[1*]Reshma Dayma

[2]Sajid Patel

[3]Dhruti Patel

Machine Learning Algorithms as a Boon for Chronic Kidney Disease Prediction

Journal of Electrical Systems

Abstract: - Chronic kidney disease (CKD) is one type of condition where kidney function damage over several month or year. In body, kidney's main task is to filter impurities and waste from blood which is flush out from body in form of urine. But because of some condition or diseases, in which the kidneys are damaged and cannot filter blood as well as it should. People with kidney disease may not feel ill or notice any symptoms in early stage but it is very serious problem as it may lead to complete failure of kidneys. Machine learning (ML) techniques are used for prediction. Here we have created machine learning model for CKD prediction. We have use three algorithms, logistic regression, support vector machine (SVM) and random forest with feature selection technique and finally applied bagging method on it. we have applied this model on chronic kidney dataset which have derived from UCI machine learning repository. This model predict person have chronic kidney disease or not.

Keywords:Chronic kidney disease (CKD), Machine Learning, logistic regression, support vector machine (SVM), random forest, bagging.

I. INTRODUCTION

Kidney's main task is to filter impurities and waste from blood and keeps the blood clean. But because of deformation or disease in the kidneys, it cannot function effectively and the result will be impure bloods which will create many other problems. Chronic kidney disease (CKD) is long term condition, in which kidney gets damaged over several month or year. It is slow but progressive disease in which first the risk of illness increases and then it damages kidney function then it reduces kidney function and finally leads kidney failure. Chronic kidney disease is long term condition. In this condition, the kidney loses function overtime. It is a slow process such that human can't understand what is going on with kidney. Symptoms of CKD may not appear until the condition is advanced. There are many factors that affect CKD like blood pressure, diabetes, hypertension etc. In advance condition dialysisor kidney transplant may be necessary. In the healthcare industry, machine learning is rapidly being applied, especially for the prediction and diagnosis of chronic kidney disease (CKD). Here are some examples of how machine learning might help in CKD prediction:

Early detection: Machine learning algorithms can analyze huge amounts of patient data, such as laboratory test results, medical history, and demographic information, to identify patients at high risk of developing CKD. This enables early intervention and treatment in order to stop or prevent the progression of the disease.

Personalized treatment: Machine learning algorithms can also be used to analyze patient data in order to discover the most effective treatments for particular patients. By reducing unneeded treatments or hospitalizations, this can assist to improve patient outcomes and reduce the cost of healthcare.

Improved accuracy: Because machine learning algorithms can analyze complicated datasets with various variables, they can predict CKD risk and progression more accurately. This can aid in the identification of patients who would otherwise be missed by traditional diagnostic techniques.

Machine learning has the potential to increase the accuracy and efficiency of CKD prediction and treatment, resulting in better patient outcomes and reduced medical expenses. Here, we have used three machine learning algorithms Logistic Regression, SVM and Random forest and also used bagging ensemble technique for chronic kidney disease predictions. we applied these techniques on chronic kidney disease dataset.

[1*,2] L.D. College of Engineering, Ahmedabad, India., [3] Bhaskaracharya Institute For Space Applications and Geo-Informatics

Corresponding author: Reshma Dayma, Computer Engineering Department, L.D. College of Engineering, Ahmedabad, India.

ceradayma@gmail.com

In this paper, section II discuss about literature review Section III is discussion about dataset. Section IV is proposed model for CKD prediction. Section V is about result and comparison, section VI is conclusion and future work.

