iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50321

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Detection of Kidney Disease Using Machine Learning

T. Sruthi[1], S. Yaswanthi[2], T. Bindu Sri[3], V. Tualsi[4], R. Prasanna Kumari[5]
[1, 2, 3, 4]*U.G. Student, Department of Information Technology, GVPCEW, Vishakapatnam, India*
[5]*Associate Professor, Department of Information Technology, GVPCEW, Vishakapatnam, India*

Abstract: ***Severe Damage to the kidneys occurs as a result of chronic kidney disease (CKD). It is a widespread health issue thatis causing many people to pass up their prime years of life. Unlikeother diseases, which may be treated if detected in the early stages, 40% of people with CKD are oblivious to the disease present. In order to determine if a patient has CKD or not, blood pressure, diabetes status and other CKD related data are obtained from participants in this study. In order to solve the issue and identify the disease at an early stage, the use of various machine learning techniques including Random Forest, Support Vector Machine and Naïve Baye's techniques are suggested in this study. In this study,the CKD dataset is used to determine if a person will be affected in the near future.***

Index Terms: Machine Learning, Chronic disease, Random Forest, Support Vector Machine, Naïve Baye's

I. INTRODUCTION

Kidney disease is a serious medical condition that affects millions of people worldwide. It can be caused by a variety of factors such as diabetes, hypertension, and other genetic or environmental factors. Early detection of kidney disease is crucial for effective treatment and management. In recent years, machine learning has emerged as a powerful tool for medical diagnosis and prediction. By leveraging large datasets and advanced algorithms, machine learning models can analyze complex patterns in medical data to accurately detect and diagnose kidney disease.In this project, we aim to develop a machine learning-based approach for early detection of kidney disease. We will use a dataset of patient information including medical history, laboratory test results, and demographic information. Using this data, we will train and validate several machine learning models to predict the risk of kidney disease in patients. The performance of these models will be evaluated using various metrics such as accuracy, precision, and recall. The ultimate goal of this project is to provide clinicians with a tool that provides them with accurate results which help them to make more informed decisions and provide better care for patients with kidney disease.

II. LITERATURE SURVEY

According to a study byHamida Ilyas, Sajid Ali, Mahvish Ponum, corresponding author, Osman Hasan, Muhammad Tahir Mahmood, Mehwish Iftikhar, and Mubasher Hussain Malik [1],chronic kidney disease (CKD) using decision tree algorithms. The authors used a dataset of 400 patients with CKD and 400 healthy individuals to train and test their models. They used two different decision tree algorithms, including C4.5 and CART, and evaluated their performance using metrics such as accuracy, precision, recall, and F1 score.The literature suggests that decision tree algorithms have been widely used in medical diagnosis and have shown promising results for the detection of CKD. Decision tree algorithms are known for their interpretability and ability to handle both categorical and continuous variables.In another study by Chronic Kidney Disease Diagnosis Using Machine Learning" by Dr.Vijayaprabakaran . K, PratheekReddy.P, Puthin Kumar Reddy.T, Munnaf . K, ReddiPrasad .G[2],The literature suggests that machine learning approaches have shown promising results for the diagnosis and detection of CKD. The paper's use of multiple algorithms for comparison provides a comprehensive evaluation of their effectiveness,they evaluated their performance using metrics such as accuracy, precision, recall, and F1 score.In a study by S. Manoharan and R. Anitha[4] decision tree-based approach for the classification of chronic kidney disease (CKD) and aims to identify the most significant risk factors for CKD. The authors used a dataset of 400 patients with CKD and 200 healthy individuals to train and test their models. They used the J48 algorithm, which is an implementation of the C4.5 algorithm, to build a decision tree model and evaluated its performance using accuracy, sensitivity, specificity, and F1 score.They identified age, blood pressure, and serum creatinine levels as the most significant risk factors for CKD. The paper provides valuable insights into the risk factors for CKD and demonstrates the potential of decision tree algorithms for CKD classification.

©IJRASET: All Rights are Reserved

In a recent study by Karthikeyan, K., Soundararajan, R., &Anitha [5] in classifying CKD patients and healthy patients. The model was also able to identify the most important features for CKD diagnosis, which were serum creatinine level, blood pressure, and age.A decision support system for physicians to identify CKD patients at an early stage, which can lead to better management of the disease and improved patient outcomes.

