Chronic Kidney Disease Identification using Random Forest and XGBoost

M.S. Abirami
Dept. of Computational Intelligence
SRM Institute of Science and Technology
Chennai, India
Corresponding author :
abiramim@srmist.edu.in

Sayak Das
Dept. of Computational Intelligence
SRM Institute of Science and Technology
Chennai, India
sd8675@srmist.edu.in

Roopal Sood
Dept. of Computational Intelligence
SRM Institute of Science and Technology
Chennai, India
rs2897@srmist.edu.in

***Abstract*— Chronic Kidney Disease (CKD) is a serious and long term condition stemming from either renal disease or flawed kidney functions. Kidney cancer, known for its lethality, holds paramount importance in patient survival, necessitating early diagnosis and accurate classification. Timely intervention and suitable therapy can impede or postpone the progression of this chronic ailment to its advanced stages, where life-saving measures like dialysis or renal transplantation become imperative. The pressing challenge in current research revolves around the development of automated tools proficient in precisely identifying Chronic kidney Disease . This paper introduces the application of XGBoost and Random Forest algorithm which efficiently and effectively detect Chronic kidney disease, offering the potential to mitigate its progression and improve patient prognosis. The efficacy of classification technology is contingent upon the quality of the dataset.**

Keywords— Chronic Kidney Disease (CKD), XGBoost, KNN, Random Forest, Scikit learn, Decision Tree Classifier, AI, Confusion Matrix

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a persistent and serious health condition that arises from either renal disease or impaired kidney functions. It poses a significant threat to public health, with its severity escalating if not detected and addressed early. Timely identification and accurate prediction of CKD are vital for implementing effective interventions that can slow or halt its progression, preventing the necessity of life-saving measures like dialysis or renal transplantation.

The artificial intelligence (AI) topic of deep learning has attracted a lot of interest lately because of its outstanding performance in a number of areas, such as autonomous systems, natural language processing, and picture recognition. Its ability to process big information and identify complex patterns, and model complex relationships has ignited interest in applying deep learning techniques to the critical task .

In the current landscape of medical research, the focus on kidney-related ailments, particularly kidney cancer, has intensified due to its lethal nature. Early diagnosis and classification of kidney diseases, including cancer, are crucial for enhancing patient survival rates. The advent of technology and the growing importance of the Internet of Medical Things (IoMT) present an opportunity to revolutionize healthcare through the integration of advanced machine learning (ML) models. This research work aims to leverage cutting-edge technology, specifically the XGBoost and Random Forest models, for predicting Chronic Kidney Disease. The IoMT platform serves as the foundation for this endeavor, allowing seamless integration of medical data and facilitating real-time monitoring of patients' health status.

To enhance the model's generalization and robustness, advanced machine learning techniques are explored. Models are fine-tuned on the dataset to leverage their feature extraction capabilities. Advanced machine learning models helps the models adapt to the specific characteristics of the forest fire prediction task, improving their performance with limited labeled data. Through the convergence of innovative technologies, this research work aspires to make significant strides in the prediction of Chronic Kidney Disease, offering a holistic future work that integrates IoMT and advanced ML models. The ultimate goal is to contribute to the advancement of medical science and improve patient outcomes in the realm of kidney health.

II. LITERATURE SURVEY

The provided journal paper discusses the challenges associated with the early identification of kidney cancer, emphasizing its lethality and the limitations of conventional clinical methods. Despite being a leading cause of cancer-related deaths, renal cancer research is considered insufficient in the current research landscape, often overshadowed by other types of cancer. The paper highlights the need for automated diagnostic tools to facilitate quick and accurate identification of kidney diseases, contributing to improved patient survival.

Studies like Guozhen chen et al. [1] addresses the critical need for early detection and classification of kidney cancer, which is crucial for patient survival and management of chronic kidney disease (CKD). It introduces an Adaptive Hybridized Deep Convolutional Neural Network (AHDCNN) aimed at efficiently and effectively identifying kidney disease subtypes. The paper emphasizes the importance of accurate classification methods and the role of data sets in enhancing classification system accuracy.

