Chronic Kidney Disease Prediction

MD.SHAKIL MIA
Department of IRE
Faculty of Cyber Physical System
Bangabandhu Sheikh Mujibur Rahman Digital University
Kaliakair, Gazipur, Dhaka, Bangladesh
1901023@iot.bdu.ac.bd

MD. MUNTASIRE MAHAMUD
Department of IRE
Faculty of Cyber Physical System
Bangabandhu Sheikh Mujibur Rahman Digital University
Kaliakair, Gazipur, Dhaka, Bangladesh
1901012@iot.bdu.ac.bd

***Abstract*—The biosciences have progressed to a greater degree and have produced vast quantities of data from Electronic Health Records. As a result, there are now acute the requirement for knowledge creation from this massive volume of data. Machine learning and data mining techniques are crucial function in the biosciences in this regard. CKD, or chronic kidney disease, is a disorder when the kidneys are injured and unable to filter as they always do, blood. a history of kidney illnesses in the family or failure, hypertension, and type 2 diabetes could result in CKS. There is long-term renal damage and a potential for increasing over time is high. Frequently occurring complications that are brought on by renal failure include anemia, cardiac problems, bone conditions,**

***Index Terms*—Kidney Disease, Chronic Kidney Disease, Kidney Function Prediction, Machine Learning, Classification Models, Random Forest, Model Evaluation, Medical Diagnostics, Decision Support Systems, Data Preprocessing, Feature Engineering, AUC Score, Confusion Matrix.**

I. Introduction

Kidney Disease Prediction refers to the kidneys' inability to perform their usual blood filtering function and other associated processes. The term ”chronic” signifies a gradual deterioration of kidney cells over an extended period. CKD is a significant form of kidney failure, where the kidneys cease their blood filtering function, leading to a buildup of excess fluids in the body. This accumulation results in elevated levels of potassium and calcium salts, leading to various health issues. High salt concentrations can cause a range of ailments, impacting factors such as blood pressure regulation, hormone activation, and red blood cell production. Excess calcium can lead to bone diseases and cystic ovaries in women. CKD can also lead to sudden illness or sensitivity to certain medications, known as Acute Kidney Injury (AKI). Elevated blood pressure resulting from CKD can contribute to heart problems and heart attacks. In many cases, CKD necessitates permanent dialysis or kidney transplants, particularly when there is a family history of kidney disease. Research indicates that nearly one in three individuals diagnosed with diabetes also suffers from CKD.

Furthermore, early identification and management of CKD can significantly improve patients quality of life. Machine learning prediction algorithms can be employed intelligently to forecast the occurrence of CKD and provide early intervention. An extensive review of the literature demonstrates the utilization of various machine learning algorithms for CKD prediction. This study aims to predict CKD by employing classifiers such as Decision Tree, Random Forest, and K-Nearest Neighbors, while also proposing the most effective prediction model.

II. Literature Survey

In [10], M. P. N. M. Wickramasinghe et al. introduce a methodology for managing the disease through a tailored diet plan. They construct classifiers using various algorithms such as Multiclass Decision Jungle, Multiclass Decision Forest, Multiclass Neural Network, and Multiclass Logistic Regression. These classifiers predict an allowable potassium range based on the patient's blood potassium levels and recommend a diet plan accordingly.

In [9], H. A. Wibawa et al. propose and evaluate the use of Kernel-based Extreme Learning Machine (ELM) to predict Chronic Kidney Disease. They compare the performance of four kernels-based ELM, including RBF-ELM, Linear-ELM, Polynomial-ELM, and Wavelet-ELM, with standard ELM. Radial Basis Function – Extreme Learning Machine (RBF-ELM) demonstrates higher prediction accuracy based on sensitivity and specificity metrics.

[3] highlights the increased risk of Cardiovascular Disease (CVD) factors associated with CKD, including hypertension, diabetes mellitus, dyslipidemia, and metabolic syndrome. U. N. Dulhare et al. extract action rules based on CKD stages and predict CKD using a Naïve Bayes model with OneR attribute selector to prevent the progression of chronic renal disease.

