Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

KIDNEY DISEASE PREDICTION USING RANDOM FOREST ALGORITHM

Mr.B.Hanumanth Rao Associate Professor, Department of Computer Science and Engineering, Potti Sriramulu Chalavadi Mallikarjunarao College of Engineering and Technology, Vijayawada, AP, India

Ms. Sri charita Chilakanti [2], Ms. Ravya Korangi [3], Mr. Sri Harsha Isukapalli [4] Student, Potti Sriramulu Chalavadi Mallikarjunarao College of Engineering and Technology, Vijayawada, AP, India

ABSTRACT

Kidney disease is one of the most dangerous illnesses in the world right now. In addition to filtering blood and eliminating waste materials, the kidney also plays a critical role in maintaining bodily functions. Patients with high blood pressure and diabetes are typically affected by kidney disease. Most of kidney diseases can affect diabetes patients and high blood pressure patients. However, immediately identifying the disease at the earliest stage can save a patient's life. In the existing system, data mining techniques are used to predict kidney disease, but the system does not give efficient accuracy and it is time-consuming. In this paper, machine learning algorithms are used such as random forest and light GBM (gradient boosting machine). These algorithms can detect the stage of a deadly disease by taking less time with a reliable CKD data set to get an accurate rate of prediction of the kidney disease. Machine learning is essential in the healthcare sector and aids in disease prediction from the dataset by using the Random Forest and Light GBM Classifier.

KEYWORDS: Random forest algorithm, Light GBM (gradient boosting machine) Classifier, CKD data set, Model comparison, classification.

I. INTRODUCTION

The prediction of kidney disease using machine learning algorithms is presented in this paper.

Due to damage, the kidneys are unable to filter blood as effectively as they should on a daily basis. A patient is more likely to develop renal disease if they have diabetes or high blood pressure. Dialysis or a kidney transplant are two options for treating renal failure. Kidney disease progresses very gradually and without any symptoms. Around the world, kidney disease can take many different forms. Therefore, the majority of doctors typically waste valuable time trying to determine whether a patient has kidney disease or not. To determine which algorithm to use in this paper, we are essentially working on "Chronic Kidney Disease" based on various performance indices. Thousands of lives could be saved worldwide if the disease could be quickly predicted and treated before patients suffer significant harm. Additionally, algorithms for machine learning can be used to find this illness. Several machine learning algorithms can be trained using patient data to identify this emerging disease. But getting the most accurate forecast in the shortest amount of time is the challenge.

Several performance assessment metrics, including the false negative rate (FNR), accuracy (ACC), precision (PRE), negative predictive value (NPV), F1 score (F1), false discovery rate (FDR), standard deviation (SD), specificity (SPE), mean absolute error (MAE), mean squared error (MSE), sensitivity (SEN), and root mean square error, are used in the previous model to properly evaluate classifiers. (RMSE). In order to obtain the most accurate prediction in the shortest amount of time, used this model.

II. RELATED WORK

[1] With multilayer perception and neural network preprocessing to fill in the blanks, Hussain and the team were able to accurately predict CKD in its early stages with an accuracy of 0.995. In the workflow, the outliers are thrown out, the best seven attributes are chosen using statistical analysis, and the principal component analysis results are used to throw out the attributes with the highest inter-co-relation. (PCA).

[2] Rady, El-Houssainy, and Anwar, Ayman Create the e- GFR dataset and the dataset algorithms. (PNN,REF,SVM,MLP) In comparison to other algorithms, the probabilistic neural network algorithm provides the highest overall classification accuracy percentage

when classifying the stages of CKD patients. In contrast, the probabilistic neural network needs 12 s to complete the analysis while the multilayer perception needs to run for at least 3 s. Based on correctly classified CKD patient stages and the amount of time needed to reach that limitation, these algorithms were contrasted in terms of classification accuracy.