The proposed AHDCNN model utilizes deep learning techniques to extract features from CT images for renal cancer detection. It integrates CNN features with a support vector machine and employs the utilization of fully convolutional networks and conditional random fields (CRFs) in kidney cancer segmentation shows promising outcomes in the early detection and diagnosis of Chronic Kidney Disease (CKD), as demonstrated through experimental procedures conducted on the Internet of Medical Things platform.

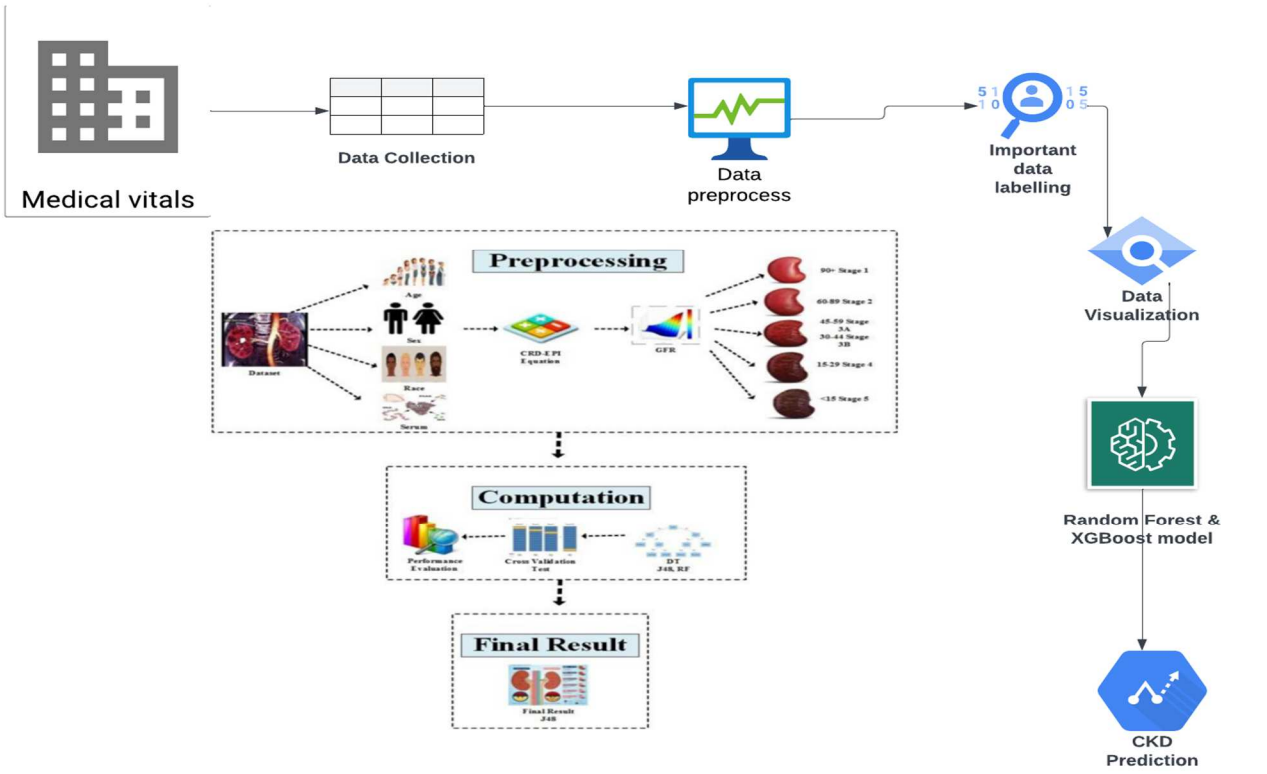

Fig 1: Chronic Kidney Disease Identification

Additionally, the paper reviews related works in the field of kidney cancer prediction, highlighting various machine learning approaches such as Neighborhood Component Analysis (NCA), Deep Neural Network (DNN), Hybrid Neural Network (HNN), and Recurrent Neural Network (RNN).

In summary, the research proposes an innovative approach using deep learning techniques for the prediction and diagnosis of chronic kidney disease. It demonstrates the potential of machine learning algorithms in improving early detection and management of kidney cancer, leveraging advancements in medical technology and data analytics.

