IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

Chronic Kidney Disease Prognosis using Machine Learning

K. Papayamma[1], D. Deeksha Sai[2], G. Mahendra Varma[3], G. Manikanta Srinivas[4], J. Bhargav Vamsi Krishna[5]
Assistant Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3,4,5]
Raghu Institute of Technology, Visakhapatnam, AP, India

Abstract: *Goal three of the UN's Sustainable Development Goal is focused on promoting good health and well-being, with a specific emphasis on addressing the challenges posed by non-communicable diseases. One of the objectives for this goal is to reduce premature mortality from non-communicable diseases by a third by the year 2030. Chronic kidney disease (CKD) is a major contributor to morbidity and mortality from non-communicable diseases, affecting between 10 and 15% of the global population. Early and accurate detection of the stages of CKD is considered vital in order to minimize the impact of the associated health complications, such as hypertension, anemia, mineral bone disorder, poor nutritional health, acid base abnormalities, and neurological complications. To this end, machine learning techniques have been used in various research studies to detect CKD at an early stage. However, previous research has not focused on specific stage prediction. In this study, both binary and multi-classification for stage prediction were carried out using Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) prediction models. Analysis of variance and recursive feature elimination were applied for feature selection, and tenfold cross-validation was used to evaluate the models. The results showed that RF based on recursive feature elimination with cross-validation had better performance than SVM and DT for stage prediction of CKD. This research has the potential to lead to earlier detection and intervention, ultimately reducing premature mortality from non-communicable diseases as outlined in the UN's Sustainable Development Goal of good health and well-being.*

Keywords: Support Vector Machine, K-nearest Neighbor, Decision tree classifier, Random Forest classifier, Naive Bayes Classifier, Machine Learning

I. INTRODUCTION

Chronic kidney disease (CKD) is a global public health issue that affects a substantial proportion of the population worldwide. With an estimated 850 million people living with CKD, the disease has reached alarming levels, and advanced forms of CKD, such as kidney failure, are often not publicly funded in many countries, leading to economic hardships and premature deaths. In 2019 alone, kidney diseases accounted for over 250,000 deaths regionwide, with a higher death rate in men than women. The age-standardized death rate due to kidney diseases varies greatly across countries, with some experiencing rates as high as 73.9 deaths per 100,000 population. Furthermore, the disease has resulted in 5.2 million years of life lost due to premature mortality, a 73% increase from 2000. These statistics highlight the urgent need for effective prevention, early detection, and treatment strategies for CKD to reduce its burden on individuals, families, and healthcare systems.

To address the challenges posed by non-communicable diseases, Goal three of the UN's Sustainable Development Goal is focused on promoting good health and well-being, with a specific emphasis on reducing premature mortality from non-communicable diseases by a third by the year 2030. Chronic kidney disease (CKD) is a major contributor to morbidity and mortality from non-communicable diseases, affecting between 10 and 15% of the global population. Early and accurate detection of the stages of CKD is considered vital in order to minimize the impact of the associated health complications. Machine learning techniques have been used in various research studies to detect CKD at an early stage. However, previous research has not focused on specific stage prediction. In this study, both binary and multi-

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

classification for stage prediction were carried out using Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) prediction models. Analysis of variance and recursive feature elimination were applied for feature selection, and tenfold cross-validation was used to evaluate the models. The results showed that RF based on recursive feature elimination with cross-validation had better performance than SVM and DT for stage prediction of CKD. This research has the potential to lead to earlier detection and intervention, ultimately reducing premature mortality from non-communicable diseases as outlined in the UN's Sustainable Development Goal of good health and well-being.

II. RELATED WORKS

Using patient data, various machine learning algorithms have been applied to accurately classify chronic kidney disease. Existing literature was examined in order to gain the necessary understanding of numerous ideas linked to the proposed design.

To predict the stages of kidney illness, Rady and Anwar [1], compared the probabilistic neural networks (PNN), multilayer perceptron (MLP), support vector machine (SVM), and radial basis function (RBF) methods. A short dataset and a limited number of features were used in the investigation. The outcome of this study demonstrates that the Probabilistic Neural Networks method has a classification accuracy percentage of 96.7%, which is the highest overall.

The development of a self-learning knowledge-based system for the detection and treatment of the first three stages of chronic renal disease has been studied by Mohammed and Beshah [2] utilising machine learning. A modest quantity of data were used in this study, and a prototype was created that allows patients to query KBS to track the delivery of recommendations. To create the rules, they employed a decision tree. The prototype's overall effectiveness has been reported to be 91% accurate.