[3] Ankit Chatorikar, Siddheshwar Tekale, Pranjal Shingavi, Sukanya Wandhekar, and others used CKD, Decision Tree, GFR, SVM, and Machine Learning. Because of the size of the data set and the missing attribute values, this study's limitations include a lower level of data strength. A machine learning model that targets chronic renal disease and has an overall accuracy of 94.99% requires millions of records with no missing values.

[4] Using Various Decision Tree Methodologies, Chronic Kidney Disease Prediction A.R.M. Alam, S. Baeha, A.S. Sianipar, D. Hartama, M. Zarlis, I.A. Pasadana, A.S. Sianipar, and A. Some of the decision tree techniques used in this study include Munandar DecisionStump, HoeffdingTree, J48, CTC, J48graft, LMT, NBTree, Random Forest, RandomTree, REPTree, and SimpleCart. The purpose of this study is to predict CKD using data mining methods. The main goal of this study is to determine the best decision tree for the prediction of CKD by comparing various decision tree techniques and using various decision tree techniques for CKD prediction.

[5] Endah W. Iiji Lestari, Taufik Asra, Ahmad Setiadi, Mahmud Safudin, Nila Hardi, and Doni Purnama Alamsyah AdaBoost algorithm implementation for predicting chronic kidney disease Algorithm testing is the methodology employed in this study. Processing data on chronic kidney disease as training data is the first step in creating a model. The model's output from the training set of data was then put to the test. The test results analysis is then contrasted. The purpose of the study is to determine whether data on chronic kidney disease are impacted. The adaboost algorithm can improve precision and accuracy.

Table 1: Existing system analysis

S.No	Title	Algorithm Used	Merits	Demerits/ Future work	Accuracy
1	kidney disease stages are predicted using data mining algorithms	Algorithm for Probabilistic Neural Networks	When compared to other algorithms, it has the highest overall classification accuracy for classifying the stages of CKD patients.	Based on the time required and correctly classified CKD patient stages, these algorithms have been compared for classification accuracy.	96.7%
2	Using a Machine Learning Algorithm, Prediction of Chronic Kidney Disease	SVM, Decision Tree, and GFR	Because of the size of the data set and the missing attribute values, the data's strength is not higher.	Millions of records with no missing values will be required. to create an accurate machine-learning model that targets chronic kidney disease.	94.99%

3	Using an Adaptive Hybridized Deep Convolutional Neural Network, Chronic Kidney Disease Prediction	Deep Learning, Convolution Neural Network, IoMT	The physical state of the body can be tracked remotely using IoMT, and medical professionals can spot anomalies.	Efficient use of the learning and taxation mechanisms is a method of double-training.	93%
4	Machine learning algorithm optimization for chronic kidney disease prediction	Gradient Boosting, Linear Discriminant Analysis, Support Vector Machine, and AdaBoost	To obtain a precise expectation rate over the presented dataset, four different algorithms were chosen.	This model is used with reliable algorithms to get highly accurate predictions.	94.5%

5	Application of various classification algorithms to the prediction of chronic kidney disease	Decision tables and K-nearest neighbour (K-NN)	Instead of using the entire set of 22 attributes in the dataset, accuracy for the prediction of a CKD case can be achieved using a chosen set of 5 attributes.	The disease prediction models were finished by applying feature selection to the attributes present in the CKD dataset.	93.4%

III. PROPOSED METHODOLOGY

In the proposed model of Chronic Kidney Disease (CKD), data sets have been utilized. In this system, it can be used by patients to know if kidney disease is present or not by inputting data such as "age", "blood pressure", "serum creatinine", "sugar", bacteria," etc. It will give some clear information about the concept of work. The data is separated into training data and testing data after it has been cleaned and processed. Using the training data, two machine learning classification algorithms are trained. The algorithms are implemented on the test data after training in order to produce predictions. In order to identify the most effective algorithm to predict chronic kidney disease in patients, the accuracy and performance of the predictions made by the two algorithms are compared in this study. The suggested model is demonstrated in the following.