Mithila rani et.al presented The research paper explores the prevalence and detection of Chronic Kidney Disease (CKD) using data mining techniques, specifically focusing on the Boruta algorithm.The Boruta algorithm, applied in this study, aids in identifying significant factors associated with CKD prediction. It works by creating shadow attributes and training a random forest classifier to assess attribute importance. Out of 24 attributes, only 7 were confirmed to be important in predicting CKD. The analysis shows that reducing the number of features can lead to a slightly lower accuracy of 99.19%, compared to 100% accuracy with all features, but significantly reduces processing time and memory load.

The research discusses the importance of factors such as hypertension, blood pressure, and specific tests like urine albumin and serum creatinine in detecting CKD. It emphasizes that a combination of tests is necessary for accurate diagnosis, especially in senior patients who are at higher risk. Furthermore, it suggests potential correlations between age and CKD-related factors.Numerical readings include statistics on attribute importance, with variables like sodium, age, packed cell volume, and hemoglobin deemed significant. Sensitivity and specificity of the model are reported as 1 and 0.98 respectively, with an out-of-bag estimate error rate of 1.08%. Confusion matrices illustrate the model's performance, with 47 correctly classified as "notCKD" and 75 as "CKD" in the reduced feature model.

In conclusion, the study underscores the economic burden of CKD and the importance of affordable diagnostic methods. Boruta Analysis emerges as a valuable tool for medical diagnosis, offering cost-effective and faster solutions. The paper suggests avenues for future research, including the application of data mining algorithms in other chronic diseases for early detection and improved patient outcomes.

The Random Forest (RF) model exhibited the most superior predictive performance, boasting an average accuracy of 99.50%, sensitivity of 98.75%, specificity of 100%, precision of 100%, F1 score of 99.35%, and an AUC of 99.38%.Significant features highlighted by SHAP analysis included hemoglobin (hemo), specific gravity (sg), serum creatinine (sc), albumin (al), packed cell volume (pcv), red blood cell count (rbcc), hypertension (htn), blood glucose random (bgr), diabetes mellitus (dm), age, sodium (sod), blood urea (bu), and blood pressure (bp).

To streamline the dataset, a reduced set with 13 selected attributes was generated, leading to the formation of six distinct datasets based on various pathological tests. RF achieved the highest classification accuracy across different datasets, scoring 99.00% with the full dataset (DB-I), 97.75% with blood and other pathological tests (DB-II), and 97.00% with urine test attributes and others (DB-III).- GB and XGB classifiers also showed good performance, with RF generally outperforming LR and SVM.

III. METHODOLOGY

The research work focuses on identifying chronic kidney disease through an advanced methodology blending machine learning models. The methodology commences with comprehensive data collection, cleaning, and augmentation, followed by the division of the dataset into training, validation, and test sets. Feature extraction is performed using XGBoost, Random Forest and other models tailored for kidney disease identification are developed. The model is trained with advanced machine learning techniques, and advanced machine learning models like XGBoost , Forest tree classifier are explored. Ensemble methods and IoT platform experiments contribute to a holistic evaluation, with metrics such as accuracy, precision, recall, and F1 score. Results are analyzed, guiding fine-tuning and optimization for the development of an efficient and accurate model. Comprehensive documentation and reporting encompassing preprocessing, model architectures, and findings conclude the project. This methodology aims to advance early chronic kidney disease detection, emphasizing the ability of machine learning in healthcare.

age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	...	38	6000	NaN	no	no	no	good	no	no	ckd
62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	...	35	7300	4.6	no	no	no	good	no	no	ckd

Table 1: Chronic Kidney Disease Dataset

The dataset which is used for this research work consists of 24 features and 1 label. All features are basically human body vital datas which are directly and indirectly related for the cause of Chronic Kidney Disease. Label is categorized into two classes ckd and notCkd. The code begins by importing necessary libraries for data visualization, including pandas, numpy, matplotlib, seaborn, and plotly. It also suppresses warning messages and sets a specific plotting style. Then, the code mounts Google Drive to access data. After loading the dataset into a DataFrame (assumed as `df`), it converts specific columns to numerical type using `pd.to_numeric` with the `errors='coerce'` parameter to handle any conversion errors gracefully. Categorical and numerical columns are then separated into two lists using list comprehensions. The code iterates over categorical columns to print unique values for each column. This step is useful for understanding the nature and diversity of categorical data in the dataset. Overall, the code sets up the environment, prepares the dataset for analysis, and provides insights into the categorical data's uniqueness, which is crucial for subsequent analysis and visualization tasks.

