Kidney Disease Prediction

Mariam Hany
Department of systems and biomedical
Cairo university faculty of engineering

Carol Emad
Department of systems and biomedical
Cairo university faculty of engineering

Mohamed Mesilhy
Department of systems and biomedical
Cairo university faculty of engineering

Mohammed Ali A
Department of systems and biomedical
Cairo university faculty of engineering

***Abstract*— Chronic kidney disease (CKD) is a widespread health issue affecting millions of individuals worldwide. Early detection of CKD is crucial for initiating timely interventions and improving patient outcomes. In this study, we aim to develop a robust machine learning model for accurately predicting the presence of CKD. By leveraging a diverse set of clinical parameters and diagnostic measurements, including blood pressure, albumin levels, serum creatinine, and other important biomarkers, our model provides a comprehensive assessment of renal health status**

We employed multiple machine learning techniques, including decision trees, logistic regression, and support vector machines using grid search optimization to get the best hyper parameters to build predictive models for CKD. The proposed models were trained and evaluated on a dataset containing the relevant clinical and diagnostic features for CKD prediction. The results demonstrate the promising performance of the developed machine learning models in accurately predicting the presence of CKD. The decision tree classifier achieved a training accuracy of 99.69% and a testing accuracy of 100%, with perfect precision and recall. The logistic regression model and support vector machine model also achieved a testing accuracy of 100%, with equally impressive precision and recall metrics.

Keywords*— *Chronic kidney disease, machine learning, prediction model, clinical parameters, Decision tree, Logistic regression Support vector machine,Grid search optimization

I. Introduction

Chronic kidney disease (CKD) is a prevalent and serious health condition affecting millions of individuals worldwide . Early detection and management of CKD are crucial for improving patient outcomes and reducing the risk of complications, such as end-stage renal disease, cardiovascular events, and mortality. Identifying individuals at risk of CKD at an early stage can enable timely interventions, leading to better treatment strategies and improved quality of life for patients.

In this study, we aim to develop good machine learning models for accurately predicting the presence of chronic kidney disease. By using a diverse set of clinical parameters and diagnostic measurements, our models provide a comprehensive assessment of renal health status. The input features in our dataset include blood pressure, albumin levels, serum creatinine, and other important biomarkers that are commonly used in the diagnosis and monitoring of kidney function. The output of our models is a binary classification of individuals as either having kidney disease or not.

Our motivation for pursuing this problem stems from the significant impact that early CKD detection can have on patient outcomes and prevent them from death. However, due to the availability of various diagnostic tools, many cases of CKD remain undiagnosed, leading to delayed treatment and increased risk of adverse health outcomes . By developing accurate machine learning models, we aim to contribute to the advancement of diagnostic tools for kidney disease, ultimately assisting healthcare professionals in identifying individuals at risk and enabling prompt intervention.

The background and significance of this research are well-established in the literature. Chronic kidney disease is a global public health concern, affecting an estimated 850 million people worldwide . Early detection and management of CKD can slow disease progression, reduce the risk of cardiovascular complications, and improve overall patient prognosis . In this context, the application of machine learning techniques to the prediction of CKD holds great promise for enhancing clinical decision-making and improving patient care.

II. RELATED WORK

Approaches Using Classification Algorithms:

M.P.N.M. Wickramasinghe et al. [1]: This work used various classification algorithms like Multiclass Decision Jungle, Multiclass Decision Forest, Multiclass Neural Network, and Multiclass Logistic Regression to predict an allowable potassium zone based on blood potassium levels and recommend a diet plan.It’s strengths is Provides a systematic methodology to manage CKD using diet recommendations based on predicted potassium levels. it’s weaknesses is Evaluated only on a single dataset, needs validation on more diverse patient populations.

H.A. Wibawa et al. [2]: Proposed and evaluated kernel-based Extreme Learning Machine (ELM) models like RBF-ELM, Linear-ELM, Polynomial-ELM, Wavelet-ELM for CKD prediction. Its strengths is Compared the performance of different kernel-based ELM models and showed RBF-ELM had the highest prediction accuracy. Its weaknesses is Limited to evaluating only ELM-based models, could compare with other popular classification algorithms as well.

Approaches Using Survival Analysis and Prognostic Modeling:

4. H. Zhang et al. [3]: Investigated the performance of Artificial Neural Network (ANN) models for survivability prediction of CKD patients. It's strengths is Focusing on an important aspect of predicting survival time for CKD patients and It's weaknesses is Considering only ANN, could compare with other survival analysis techniques like Cox regression.

