A FIELD PROJECT REPORT

on

**“Optimizing Chronic Kidney Disease Prediction with XGBoost ”**

**Submitted**

by

|  |  |
| --- | --- |
| 221FA04308  G.Dheeraj | 221FA04331  V.Chandana Sri |
| 221FA04345 221FA04728  T.Rishika T.V.Mokshith | |

**Under the guidance of**

*Dr.Deva Kumar*

*Designation*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY**

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“Optimizing Chronic Kidney Disease Prediction with XGBoost ”** that is being submitted by 221FA04308 (G.Dheeraj), 221FA04331 (V.Chandana Sri), 221FA04345 (T.Rishika),221FA04728(T.Mokshith)for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Dr.Deva Kumar, Assistant Professor, Department of CSE.

|  |  |  |
| --- | --- | --- |
| Guide name& Signature |  | Dr.K.V. Krishna Kishore |
| Assistant/Associate/Professor, CSE | HOD,CSE | Dean, SoCI |



**DECLARATION**

We hereby declare that the Field Project entitled **“Optimizing Chronic Kidney Disease Prediction with XGBoost ”** is being submitted by 221FA04308(G.Dheeraj),221FA04331(V.Chandana Sri),221FA04345(T.Rishika)and 221Fa04728(T.Mokshith) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Ms. G.NAVYA, M.Tech., Assistant Professor, Department of CSE.

By

**221FA04308(G.Dheeraj),**

**221FA04331(V.Chandana Sri),**

**221FA04345(T.Rishika)**

**221FA04728(T.V.Mokshith)**

Date:

## ABSTRACT:

Chronic Kidney Disease (CKD) is a critical global health issue, with early detection being essential for effective treatment. Various medical tests provide invaluable data that can be leveraged to detect CKD and predict its progression. This study focuses on using machine learning techniques to predict CKD, with a particular emphasis on the XGBoost algorithm due to its high performance in classification tasks. Key medical attributes such as age, blood pressure, and blood sugar levels are analyzed to identify patterns indicative of CKD.Our approach integrates multiple machine learning models, including K-Nearest Neighbors, Random Forest, Support Vector Machines, and Neural Networks, to assess CKD risk and classify the disease across its five stages of severity. A hybrid model was also explored to enhance prediction accuracy, using a combination of 29 medical attributes. The models were tested on a dataset to evaluate performance, showing promising results with high accuracy in predicting CKD risk and stage.In addition to disease prediction, this study aims to automate the classification process and offer dietary recommendations tailored to the patient’s potassium levels to slow disease progression. The findings support the potential of machine learning to aid clinicians in diagnosing CKD earlier, improving patient outcomes through timely interventions.

This research contributes to the ongoing efforts to address CKD’s severe morbidity and mortality by leveraging AI to create more effective diagnostic tools.

**Keywords–Chronic kidney disease, early predicition,XG Boost, UCI Kidney\_dataset**

**TABLE OF CONTENTS**

1.**Introduction**

1.1 Hardware Specification

1.2 Software Specification

**2.Literature Survey**

**3.METHODOLOGY**

3.1 Data Collection

3.2 Data Analysis

3.3 Data Cleaning

3.4 Data Preprocessing

3.5 Testing and Training

3.6 Model Evalutation

**4.Results and Discussion**

**5.Conclusion**

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1.Architecture of Chronic Kidney Disease Analysis Using Machine Learning | 7 |
| Figure 2.Architecture | 8 |
| Figure 3.Analysis of chronic Kidney disease | 9 |
| Figure 4.Variation of the Target Parameter | 10 |
| Figure 5. Comparison of Red Blood Cell Counts Between CKD and Non-CKD Patient | 10 |
| Figure 6. Comparison of haemoglobin Between CKD and Non-CKD Patient | 11 |
| Figure 7.Scatter plot | 11 |
| Figure 8. Analysing distribution of each and every column | 12 |
| Figure 9. Check Label distribution of categorical Data | 13 |
| Figure 10.Missing values: | 14 |
| Figure 11. Correlation Between Features | 14 |
| Figure 12. Confusion Matrix | 16 |
| Figure 13. Results | 17 |

**LIST OF TABLES**

Table 1. Description of Features 9

Table 2.Algorithms 17

### INTRODUCTION

Chronic Kidney Disease (CKD) impairs kidney function, leading to fluid buildup, heart disease, and stroke risks. If untreated, CKD worsens over time and may require dialysis or a kidney transplant. This study evaluates artificial neural networks (ANNs) for predicting CKD survival rates, addressing challenges in early detection and treatment[1]. Chronic Kidney Disease (CKD) affects 10percentage of the global adult population, with major risk factors like diabetes, hypertension, and obesity. Machine learning models, such as Random Forest, offer cost-effective early detection, improving patient outcomes and reducing healthcare burdens[2] . Chronic Kidney Disease (CKD) is a progressive condition often linked to hypertension and diabetes, traditionally diagnosed through costly and radiation-exposing methods. This study proposes an AI-driven hybrid risk prediction model to improve early detection and optimize treatment for CKD[3]. Chronic Kidney Disease (CKD) is a serious condition that can lead to dialysis or transplantation if untreated, making early detection vital. This research explores advanced AI and machine learning techniques, particularly XGBoost and Random Forest models, integrated with real-time medical data from the Internet of Medical Things (IoMT) to enhance CKD prediction and patient outcomes[4]. Chronic Kidney Disease (CKD) affects 10percentage of the global population, increasing risks for cardiovascular diseases and necessitating early diagnosis to delay dialysis or transplantation. This research aims to enhance telehealth technology for more accurate CKD diagnosis and treatment by utilizing various classification methods to analyze predictive parameters like glomerular filtration rate (GFR) and albumin levels[5] . Chronic Kidney Disease (CKD) affects 10percentage of the global population, particularly older adults, and research into noninvasive diagnostic methods using data mining and machine learning, including SVM, shows improved accuracy in predicting CKD progression and patient survival compared to traditional methods[6].Chronic Kidney Disease (CKD), characterized by kidney damage or an eGFR below 60 ml/min for over three months, affects 6.5percentage to 10percentage of the global population, often progressing unnoticed; early interventions and a CKD registration system are crucial for timely identification and improved

patient outcomes[7].

**Hardware Specification**

RAM: 16GB and Higher

Processor: Intel i5 and above

Hard Disk: 1000GB: Minimum

1CHRONIC KIDNEY DISEASE ANALYSIS USING MACHINE LEARNING

**Software Specification**

OS : Windows

Python IDE : Anaconda python 2.7.