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2$$

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

The top equation shows the Taylor Series expansion of the function f(x) around a base point a. This is a way of approximating a function with a polynomial centered around a specific input value.

The bottom equation shows part of the XGBoost objective function. In gradient boosting, the goal is to trim down a loss function over a set of training data. The loss function measures how different the predictions of a model are from the actual values. The XGBoost objective function combines a loss function, represented by l(yi,y ^(t−1)) (where i indexes the training sample, yi is the true value for that sample, and y^(t−1) is the prediction from the model at the previous iteration), with a regularization term, represented by Ω(ft). The regularization term penalizes the complexity of the model to prevent overfitting.

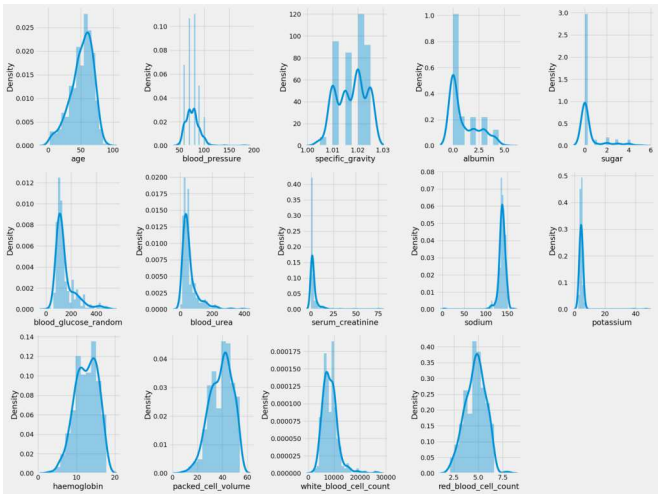

Fig 2: Distribution of all Numerical Features

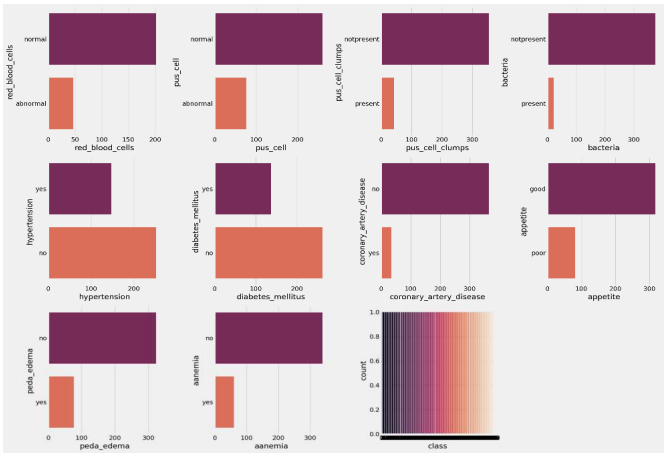

Fig 3: Distribution of all categorical Features

The code first creates subplots for visualizing the distribution of numerical and categorical columns separately. For numerical columns, it iterates over each column, plotting a distribution plot using Seaborn's `sns.distplot`. Similarly, for categorical columns, it plots a count plot using `sns.countplot`. The plots are arranged in a grid structure to fit the figure appropriately. Afterward, a heatmap of the correlation matrix is generated using Seaborn's `sns.heatmap`, annotating the correlations and displaying them color-coded. The heatmap is saved as a PNG file named 'corr1.png'.
Next, the code checks for missing values in the DataFrame using `df.isna().sum().sort_values(ascending=False)` to identify the columns with null values. It then defines two functions for