[11] emphasizes the importance of evaluating the condition of late-stage CKD patients, as it significantly impacts the appropriate care and treatment. H. Zhang et al. investigate the performance of Artificial Neural Network (ANN) models for predicting the survivability of CKD patients.

In [1], the paper discusses how dialysis or kidney transplant remains the primary treatment for End Stage Renal Disease (ESRD) patients. Early prediction of CKD and proper treatment through diet can slow down or halt disease progression.

J. Aljaaf et al. conclude that applying machine learning algorithms with predictive analytics offers an intelligent solution for early disease prediction.

[7] introduces data mining models that employ ensemble techniques like Boosting to enhance prediction. AdaBoost and LogitBoost are commonly used to compare classification algorithm performance in detecting CKD. Arif-Ul-Islam et al. analyze the performance of boosting algorithms, employing the Ant-Miner machine learning algorithm along with Decision trees to derive rules.

[5] underscores the significance of data mining in making decisions based on chronic disease datasets, particularly when dealing with large amounts of structured, unstructured, and semi-structured data. G. Kaur et al. predict chronic kidney disease using various data mining algorithms in a Hadoop environment, including classifiers like K-Nearest Neighbor (KNN) and Support Vector Machine (SVM).

In [8], Nusrat Tazin et al. develop a classification model for predicting the transitional stage of Kidney disease from stages 3 to 5. They use Decision trees, K-nearest neighbor, Naïve Bayes, and Artificial neural networks to create a classification model with a selected set of attributes.

[6] discusses the role of creatinine, sodium, and urea levels in blood in predicting patient survival or the need for kidney transplantation during dialysis. V. Ravindra et al. employ the K-means algorithm to uncover relationships between these CKD parameters and patient survival, demonstrating that clustering predicts patient survival during dialysis.

In [2], R. Devika et al. examine the performance of Naïve Bayes, K-Nearest Neighbor (KNN), and Random Forest classifiers for CKD prediction based on accuracy, precision, and execution time.

In [4], the paper addresses the chronic kidney damage caused by diabetes, a slow but significant process. High blood glucose levels disrupt kidney function. Bharathi et al. apply association rule mining to predict diabetes mellitus in a dataset by generating summarization rules.

III. CKD PREDICTION USING MACHINE LEARNING MODELS

Knowledge discovery is an important application of datamining which involves various stages of processing. The application of datamining algorithms are facilitated by preprocessing the data collected from multiple sources. Data preperation or preprocessing involves cleaning, extracting and transforming data to suitable formats. The key factors of knowledge representation are identified from a larger feature set. Later various classification or pattern evaluation algorithms are applied for knowledge discovery. Three machine learning algorithms namely Decision tree, Random Forest and Support Vector machines are used to predict the early occurence of CKD. The goodness of each algorithm is analysed. The model with high accuracy is derived from the below process. The system architecture is given in Figure 1.

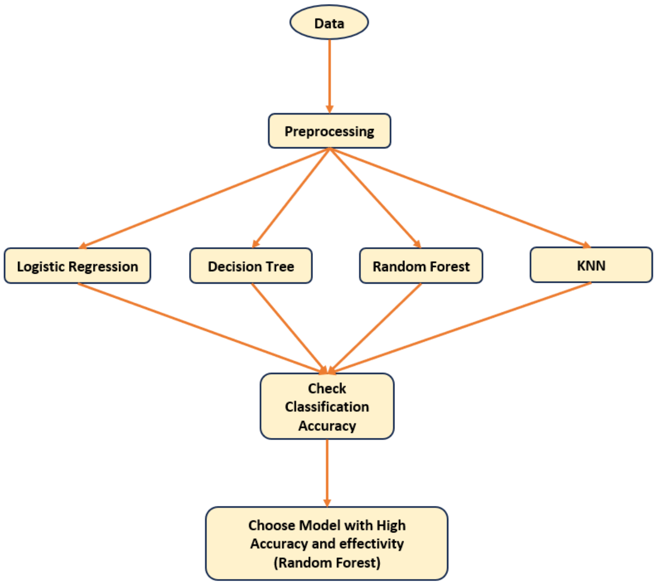

Data
Preprocessing
Logistic Regression
Decision Tree
Random Forest
KNN
Check Classification Accuracy
Choose Model with High Accuracy and effectivity (Random Forest)

Fig. 1. CKD Prediction Using Machine Learning Models

A. *LogisticRegression Classifier:*

Logistic Regression is a popular machine learning algorithm used for binary classification tasks. It predicts the probability of an input belonging to one of two possible classes, typically denoted as 0 or 1. The model is trained on labeled data to learn the relationship between input features and the binary outcome. It's widely used for tasks such as spam detection, medical diagnosis, and sentiment analysis due to its simplicity and interpretability. The model's output is a probability score that can be used to make binary predictions.