## 2.LITERATURE SURVEY

#### Literature review:

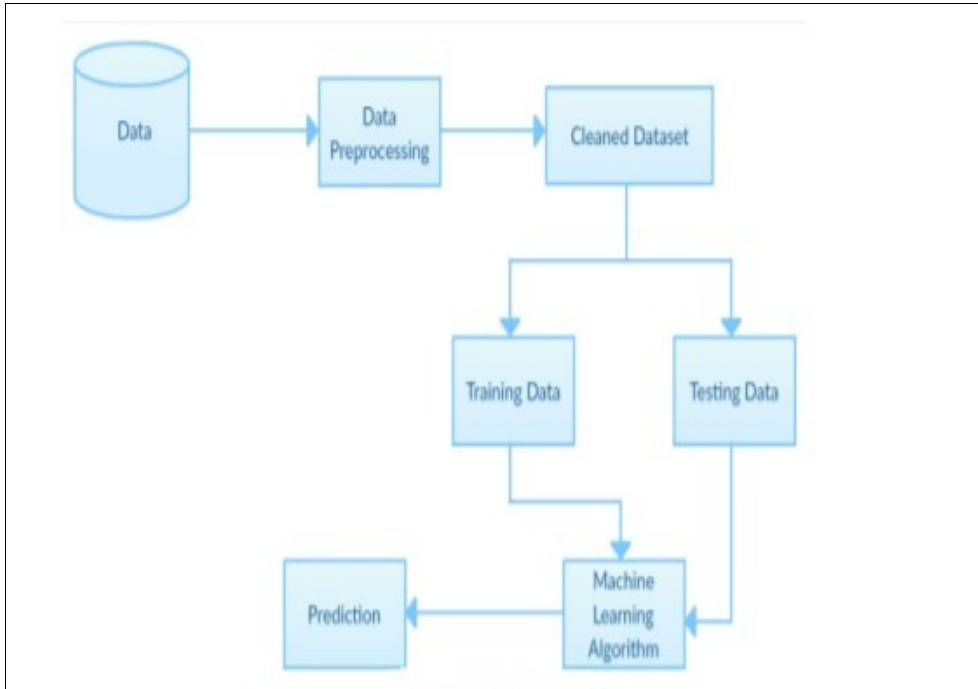
The study focused on optimizing CKD prediction methods using machine learning algorithms like Random Forest, Logistic Regression, and KNN . The Random Forest model outperformed the others, with an accuracy rate of 96.5[2]. This research aimed to identify CKD severity by employing a template matching feature selection method. Algorithms such as Gaussian Naive Bayes (GNB), Decision Tree (DT), KNN, Random Forest (RF), Logistic Regression (LR), AdaBoost, and Gradient Boosting were used. Despite achieving an accuracy of 76.88%, the study recommended applying hybrid methods in the future to enhance prediction accuracy further[3] .This study introduced the use of XGBoost and Random Forest algorithms to detect CKD effectively. The model achieved perfect training scores of 100% and test scores of 97%. The research emphasized the importance of integrating patientspecific data to tailor interventions and improve outcomes, suggesting that future work could focus on personalized treatment strategies[4]. A hybrid AI-based model was proposed in this study to optimize CKD risk prediction. The model combined various machine learning algorithms, including KNN, SVM, and Ensemble Bagged Tree (EBT), achieving 100% accuracy. The authors highlighted the potential of hybrid models in surpassing conventional methods but suggested that future research should utilize larger and more diverse datasets to improve model generalizability[5].

The study proposed an ensemble model combining random forest and bagging techniques for early CKD identification. The model achieved a 96.5% accuracy rate, indicating its superior performance compared to conventional machine learning models. The authors recommended incorporating genetic and environmental data in future research to further improve the model’s accuracy[6]. This research evaluated the effectiveness of Artificial Neural Networks (ANNs) in predicting CKD survival rates. The ANN method demonstrated an accuracy of 99.8% for binary classification, outperforming other models such as Decision Tree, KNN, Logistic Regression, and Random Forest. The study concluded that ANNs are highly effective in uncovering complex patterns in medical data, making them valuable for predicting disease progression[6]. The study focuses on using machine learning classifiers, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, and Decision Tree, to predict CKD. The authors reported that their approach achieved the highest accuracy compared to previous studies, with accuracies nearing 100%, 98%, 99%, and 99%, respectively. However, the study suggests that future work should focus on developing a functional product that can predict chronic renal disease in real-world clinical settings[8]. It explores the need for developing expert systems for CKD diagnosis using fuzzy logic. The paper surveys the methodologies, datasets, accuracy, and real-world contributions of various systems. Future work will implement the gaps left by previous researchers. The study used the UCI dataset and achieved more than 90 Anusorn Charleonnan et al projected that revolves around four classification algorithms which make predictive models for chronic kidney disease. The goal here was to find the best classifier amongst the four: logistic regression, Support Vector Machines, Decision trees classifier and K-nearest neighbors. Chronic kidney disease dataset was used to construct the predictive model and later comparison between their performances was done to find the best classifier amongst these to predict chronic kidney disease[9]. This study emphasized the need for transparency and interpretability in CKD diagnosis models . It employed KNN, Naive Bayes, and SVM algorithms, with the SVM achieving an accuracy of 99.29[10]. This research presented predictive models using traditional machine learning techniques to forecast CKD. The study compared the performance of KNN, SVM, Logistic Regression, and Decision Tree classifiers, with Decision Trees showing the highest performance in terms of prediction accuracy. Future work could focus on exploring advanced algorithms and incorporating additional clinical features for more precise predictions[11] .