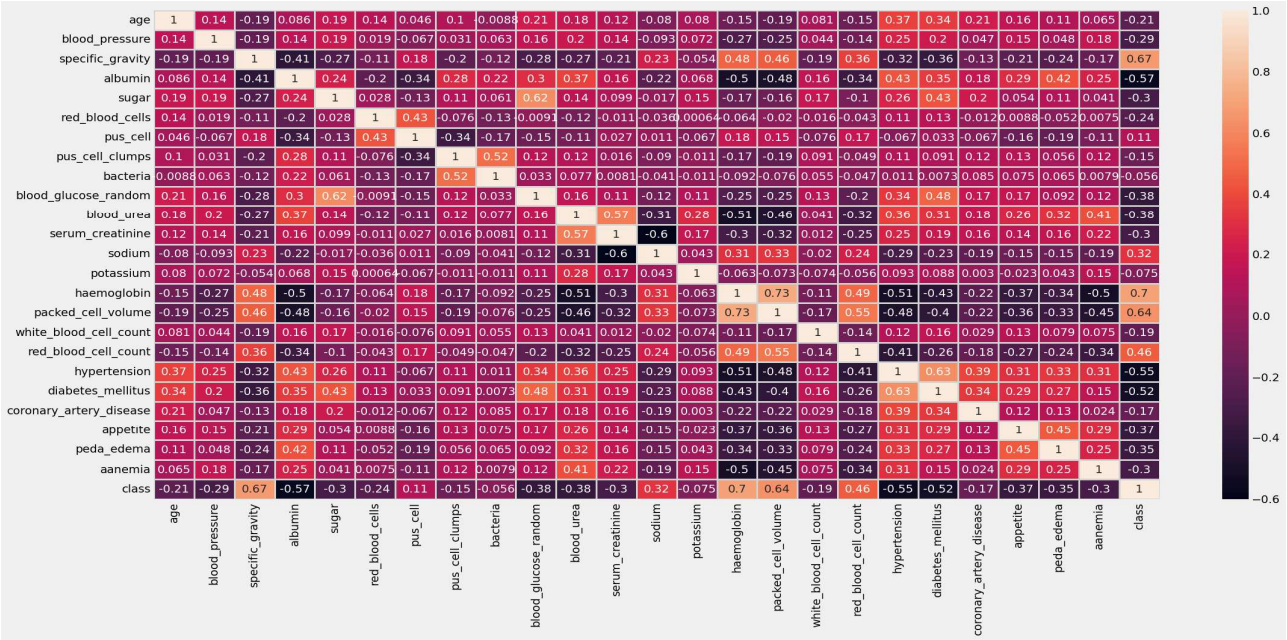

Fig 4: Heatmap of all Features of the Dataset

imputing missing values: `random_value_imputation` for numerical columns and 'red_blood_cells' and 'pus_cell' in categorical columns, and `impute_mode` for the rest of the categorical columns. Random sampling is used to fill null values in numerical columns, while mode imputation is used for categorical columns. The missing values are replaced accordingly.

Label encoding is applied to categorical columns using Scikit-learn's `LabelEncoder`, converting categorical values into numerical labels. Finally, correlations with the target variable 'class' are computed and displayed for both numerical and categorical features. This comprehensive approach ensures the dataset is prepared for further analysis and modeling by handling missing values and encoding categorical features appropriately.

The code begins by importing necessary libraries for scaling features, splitting data, and implementing machine learning algorithms. It uses `MinMaxScaler` from Scikit-learn to scale up the features to a specified range. The scaled features are stored in `new_features` after transforming the original feature matrix `X`. The data is then split into training and test sets using `train_test_split` from Scikit-learn, with 75% of the data used for training and 25% for testing, maintaining consistency with a random state of 67 for reproducibility.

Five classification models are trained and evaluated: K-Nearest Neighbors (KNN), Random Forest Classifier, Decision Tree Classifier, Gradient Boost Classifier and XGBoost. For each model, the training and test accuracies are printed along with a classification report, which provides metrics like precision, recall, and F1-score for each class. Additionally, confusion matrices are plotted using a custom function `plot_confusion_matrix`. After fitting each model to the training data, the accuracies and classification reports for the test data are printed out. These metrics provide insights into how precise each model performs on unseen data. The final step involves displaying the confusion matrices for each model, allowing for a visual examination of the model's performance in classifying different classes.