II. LITERATURE REVIEW

For chronic kidney disease predictions in health care, various algorithms and machine learning approaches are used for early CKD prediction. Using a dataset of patients, researchers used various ML techniques, including logistic regression, decision tree, and support vector machine, XGBoost algorithms etc to predict CKD using various factors that affect chronic kidney disease. Chronic Kidney Disease Prediction Using Machine Learning Techniques [1] uses three algorithms used for chronic kidney disease prediction i.e. logistic regression, random forest and SVM algorithm. These algorithms predict that a person is having a chronic kidney disease or not. This can help the medical practitioners for the early prediction of CKD with 97.23%accuracy. The model can be further tuned by applying feature selection methods to increase the performance of the prediction. Prediction of Chronic Kidney Disease - A Machine Learning Perspective [2] author used Artificial neural network, logistic regression, linear support vector machine (LSVM), K-Nearest neighbors and random Forest algorithms on chronic kidney disease dataset and predict person have chronic kidney disease or not. Here among this all algorithms (LSMV) give highest accuracy 98.86% with SMOTE (synthetic Minority oversampling technique) with full features.

A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease [3] used AdaBoost algorithm with information-gain-based feature selection method for kidney disease prediction. In last author predict AdaBoost give 96.4% accuracy with feature Exploring

A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease [4] used Deep Neural Network, with Recursive Feature Elimination feature selection method for chronic kidney disease prediction. This model has obtained an accuracy of 100%. The proposed approach could be a useful tool for nephrologists in detecting CKD. The limitation of the proposed model was that it had been tested on small data sets.

Chronic kidney disease prediction using machine learning Techniques [5] used Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT), with 10 fold cross validation for chronic kidney disease prediction.SVM and RF produced the highest accuracy of 99% for binary class data. Only binary class data is used for prediction.

Early Prediction of Chronic Kidney Disease Using Machine Learning Algorithms for Small and Imbalanced Datasets [6] used Decision tree (DT), random forest and Synthetic minority oversampling technique (SMOTE) on chronic kidney disease dataset. And here result compare with another algorithm. Here DT model with the SMOTE technique improves the results with 98% accuracy.

Back propagation Neural Network-Based Machine Learning Model for Prediction of Blood Urea and Glucose in CKD Patients [7]. In this paper, author used back propagation neural network-based machine learning model for CKD prediction, which will be helpful for diabetes patients with Chronic Kidney Disease (CKD) to monitor blood urea 98% and glucose with 95.96% accuracy.

Development and Validation of an Insulin Resistance (IR) Model for a Population with Chronic Kidney Disease Using a Machine Learning Approach [8]. Here, authors used various algorithms on chronic kidney disease dataset. Algorithms Random forest (RF), eXtreme Gradient Boosting (XGboost), logistic regression algorithms, and deep neural learning (DNN) which author applied on dataset. ML algorithms, particularly RF, can help predict IR in patients with CKD.

Chronic Kidney Disease Prediction Using Machine learning [9] used Apriori association algorithm for CKD prediction. And compare result with greedy stepwise search method, k-nearest neighbor, naive- Bayes algorithms on chronic kidney disease dataset. Which predict person have chronic kidney disease or not. Best result can be achieved using Apriori associative algorithm with IBK classifier for 97% accuracy.

Chronic Kidney Disease Prediction using Machine Learning Algorithms [10]. Here, authors implementmany algorithms on dataset, Random Forest Classifier, SVM and Decision Tree classifier, KNN. Here Logistic regression produces 98 % accuracy as compare to others

III. DATASET FOR MODEL TRAINING

This CKD dataset is derived from UCI machine learning repository. The CKD dataset typically includes data on people's age, gender, ethnicity, and medical history, including if they have diabetes, hypertension, or other medical conditions. The information also contains the outcomes of laboratory tests, including those for blood pressure, serum creatinine levels, and albuminuria etc.

The dataset is separated into two categories: numerical variables and categorical variables. Laboratory test results, such as serum creatinine levels, are continuous variables in the numerical variables. Categorical variables include demographic information, medical history, and other binary factors, such as whether the patient has hypertension or not. Dataset have 400 patients record. It includes 25 features. Information of features is given in below table.

A. Data preprocessing

To use the CKD dataset for machine learning, the data needs to be preprocessed, which involves cleaning, Data Transformation, Data Reduction, Data Integration and Data Discretization to put it into a suitable format for analysis. This cover dealing with missing data, normalizing and standardizing numerical variables, and encoding categorical variables. Data preprocessing is an important stage in machine learning.