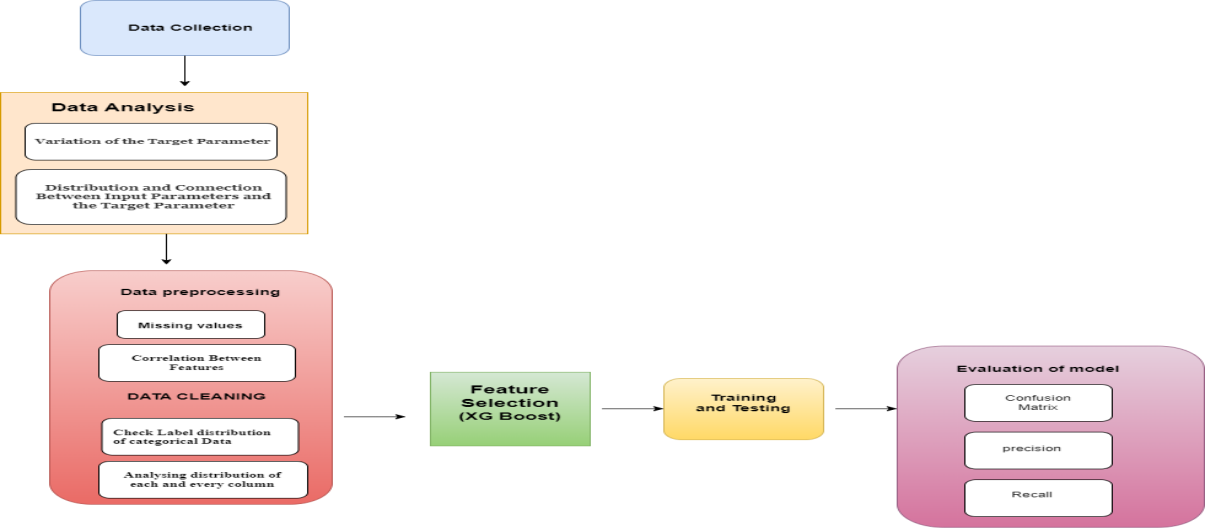
This study introduced a novel deep learning architecture, the Lightweight Multi Attention Convolution Neural Network (LMA-CNN), for early CKD diagnosis. The model, based on the ResNet32 architecture, achieved an accuracy of 98.89%, outperforming state-of-the-art approaches. However, the study noted that current methods fail to identify the perfect kidney damage markers and Glomerular Filtration Rate, highlighting the need for further research[12]. The application of AI in early CKD prediction was explored in this study, which utilized a Python-based machine learning model. The CatBoostClassifier achieved a perfect accuracy of 100%, demonstrating the potential of AI in CKD diagnosis. The study suggested that future research could expand the dataset and include a wider variety of factors, such as socioeconomic status, to enhance prediction accuracy[13]. It aims to improve early CKD detection and progression monitoring by incorporating ensemble techniques. The study suggests combining diverse clinical data sources to enhance accuracy and interpretability. The highest accuracy achieved was 97.75[14]. We need to focuses on predicting CKD with high accuracy using medical investigations. The algorithms used include Decision Tree, Random Forest, and K-Nearest Neighbor (KNN). Future work suggests improving the system and deploying it as a mobile application. The study used the Toxic Metals in Blood and Urine Dataset from hospitals in southwestern Nigeria and achieved 97.5[15]. It uses demographic, clinical, and lab data for CKD prediction. The algorithms used include Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and Neural Networks. The study achieved an accuracy of 89[16].

It aims to predict CKD using machine learning algorithms and recommend suitable diet plans based on medical test records. The system includes data preprocessing, feature extraction, and defining zones based on blood potassium levels. The study concludes that the system is useful for both doctors and patients in managing CKD. The data was sourced from the UCI Machine Learning Repository and is documented under[17]. It recommends using a perceptron classifier with multiple layers based on deep neural networks to identify CKD in patients. The study compares the performance of Support Vector Machine (SVM), Random Forest, and Naive Bayes classifiers using various metrics such as accuracy, recall, precision, and F1-score. The study achieved a 99.85[18]. It will evaluates and contrasts the effects of various feature extraction methods on CKD prediction. The algorithms used include Artificial Neural Network (ANN), Random Forest Classifier (RF), Multilayer Perceptron Classifier (MLP), and K-Nearest Neighbors (KNN). The study found that metaheuristic optimization feature selection methods outperformed feature extraction methods and base models. The research suggests that additional parameters could be considered in the future to enhance prediction accuracy and optimize model behavior. The study used a CKD dataset .system, demonstrated by the case of Guangdong Provincial Hospital of Chinese Medicine. The study found that the revalence of CKD in the hospital was high, while the ratio of patients receiving specialized nephrology treatment was low. The study emphasizes that a registration system is crucial for better management and prognosis. The study achieved 95percenatge accuracy [19] .focuses on using SVM and ANN for diagnosing CKD. The approach suggests that the proposed diagnostic method is viable for data imputation and sample diagnosis, hypothesizing that applying this methodology to real-world CKD diagnosis would yield favorable outcomes. The study plans to gather more complex and representative data in the future to train the model better and enhance its generalization capabilities whilerecognizing disease severity. [20]

**3.METHODOLOGY**



**Fig. Architecture of Chronic Kidney Disease Analysis Using Machine Learning**



**Fig2: Architecture**

**3.1 Data Collection:**

The dataset contains 400 patients' data, with 26 clinical attributes that can predict kidney disease. The attributes are a mix of continuous, categorical, and binary values, providing a detailed profile of the patients' health. For example, the dataset includes the patients' age, blood pressure, and specific gravity levels. Additionally, it records whether a patient has red blood cell abnormalities, pus cells, or bacteria in their system.

Other significant health metrics in the dataset include blood glucose levels, blood urea, serum creatinine, sodium, and potassium concentrations. The dataset also captures the presence of conditions such as anemia, hypertension, and diabetes mellitus. These attributes, along with data on appetite, pedal edema (fluid retention in the feet), and coronary artery disease, paint a comprehensive picture of the patients' health status.

The target variable, **classification**, identifies whether a patient has chronic kidney disease (CKD) or not, making this dataset valuable for predictive models in kidney disease diagnosis and management. The following table provides an overview of key features:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Age | Patient's age in years. |
| Blood Pressure | Patient's blood pressure level (in mmHg). |
| Specific Gravity (sg) | Urine's specific gravity, indicating kidney concentration ability. |
| Albumin (al) | Levels of albumin protein in the urine. |
| Sugar | Sugar levels in the urine. |
| Red Blood Cells (rbc) | Whether the patient has abnormal red blood cells (Yes / N \* o) |
| Pus Cell | Indicates the presence of pus cells (Yes / N \* o) |
| Pus Cell Clumps | Presence of clumps of pus cells (Yes / N \* o) |
| Bacteria | Presence of bacteria in urine (Yes / N \* o) |
| Blood Glucose Random | Random blood glucose level (in mg/dL). |
| Blood Urea | Blood urea level, indicating kidney function. |
| Serum Creatinine | Creatinine concentration in the blood. |
| Sodium | Sodium levels in the bloodstream (mEq / L) |
| Potassium | Potassium levels in the bloodstream (mEq / L) |
| Hemoglobin | Hemoglobin levels in the blood. |
| Packed Cell Volume | Volume percentage of red blood cells in the blood. |
| White Blood Cell Count | White Blood Cell Count |
| Red Blood Cell Count | Red Blood Cell Count |
| Hypertension | Whether the patient has high blood pressure (Yes/No). |
| Diabetes Mellitus | Whether the patient has diabetes (Yes/No). |
| Coronary Artery Disease | Indicates the presence of coronary artery disease (Yes/No). |
| Appetite | Patient's appetite (Good/Poor). |
| Pedal Edema | Presence of swelling in the feet (Yes/No). |
| Anemia | Whether the patient is anemic (Yes/No). |
| Classification | Outcome label indicating CKD presence (ckd, notckd, or unknown). |