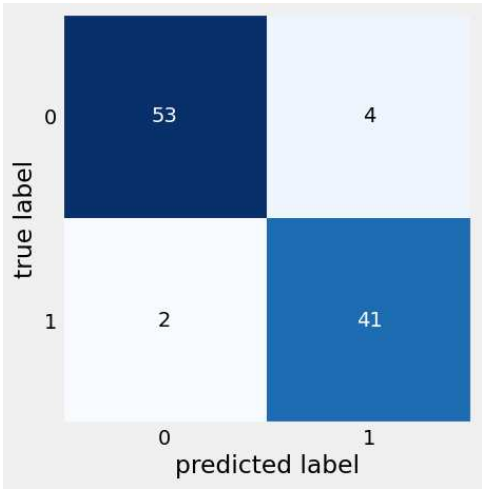

Fig 5: Confusion Matrix of KNN model

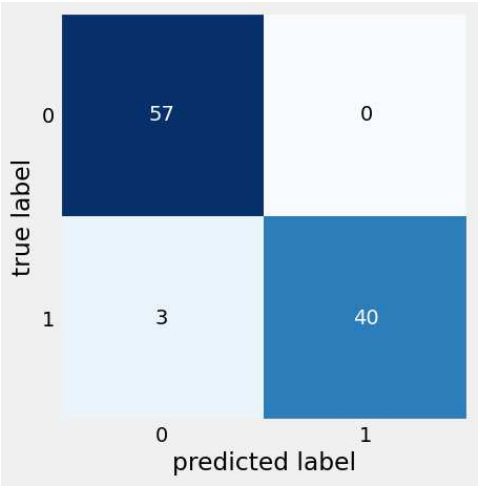

Fig 6: Confusion Matrix of Random Forest model

The Random Forest, Gradient Boost, and XGBoost classifiers achieved perfect training accuracy but showed slightly lower test accuracies of 0.97, 0.97, and 0.95, respectively, indicating a minor drop in performance on unseen data. Precision, recall, and F1-score metrics were high for both classes, suggesting few false positives and robust performance in identifying instances of both classes. Despite the slight decrease in test accuracy, the Random Forest Classifier displayed strong overall performance in classifying the dataset, with high precision, recall, and F1-scores across all classes.

IV. RESULTS AND DISCUSSION

Chronic Kidney disease Identification using various advanced machine learning models have revolutionized result and implementation in the filed of medical diagnosis. This research work's performance was measured by three metrics: precision, recall and f1-score. Early Detection of CKD can be very helpful in order to enable timely intervention and personalized treatment plans.

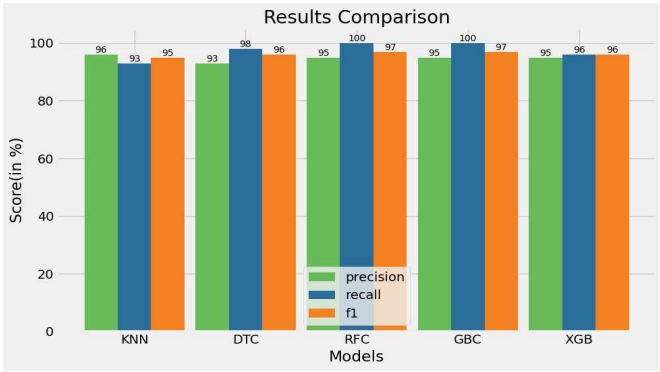

Fig. 7: Performance Measures of Various Predictive Models

The K-Nearest Neighbors (KNN) classifier attained a training accuracy of 97.67% and a test accuracy of 94.00%. Precision, recall, and F1-score for both negative (class 0) and positive (class 1) instances were reported. For class 0, precision was 96%, recall was 93%, and F1-score was 95%. For class 1, precision was 91%, recall was 95%, and F1-score was 93%. These metrics collectively indicate the classifier's balanced performance in distinguishing between the two classes. With an overall accuracy of 94%, the KNN model demonstrated effectiveness in accurate predictions on the test dataset.