B. *Random Forest Classifier:*

Random forest algorithm constructs multiple decision trees to act as an ensemble of classification and regression process. A number of decision trees are constructed using a random subsets of the training data sets. A large collection of decision trees provide higher accuracy of results. The runtime of the algorithm is comparatively fast and also accommodates missing data. Random forest randomizes the algorithm and not the training data set. The decision class is the mode of classes generated by decision trees.

C. *DecisionTree Classifier:*

A Decision Tree classifier is a popular machine learning algorithm used for both classification and regression tasks. It is a predictive modeling tool that learns to make decisions by partitioning the input feature space into regions and assigning a class label to each region. Decision Trees are known for their simplicity and interpretability, making them valuable in various domains. Here is a brief description of the Decision Tree Classifier:

D. SVC Classifier

The Support Vector Machine (SVM) classifier is a powerful and versatile machine learning algorithm used for both binary and multi-class classification tasks. It is known for its ability to find the optimal hyperplane that maximally separates different classes in the feature space. SVM classifiers are versatile and offer a robust approach to classification problems, particularly when data separation is not straightforward. They are valued for their ability to handle both linear and non-linear classification tasks effectively.

*E. **KNN Classifier:***

KNN (K-Nearest Neighbors) Classifier is a simple and effective algorithm for classification tasks in machine learning. It works by finding the k nearest neighbors to a new data point in the training dataset and assigning the class that is most common among those neighbors as the prediction for the new data point. The value of k is a hyperparameter that can be tuned to optimize performance. KNN is a non-parametric algorithm and can work well with small datasets, but its performance can degrade with high-dimensional data.

IV. DATA SET

The Chronic Kidney Disease dataset is a well-known dataset used in machine learning and data mining research. It contains data on patients with chronic kidney disease, including demographic information, laboratory test results, and diagnosis information.

The dataset consists of dataset there are 400 rows and 25 columns which means a total of 9600 data. The input features include age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, bacteria, glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes, coronary artery disease, appetite, pedal edema, anemia, and classification.

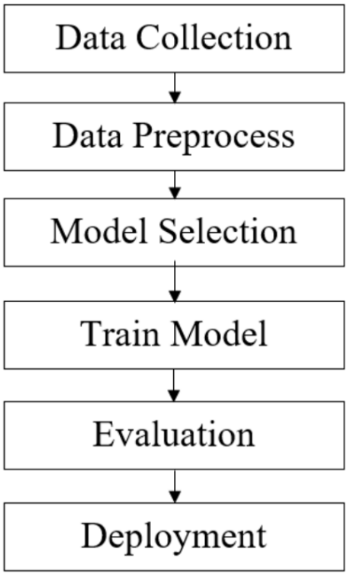
Here, we don't need all data for CKD (Chronic Kidney Disease) prediction. We only use the most important 8 features for CKD (Chronic Kidney Disease) prediction, which is given below in Figure 2:

V. WORKING PROCEDURE

Here At first we take input Kidney data.csv file as a dataset; In output we use the most accurate and efficient machine learning algorithm for predicting CKD;

-

Step 1: Input data
Step 2: Preprocess the data
Step 2.1: Convert Categorical values to numerical values
Step 2.2: Replace numerical missing values by Mean
Step 2.3: Replace Categorical missing values by Mode
Step 3: Construct Random Forest Model.
Step 3.1: Construct Decision Tree Model.
Step 3.2: Construct KNN (K-Nearest Neighbors) Model.
Step 4 : Check the accuracy of the constructed models using confusion matrix.