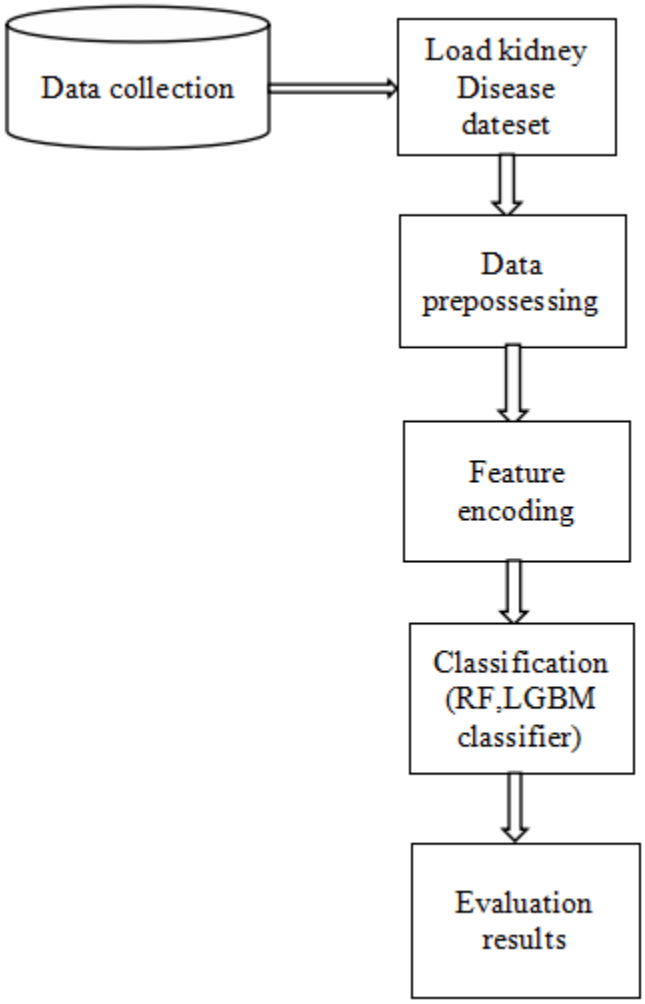

Fig.1: The suggested kidney disease prediction model

3.1 Random forest:

Random Forest is a part of the supervised learning strategy. It is applicable to machine learning (ML) algorithms that combine regression and classification. Its foundation is the idea of ensemble learning, a method for combining different classifiers to handle complex issues and enhance model performance. In order to increase accuracy, the classifier Random Forest averages several decision trees that were applied to different subsets of the input data set. The random forest chooses the result based on the votes of the majority of predictions using predictions from each tree.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

3.2 Light GBM Classifier:

The Light GBM gradient-boosting system makes use of a tree-based learning method. Light GBM grows trees vertically, as opposed to other algorithms that grow trees horizontally, which translates to Light GBM growing trees leaf-wise as opposed to other algorithms growing levels-wise The leaf with the greatest delta loss will be chosen to grow. A leaf-wise method can reduce loss more than a level-wise method when growing the same leaf.

3.3 DATASET DESCRIPTION

Step 1: There are 25 features and 1 class label for every chronic kidney disease record, and the features are age, bp, sugar, serum creatinine sodium, hemoglobin etc.