The Decision Tree Classifier (DTC) achieved perfect training accuracy of 100% and 95% test accuracy. Precision, recall, and F1-score were reported for both negative (class 0) and positive (class 1) instances. For class 0, precision was 93%, recall was 98%, and F1-score was 96%. For class 1, precision was 97%, recall was 91%, and F1-score was 94%. These metrics collectively demonstrate the classifier's strong performance in differentiating between the two classes. With an overall accuracy of 95%, the Decision Tree model exhibited robust predictive power on the test dataset.

The Random Forest Classifier achieved 100% training accuracy and a high 97% test accuracy. Precision, recall, and F1-score were reported for both negative (class 0) and positive (class 1) instances. For class 0, precision was 95%, recall was 100%, and F1-score was 97%. For class 1, precision was 100%, recall was 93%, and F1-score was 96%. These metrics collectively demonstrate the classifier's robust performance in distinguishing between the two classes. With an overall accuracy of 97%, the Random Forest model showcases strong predictive capability on the test dataset.

XGBoost classifier is performing exceptionally well! The test accuracy of 0.95 means it's highly effective at classifying new data. The classification report confirms this, showing excellent precision of 95%, recall 96%, & 93% and f1-scores 96% & 94% across both classes. However, the perfect training accuracy raises a potential concern of overfitting. This means the model might be overly-tuned to the training data, so keep an eye on its performance with a separate validation dataset to ensure it generalizes well to unseen examples.

	Model	Training_Score	Test_Score
2	Random Forest Classifier	1.000000	0.97
3	Gradient Boosting Classifier	1.000000	0.97
4	XGBoost	1.000000	0.95
0	KNN	0.976667	0.94
1	Decision Tree Classifier	0.976667	0.93

Table 2: Training & Test Accuracies of Various Models

The table summarizes the performance of different machine learning models. XGBoost, Random Forest Classifier and Gradient Boosting Classifier achieved perfect training scores of 100% and test scores of 97%. XGBoost demonstrated a slightly lower test score of 95% despite a perfect training score. KNN achieved a training accuracy of about 97.67% with a corresponding test accuracy of 94%. The Decision Tree Classifier showed a test score of 93% with a similar training accuracy to KNN. Overall, Random Forest Classifier and Gradient Boosting Classifier performed best, closely followed by XGBoost, KNN, and the Decision Tree Classifier.

Evaluation indicators after developing the confusion matrix:

$$Accuracy\ Score = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision\ Score = \frac{TP}{TP+FP}$$

$$Recall\ Score = \frac{TP}{TP+FN}$$

$$F1\ Score = \frac{2*(Precision*Recall)}{Precision+Recall}$$

Model	Precision	Recall	F1-score	Accuracy
KNN	0.96	0.93	0.95	0.94
Decision Tree Classifier	0.93	0.98	0.96	0.93
Random Forest Classifier	0.95	1.00	0.97	0.97
Gradient Boost Classifier	0.95	1.00	0.97	0.97
XGBoost	0.95	0.96	0.96	0.95

Table 3: Performance Measures of all Models

A. Advanced ML Models: Explore and implement advanced machine learning models to enhance the accuracy and reliability of CKD predictions, ensuring robust performance in diverse healthcare scenarios.

B. Feature Processing and conversion: Develop data preprocessing environment to process features dimensions and converting categorical data into numerical form improving the efficiency and accuracy of the CKD prediction system.

C. Multi model usage: Implementing multiple machine learning models that not only identifies chronic kidney disease (CKD) but also provides insights also considering individual health data, lifestyle factors, and genetic predispositions.

V. CONCLUSION AND FUTURE WORK

In conclusion, this research work has sought to address the critical issue of Chronic Kidney Disease (CKD) identification through the implementation of advanced Machine Learning (ML) models, with a specific focus on the novel XGBoost, Random forest. The project was driven by the recognition of the life-threatening nature of CKD, the challenges in early detection, and the potential of cutting-edge technology to enhance diagnostic accuracy. The highest accuracy is of the Random forest & gradient boost model which is 97% both. The project's primary objectives were to develop a robust ML-based system for early CKD identification, leverage the capabilities of the XGBoost and Random Forest, and contribute to the evolving landscape of healthcare through the integration of intelligent solutions and webapp functionality.