B. Feature selection

Feature selection is the process of selecting a subset of relevant features or variables from a larger set of features in a dataset for use in a machine learning model. In machine learning, features or variables are the measurable characteristics or attributes of the data that are used to make predictions or classifications. Feature selection helps to identify the most important features that have the most impact on the model's performance, while reducing the computational complexity of the model by removing irrelevant or redundant features. We used recursive feature elimination technique for selection. Recursive Feature Elimination (RFE) is a feature selection technique in machine learning that works by recursively removing features and building a model on the remaining features until the desired number of features is achieved. The RFE algorithm starts by building a model using all of the available features and then eliminates the least important features based on the feature importance ranking provided by the model. The process is repeated until the desired number of features is reached or until a predefined threshold is met. RFE is often used with linear regression and support vector machine models.

TABLE I. Results of Different algorithm for CKD

Sr No	Feature Name	Full Form	Information	Sr No	FeatureName	Full Form	Information
1.	Ag	Age	Age (numerical) ageinyears	**2.**	Sod	Sodium	Sodium (numerical) sod in mEq/L
3.	Bp	Bloodpressure	BloodPressure (numerical) bpinmm/Hg	**4.**	Pot	Potassium	Potassium (numerical) pot in mEq/L
5.	Sg	Specificgravity	Specific Gravity (nominal) Sg (1.005, 1.010, 1.015, 1.020, 1.025)	**6.**	Hemo	Hemoglobin	Hemoglobin (numerical) hemo in gms
7.	Al	Albumin	Albumin (nominal)	**8.**	Pcv	Packed	Packed Cell

			al-(0,1,2,3,4,5)			cell volume	Volume (numerical)
9.	Su	Sugar	Sugar(nominal) su-(0,1,2,3,4,5)	**10.**	Wc	White Blood Cell blood cell count	White Blood Cell Count (numerical) wc in cells/cumm
11.	Rbc	Redblood cell	RedBloodCells(nominal) rbc-(normal, abnormal)	**12.**	Rc	Red blood cell count	Red Blood Cell Count (numerical) rc in millions/cmm
13.	Pc	Pus cell	Pus Cell (nominal) pc - (normal, abnormal)	**14.**	Htn	Hyper-tension	Hypertension (nominal) htn - (yes, no)
15.	Pcc	Pus cell clumps	Pus Cell clumps(nominal) pcc - (present, not present)	**16.**	Dm	Diabetes mellitus	Diabetes Mellitus(nominal) dm - (yes, no)
17.	Ba	Bacteria	Bacteria (nominal) ba - (present, not present)	**18.**	Cad	Coronary artery disease	Coronary Artery Disease (nominal) cad - (yes, no)
19.	Bgr	Blood glucose random	Blood Glucose Random (numerical) bgr in mgs/dl	**20.**	Appet	Appetite	Appetite (nominal) appet - (good, poor)
21.	Bu	Blood urea	Blood Urea (numerical) bu in mgs/dl	**22.**	Pe	Pedal edema	Pedal Edema (nominal) pe - (yes, no)
23.	Sc	Serum creatinine	Serum Creatinine (numerical) sc in mgs/dl	**24.**	Ane	Anemia	Anemia (nominal) ane - (yes, no)
25.	Classification	Classification	Classification (nominal) classification (ckd, notckd)				

IV. **PROPOSED MODEL**

In this proposed methodology, we have applied hybrid model on chronic kidney disease dataset. Work flow of the system is as per Fig. 1. In first step, import necessary libraries which provide various tools, algorithms, and functions that can be used to build and train machine learning models.

In second step load dataset and then describe dataset.Data describe means describing information about data means how many rows, columns, dataset mean, mode, standard deviation etc.