J. Aljaaf et al. [4]: Concluded that machine learning with predictive analytics is an intelligent solution for early CKD prediction. It's strengths: Highlights the potential of data-driven predictive modeling for CKD and It's weaknesses is Does not provide details of the specific models and their performance evaluation.

Approaches Using Ensemble Methods and Rule Extraction: Arif-Ul-Islam et al. [5]: Analyzed the performance of boosting algorithms (AdaBoost, LogitBoost) and used Ant-Miner with Decision Trees to derive rules for CKD detection.It's strengths is Exploring the benefits of ensemble methods and rule-based interpretability and It's weaknesses is Focusing only on a limited set of algorithms, could expand the analysis to other ensemble techniques.

G. Kaur et al. [6]: Predicted CKD using KNN and SVM classifiers in a Hadoop environment, highlighting the role of big data.it's strengths is Demonstrates the application of data mining in a big data framework for prediction and It's weaknesses is Could explore more advanced big data techniques beyond just KNN and SVM.

the state-of-the-art in CKD prediction and management using data mining and machine learning includes a diverse set of approaches such as Classification algorithms [1]for predicting disease stages, potassium levels, and recommending dietary interventions.Survival analysis[7] and prognostic modeling to estimate patient outcomes and support clinical decision-making Ensemble methods and rule extraction techniques[8] to improve prediction performance and provide interpretable insights .Leveraging big data technologies to handle large-scale CKD datasets.The strengths of these approaches lie in their ability to provide automated, data-driven tools for early CKD detection, disease progression monitoring, and personalized management. The weaknesses include the need for more comprehensive evaluations, comparisons across a wider range of algorithms, and validation on diverse patient populations. Integrating these techniques into clinical workflows and assessing their real-world impact remains an important area for further research.

III. DATASET AND FEATURES

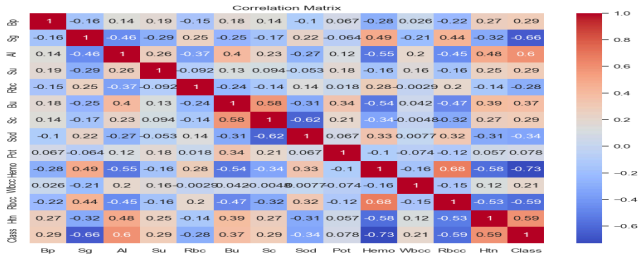
It is possible to use machine learning to make predictions about chronic renal disease by downloading a dataset from the Kaggle competition. The dataset contained information on a total of 400 different patients' records. The ages of the people involved, bacteria, serum creatinine, white blood cell count, potassium, albumin, and red blood cell count are also included on the list of 14 factors. Patients frequently exhibit erratic and unpredictable patterns regarding their blood glucose and urea levels, as well as their classification, appetite, and packed cell volume. Diabetes and high blood pressure are the two primary contributors to chronic kidney disease (CKD) [9]. We should prepare ourselves for high blood sugar levels as a natural consequence of the damage that diabetes causes to our many organs. It is of the utmost importance that the patient's condition is ascertained as quickly as possible. Within the scope of this study, several different approaches to machine learning were modified to forecast the illness. The following table of subset of our dataset used .

	Bp	Sg	Al	Su	Rbc	Bu	Sc	Sod	Pot	Hemo	Wbcc	Rbcc	Htn	Class
0	80.0	1.020	1.0	0.0	1.0	36.0	1.2	137.53	4.63	15.4	7800.0	5.20	1.0	1
1	50.0	1.020	4.0	0.0	1.0	18.0	0.8	137.53	4.63	11.3	6000.0	4.71	0.0	1
2	80.0	1.010	2.0	3.0	1.0	53.0	1.8	137.53	4.63	9.6	7500.0	4.71	0.0	1
3	70.0	1.005	4.0	0.0	1.0	56.0	3.8	111.00	2.50	11.2	6700.0	3.90	1.0	1
4	80.0	1.010	2.0	0.0	1.0	26.0	1.4	137.53	4.63	11.6	7300.0	4.60	0.0	1
5	90.0	1.015	3.0	0.0	1.0	25.0	1.1	142.00	3.20	12.2	7800.0	4.40	1.0	1
6	70.0	1.010	0.0	0.0	1.0	54.0	24.0	104.00	4.00	12.4	8406.0	4.71	0.0	1
7	76.0	1.015	2.0	4.0	1.0	31.0	1.1	137.53	4.63	12.4	6900.0	5.00	0.0	1
8	100.0	1.015	3.0	0.0	1.0	60.0	1.9	137.53	4.63	10.8	9600.0	4.00	1.0	1
9	90.0	1.020	2.0	0.0	0.0	107.0	7.2	114.00	3.70	9.5	12100.0	3.70	1.0	1