age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wbcc
48	80	1.02	1	0		normal	notpreser	notpreser	121	36	1.2			15.4	44	7800
7	50	1.02	4	0		normal	notpreser	notpresent		18	0.8			11.3	38	6000
62	80	1.01	2	3	normal	normal	notpreser	notpreser	423	53	1.8			9.6	31	7500
48	70	1.005	4	0	normal	abnormal	present	notpreser	117	56	3.8	111	2.5	11.2	32	6700
51	80	1.01	2	0	normal	normal	notpreser	notpreser	106	26	1.4			11.6	35	7300
60	90	1.015	3	0			notpreser	notpreser	74	25	1.1	142	3.2	12.2	39	7800
68	70	1.01	0	0		normal	notpreser	notpreser	100	54	24	104	4	12.4	36	
24		1.015	2	4	normal	abnormal	notpreser	notpreser	410	31	1.1			12.4	44	6900
52	100	1.015	3	0	normal	abnormal	present	notpreser	138	60	1.9			10.8	33	9600
53	90	1.02	2	0	abnormal	abnormal	present	notpreser	70	107	7.2	114	3.7	9.5	29	12100
50	60	1.01	2	4		abnormal	present	notpreser	490	55	4			9.4	28	
63	70	1.01	3	0	abnormal	abnormal	present	notpreser	380	60	2.7	131	4.2	10.8	32	4500
68	70	1.015	3	1		normal	present	notpreser	208	72	2.1	138	5.8	9.7	28	12200
68	70						notpreser	notpreser	98	86	4.6	135	3.4	9.8		
68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000
40	80	1.015	3	0		normal	notpreser	notpreser	76	162	9.6	141	4.9	7.8	24	3800
47	70	1.015	2	0		normal	notpreser	notpreser	99	46	2.2	138	4.1	12.6		
47	80						notpreser	notpreser	114	87	5.2	139	3.7	12.1		
60	100	1.025	0	3		normal	notpreser	notpreser	263	27	1.3	135	4.1	12.7	37	11400
62	60	1.015	1	0		abnormal	present	notpreser	100	31	1.6			10.3	30	5300
61	80	1.015	2	0	abnormal	abnormal	notpreser	notpreser	173	148	3.9	135	5.2	7.7	24	9200
60	90						notpreser	notpresent		180	76	4.5		10.9	32	6200

Fig.2 :Data set

Step 2:

Cleaning the data:

The data may be incomplete and contain a lot of useless information. Data cleaning is completed to handle this portion. Missing data handling, data analysis, feature engineering, handling noisy data, etc. are all included.

Missing Data:

When some data is missed, this circumstance occurs. There are several ways to handle it.

Among them are:

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

1. Ignore the tuples: This strategy only works when the dataset at hand is sizable and a tuple contains multiple missing values.

2. Adding the Missing Values: There are several methods for completing this task. The most likely value to manually fill in the missing values is attribute mean.

Step 3:

The obtained data from stage is taken into consideration then data is trained using the classification algorithm and obtained result is analyzed. In order to improve accuracy, the obtained data is also trained using machine learning algorithms like RF and LGBM.

3.4 PREPROCESSING:

In machine learning, data preprocessing prepares the metadata to make it appropriate for the model. It is the first and most crucial stage in the development of a model. In this study, feature encoding is utilized to convert categorical data from a dataset into numerical data.

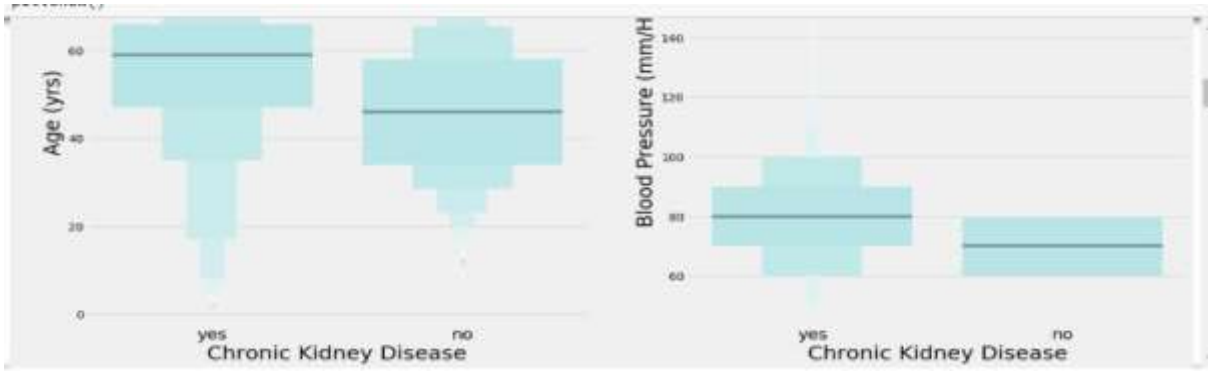

Fig.3: Distribution of categorical data

3.5 FEATURE ENCODING

Feature encoding is the process of transforming categorical data in a dataset into numerical data. Feature encoding is essential because the majority of machine learning models can only take into account numerical data and not written data.