Third step is data preprocessing in this step, various data preprocessing techniques like finding missing values, mapping of categorical data into numerical data, correlation analysis are applied. First we check missing values available or not in dataset. In this dataset there are many columns having missing values. So, we replaced these

missing values with 0. Then map the categorical data into numerical data using hot encoding method. Then check correlation analysis that shows how predictor variable is correlated with target variable.

Fourth step is splitting dataset. Split dataset means dividing dataset into two or more part. It is important part because it reduces over fitting problem. Over fitting is problem occur with machine learning models in this problem machine learning models give accurate result for training data but not give accurate result for new or unseen data. So, it is necessary. Here, we divided CKD dataset into two-part training and testing part. Set ratio of training and testing, 80% data for training part and 20% for testing part.

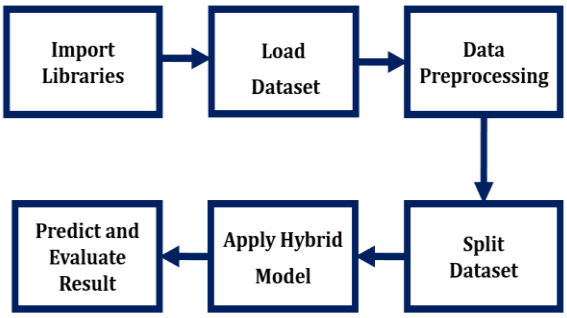

Figure 1. Work Flow

A. CKD Prediction Model

Here we have created hybrid model for chronic kidney disease prediction using machine learning. A hybrid model in machine learning refers to a model that combines multiple machine learning techniques or approaches to achieve better performance than using a single technique. In this model, three machine learning algorithms logistic regression, Random Forest, SVM and bagging method is used for chronic kidney disease prediction. In first step, three machine learning algorithms are applied on CKD dataset. Then bagging method is applied on output of three algorithms and in last, prediction is made whether person have chronic kidney disease or not. Fig. 2 shows Hybrid Model for CKD prediction.

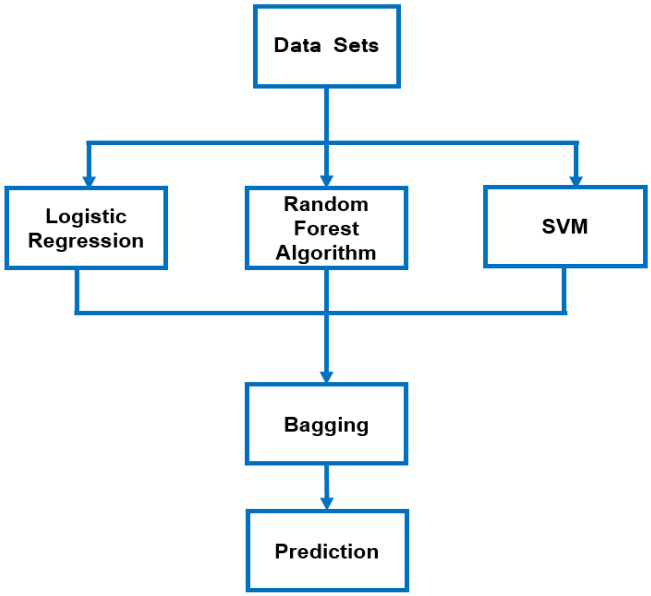

Figure 2. Hybrid Model

B. *Logistic Regression:* It is supervised machine learning algorithm[11].

It is used for classification problem. Main task of this algorithm is to predict instance belonging to given class or not. This algorithm identifies relationship between dependent and independent variables. It is decision making algorithm. It is used for binary classification problem like where output is binary like 0 or 1, yes or no etc[11]. Here logistic regression takes output of linear regression and give this output to sigmoid function which compute probability of instance belonging to which class.

C. Random Forest: It is supervised machine learning algorithm. This algorithm is used for both classification and regression. This algorithm is combination of multiple decision trees[12]. In last it combines result of decision tree using ensemble method. Decision tree is tree like structure. In decision tree root node represent dataset. Internal nodes represent rules and leaf nodes represent output of algorithm.