III. RULE BASED SENTIMENT ANALYSIS

A dataset can be seen as a collection of data objects, and data pre- processing is the stage where the data is encoded so that the computer can readily analyse it. There is a possibility that the dataset contains missing values. So, the missing values must be dealt with first; they can either be calculated or removed from the dataset. Filling in missing values with the mean, median, mode, or constant value of the corresponding feature is the most typical approach to this problem. We must transform categorical data of object type into float64 type for analysis. The categorical attributes' null values are replaced with the value that appears the most frequently in that attribute column i.e. mode. The process of label encoding converts category attributes into numerical values by converting each distinct attribute value to an integer representation. The characteristics are converted to int type automatically as a result. The pandas package is helpful for data preprocessing in theseprocedures.

Feature selection is a technique that involves computationally choosing the features that have the greatest impact on the output or prediction variable. This is due to the chance in which the learning model may perform less effectively when using additional attribute columns. The classifier algorithm with feature selection improves performance and shortens the model's execution time. It isessential for any predictive modelling and involves the process of automatically or manually choosing the features that have the greatest impact on the prediction variable or output that interests you.

IV. METHODOLOGY

Early CKD diagnosis and treatment are highly preferred because they can help to avert negative outcomes. Recent years have seen a rapid increase in the use of machine learning metods for illness diagnosis and early symptom identification. Usingmachine learning classification algorithms on the dataset compiled from the medical records of those who have the condition, this study aims to predict the various stages of CKD using the same motivation. In this paper ,We have specifically used the Random Forest , Naïve Bayes and Support Vector Machine.

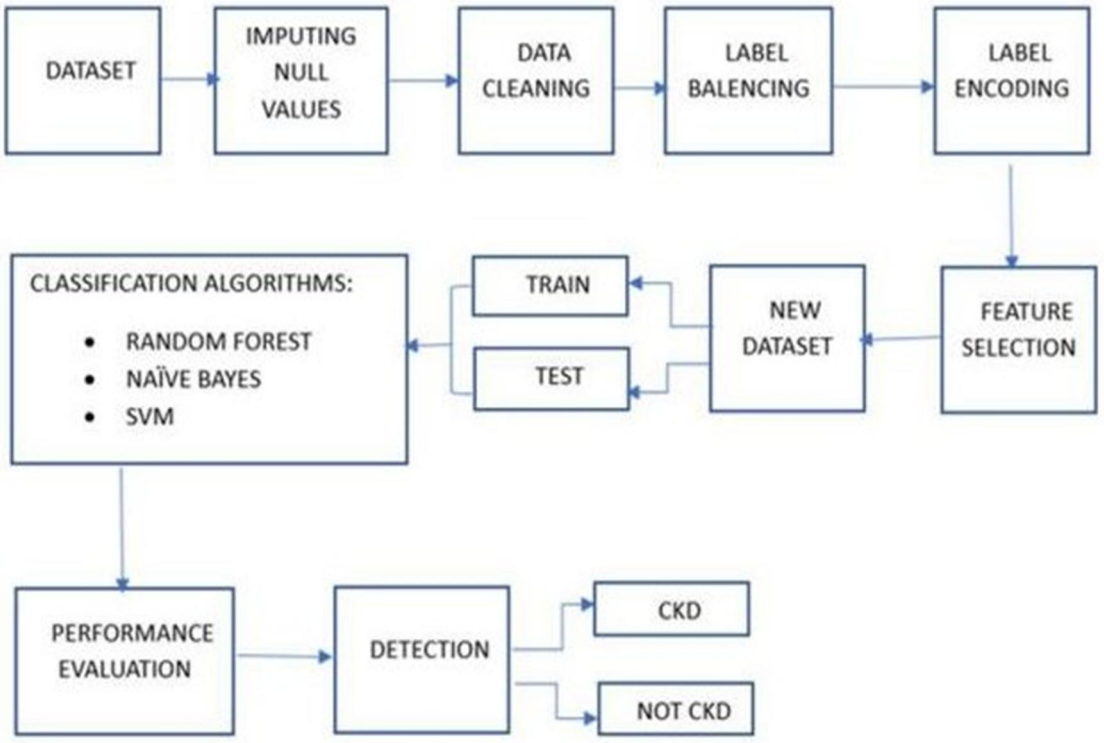


Fig i. Content Diagram of the proposed System.