3.6 MODEL COMPARISION

In this study the goal of comparing two algorithms is:

A) Better performance

When there is model comparison and selection, machine learning operates more effectively. Choosing the algorithms that work best for the data is the main objective.

B) Longer lifetime

High performance may not last long if the selected model is tightly correlated with the training data and unable to interpret unknown input. The key to ensuring that predictions are accurate over time and that little retraining is necessary is to find a model that understands underlying data patterns.

C) Speedy process

D) It is needed to advance quickly and with the maximum accuracy. The machine learning solutions must be configured using a number of parameters.

E) Easier retraining data

Minute details and metadata are recorded when models are reviewed and prepared for comparisons, and they are useful during retraining.

VI. PERFORMANCE ESTIMATION

In this section, the data set used for the classification process, aspects that are extracted from the reviews, and performance parameters of the classifiers such as precision, recall, accuracy, and F-score values that are obtained by the LGBM and RF are discussed. A healthy comparison has been made between the proposed method and the other existing classification techniques.

4.1ACCURACY

One metric for assessing classification models is accuracy. Accuracy is the proportion of predictions that our model correctly predicted. Accuracy can also be determined in terms of positives and negatives for binary classification, as shown below:

Accuracy=TP+TNTP+TN+FP+FN

Where

TP = True Positives,

TN = True Negatives,

FP = False Positives, and

FN = False Negatives.

1. True Positive (TP): Positive and anticipated positive values.

2. Values that are predicted to be positive but are actually negative are known as false positives (FP).

3. False Negatives (FN) are positive values that are expected to be negative.

4. Values that are predicted to be negative and are therefore true negatives (TN).

4.2 CONFUSION MATRIX

When describing how well a classification model performs on a set of test data for which the true values are known, a table known as a confusion matrix is frequently used. The confusion matrix shows the confusion of the classification model during prediction. It shows the TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) values for the testing dataset.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

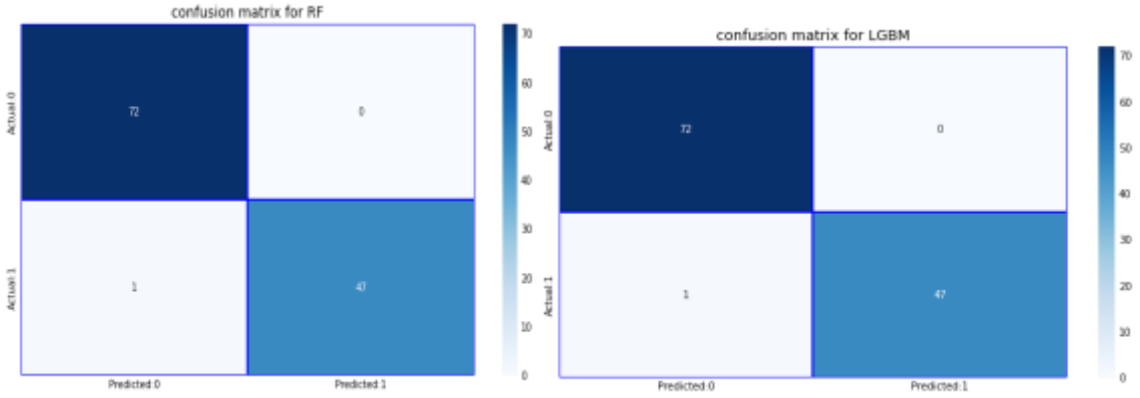

Fig.4: Matrix of Confusion for RF and LGBM

V. PERFORMANCE MEASURES

To describe classification performance, "accuracy" or "error" in predictions is frequently used. Although specificity and precision are computed differently, accuracy is typically used interchangeable with them. It is determined by comparing the proportion of correctly classified samples to all samples. The accuracy is established by

TP+TN/P+N

Where TP=True Positives, TN=True Negatives, P=Positives, N=Negatives.