D. Support Vector Machine:Support Vector Machines (SVM) [13] is a supervised machine learning algorithm which can be used for classification and regression analysis. SVM is a non-probabilistic binary linear classification algorithm that uses a hyperplane to separate the data into two classes. In SVM, the algorithm tries to find a hyperplane that separates the data points in the feature space with the maximum margin between the two classes. The margin is the distance between the hyperplane and the closest data points in each class. The hyperplane that maximizes the margin is considered the optimal hyperplane, which provides better generalization ability. SVM works by transforming the data into a higher-dimensional space using a kernel function. The transformed data can then be separated by a hyperplane in this new space. SVM uses a soft margin to allow for some misclassification of data points, which helps to improve the generalization ability of the algorithm.

E. Bagging: Bagging (Bootstrap Aggregation) is an ensemble learning method in machine learning, which combines multiple models (often decision trees) to improve the accuracy and robustness of the predictions. Bagging is a form of parallel ensemble method where multiple models are trained independently on different subsets of the training data. The basic idea of bagging is to create a set of independent training datasets by randomly sampling from the original training dataset with replacement (i.e., bootstrapping). Each of these datasets is used to train a separate model. During the training process, the models are allowed to use a subset of the features (randomly selected) to improve their diversity. Once the models are trained, their predictions are combined using a simple voting or averaging scheme to make the final prediction. This approach helps to reduce the variance of the individual models and improve the overall accuracy of the ensemble. The main advantage of bagging is that it can reduce over fitting by increasing the diversity of the models. It also allows for parallel training of the models, which can significantly speed up the training process.

V. EXPERIMENTS AND RESULTS

Here we have used three machine learning algorithms for hybrid model and then applied bagging ensemble method on result to make prediction. We have used recursive feature elimination technique for feature selection. Here we have used 5 parameters for model evaluation. Accuracy, precision, recall and F-1 score and confusion matrix.

A. *Accuracy:* Accuracy measures the proportion of correct predictions made by the model. It is calculated by dividing the number of correct predictions by the total number of predictions made.

1) *Formula:*

$$\frac{\text{(True positives + True negatives)}}{\text{(True positives + False Positives + True Negatives + False Negatives)}}$$

B. *Precision:* Precision measures the proportion of true positive predictions out of all the positive predictions made by the model. It is calculated by dividing the number of true positive predictions by the total number of positive predictions made.

1) Formula:

$$\frac{\text{True positives}}{\text{(True positives + False Positives)}}$$

C. Recall: Recall measures the proportion of true positive predictions out of all the actual positive instances in the dataset. It is calculated by dividing the number of true positive predictions by the total number of positive instances in the dataset.

1) Formula:

$$\frac{\text{True positives}}{\text{(True Positives + False Negatives)}}$$

D. F1 score: F1 score is a harmonic mean of precision and recall. It is a useful metric when both precision and recall are important. It is calculated by taking the harmonic mean of precision and recall.

*1) Formula: 2 * ((Precision * Recall) / (Precision + Recall))*

E. Terms used in evaluation parameters:

1) True Positive (TP): It refers to the cases where the model correctly predicts the positive class (condition present) when the actual ground truth is indeed positive.

2) False Positive (FP): It occurs when the model incorrectly predicts the positive class (condition present) when the actual ground truth is negative (condition absent).

3) False Negative (FN): It happens when the model incorrectly predicts the negative class (condition absent) when the actual ground truth is positive (condition present).

4) True Negative (TN):It refers to the cases where the model correctly predicts the negative class (condition absent) when the actual ground truth is indeed negative.