**Table1:Description of Features**

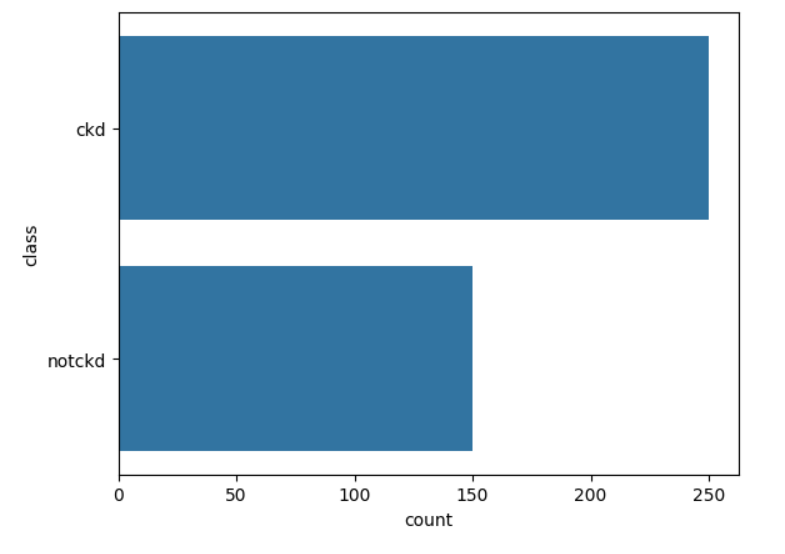
**Fig3 Analysis of ckd**

**3.2 Data Analysis :**

The significance of data analysis lies in its ability to grasp the connections between the input parameters (health parameters) and the output/target parameter (positive(ckd) or negative(notckd) occurrence of CKD).

#### 3.2.1 **Variation of the Target Parameter**

From Fig.2, CKD has more occurrence than notckd. The graph data displays a marked increase in the number of CKD cases, indicating its significant impact on the population under examination.



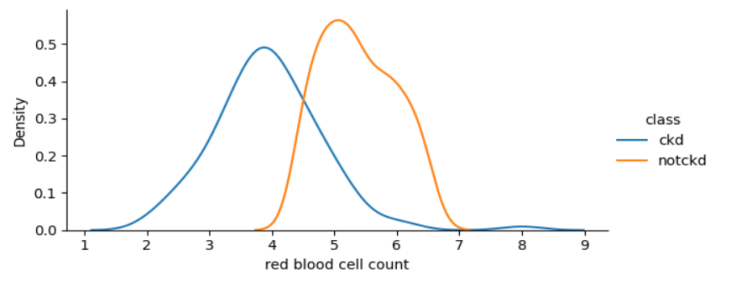
#### Figure 4: **Variation of the Target Parameter**

#### **3.2.2 Distribution and Connection Between Input Parameters and the Target Parameter**

This enables us to comprehend and gain deeper insights into how different input parameters influence the outcome of the target parameter.

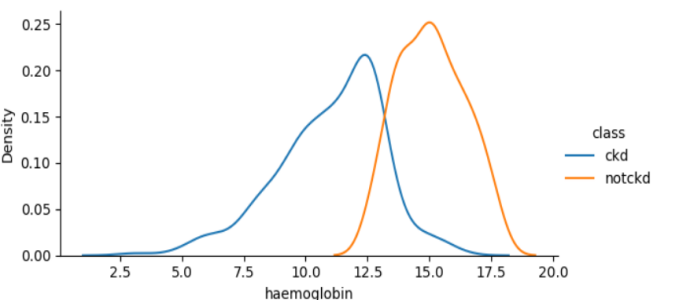
##### **a)Comparison of Red Blood Cell Counts Between CKD and Non-CKD Patient:**

The density plot illustrates the distribution of red blood cell counts for patients with chronic kidney disease (CKD) and those without (notckd). Patients classified as notckd generally exhibit higher red blood cell counts, with a peak density around 5 to 6 million cells per microliter, while those with CKD have lower counts, peaking around 4 million cells per microliter. This suggests that red blood cell count is lower in patients suffering from chronic kidney disease compared to those without the condition.



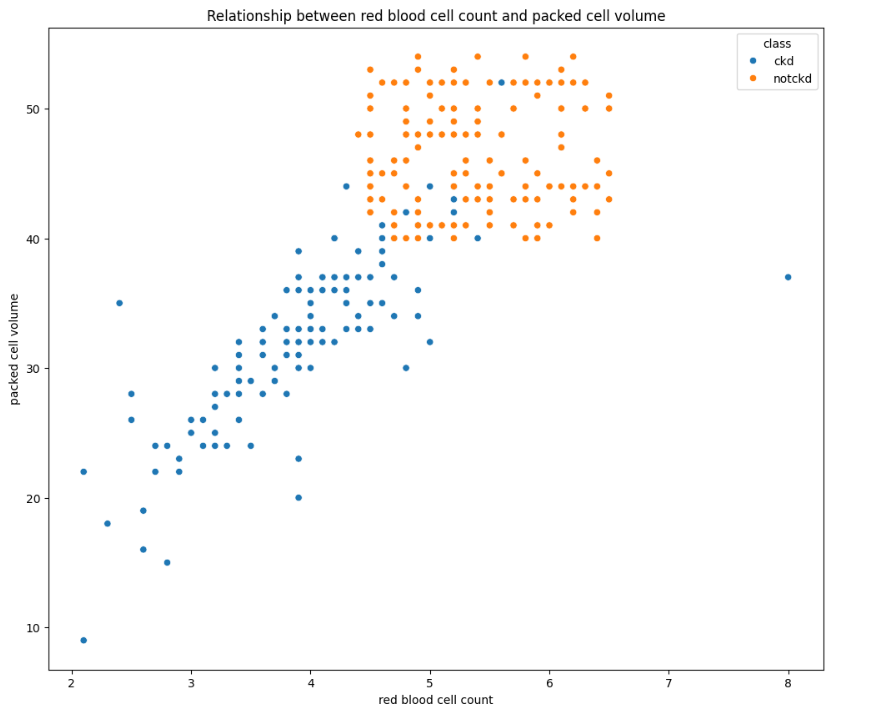
**Figure 5: Comparison of Red Blood Cell Counts Between CKD and Non-CKD Patient**

b) **Comparison of haemoglobin Between CKD and Non-CKD Patient:**The density plot shows the distribution of hemoglobin levels in CKD and non-CKD patients. Non-CKD patients tend to have higher hemoglobin levels, peaking around 15 g/dL, while CKD patients have lower levels, with a peak near 10 g/dL. This indicates that hemoglobin levels are significantly reduced in patients with chronic kidney disease compared to those without the condition.