Additional metrics for categorization performance include sensitivity, specificity, recall, and precision. Sensitivity (also known as recall) and precision assess the "True Positive Rate" for a binary classification task, which measures the likelihood of making the correct prediction in a "positive or true" case. (e.g., in an attempt to predict disease, the disease is correctly predicted for a patient who truly has this disease).

Sensitivity = TP/TP+FN

Precision = TP/TP+FP

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Specificity describes the probability of making the correct prediction in a "false or negative" instance, or the "true negative rate," for a binary classification problem. (e.g., in an attempt to predict disease, no disease is predicted for a healthy patient).

Specificity = TN/TN+FP

VI. RESULTS AND DISCUSSION

After the data set underwent a successful evaluation process to determine whether a patient has kidney disease or not, a sizeable amount of data was divided into training and testing. By virtue of classification, performance indicators are employed to support various algorithmic approaches. A person is classified as positive (0) when they exhibit symptoms of renal illness, if they do not, have the symptoms they are classified negative (1). Out of the two algorithms, Random Forest yields the best results.

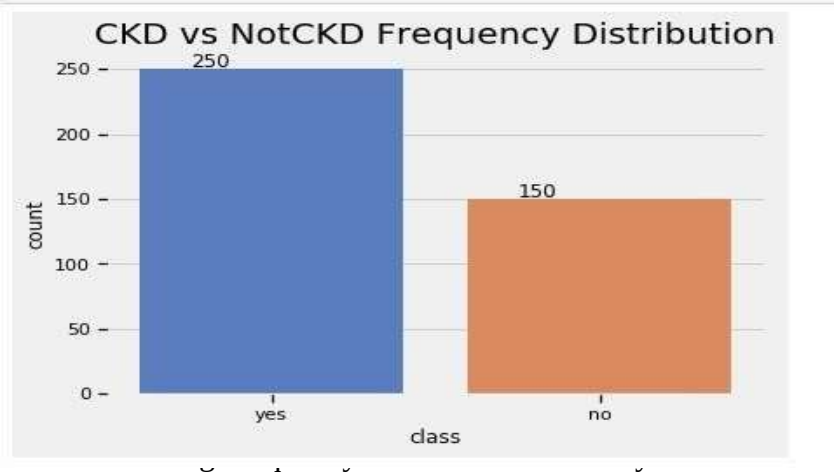

Fig.5: Frequency distribution of kidney disease

When compared to the Light GBM classifier, Random Forest algorithm has the highest accuracy. The graph below compares the models of two algorithms.