Comparison of results with different machine learning algorithms with our hybrid CKD model based on Evaluation parameter in below table 1. Accuracy of Logistic regression algorithms is 94%, random forest algorithm is 98 % and SVM algorithm is 97%. We have evaluated our model with four parameters Accuracy, precision, recall and F-1 score values of this parameters are 99.375 %, 96.0%, 99.375% and 99.064%. We compare our model result AdaBoost algorithm with information gain based feature selection techniques applied for feature selection. In the end, it predicts result person have CKD or not with selected features. This algorithm accuracy precision recall and F-1 score values are given in table 2.

TABLE I. Results of Different algorithm for CKD

Sr No	**Methods**	**Accuracy**	**Precision**	**Recall**	**F-1 score**
1.	Logistic Regression	94.375	93.777	94.375	94.068
2.	Random Forest	98	98.44	99.04	98.75
3.	SVM	97.812	97.243	97.81	97.5
4.	CKD Prediction Model	99.375	98.756	99.375	99.064

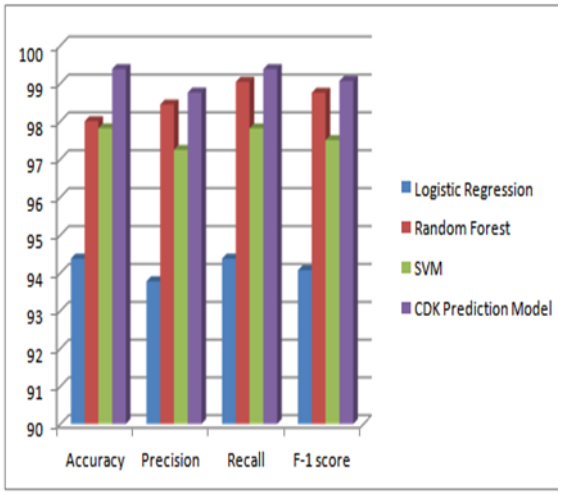

Figure 3. Results of Different algorithm for CKD

TABLE I. Results of Different algorithm for CKD

Sr No	Methods	Accuracy	Precision	Recall	F-1score
1.	AdaBoost	96.4	96	96.8	97
2.	CKD Prediction Model algorithm	99.375	98.756	99.375	99.064

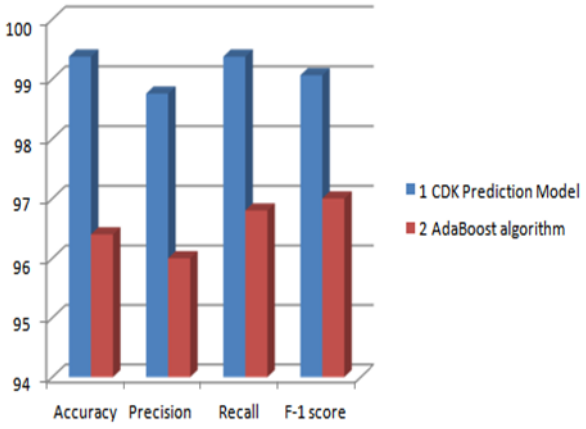

Figure 3. Comparison CKD Model with Adaboost

Next,we have compared various algorithms on CKD dataset with feature selection and without feature selection. Without feature selection our model gives 97% accuracy. Then we applied recursive feature elimination feature selection technique on our model. So with feature selection our model gives 99% accuracy. Then we applied linear regression model on our dataset with or without feature selection. Linear regression with REF gives 63% accuracy and without REF gives 79% accuracy. Other values of evaluation parameters are given in below table.3.

TABLE I. COMPARISON OF CKD MODELS WITH AND WITHOUT RECURSIVE FEATURE ELIMINATION

Sr. No.	Methods	Accuracy	Precision	Recall	F-1score
1.	CKDPrediction ModelwithRFEfeatureselection	99.375	98.756	99.375	99.064
2.	CKDPrediction Model withoutfeatureselection	97.23	97.20	97.0	96.23
3.	Linearregression withoutRFE	79.375	98.105	79.375	87.08
4.	Linearregression withRFE	63.43	40.24	63.43	49.24

VI. CONCLUSION

Our new method is based on machine learning algorithms. Three machine learning algorithms logistic regression, random forest and SVM with bagging ensemble method are used for this approach. This approach is applied on chronic kidney disease dataset which predict whether the person have chronic kidney disease or not. This proposed approach improves the accuracy of prediction. Without feature selection this model gives 97% accuracy and with recursive feature elimination feature selection technique, it gives 99.375% accuracy.