After measuring the Accuracy of these models in the result section, we show the accuracy and confusion matrix for all measured algorithms, then we take Random Forest Classifier Model for prediction CKD in this system based on higher accuracy and efficiency.

VI. RESULTS AND DISCUSSION

The models have been constructed using 70% training data of the original CKD dataset. Constructed models have been validated using test data, which is 30

The comparative analysis of classification algorithms is done based on the performance factors of classification accuracy, precision, and F1-score. The classification algorithm is applied, and the results are based on the following terms:

1) True positives (TP): These are the cases in which CKD is predicted (they have the disease).
2) True negatives (TN): If predicted no-CKD, and they don't have the disease.
3) False positives (FP): If predicted CKD, but they don't have the disease.
4) False negatives (FN): If predicted no-CKD, but they do have the disease.
5) Confusion Matrix: A summary of prediction results on a classification problem. The following table shows the confusion matrix obtained for each classification model.
6) Accuracy: Accuracy is decomposed as the fraction of unerringly classified records and the total number of records present in the dataset.

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

(i) Precision: Precision is the capability of a classification model to return only pertinent instances. It is the number of units correctly predicted as faulty.

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

(ii) Recall: It is the capability of a classification model to identify all pertinent instances. It is also called as True Positive Rate (TPR).

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

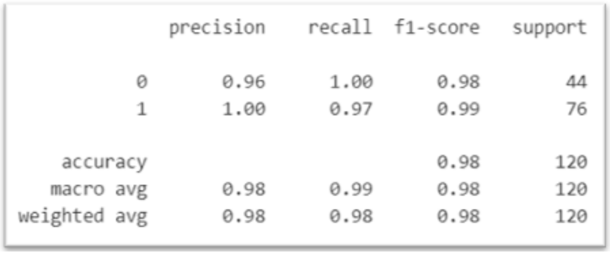
(iii) F-measure: It is a single bar that amalgamates recall and precision using the harmonic mean.

$$F1 = 2* \frac{Precision * recall}{precision + recall}$$

(iv) Receiver Operating Characteristic (ROC) curve: A Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model. It is commonly used in machine learning to evaluate and compare the performance of different models. The ROC curve provides a useful tool for evaluating and comparing the performance of binary classification models and optimizing the trade-off between false positives and false negatives.
ROC curve: It maps the true positive rate (TPR) against the false positive rate (FPR) as a function of the model's threshold for classifying a positive.
The ROC curve provides a useful tool for evaluating and comparing the performance of binary classification models and optimizing the trade-off between false positives and false negatives.

A. *Accuracy of Logistic Regression :*

The Accuracy of Random Forest Classifier is 0.985 that means 98.5%. Here is the confusion matrix of Random Forest in Figure 4:



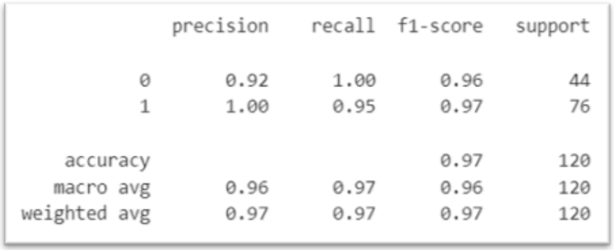
```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        44
           1       1.00      0.97      0.99        76

    accuracy                           0.98       120
   macro avg       0.98      0.99      0.98       120
weighted avg       0.98      0.98      0.98       120
```

B. *Accuracy of Decision Tree:*

The Accuracy of Decision Tree Classifier is 0.99 that means 99%.Here is the confusion matrix of Decision Tree Classifier in Figure 5:



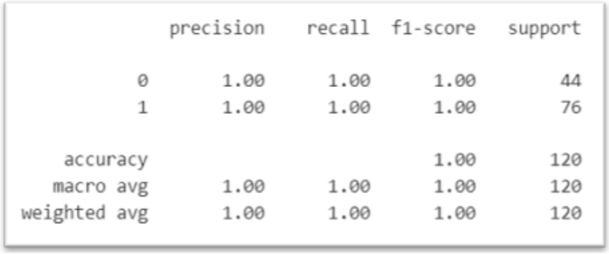
```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        44
           1       1.00      0.95      0.97        76

    accuracy                           0.97       120
   macro avg       0.96      0.97      0.96       120
weighted avg       0.97      0.97      0.97       120
```