Fig:ChIronic Kidney Disease Dataset

A. Data preProcessing

To prepare the dataset for analysis, several preprocessing steps were performed. The details of each step are outlined below. Data Cleaning[1] which is The dataset was checked for missing values and duplicates. Fortunately, no missing values or duplicates were found, ensuring the integrity of the dataset. Handling Outliers[2] as Outliers in the continuous value columns were addressed to prevent their undue influence on the analysis. Interquartile Range (IQR) method was performed but the resultant accuracy dropped as the data is not as that big to drop data from ;it was observed that overfitting takes place by dropping outliers as the accuracy changes significantly each time the data model is trained. Feature Selection[3]Which is an important step to ensure that only the most informative variables are included in the model development process. Correlation matrix was plotted to observe the correlation between features and it was observed that all the features are relevant to the problem.



Categorical Variable Handling[4] in the dataset were investigated for inconsistencies and visualized to gain insights into their distributions. One categorical variable, 'Htn' (hypertension), had an unexpected value of 0.37, which was deemed inconsistent. Rows with this value in the 'Htn' column were removed to maintain data integrity.Addressing Class Imbalance[5] to address the class imbalance issue in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generates synthetic samples for the minority class to balance the class distribution. By resampling the data using SMOTE, we aimed to improve the performance of the predictive models. Standardization[6] to ensure that all features are on a similar scale and to prevent any bias in the modeling process, the features were standardized using the StandardScaler. This transformation scales the features to have zero mean and unit variance. After applying the *Data preProcessing* Data that has been converted into a format that a machine can understand can be comprehended quickly and easily by the machine. The term "dataset" is used to refer to a collection of individual data elements . Criteria such as the mass or time at which an event is guaranteed can be used to facilitate the identification and assurance of fundamental properties of data items. There is a good chance that the dataset contains missing values; these can either be calculated or removed. The value of the mean, median, or mode of the associated characteristics can be used to fill in the blanks when dealing with missing data . This is the most common method for dealing with missing data. A conversion from object-typed numerical numbers to float 64 values is required before analysis can be performed. When dealing with categorical attributes that contain null values, the value that appears in the attribute column the most frequently is substituted for the null value. The transformation of categorical data into numeric properties can be accomplished through the use of label encoding; this involves giving each attribute value its integer value. As a direct consequence of this, an int data type will be generated immediately. Calculations are made in advance to determine the mean values of each column, and those values are then used to fill in any gaps in the respective attribute column. It is possible to calculate the mean value for each column by utilizing the classifier function. After the data has been replaced and encoded, it needs to go through the processes of training, verification, and testing. Our algorithms acquire the knowledge necessary to construct a model through the process of learning from the data that we provide for them. The validation portion of the dataset is utilized by us to check the accuracy of the multiple model fits that we have created and to enhance the model but in case our data we cheek if there are categorical and missing values and we find that there is no any thing of these as data is cleaned in the kaggel.

IV. *Methods and Algorithm*

A method of artificial intelligence called machine learning enables learners to process information without having to be explicitly programmed. It focuses on producing computer programmers who can change in response to fresh data. It can be classified as either supervised or unsupervised . It all comes down to combining the proper characteristics to create frameworks that achieve the proper objectives. Examples of these tasks include multidimensional and multi classification, predictive clustering, and parametric modeling .
Three main steps are involved in the proposed methodology: preprocessing of the data, training of the models, and model selection (Fig. 1).