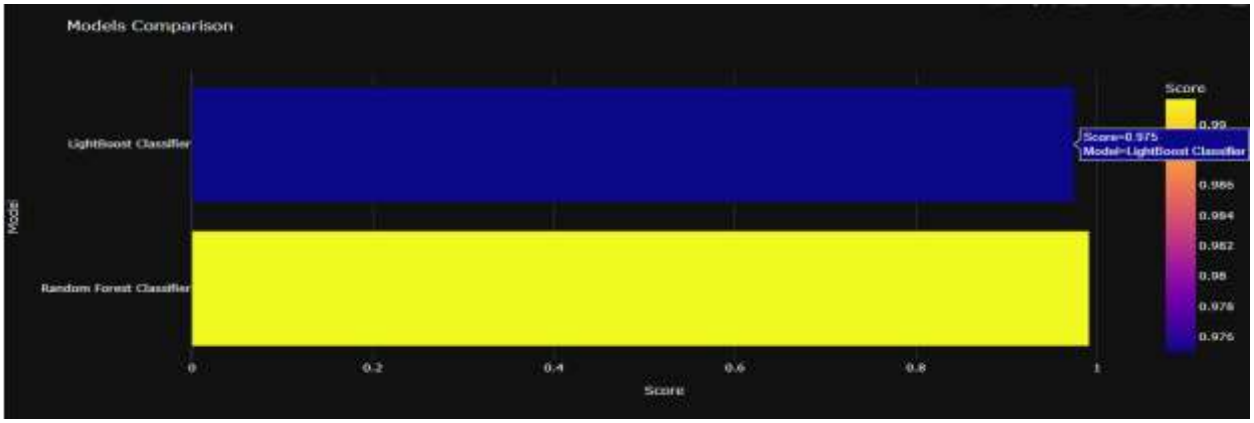

Fig.6: Model comparison of kidney disease

VII. CONCLUSION AND FUTURE WORK

To obtain a precise expectation rate over the introduced data set in this paper, two different algorithms were chosen. Comparing all of the methods used, the random forest algorithm (accuracy rate: 99.10%) yielded the best results, while LGBM (97.05%) scored poorly. In addition, the LGBM classifier takes longer than RF to provide a prediction and the highest score that can be predicted. Since a precise pace of expectation is unquestionably dependent on the preprocessing strategy, the preprocessing techniques must be handled carefully to produce recognized results precisely.

VIII. REFERENCE:

[1] Alghamdi, Raghad A. Makawi, Eman A. Albiety, Tayeb Brahimi, Akila Sarirete: Sensor Based Human Activity Recognition Using Adaboost Ensemble Classifier, In: Procedia Computer Science, Volume 140, (2018), Pages 104-111, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.10.298.

[2] Manogaran, G., Priyan, M.K., and Varatharajan, R. An improved SVM algorithm and LDA for big data categorization of ECG signals in cloud computing. 77, 10195–10215 Multimed Tools Appl (2018). https://doi.org/10.1007/s11042-017-5318-1

[3] "Symptoms, treatment, causes, & prevention of chronic kidney disease" American Kidney Fund, (Accessed on July 31, 2020), http://www.kidneyfund.org/kidneydisease/chronic-kidney-disease-ckd/.

[4] Predicting Chronic Kidney Disease Survival Using Artificial Neural Networks, 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, pp. 1351–1356. Additionally present were C. Hung, W. C. Chu, P. Chiu, C. Y. Tang, and H. Zhang.

[5]A Web Based Application for Agriculture: "Smart Farming System," International Journal of Emerging Trends in Engineering Research, July 2020, ISSN: 2347-3983, by F. M. J. M. Sharman, M. Asaduzzaman, P. Ghosh, M. D. Sultan, and Z. Tasnim.

[6] Boosting Classifiers, Ant-Miner, and J48 Decision Tree for Rule Induction and Chronic Kidney Disease Prediction, 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, pp. 1-6, 2019.

[7] P. Ghosh, M. H. Sadek, M. A. Kazi, S. Shultana, F. M. J. M. Shamrat, "Implementation of Machine Learning Algorithms to Predict the Prognosis Rate of Kidney Disease" is a paper presented at the 2020 IEEE International Conference on Innovation in Technology.

[8] "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods" is published in the 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, pp. 1-4.

[9] On the Application of Machine Learning to Forecasting Cancer Outcome, P. Ghosh, M.Z. Hasan, O.A. Dhore, A.A. Mohammad, and M. I. Jabiullah, Proceedings of the International Conference on Electronics and ICT - 2018, Dhaka, Bangladesh, November 25–26, 2018, pp. 60.

[10] An analysis on breast disease prediction using machine learning approaches was published in the International Journal of Scientific & Technology Research, Volume 9, Issue 02, February 2020, ISSN: 2277-8616, pp. 2450–2455. It was written by F. M. Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M. Sazzadur Rahman, Imran Mahmud, and Rozina Akter.