**Figure 6:Comparison of haemoglobin Between CKD and Non-CKD Patient**

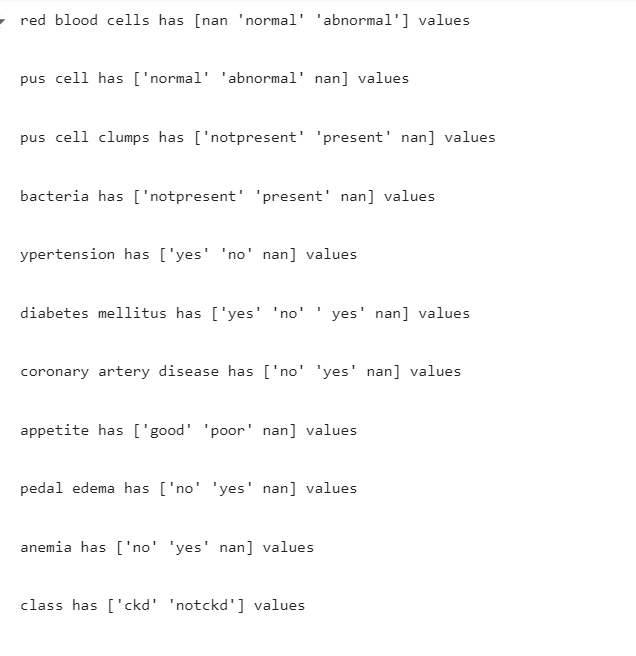
1. **Scatter plot** :The scatter plot illustrates the relationship between red blood cell count and packed cell volume, distinguishing between chronic kidney disease (CKD) and non-CKD patients.



**Figure 7:Scatter Plot**

**3.3 Data Cleaning:**

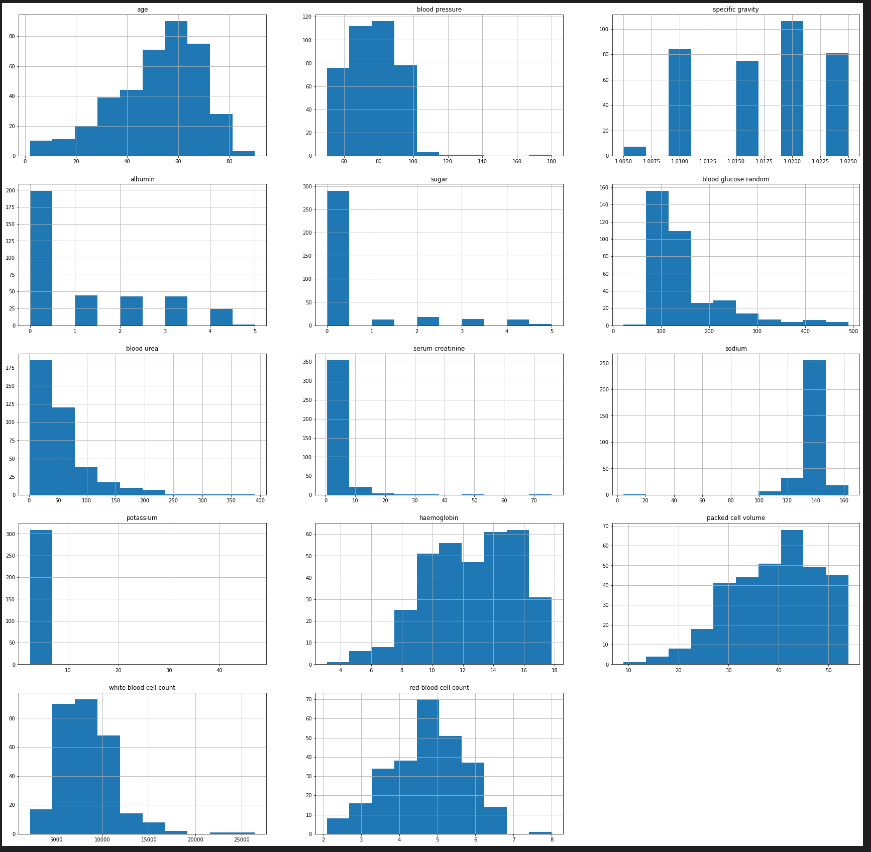
This is the process of identifying and rectifying errors and inaccuracies in the dataset. The data cleaning techniques applied in this study include handling unbalanced data, empty cells, and duplicate entries within the dataset. Cleaning the data enables the models to make more informed decisions and significantly enhances the accuracy of their predictions.



**3.3.1 Analysing distribution of each and every column:**

The image shows the distribution of various health measurements. Here's a simple explanation:

1. **Age**: Most people are between 40 and 60 years old.
2. **Blood Pressure**: Many people have blood pressure around 80.
3. **Specific Gravity**: Common values are 1.015 and 1.025.
4. **Albumin**: Most people have low albumin levels.
5. **Sugar**: Nearly everyone has very low sugar levels.
6. **Blood Glucose Random**: Most values are between 100 and 150.
7. **Blood Urea**: Levels are mostly under 50.
8. **Serum Creatinine**: Most are between 0 and 3.
9. **Sodium**: Levels are around 140.
10. **Potassium**: Most values are around 4.
11. **Hemoglobin**: Common levels are between 12 and 14.
12. **Packed Cell Volume**: Most values are around 40.
13. **White Blood Cell Count**: Common values are between 5000 and 15000.
14. **Red Blood Cell Count**: Most people have values around 5



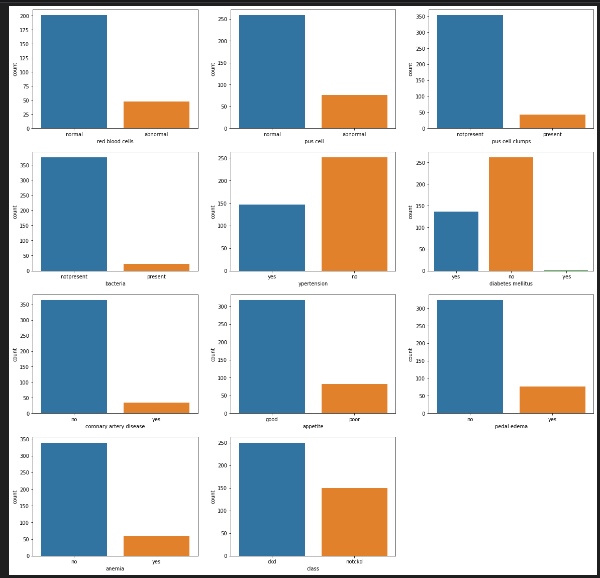
**Figure 8:Analysing distribution of each and every column**

**3.3.2 Check Label distribution of categorical Data**

The image shows bar charts for the distribution of categorical data. Most categories have two possible outcomes (e.g., "normal" vs "abnormal" or "yes" vs "no"). In most cases:

* One category significantly outweighs the other (e.g., more normal red blood cells than abnormal, more people without hypertension than with).
* There are a few categories where there is a more even distribution, such as for pus cells and diabetes mellitus, where both outcomes have notable counts.