A. Algorithms

1) *SVM (Support vector Machine) Algorithm:* SVM stands for Support Vector Machine, which is a supervised algorithm in machine learning which is used for classification and regressionanalysis. The main idea behind SVM is to find a hyperplane in a high- dimensional space that separates the data into different classes.
2) *Random Forest Algorithm:* The random forest is an supervised algorithm used for classification. The random forest algorithm generates decision treeson data samples provided and then finds the prediction from each of them and selects the best out of the results.
3) *Naïve Baye's Algorithm:* Naive Bayes classifiers are a set of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. all the set of features being classified are independentof each other.
4) *Bayes' Theorem:* Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge.
   It depends on the conditional probability. The formula for Bayes' theorem is given as:

P(A|B)=P(B|A)P(A) P(B) P(A|B):

(conditional probability) Probability of occurrence of event A given the event B is true P(A) and P(B): Probabilities of occurrence of event A and B respectively. P(B|A):Probability of the occurrence of event B given the event A is true.

A. Data Collection

For this project, the CKD Dataset of 400 samples with 25 qualitiesare included. 25 attributes are present, of which 11 are numerical, 13 are category, and 1 is a class attribute.

The dataset is missing some of the data values. The patient's information such as age, blood pressure, red blood cells, white blood cells,sugar , hemoglobin, and more are present in the data collected.

V. DATA PRE-PROCESSING

Data pre-processing is an important step in preparing a dataset foranalysis, as it can help improve the quality of the data and enhance the accuracy of the analysis. In the case of a kidney disease dataset, the following are some common pre-processing steps thatmay be undertaken:

1) *Data Cleaning:* This involves identifying and handling missing, incomplete, or inconsistent data points. For instance, missing values can be imputed using techniques such as mean or median imputation, or dropped altogether.
2) *Outlier Detection:* This involves identifying and handlingoutliers, which are data points that deviate significantly from the rest of the data. Outliers can be handled using techniques such asremoving them or replacing them with more appropriate values.
3) *Feature Selection:* This involves selecting the most relevant features or variables that are likely to have a significant impact on the analysis. This is achieved by making use of mutual info regression which is an statistical test which provides us with the contribution of each data.

A. Mutual Information Regression(MIR)

Mutual information (MI) regression is a machine learning technique that uses mutual information to quantify the relationship between two variables. It can be used for feature selection, dimensionality reduction, and model selection, among other things. MI regression is particularly useful in scenarios where the relationship between variables is nonlinear or complex.

1) *Data Transformation:* This involves transforming the data to amore suitable format for analysis. For example, categorical data can be transformed into numerical data using one-hot encoding orlabel encoding.

Attributes	Type
Age	Numeric
Blood Pressure	Numeric
Specific Gravity	Numeric
Albumin	Numeric
Sugar	Numeric
Red Blood Cells	Nominal
Pus Cell	Nominal
Pus Cell clumps	Nominal
Bacteria	Nominal
Blood Glucose Random	Numeric
Blood Urea	Numeric
Serum Creatinine	Numeric
Sodium	Numeric
Potassium	Numeric
Hemoglobin	Numeric
Packed Cell Volume	Numeric
Red Blood Cell count	Numeric
White Blood Cell Count	Numeric
Hypertension	Nominal
Diabetes Mellitus	Nominal
Coronary Artery Disease	Nominal
Appetite	Nominal
Pedal Edema	Nominal
Anemia	Nominal
Class	Class

Fig: List of attributes in the dataset

B. Metrics

Metrics of model performance are used to evaluate how well a model is performing on a specific task or problem. There are several common metrics that can be used, depending on the type of problemand the goals of the analysis.

1) *Accuracy:* This is the most basic metric used to evaluate a model's performance. It is simply the percentage of correctly classified instances out of all the instances.
2) *Precision:* This metric measures the proportion of truepositives (correctly predicted positives) out of all the positive predictions made by the model.
3) *Recall:* This metric measures the proportion of true positives out of all the actual positive instances in the data.
4) *F1 Score:* This is a weighted harmonic mean of precision and recall, which takes into account both measures to provide an overall score for the model's performance.