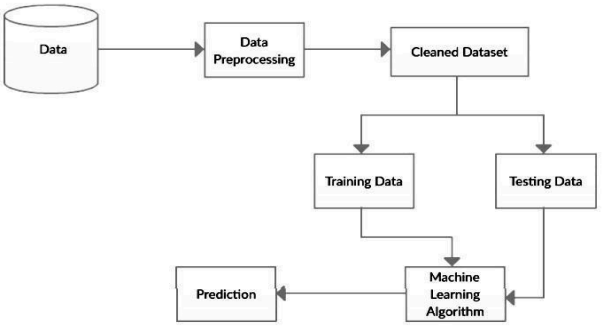


Figure 1. Proposed methodology

Classifiers:

A . *Decision Tree*

The most essential components of a decision tree are the tree's trunk, its nodes, and its branches. It is a graphical representation of a particular decision situation that is included in predictive models. In fields of medicine with a large number of factors to take into consideration, the use of decision trees has become increasingly common. Out of all the different machine learning techniques, decision trees are by far the most effective . These unmistakably reflect important facets of the data collection process that took place earlier. They also have the potential to produce the characteristic that has the greatest influence on the lives of the vast majority of people. Entropy is the foundation upon which the decision tree is constructed, and the information gained from the dataset demonstrates just how essential it is. The use of decision trees comes with a variety of drawbacks, the most notable of which is overfitting and a greedy strategy . Because it required a large number of nodes to divide the data, using a decision tree to split datasets aligned to axes led to overfitting. This was because the tree required a large number of nodes. According to J48, it is impossible to generate exponentially more trees using a dynamic approach as opposed to a greedy approach because of the practical difficulties involved .

The key mathematical equations used in decision trees include:

- **Entropy**: $H(p) = -p\log_2(p) - (1-p)\log_2(1-p)$
- **Information Gain**: $IG(D, A) = H(D) - \sum_{v \in \text{values}(A)} \frac{|D_v|}{|D|} H(D_v)$
- **Gini Impurity**: $G(p) = 1 - p^2 - (1-p)^2$
- **Mean Squared Error** (for regression): $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- **Cost Complexity Pruning**: $R_\alpha(T) = R(T) + \alpha \cdot |\text{leaves}(T)|$

B. Support Vector Machine

A linear model for classification and regression is Support Vector Machine (SVM) that can be used to solve both linear and non linear problems. The algorithm classifies data using a hyperplane. In this algorithm, each data item will be plotted as a point in n-dimensional space (where n is the number of features) with the value of each feature being the value of a particular coordinate. Classification will be performed by finding the right hyper-plane which can differentiate the two classes efficiently.

$$J(w,b) = \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y^{(i)}(w^T x^i + b)) + \lambda||w||^2$$

***C**. Logistic Regression (LG):*

Logistic Regression predicts the likelihood of a binary outcome based on input features. It calculates coefficients for these features using maximum likelihood estimation, then uses a function to convert these into probabilities between 0 and 1 which is the logistic function

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m} [y^i \, log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))]$$

V. Result and Discussion

Parameter Selection :

1. *Decision Tree:*

-Criterion: which measures quality of split to get best split value and future.Chosen Criterion are

Information Gain *which is a measure of reduction of uncertainty ,**Gini** and **entropy*** both are measure of impurity . (almost all available choices were left)

-Max depth: 3:6 as the data is not that big and won't need more that 6 levels of depth

-Splitter: The strategy used to choose the split at each node. Can be "best" to choose the best split or "random" to choose the best random split. Both were kept to avoid overfitting

-Min samples leaf: minimum number of samples required for a node to be split further.It is set to be between 1 to 7 not making it too big to control the complexity of the model and avoid overfitting.

-Min samples split:The minimum number of samples required to split an internal node. also as the min samples leaf was kept between 1:7 to avoid overfitting

-Max features:The parameter controlling the number of features to consider when looking for the best split. The chosen values were : Auto to consider all features and find the best split and None as the other options are suitable for high dimensional datasets

2. *Logistic Regression:*

-C (Regularization Strength):A smaller C value indicates stronger regularization, meaning stronger penalties for large coefficients, which can help prevent overfitting.All values were kept to keep grid search chose the appropriate one

-Penalty:parameter specifies the type of regularization penalty to use in logistic regression. It can be either 'l1' for L1 regularization (Lasso) or 'l2' for L2 regularization (Ridge).L2 only was kept as from the correlation matrix plotted at the beginning it can be seen that most of the features are relevant to the problem and L1 takes an aggressive approach by zeroing some features.