C. *Accuracy of Random Forest :*

The Accuracy of Random Forest Classifier is 1.00 that means 100%.Here is the confusion matrix of Random Forest Classifier in Figure 5:



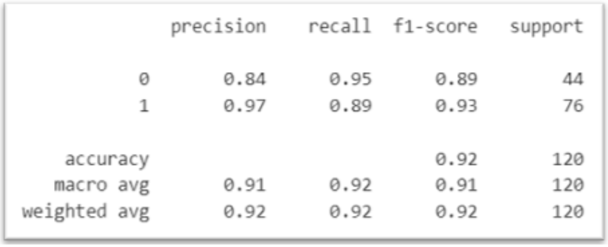
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        44
           1       1.00      1.00      1.00        76

    accuracy                           1.00       120
   macro avg       1.00      1.00      1.00       120
weighted avg       1.00      1.00      1.00       120
```

D. *Accuracy of KNN :*

The Accuracy of Decision Tree Classifier is 0.925 that means 92.5%. Here is the confusion matrix of KNN Classifier in Figure 7:



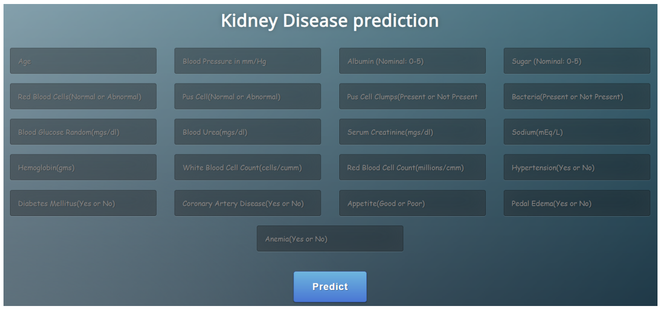
```
              precision    recall  f1-score   support

           0       0.84      0.95      0.89        44
           1       0.97      0.89      0.93        76