VI. COMPARISON OF ALGORITHMS

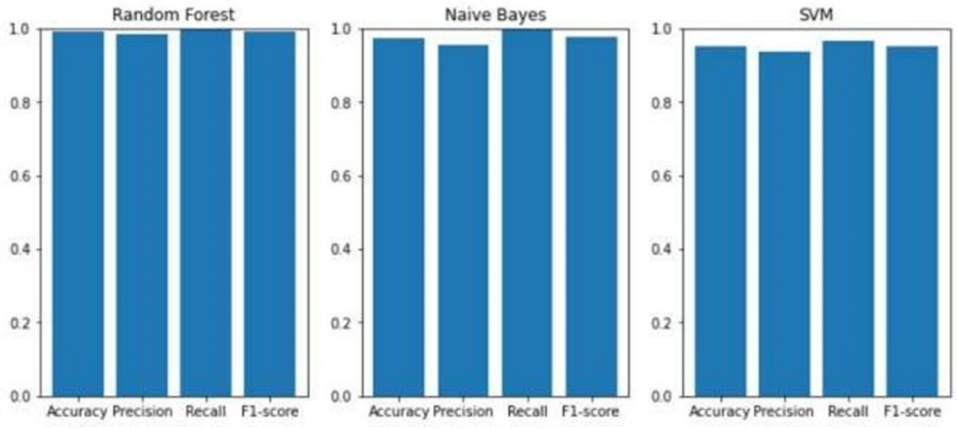


Fig ii. Bar plot for accuracy comparison for different classifiers

VII. DATAFLOW DIAGRAM

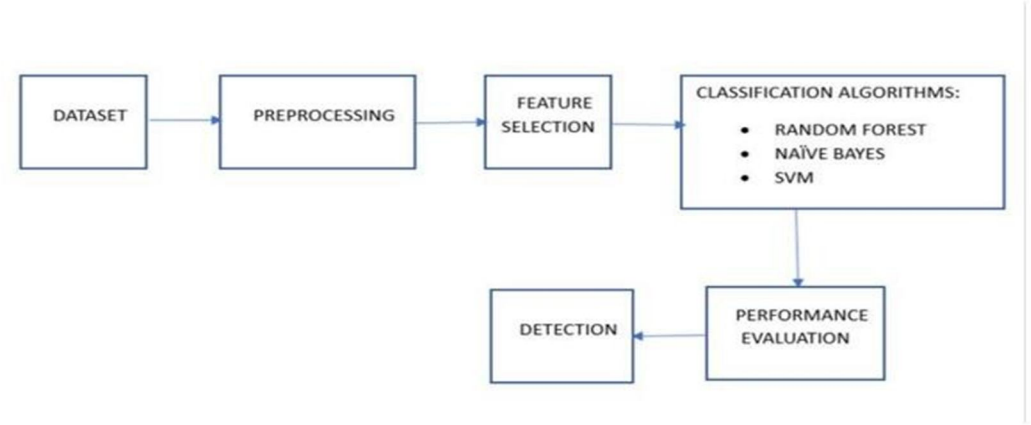
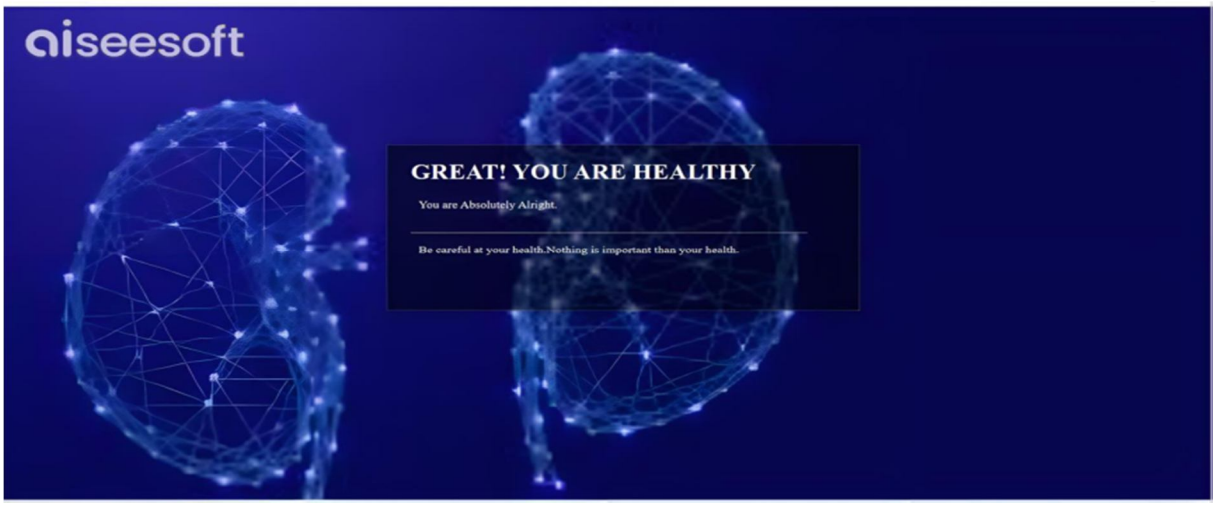


Fig iii . Operational Flow for the proposed system.