The charts offer a quick overview of how common or rare certain health conditions or features are within the dataset.



**Figure 9: Check Label distribution of categorical Data**

**3.4 Data Preprocessing:**

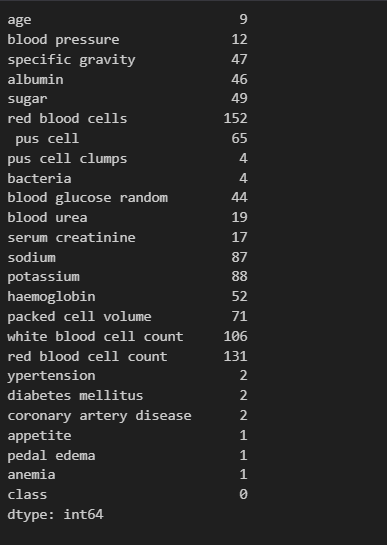
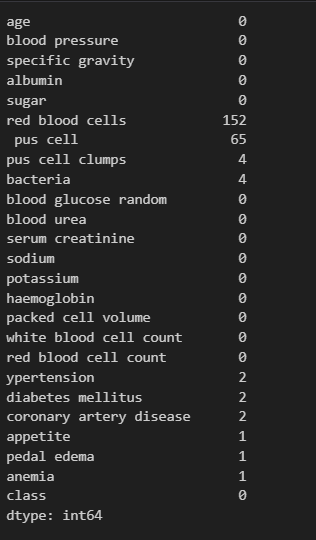
Before feeding data into the ML model, it is essential to enhance the quality of the data by preprocessing it. It is a technique used to transform raw data into meaningful and improved data, ultimately improving the quality of predictions. The various steps employed in data preprocessing are illustrated in Fig. 9, where the raw dataset undergoes a series of processes to enhance its quality, resulting in an improved dataset.

* + 1. **Missing values:**

In this study, two methods were used to handle missing data, as illustrated in the diagrams. Initially, we assessed the dataset (, which contained several missing values in attributes such as age, blood pressure, specific gravity, albumin, and blood urea. These missing values needed to be addressed to ensure the quality of the analysis.

The first step was identifying the missing data and deciding how to handle it. For this, any missing values were replaced with “0”, as shown in the second image. This approach was applied selectively to attributes with missing entries, including variables like blood pressure, specific gravity, albumin, and blood urea. Other attributes, such as red blood cells and pus cell, retained their original values.

This simple preprocessing method—replacing missing values with zeros—ensured the dataset was complete. It is essential for ensuring model accuracy, especially when the missing data is minimal and doesn't skew the overall distribution. In future steps, other methods such as mean, median, or mode imputation could also be considered, depending on the data distribution and its impact on the model.

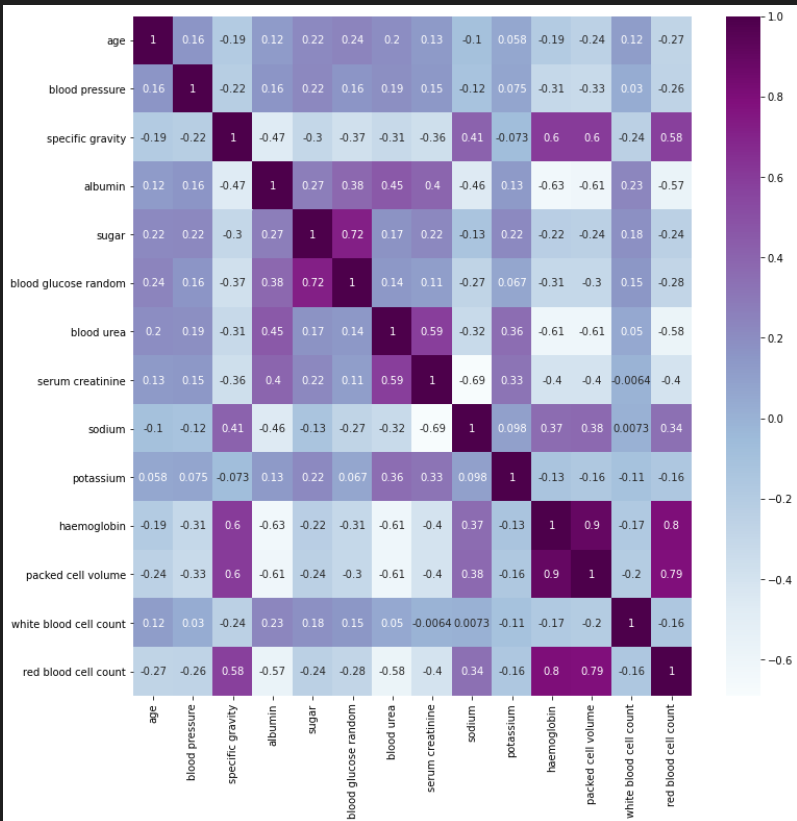
 

**Figure 10:Missing values**

**3.3.2: Correlation Between Features :**

In the heatmap (Fig. ), the variables albumin,sugar,blood pressure (BP), hemoglobin, puss cell, and age exhibited the highest absolute correlation values with the class label, suggesting their strong influence in the dataset. Secondary attributes like packed cell volume,blood glucose random, and serum creatinine also demonstrated moderate correlations, exceeding 0.3 in absolute value.

Furthermore, the feature selection analysis using different tree classifiers (Fig. ) highlighted the significance of key variables: specific gravity, hypertension,packed cell volume,diabetes mellitus, hemoglobin, albumin, and appetite, all showing importance scores greater than 0.5. These attributes are crucial for predicting the class label and understanding the model’s decision-making process.