3. *SVC:As the logistic regression model.*

A. Primary Metrics:

It's used for evaluating purposes which gives deeper details than the accuracy that is extremely beneficial in such problem due to it's cricility as medical diagnosis problem . It contains serval parameters which are:

Accuracy of the constructed classifier model can be calculation using the following equation.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where,

TP = Observation is positive an predicted is also positive
TN=Observation is negative and predicted is also negative
FP = Observation is negative but predicted is positive
FN = Observation is positive but predicted is negative

2. *Precision : ratio of true positive to all positive predictions (true and false).*

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

1. *Recall: ratio of positives predictions to actual positive data samples .*

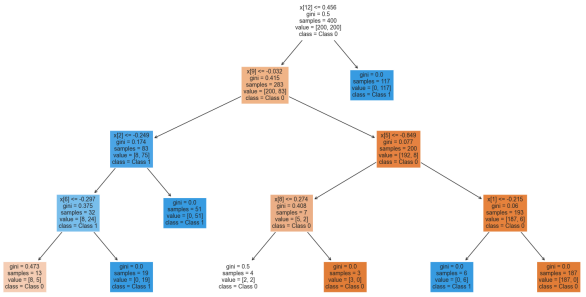
$$Recall = \frac{True\ Positives\ (TP}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

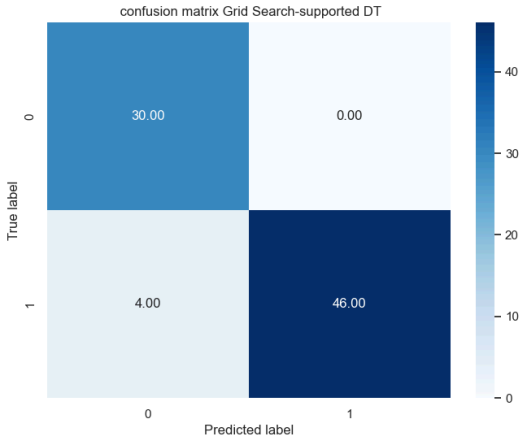
2. *F1-score:is the balance of precision and recall .*

$$f1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

B. Results with visualization :

1. Grid Search-supported Decision Tree:



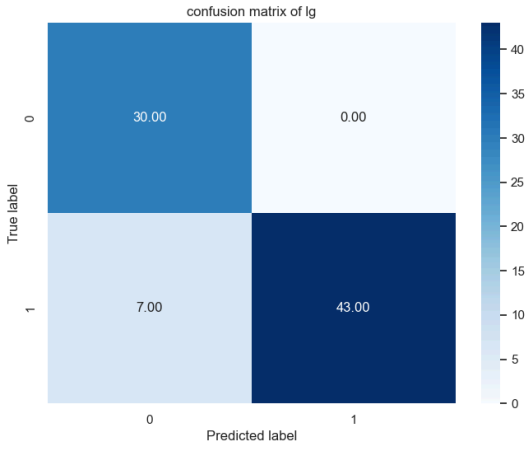



```
Training Accuracy of DTC is 0.9825
Testing Accuracy of DTC is 0.95
              precision    recall  f1-score   sup

           0       0.88      1.00      0.94
           1       1.00      0.92      0.96

    accuracy                           0.95
   macro avg       0.94      0.96      0.95
weighted avg       0.96      0.95      0.95
```

2. Grid Search-supported Logistic Regression:

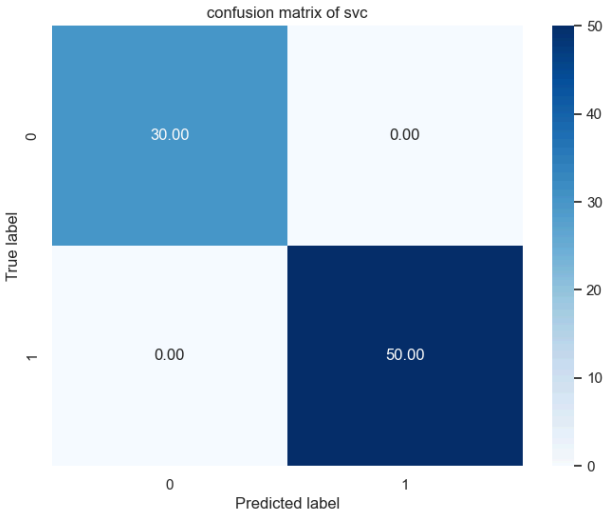



```
Training Accuracy of lg is 0.97
Testing Accuracy of lg is 0.9625
              precision    recall  f1-score   support

           0       0.81      1.00      0.90        30
           1       1.00      0.86      0.92        50

    accuracy                           0.91        80
   macro avg       0.91      0.93      0.91        80
weighted avg       0.93      0.91      0.91        80
```

2. 3.Grid Search-supported SVC:




```
Training Accuracy of SVC is 0.9875
Testing Accuracy of SVC is 1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        30
           1       1.00      1.00      1.00        50

    accuracy                           1.00        80
   macro avg       1.00      1.00      1.00        80
weighted avg       1.00      1.00      1.00        80
```