VIII. EXPERIMENTAL RESULT

Output (Detected CKD)



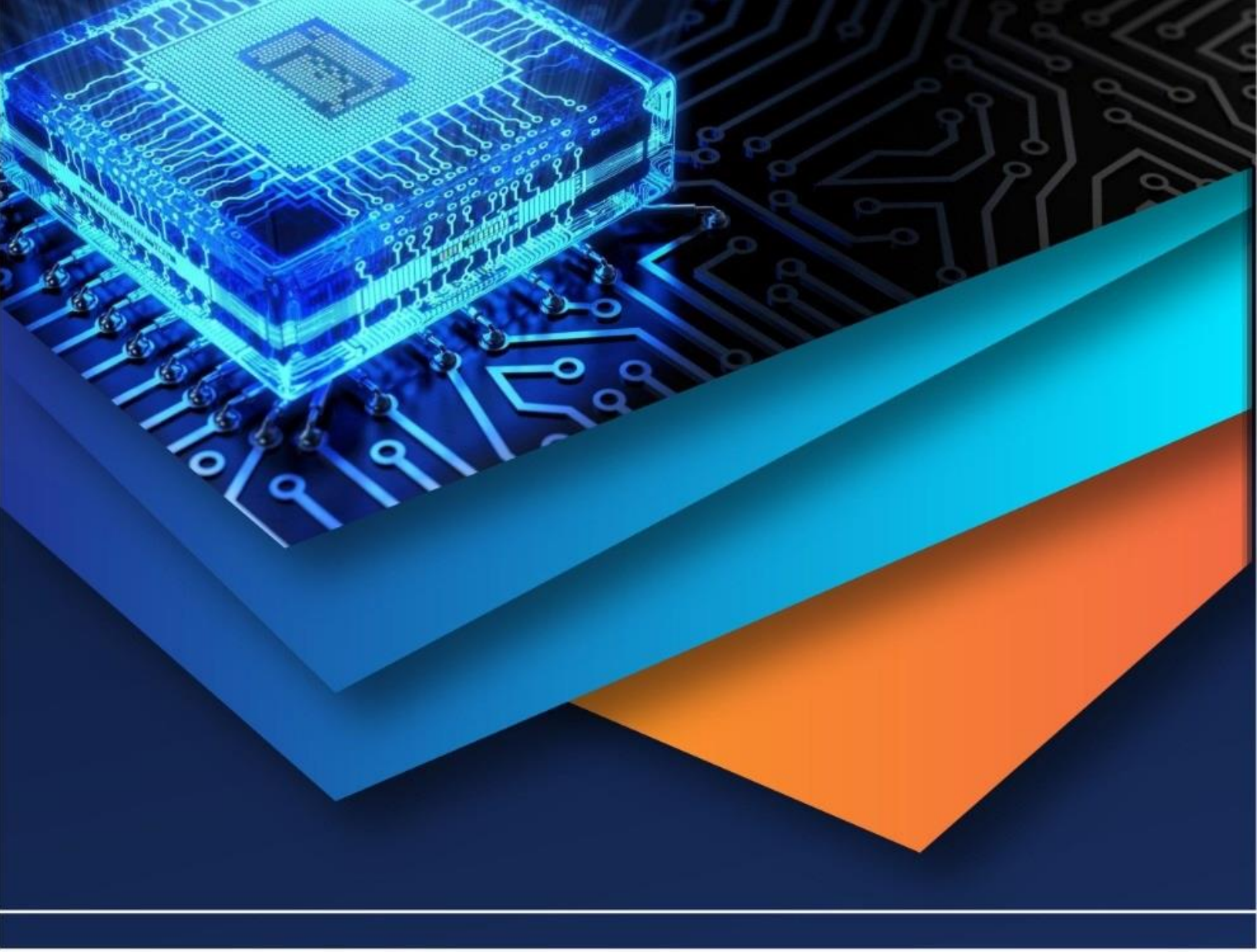
Output (No - CKD)