    accuracy                           0.92       120
   macro avg       0.91      0.92      0.91       120
weighted avg       0.92      0.92      0.92       120
```

E. *Prediction Result:*

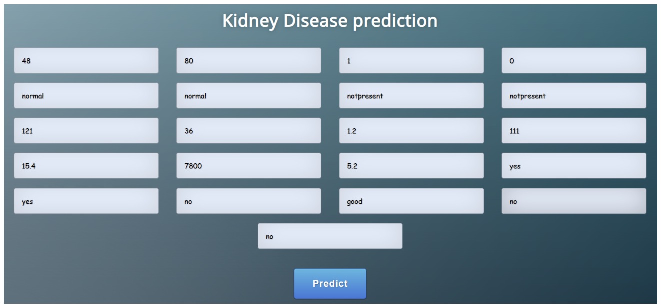
In our system we use Random Forest Classifier model because It can perform both regression and classification tasks. A random forest produces good predictions that can be understood easily. It can handle large datasets efficiently.Random forest improves on bagging because it decorrelates the trees with the introduction of splitting on a random subset of features. After selecting Random Forest Classifier model, we can start predicting a patient have Chronic Kidney Disease or not.Here is the example of prediction:
At first when we want to predict any patient's disease,then run the app.py code and it provide a IP address http://127.0.0.1:5000/ then show a web page;



Age	Blood Pressure in mm/Hg	Albumin (Nominal: 0-5)	Sugar (Nominal: 0-5)
Red Blood Cells(Normal or Abnormal)	Pus Cell(Normal or Abnormal)	Pus Cell Clumps(Present or Not Present	Bacteria(Present or Not Present)
Blood Glucose Random(mgs/dl)	Blood Urea(mgs/dl)	Serum Creatinine(mgs/dl)	Sodium(mEq/L)
Hemoglobin(gms)	White Blood Cell Count(cells/cumm)	Red Blood Cell Count(millions/cmm)	Hypertension(Yes or No)
Diabetes Mellitus(Yes or No)	Coronary Artery Disease(Yes or No)	Appetite(Good or Poor)	Pedal Edema(Yes or No)
	Anemia(Yes or No)		

Fig. 2. First page when we open http://127.0.0.1:5000/

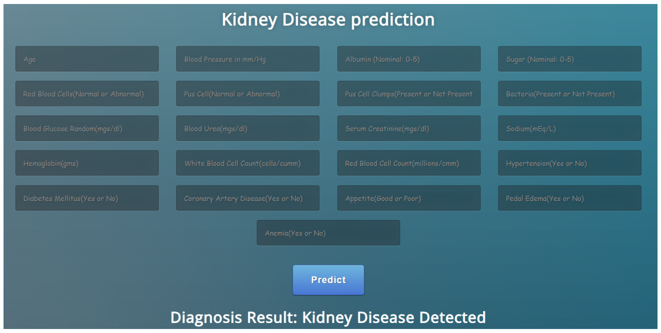
Now we have to fill all of the required field of patient's information:
1. Firstly, we have to insert data (such as Age, Blood Pressure, Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell Clumps, Bacteria, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, Red Blood Cell Count, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia) in all required field of the form for predicting CKD (Chronic Kidney Disease):



48	80	1	0
normal	normal	notpresent	notpresent
121	36	1.2	111
15.4	7800	5.2	yes
yes	no	good	no
	no		

Fig. 3. Information of CKD patient

2. Here, we insert all required data(Age, Blood Pressure, Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell Clumps, Bacteria, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, Red Blood Cell Count, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia) in the form to see the prediction result.



Age	Blood Pressure in mm/Hg	Albumin (Nominal: 0-5)	Sugar (Nominal: 0-5)
Red Blood Cells(Normal or Abnormal)	Pus Cell(Normal or Abnormal)	Pus Cell Clumps(Present or Not Present	Bacteria(Present or Not Present)
Blood Glucose Random(mgs/dl)	Blood Urea(mgs/dl)	Serum Creatinine(mgs/dl)	Sodium(mEq/L)
Hemoglobin(gms)	White Blood Cell Count(cells/cumm)	Red Blood Cell Count(millions/cmm)	Hypertension(Yes or No)
Diabetes Mellitus(Yes or No)	Coronary Artery Disease(Yes or No)	Appetite(Good or Poor)	Pedal Edema(Yes or No)
	Anemia(Yes or No)		

Diagnosis Result: Kidney Disease Detected

Fig. 4. When Press the Predict button

After given all required data, we press the submit button, the system shows the output of CKD (Chronic Kidney Disease) that is patient have chronic kidney disease or not.

Diagnosis Result: Kidney Disease Detected

Fig. 5. Result looks Like This

It is trained using labeled data, learning patterns and relationships between input parameters and output labels.The input features are passed through the trained model's algorithms or equations. The model generates a predicted output value based on the input features. The predicted output represents the model's estimation or classification for the given input. This prediction process allows your model to make accurate predictions based on the learned patterns from the training data.

VII. FUTURE SCOPE

In the future, the field of chronic kidney disease (CKD) prediction holds immense potential for advancements that can significantly improve patient care and public health outcomes. Several key directions can be explored to enhance the accuracy and effectiveness of CKD prediction models while ensuring transparency and integration with healthcare systems.

One important avenue for future research is the refinement of feature selection techniques. By identifying and incorporating the most relevant and informative features from the dataset, the prediction models can be further optimized for accurate CKD diagnosis.

This involves exploring advanced feature selection algorithms and techniques that consider both clinical and genetic factors, enabling a more comprehensive understanding of CKD risk factors.