VII. FUTURE SCOPE

This approach predict person have CKD or not using various feature which are more related to our target variable. But there are many parameters like water intake, alcohol, hereditary etc which also affects possibilities of chronic kidney disease. So, using this type of data too, we can better predict person have CKD or not. Another future scope is use of very large dataset to train model, apply various feature selection techniques to build prediction model to further improve model performance.

VIII. DECLARATIONS

Conflict of Interest: The authors declare that they have no competing interests.

Ethical Approval: This research did not involve human participants or animals. The results were obtained through simulations and repeated testing to ensure the final values were reliable.

REFERENCES

[1]. Pal, S. Chronic Kidney Disease Prediction Using Machine Learning Techniques. *Biomedical Materials & Devices* (2022). https://doi.org/10.1007/s44174-022- 00027-y.

[2]. P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in IEEE Access, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[3]. Ebiaredoh-Mienye, S.A.; Swart, T.G.; Esenogho, E.; Mienye, I.D. A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic KidneyDisease. Bioengineering 2022, 9, 350. https://doi.org/10.3390/bioengineering9080350.

[4]. Singh, V.; Asari, V.K.; Rajasekaran, R. A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. Diagnostics , 12,116. https://doi.org/10.3390/diagnostics12010116.

[5]. S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh, Chronic Kidney Disease Prediction using Machine Learning Models. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019.

[6]. Silveira, A.C.M.d.; Sobrinho, Á.; Silva, L.D.d.; Costa, E.d.B.; Pinheiro, M.E.; Perkusich, A. Exploring Early Prediction of Chronic Kidney Disease Using Machine Learning Algorithms for Small and Imbalanced Datasets. Appl. Sci. 2022, 12, 3673. https://doi.org/10.3390/app12073673.

[7]. J. Parab, M. Sequeira, M. Lanjewar, C. Pinto and G. Naik, "Backpropagation Neural Network-Based Machine Learning Model for Prediction of Blood Urea and Glucose in CKD Patients," in *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1- 8, 2021, Art no. 4900608, doi: 10.1109/JTEHM.2021.3079714.

[8]. I. A. Pasadana et al: Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques. Journal of Physics Conference Series, Aug 2019, DOI: 10.1088/1742-6596/1255/1/012024.

[9]. Pooja Mane , Naresh Thoutam, Neha Tiwari, Gauri Mandlik and Nutan Pandey, Chronic Kidney Disease Prediction Using Machine learning.International nternational Journal of Research Publication and Reviews Vol (2) Issue (7) (2021) Page 60-66 ISSN 2582-7421.

[10]. Kallu Samatha, Muppidi Rohitha Reddy, Pattan Faizal Khan, Rayapati Akhil Chowdary, PVRD Prasada Rao, "Chronic Kidney Disease Prediction using Machine Learning Algorithms.", International Journal of Preventive Medicine and Health (IJPMH) ISSN: 2582-7588(Online), Volume-1 Issue-3, July 2021.

[11]. SandroSperandei, “Understanding logistic regression analysis”, Biochemia Medica 2014;24(1):12–8, http://dx.doi.org/10.11613/-BM.2014.003

[12]. Leo Breiman, “Random Forests”, Machine Learning, 45, 5–32, 2001,_c 2001 Kluwer Academic Publishers. Manufactured in The Netherlands., http://dx.doi.org/10.11613/-BM.2014.003.

[13]. Mariette Awad, Rahul Khanna, “Support Vector Machines for Classification”, https://www.researchgate.net/publication/300723807, OI:10.1007/978-1-4302-5990-9_3.