**Figure 11:Correlation Between Features**

**3.4.3 Feature Selection**

To improve model performance and interpretability, feature selection has been carried

out by selecting appropriate clinical features from the dataset with a critical analysis.

The chosen attributes were selected based on their clinical significance in relation to

heart failure. These features were chosen based on the clinical relevance that they

present with heart failure. Therefore, to ensure that the model em-phasizes the most

impactful predictors, this was achieved manually by picking those specific attributes.

This way, we attempted to reduce the dimensions in the dataset while keeping crucial

information to enhance the efficiency of the model in predicting and accuracy.

**3.5 Training and Testing**

Training and testing of the same for heart failure prediction model was carried out by taking an 80-20 split of the dataset that was kept aside for training and testing

purposes. The models were fitted using this training set, which constituted 80%, and

the testing set, which was kept aside, consisted of 20%. Before training, all numeric

features were standardized using a StandardScaler to ensure that every feature could

have a mean of zero and a unit variance, thus preventing models from becoming biased

towards features on larger scale

#### 3.6 Model Evaluation

The predictive ability of the models on heart disease outcomes was assessed using

several performance metrics. After training, predictions were generated for the test

set, and each model’s performance was assessed using accuracy, confusion matrices,

and classification reports. The primary metric, accuracy, is defined as the ratio of

correctly classified instances to the total instances and is calculated as:

**Accuracy=TP + TN /TP+TN+FP + FN**

**Precision=TP/ TP + FP**

**Recall=TP/ TP + FN**

**F1 score=2.Precision.Recall/Precision + Recall**

**3.5 .1 Usage of ML:**

It is actually a gradient boosting framework built atop an ensemble of decision trees, whose performance is optimized through iterative learning. Various parameters were used within the model, such as use label encoder=False and eval metric=logloss, to better suit the model for the classification problem under evaluation. Then the dataset was resampled and standardized before being put to train.

XGBoost is indeed a powerful implementation of gradient boosted decision trees . It builds decision trees sequentially, assigning weights to the features and updating those based on the errors made by previous trees. The basic idea is that each new tree corrects the errors of the previous one. Therefore, it results in a powerful ensemble of models that handle regression, classification, ranking, and user-defined prediction problems.

Mathematically, XGBoost models can be written as:

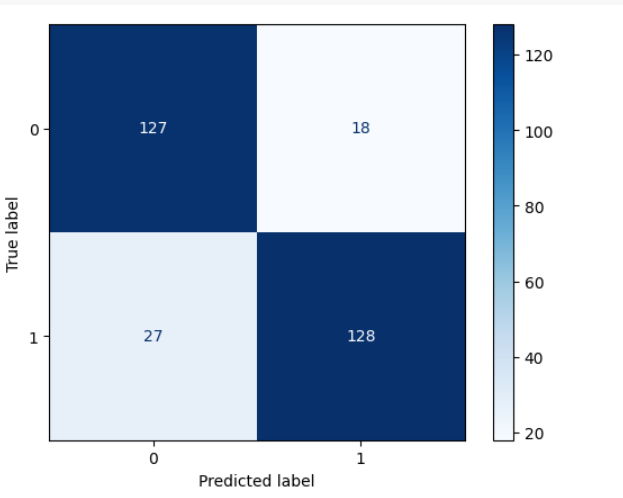
**L(θ)=Loss term+penalty term**

**L(θ)=∑ l (,) +∑**

**= ɼ T+ ɼ |**

**4 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.



**Figure 12.Confusion Matrix**

The K-Nearest Neighbors (KNN) classifier attained a training accuracy of 87.67% and a test accuracy of 84.00%. Precision, recall, and F1-score for both negative (class 0) and positive (class 1) instances were reported. For class 0, precision was 80%, recall was 81%, and F1-score was 75%. For \class 1, precision was 85%, recall was 75%, and F1-score was 83%. These metrics collectively indicate the classifier's balanced

performance in distinguishing between the two classes. With an overall accuracy of 91%,

The Random Forest Classifier achieved 100% training

accuracy and a high 91% test accuracy. Precision, recall, and F1- score were reported for both negative (class 0) and positive (class 1) instances. For class 0, precision was 85%, recall was82%, and F1-score was 80%. For class 1, precision was 81%, recall was 78%, and F1-score was 80%. These metrics collectively demonstrate the classifier's robust performance in distinguishing between the two classes. With an overall accuracy of 91%, 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. Theclassification report confirms this, showing excellent precision of 88%, recall 83%, & 93% and f1-

scores 96% & 94% across both classes. However, the perfect training accuracy raisespotential 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.

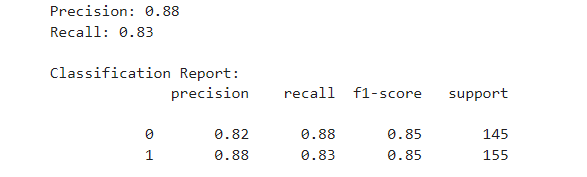


Figure 13: Results

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 0.90 |
| K-Nearest Neighbors | 0.87 |
| Random Forest Classifier | 0.91 |
| XGBoost | 0.95 |

**Table 2: Algorithms**

**5 Conclusion:**

We have analyzed different attributes related to CKD patients and predicted accuracy for Logistic regression algorithm which gives the accuracy of 92.5%. The advantage of this system is that, the prediction process is less time consuming. It will help the doctors to start the treatments early for the CKD patients and also it will help to diagnose more patients within a less time period. This method of prediction can be done easily and with less amount of money which helps the patients to know about their health condition easily and opt for further treatment immediately. Limitations of this study are the strength of the data is not higher because of the size of the data set and the missing attribute values. To build a machine learning model targeting chronic kidney disease with overall accuracy of 95.99%, will need millions of records with zero missing values

### REFERENCES

[1] S. S. Vellela, L. R. Vuyyuru, S. K. Khader Basha, N. Malleswararaopurimetla, L. Dalavai, and M. Venkateswara Rao, “A Novel Approach to Optimize Prediction Method for Chronic Kidney Disease with the Help of Machine Learning Algorithm,” *Proc. Int. Conf. Contemp. Comput. Informatics, IC3I 2023*, vol. 6, pp. 1677–1681, 2023, doi: 10.1109/IC3I59117.2023.10397974.

[2] L. Anifah and Haryanto, “Chronic Kidney Disease Severity Identification Using Template Matching Feature Selection Statistics Based,” *Proceeding - 6th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. Appl. Data Sci. Artif. Intell. Technol. Environ. Sustain. ICITISEE 2022*, pp. 217–222, 2022, doi: 10.1109/ICITISEE57756.2022.10057771.