VI. Conclusions and Future Work

Methodology[1]The proposed methodology involved three main steps: data preprocessing, model training, and model selection. Several machine learning algorithms were evaluated, including Decision Trees, Support Vector Machines (SVM), and Logistic Regression.Parameter Selection[2] For each algorithm, the report explored

various hyperparameter settings to optimize the models and avoid overfitting, such as depth, splitting criteria, and regularization strength.Evaluation Metrics[3] Given the medical diagnosis context, the report focused on evaluating the models using precision, recall, and F1-score in addition to overall accuracy.Results[4] Decision Tree: The report showed grid search-optimized decision tree models with reasonable performance but with a little bit overfitting which isas predicted as the data is not complex to need such a complex model . Logistic Regression: The grid search-supported logistic regression model performed the best with reasonable training and testing accuracy and no signs of overfitting SVM: The grid search-supported SVM (SVC) exhibited an overfit as the test accuracy is much higher than the train accuracy. Comparison: While almost all three algorithms showed promising results, the logistic regression model appeared to be the highest performing with minimal overfit . This is likely because these algorithms were better able to capture the underlying relationships in the data without overfitting, compared to the SVC approach and DTC..

For future work, if more time, team members, or computational resources were available, the following could be explored:Ensemble Methods: Explore combining multiple models (e.g., stacking, bagging, or boosting) to leverage the strengths of different algorithms and potentially achieve better overall performance.Deep Learning: Given the advances in neural networks and their ability to handle complex, nonlinear relationships, experimenting with deep learning architectures (e.g., convolutional neural networks, recurrent neural networks) could be beneficial, especially for more complex medical diagnosis tasks. Model Interpretability: Investigate techniques to improve the interpretability of the models, such as feature importance analysis or using explainable AI methods. This could provide valuable insights for the medical domain and help build trust in the model's decisions. Handling Imbalanced Data: If the dataset has a significant class imbalance, explore techniques to address this, such as oversampling, undersampling, or class weighting, to ensure the models are not biased towards the majority class. External Validation: Test the models on independent, external datasets to assess their generalization capabilities and robustness in real-world clinical settings.

By exploring these areas, the team could potentially further enhance the performance, reliability, and practical applicability of the machine learning models for medical diagnosis tasks.

REFERENCES:

1. M. P. N. M. Wickramasinghe, D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," *2017 IEEE Life Sciences Conference (LSC)*, Sydney, NSW, 2017, pp. 300-303.
2. H. A. Wibawa, I. Malik and N. Bahtiar, "Evaluation of Kernel-Based Extreme Learning Machine Performance for Prediction of Chronic Kidney Disease," *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, 2018, pp. 1-4
3. U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier," *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Chennai, 2016, pp. 1-5.
4. H. Zhang, C. Hung, W. C. Chu, P. Chiu and C. Y. Tang, "Chronic Kidney Disease Survival Prediction with Artificial Neural Networks," *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, 2018, pp. 1351-1356
5. J. Aljaaf *et al*., "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," *2018 IEEE Congress on Evolutionary Computation (CEC)*, Rio de Janeiro, 2018, pp. 1-9.
6. Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-6.
7. G. Kaur and A. Sharma, "Predict chronic kidney disease using data mining algorithms in hadoop," *2017 International* Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 973-979.
8. N. Tazin, S. A. Sabab and M. T. Chowdhury, "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique," 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Dhaka, 2016, pp. 1-6.
9. V. Ravindra, N. Sriraam and M. Geetha, "Discovery of significant parameters in kidney dialysis data sets by K-means algorithm," *International Conference on Circuits, Communication, Control and Computing*, Bangalore,2014,pp.452-454.
10. R. Devika, S. V. Avilala and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 679-684.
11. P. Panwong and N. Iam-On, "Predicting transitional interval of kidney disease stages 3 to 5 using data mining method," *2016 Second Asian Conference on Defence Technology (ACDT)*, Chiang Mai, 2016, pp.

145-150.

libraries used:

- *Pandas*
- *Scikit-learn*
- *Numpy*
- *Seaborn*
- *Matplotlib*
- *Imbalanced-learn*

Contribution:

Pre-processing	Mohamed Mesilhy .
Model selection and evaluation and Preparing the paper	MOHAMMED ALI A ,Mariam Hany and C Emad .