Furthermore, the utilization of advanced machine learning algorithms can offer new opportunities for improving CKD prediction models. Algorithms such as deep learning, ensemble methods, and hybrid models can be explored to harness the full potential of the available data and extract complex patterns and relationships. By leveraging these advanced techniques, the accuracy and performance of CKD prediction algorithms can be enhanced.

VIII. CONCLUSION

In this research paper, we have presented an algorithm for predicting Chronic Kidney Disease (CKD) at an early stage. The dataset used in this study contains input parameters collected from CKD patients, and various models have been trained and validated using these parameters. Specifically, Random Forest Classifier models have been constructed to carry out the diagnosis of CKD. The performance of these models has been evaluated based on their accuracy in predicting CKD. Through rigorous experimentation and analysis, the results of our research have demonstrated that the Random Forest Classifier model outperforms other models in predicting CKD. This indicates that the Random Forest approach is particularly effective for early detection and diagnosis of CKD. In addition to accuracy, further comparisons can be made based on other factors such as the execution time and the selection of the feature set. These aspects are crucial for evaluating the practicality and efficiency of the prediction algorithm. By considering these factors, we can further enhance the potential application of the algorithm in real-world scenarios.

IX. ACKNOWLEDGMENT

We would like to thanks our honorable teacher Nurjahan Nipa Mam, for their patient instruction, passionate support, and constructive criticisms of this study effort. I'd want to thank again sir and mam for guidance and help in keeping our development on track and for giving us all the help and guidance we needed.We are really grateful to them for their kind support.Their valuable suggestions were very helpful.

References

[1] Ahmed J Aljaaf, Dhiya Al-Jumeily, Hussein M Haglan, Mohamed Alloghani, Thar Baker, Abir J Hussain, and Jamila Mustafina, ”Early prediction of chronic kidney disease using machine learning supported by predictive analytics,” in 2018 IEEE congress on evolutionary computation (CEC), pages 1–9. IEEE, 2018.

[2] R Devika, Sai Vaishnavi Avilala, and V Subramaniyaswamy, ”Comparative study of classifier for chronic kidney disease prediction using naive bayes, knn and random forest,” in 2019 3rd International conference on computing methodologies and communication (ICCMC), pages 679–684. IEEE, 2019.

[3] Uma N Dulhare and Mohammad Ayesha, ”Extraction of action rules for chronic kidney disease using na¨ıve bayes classifier,” in 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pages 1–5. IEEE, 2016.

[4] Murari Devakannan Kamalesh, K Hema Prasanna, B Bharathi, R Dhanalakshmi, and R Aroul Canessane, ”Predicting the risk of diabetes mellitus to subpopulations using association rule mining,” in Proceedings of the International Conference on Soft Computing Systems: ICSCS 2015, Volume 1, pages 59–65. Springer, 2016.

[5] Guneet Kaur and Ajay Sharma, ”Predict chronic kidney disease using data mining algorithms in hadoop,” in 2017 international conference on inventive computing and informatics (ICICI), pages 973–979. IEEE, 2017.

[6] BV Ravindra, N Sriraam, and M Geetha, ”Discovery of significant parameters in kidney dialysis data sets by k-means algorithm,” in International Conference on Circuits, Communication, Control and Computing, pages 452–454. IEEE, 2014.

[7] Shamim H Ripon et al, ”Rule induction and prediction of chronic kidney disease using boosting classifiers, ant-miner and j48 decision tree,” in 2019 international conference on electrical, computer and communication engineering (ECCE), pages 1–6. IEEE, 2019.

[8] Nusrat Tazin, Shahed Anzarus Sabab, and Muhammed Tawfiq Chowdhury, ”Diagnosis of chronic kidney disease using effective classification and feature selection technique,” in 2016 international conference on medical engineering, health informatics and technology (MediTec), pages 1–6. IEEE, 2016.

[9] Helmie Arif Wibawa, Indra Malik, and Nurdin Bahtiar, ”Evaluation of kernel-based extreme learning machine performance for prediction of chronic kidney disease,” in 2018 2nd International conference on informatics and computational sciences (ICICoS), pages 1–4. IEEE, 2018.

[10] MPNM Wickramasinghe, DM Perera, and KADCP Kahandawaarachchi, ”Dietary prediction for patients with chronic kidney disease (CKD) by considering blood potassium level using machine learning algorithms,” in 2017 IEEE Life Sciences Conference (LSC), pages 300–303. IEEE, 2017.

[11] Hanyu Zhang, Che-Lun Hung, William Cheng-Chung Chu, Ping-Fang Chiu, and Chuan Yi Tang, ”Chronic kidney disease survival prediction with artificial neural networks,” in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1351–1356. IEEE, 2018.