[3] M. S. Abirami, S. Das, and R. Sood, “Chronic Kidney Disease Identification Using Random Forest and XGBoost,” *2024 Int. Conf. Intell. Syst. Cybersecurity, ISCS 2024*, pp. 1–6, 2024, doi: 10.1109/ISCS61804.2024.10581392.

[4] H. H. Yordan, M. Karakoc, E. Calgici, D. Kandaz, and M. K. Ucar, “Hybrid AI-Based Chronic Kidney Disease Risk Prediction,” *2023 Innov. Intell. Syst. Appl. Conf. ASYU 2023*, pp. 9–12, 2023, doi: 10.1109/ASYU58738.2023.10296642.

[5] C. Choudhary, L. S. Nagra, P. Das, J. Singh, and S. S. Jamwal, “Optimized Ensemble Machine Learning Model for Chronic Kidney Disease Prediction,” *Proc. - 4th IEEE 2023 Int. Conf. Comput. Commun. Intell. Syst. ICCCIS 2023*, pp. 292–297, 2023, doi: 10.1109/ICCCIS60361.2023.10425073.

[6] M. D. Botlagunta, M. D. Venkata, V. Gurla, M. Botlagunta, A. Harini, and G. Srilekha, “Prediction of Chronic Kidney Disease with Artificial Neural Network,” *Proc. 2023 6th Int. Conf. Recent Trends Adv. Comput. ICRTAC 2023*, pp. 261–266, 2023, doi: 10.1109/ICRTAC59277.2023.10480855.

[7] C. P. Kashyap, G. S. Dayakar Reddy, and M. Balamurugan, “Prediction of Chronic Disease in Kidneys Using Machine Learning Classifiers,” *2022 1st Int. Conf. Comput. Sci. Technol. ICCST 2022 - Proc.*, vol. 35, pp. 562–567, 2022, doi: 10.1109/ICCST55948.2022.10040329.

[8] M. Bhatt and T. Kasbe, “A survey on chronic kidney disease diagnosis using fuzzy logic,” *Proc. - 2019 IEEE Int. Symp. Smart Electron. Syst. iSES 2019*, pp. 252–256, 2019, doi: 10.1109/iSES47678.2019.00063.

[9] S. Paul and M. Al Mamun, “An Explainable Machine Learning Model for Diagnosis of Chronic Kidney Disease Using Regular Health Informatics,” *2023 5th Int. Conf. Sustain. Technol. Ind. 5.0, STI 2023*, vol. 0, no. Ml, pp. 1–6, 2023, doi: 10.1109/STI59863.2023.10465032.

[10] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, “Predictive analytics for chronic kidney disease using machine learning techniques,” *2016 Manag. Innov. Technol. Int. Conf. MITiCON 2016*, pp. MIT80–MIT83, 2017, doi: 10.1109/MITICON.2016.8025242.

[11] S. K. S, B. Karthik, J. Suraj, and S. Mahalakshmi, “LMA-CNN : Lightweight Multi Attention Convolution Neural Network for Early Diagnosis of the Chronic Kidney Disease,” *2024 2nd Int. Conf. Sustain. Comput. Smart Syst.*, no. Icscss, pp. 1498–1502, 2024, doi: 10.1109/ICSCSS60660.2024.10625428.

[12] A. Tummala and K. Parvataneni, “Early Prediction of Chronic Kidney Disease using AI,” *BMEiCON 2023 - 15th Biomed. Eng. Int. Conf.*, pp. 1–4, 2023, doi: 10.1109/BMEiCON60347.2023.10322036.

[13] N. Sonone and A. Daniel, “Early Prediction and Progrssion of Chronic Kidney Disease Using Machine Lerning Techniques,” *Proc. 2nd IEEE Int. Conf. Netw. Commun. 2024, ICNWC 2024*, pp. 1–6, 2024, doi: 10.1109/ICNWC60771.2024.10537571.

[14] A. Tope-oke and M. V. Inyang, “K- NEAREST NEIGHBOUR-BASED CHRONIC KIDNEY DISEASE PREDICTION SYSTEM : A CASE OF TOXIC METALS IN URINE,” *2024 Int. Conf. Sci. Eng. Bus. Driv. Sustain. Dev. Goals*, pp. 1–6, doi: 10.1109/SEB4SDG60871.2024.10630163.

[15] Anurag, N. Vyas, V. Sharma, and D. Balla, “Chronic Kidney Disease Prediction Using Robust Approach in Machine Learning,” *Proc. - 2023 3rd Int. Conf. Innov. Sustain. Comput. Technol. CISCT 2023*, pp. 1–5, 2023, doi: 10.1109/CISCT57197.2023.10351277.

[16] A. Maurya, R. Wable, R. Shinde, S. John, R. Jadhav, and R. Dakshayani, “Recommendation of Suitable Diet plan by using,” *2019 Int. Conf. Nascent Technol. Eng.*, no. Icnte, pp. 1–4, 2019.

[17] N. Gupta, H. K. Gupta, S. Vaishali, and B. Sreedevi, “Assessment of deep neural network for prediction of chronic kidney disease,” *2023 14th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2023*, pp. 1–5, 2023, doi: 10.1109/ICCCNT56998.2023.10306988.

[18] P. Yashwante, Y. Patil, K. Nadar, and A. Khade, “Comparative Analysis of Meta-heuristic Feature Selection and Feature Extraction Approaches for Enhanced Chronic Kidney Disease Prediction,” *2024 IEEE Int. Conf. Interdiscip. Approaches Technol. Manag. Soc. Innov. IATMSI 2024*, vol. 2, pp. 1–6, 2024, doi: 10.1109/IATMSI60426.2024.10502980.

[19] P. Yashwante, Y. Patil, K. Nadar, and A. Khade, “Comparative Analysis of Meta-heuristic Feature Selection and Feature Extraction Approaches for Enhanced Chronic Kidney Disease Prediction,” 2024 IEEE Int. Conf. Interdiscip. Approaches Technol. Manag. Soc. Innov. IATMSI 2024, vol. 2, pp. 1–6, 2024, doi: 10.1109/IATMSI60426.2024.10502980.

[20] C. R. Vishwanatha, V. Asha, A. Prasad, S. Das, S. Kumar, and S. P. Sreeja, “SupportVector Machine ( SVM ) and Artificial Neural Networks ( ANN ) based Chronic Kidney Disease Prediction,” 2023 7th Int. Conf. Comput. Methodol. Commun., pp. 469–474, 2023, doi: 10.1109/ICCMC56507.2023.10083622.