V CONCLUSION

Machine Learning can be beneficial in the field of medical domain.. It describes about the proposal of machine learning models to extract classification knowledge for aid of kidney diseases in clinical decision system and presents a framework of the tool various tools used for analysis. Sometimes the situation occurs when you need the doctor's help immediately, but they are not available due to some reason. In our project, we have designed a new detection system, which is an online system, and various patients from any locations can view it. Our system comprises of main components such as patient login, enter user details in the System, and, suggestions for the user. The application takes the input of various symptoms from the patient, does the analysis of the entered symptoms, and gives appropriate disease detection. Our system allows the users to get analysis on the symptoms they give for detecting the kidney disease they are suffering from. Our system provide the user with the analysis of the data regarding their symptoms and also detect is the patient is suffering from any kidney disease This paper is worked on prediction of CKD . A method is used for feature selection purpose.Out of the total 25 attirbutes only the top attributes are selected for the prediction purpose.The prediction isdone using the machine learning algorithms, Naïve Baye's , Random forest and nearest Neighbour.The main focus is to predict the condition of patient i.e. CKD OR NOT CKD using very few attributes with high accuracy rate.Here in this study we obtainedan accuracy of about 99.2 percentage.

REFERENCES

[1] Hamida Ilyas, Sajid Ali, Mahvish Ponum, corresponding author, Osman Hasan, Muhammad TahirMahmood, Mehwish Iftikhar, and Mubasher Hussain Malik, Chronic kidney disease diagnosis using decision tree algorithms, Published online 2022 Aug 9.
[2] Dr. Vijayaprabakaran . K , Pratheek Reddy.P, Puthin Kumar Reddy.T, Munnaf . K , ReddiPrasad .G, CHRONIC KIDNEY DISEASE DIAGNOSIS USING MACHINE LEARNING, International Research Journal of Engineering and Technology (IRJET),Volume: 08 Issue: 06 | June 2021.
[3] Reshma S , Salma Shaji , S R Ajina , Vishnu Priya S R, Janisha A, Chronic Kidney Disease Predictionusing Machine Learning, Volume & Issue :Volume 09,Issue o7 July 2020.
[4] Manoharan, S., & Anitha, R. (2015). Classification of Chronic Kidney Disease using Decision Tree Algorithm. International Journal of Innovative Research in Computer and Communication Engineering, 3(8),8049-8055.
[5] Karthikeyan, K., Soundararajan, R., & Anitha, R. (2015). A decision tree approach for early diagnosis of chronickidney disease. Journal of Medical Systems, 39(11), 171.
[6] Choudhary, P., & Kotecha, K. (2021). Detection of Chronic Kidney Disease using Decision Tree and KNN.International Journal of Computer Applications, 179(19), 32-36. [Link:https://www.ijcaonline.org/archives/volume179/number19/33795-2021957275]
[7] Kumar, V., & Singh, A. (2019). Comparative Analysis of Machine Learning Algorithms for Chronic Kidney Disease Prediction. International Journal of Innovative Technology and Exploring Engineering, 8(10), 1523-1530. [Link:https://www.ijitee.org/wpcontent/uploads/papers/v8i10/J45430681019.pdf]
[8] Moeini, M., & Mohammadi, M. (2019). Early Diagnosis of Chronic Kidney Disease using KNN and SVMClassification. International Journal of Health System and Disaster Management, 7(3), 110-115. [Link:http://www.ijhsdm.org/article_79181_0.html]
[9] Nourani, M., Rastgoo, M. N., & Ghaffari, A. (2019). A Comparative Study of Machine Learning Algorithms for Diagnosis of Chronic Kidney Disease. International Journal of Intelligent Systems and Applications in Engineering,7(1),15-23. [Link:http://www.insaeng.org/article_90150_0.html]
[10] Prasad, K. S., & Rajput, R. S. (2021). Machine Learning Models for Prediction of Chronic Kidney Disease.International Journal of Innovative Research in Science, Engineering and Technology, 10(5), 4295-4301. [Link:https://www.ijirset.com/upload/2021/may/428_IJIRSET_K.S.%20Prasad.pdf]
[11] Arora, A., Mishra, M., & Tiwari, P. (2021). Performance Analysis of Machine Learning Algorithms for ChronicKidney Disease Prediction. Journal of Medical Systems, 45(6), 1-13. [Link: https://doi.org/10.1007/s10916-021-01798-7]
[12] Huang, J., & Fang, Y. (2020). A Comparative Study of Machine Learning Algorithms for the Diagnosis of Chronic Kidney Disease. Journal of Medical Systems, 44(10), 1-10. [Link :https://doi.org/10.1007/s10916-020-01649-y]
[13] Kadian, S., & Sood, S. K. (2019). Chronic Kidney Disease Detection using Machine Learning Algorithms: A Comparative Study. International Journal of Computer Science and Network Security, 19(5), 108-116. [Link:https://doi.org/10.1109/RAICS49824.2019.8989308]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)