accuracy and performance.

SVM: Support Vector Machines (SVMs) are supervised machine learning techniques used to solve regression and classification problems. In a high-dimensional space, it is useful for locating a hyperplane that optimally divides data points into different classes and maximizes the margin between them. SVM is efficient at managing interactions that are both linear and non-linear in data.

Decision Tree: Regression and classification tasks are both handled by the flexible machine-learning technique known as decision trees. It creates a structure resembling a tree by recursively dividing the dataset into subsets according to the most important attribute at every node. Decision trees manage non-linear relationships well and are simple to interpret, and are fundamental components in ensemble methods like Random Forests.

Random Forest: During training, the Random Forest ensemble learning technique creates a large number of decision trees. It improves accuracy and mitigates overfitting by combining predictions from multiple trees. The algorithm's key strengths lie in its robustness, capability to handle high-dimensional data, and effective feature importance estimation, making it widely used in diverse fields such as finance, healthcare, and remote sensing.

Adaboost: AdaBoost, also known as adaptive boosting, is an ensemble learning method that combines weak learners to create a robust model. It iteratively adjusts the weights of misclassified samples, emphasizing difficult-to-classify instances. AdaBoost is particularly effective in improving accuracy, and its versatility extends to various domains, such as face detection and text categorization, making it a widely utilized algorithm in machine learning.

Performance Metrics:

Accuracy

Training accuracy and testing accuracy are crucial metrics in evaluating the performance of machine learning models.

ISSN: 2057-5688

Training accuracy measures how well a model performs on the data it was trained on, providing insights into its ability to capture patterns. Testing accuracy, on the other hand, assesses the model's performance on new, unseen data, indicating its ability to generalize beyond the training set for real-world predictions.

ISSN: 2057-5688

Table 2: Result of Accuracies

ALGORITHM	TRAINING SET	TESTING SET
Decision Tree	0.97	0.96
Support Vector Machine	0.99	0.98
Random Forest	1.0	0.98
Ada Boost	1.0	1.0

Precision: Precision, described as the proportion of all predicted positives to actual positive predictions, is a metric used to quantify how accurate positive predictions made by a model are. The statement highlights the significance of limiting false positives in certain contexts and highlights the model's capacity to prevent erroneous productive classifications.

Recall: Sensitivity, also known as true positive rate, gauges a model's capacity to identify and capture every relevant instance of a positive class. The model's ability to prevent false negatives and accurately identify the majority of positive situations was demonstrated by calculating the ratio of true positive predictions to all real positives.

Jaccard Score: By calculating the ratio of an element's intersection to its union, a Jaccard score is calculated to determine how similar two sets are to one another**.** In machine learning, it is commonly used to evaluate the accuracy of classification models, especially in scenarios where imbalanced class distribution exists.

Log Loss: Log Loss, or logarithmic loss, evaluates the accuracy of a probabilistic classifier's predicted probabilities. It penalizes predictions that are less confident and farther from the true labels. The goal is to minimize log loss for optimal performance.

Table 3: Performance Metrics

	Model	Accuracy	F1 Score	Precision	Recall	Jaccard Score	Log Loss
0	Decision	96.67	00.96667	00.96667	00.96667	0.846154	3.00364

ISSN: 2057-5688

	Tree						
1	Random Forest	98.33	0.983333	0.983333	0.983333	0.967213	0.600728
2	SVM	98.33	0.983333	0.983333	0.983333	0.967213	0.600728
3	AdaBoost	100	1	1	1	1	2.22045e-16

Results:

The most effective production algorithm for early CKD prediction was offered by this system. The input parameters gathered from CKD patients are displayed in the dataset, and models are trained using these parameters to identify the optimal algorithm for CKD prediction. To diagnose CKD, learning models like Support Vector Machine (SVM), Random Forest, and Decision Tree, and Adaboost classifiers are built. Model performance is measured using a range of comparison metrics, one of which is accuracy. The study's findings demonstrated the superiority of the Adaboost classifier over the other models when all of the matrix was taken into account.