A. Personalized Medicine - Developing models that consider individual patient characteristics, genetic factors, lifestyle, and environmental influences for personalized CKD risk assessment and treatment plans.

B. Remote Patient Monitoring - Designing applications and wearable devices that enable remote monitoring of patients with CKD, allowing healthcare professionals to track key indicators and intervene as needed. Integrating patient-specific data to tailor interventions and medication regimens for better outcomes.

C. Disease Progression Prediction - Build an AI model capable of predicting the progression of diseases detected through patient data. This information could guide treatment plans and help healthcare providers anticipate and address potential complications.

REFERENCES

[1] Guozhen chen et al. Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform , IEEE Access 1109, may 18, 2020

[2] PEDRO A. MORENO-SÁNCHEZ “Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and Evaluation of an Explainable AI Model ” IEEE Access 3rd April, 2023

[3] Sagar Dhanraj Pande et al. “Multi-Class Kidney Abnormalities Detecting Novel System Through Computed Tomography” IEEE Access vol 45 no 2nd June, 2020

[4] Ping Liang et al. “Deep Learning Identifies Intelligible Predictors of Poor Prognosis in Chronic Kidney Disease”, EMB vol 27 , 7th July, 2023

[5] MD. RASHED-AL-MAHFUZ et al. “Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening” IEEE Healthcare vol 11 no 25 15th April, 2021.

[6] Maithaili desai “Early Detection and Prevention of Chronic Kidney Disease” vol 21 no 70 3rd May, 2021

[7] Asif Salekin et al . “Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes”

2016 IEEE International Conference on Healthcare Informatics access vol 22 no 17 5th March, 2021

[8] Dr. Razib Hayat Khan et al. “A Comparative Analysis of Machine Learning Approaches for Chronic Kidney Disease Detection” 023 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE) vol 23 no 31 May 2023

[9] Dina saif et al. “Deep-kidney: an effective deep learning framework for chronic kidney disease prediction” Vol 21 no 13, March 2019

[10] Rahul Sawhney et al. “A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease.” Elsevier - Decision Analytics Journal 6 vol 21 no 18 (2023) 100169.

[11] Md. Ariful Islam et al. "Chronic kidney disease prediction based on machine learning algorithms" J Pathol Inform. 2023; vol 14: 100189.

[12] Dibaba Adeba Debal et al. "Chronic kidney disease prediction using machine learning techniques. Vol 21 Article number: 109 may (2022) "

[13] Hira Khalid et al. "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease Published vol 25 no 4, 14 Mar 2023 "

[14] Imesh Udara et al. "Chronic Kidney Disease Prediction Using Machine Learning Methods vol 28 no 21, 28-30 July 2020"

[15] Deema Mohammed Alsekait et al. "Toward Comprehensive Chronic Kidney Disease Prediction Based on Ensemble Deep Learning Models, vol 20 no 21, 20 March 2023".

[16] Qiong Bai et al. "Machine learning to predict end stage kidney disease in chronic kidney disease, vol 12, Article number: 8377 (2022)"

[17] Nikhila et al. "Chronic Kidney Disease Prediction using Machine Learning Ensemble Algorithm vol 15 no 19, 19-20 February 2021"

[18] Reshma S et al. "Chronic Kidney Disease Prediction using Machine Learning, Vol. 9 Issue 07, July-2020"

[19] Marwa Almasoud et al. "Detection of Chronic Kidney Disease Using Machine

Learning Algorithms with Least Number of Predictors, vol 8 issue 21, march 2013"

[20] Chamandeep Kaur et al. "Chronic Kidney Disease Prediction Using Machine Learning, ournal of Advances in Information Technology, Vol. 14, No. 2, 2023"

[21] Md. Mehedi Hassan et al. "A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records, vol 12 no 18, 22 February 2023.