Model	Accuracy
Decision Tree	0.9666666666666667
Random Forest	0.9833333333333333
SVM	0.9833333333333333
AdaBoost	1.0

CONCLUSION

Utilising machine learning models like AdaBoost, Random Forest, Decision Trees, and Support Vector Machines (SVM), the study aimed to predict ChronicKidney Disease (CKD). AdaBoost emerged with the highest accuracy, showcasing its potential for early CKD detection and efficient healthcare resource allocation. Feature importance analysis identified critical attributes influencing CKD prediction, aiding medical professionals in understanding key risk factors. AdaBoost superior accuracy positions it as the most effective model, outperforming others, including Random Forest. However, the 'best' algorithm choice depends on various factors, such as dataset characteristics and analysis objectives. Crucial features influencing CKD prediction, such as pcv, hemo, rbcc, sg, and al, were identified through feature importance exploration.

ISSN: 2057-5688

The top five features contributing to Chronic Kidney Disease (CKD) prediction based on all algorithms

Table 4: Key Features Conclusion

S. No	Feature	Mean Importance
1	pcv	0.303529
2	rbcc	0.096930
3	hemo	0.090611
4	sg	0.086413
5	al	0.086323

FUTURE ENHANCEMENT

1. Real-time Monitoring System: Building a user-friendly website or application that integrates these models can facilitate continuous health monitoring. This allows for timely intervention and personalized healthcare recommendations, enhancing patient outcomes and overall healthcare delivery.
2. Ensemble Model Integration: Exploring ensemble techniques to blend predictions from multiple models may yield superior results, providing a robust and well-rounded solution for CKD prediction.

REFERENCES:

[1]Baisakhi Chakraborty, "Development of Chronic Kidney Disease Prediction Using Machine Learning", International Conference on Intelligent Data Communication Technologies, 2019.

[2]J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Translational Medicine, vol. 17, no. 1, p. 119, 2019.

ISSN: 2057-5688

[3]A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in 2018 IEEE Congress on Evolutionary Computation (CEC).
IEEE, 2018, pp. 1–9

[4] Aljaaf AJ, Al-Jumeily D, Haglan HM, et al. Early prediction of chronic kidney disease using machine learning supported by predictive analytics. 2018 IEEE Congress on Evolutionary Computation (CEC). IEEE; 2018. p. 1–9. 3.

[5]Almasoud M, Ward TE. Detection of chronic kidney disease using machine learning algorithms with the least number of predictors. IntJ Soft Comput Appl 2019;10.

[6]J.A. Vassalotti, et al., Practical Approach to Detection and Management of Chronic Kidney Disease for the Primary Care Clinician, The American Journal of Medicine, vol. 129, no. 2, 2016.

[6]Snegha, "Chronic Kidney Disease Prediction using Data Mining", International Conference on Emerging Trends, 2020.

[7]Baisakhi Chakraborty, "Development of Chronic Kidney Disease Prediction Using Machine Learning", International Conference on Intelligent Data Communication Technologies, 2019

[8]Siddeshwar Tekale, "Prediction of Chronic Kidney Disease Using Machine Learning, International Journal of Advanced Research in Computer and Communication Engineering, 2018

[9]Nishanth and T. Thiruvaran, "Identifying Important Attributes for Early Detection of Chronic Kidney Disease," in IEEE Reviews in Biomedical Engineering, vol. 11, pp. 208-216, 2018.

[10] M. J. M. Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi, S. Shultana "Implementation of Machine Learning Algorithms to Detec t the Prognosis Rate of Kidney Disease" 2020 IEEE International Conference for Innovation in Technology, 2020.

[11]Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in 2017 IEEE 17th International Conference on Bioinformatics and

ISSN: 2057-5688

Bioengineering (BIBE), pp. 291–296, Washington, DC, USA, 2017.

[12]S. Krishnamurthy, K. KS, E. Dovgan, et al., “Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan,” Healthcare, vol. 9, no. 5, p. 546, 2021.