Using a dataset of 400 observations, Salekin and Stankovic [3] evaluated classifiers like K-NN, RF, and ANN. Five features were chosen for the study's model creation after the implementation of wrapper feature selection. With an RMSE of 0.11 and the greatest classification accuracy of 98%, RF. S. The "Prediction of Chronic Kidney Disease Using Machine Learning Algorithm" study by Tekale et al. [4] used a dataset with 400 occurrences and 14 characteristics. Both decision trees and support vector machines were employed. The dataset has undergone preprocessing, and the 25 features have been decreased to 14. With a 96.75% accuracy rate, SVM is considered as a superior model.

In order to determine the best model for BCD prediction based on various performance evaluations, Vinod [5] evaluated seven supervised machine learning algorithms, including K-Nearest Neighbor, Decision Tree, Support vector Machine, Random Forest, Neural Network, Nave Bayes, and Logistic Regression. The final outcome shown that, with 97% accuracy, k-NN outperformed all other models on the BCD dataset.

According to their performance, Xiao et al. [6] compared the models and advocated utilising logistic regression, Elastic Net, lasso regression, ridge regression, support vector machine, random forest, XGBoost, neural network, and k-nearest neighbour to predict the progression of chronic kidney disease. They graded the outcome as mild, moderate, or severe using 551 individuals' history data with proteinuria and 18 characteristics. They have concluded that, with an AUC of 0.873, sensitivity, and specificity of 0.83 and 0.82, respectively, log regression performed better.

In the context of the Internet of Things, Alsuhibany et al. [7] introduced an ensemble of deep learning-based clinical decision support systems (EDL-CDSS) for the diagnosis of CKD. The method that is being discussed uses an ensemble of three models—deep belief network (DBN), kernel extreme learning machine (KELM), and convolutional neural network with gated recur-rent unit—and uses an adaptive synthetic (ADASYN) strategy for outlier detection (CNN-GRU).

According to the reviews mentioned above, a number of studies have been done on the use of machine learning approaches to predict chronic kidney disease. The size, calibre, and timing of data collection are only a few of the many criteria that are crucial to enhancing model performance. This study focuses on the prediction of chronic kidney disease using machine learning models based on the large and more recent dataset obtained from St. Paulo's Hospital in Ethiopia with five classes: notckd, mild, moderate, severe, and ESRD, as well as binary classes: ckd and notckd. Because the sort of therapy to be given is based on the stages, the majority of previously completed researches concentrated on two classes, which makes treatment suggestions challenging.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

III. DATASET ANALYSIS

The UCI Repository for renal disease provided the Training data set. The values in the dataset are actual test results that were received, and the dataset was gathered from several hospitals in Andhra Pradesh. The dataset has a total of 24 properties, but preprocessing revealed that just 6 of them are crucial for establishing that association. There are 400 samples in all. This will enable accurate early identification of chronic kidney disease using the forecasts.

TABLE I: LIST OF DATASET

Attribute	Representation	Information Attribute	Description
1	Age	Age	Numerical
2	Blood Pressure	Bp	Numerical
3	Specific Gravity	Sg	Nominal
4	Albumin	Al	Nominal
5	Sugar	Su	Nominal
6	Red Blood Cell	Rbc	Nominal
7	Pus Cell	Pc	Nominal
8	PusCellClumps	Pcc	Nominal
9	Bacteria	Ba	Nominal
10	Blood Glucose Random	Bgr	Numerical
11	Blood Urea	Bu	Numerical
12	Serum Creatinium	Sc	Numerical
13	Sodium	So	Numerical
14	Potassium	Pot	Numerical
15	Haemoglobin	hemo	Numerical
16	Packed Cell Volume	Pcv	Numerical
17	White Blood cell Count	Wc	Numerical
18	Red Blood Cell count	Rc	Numerical
19	Hypertension	Htn	Nominal
20	Diabetes Mellitus	Dm	Nominal
21	Coronary Artery Disease	Cad	Nominal
22	Appetite	appet	Nominal
23	Pedal Edema	Pe	Nominal
24	Anemia	Ane	Nominal
25	Class	Class	Nominal

TABLE II: MODEL TRAINING DATASET

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	36.0	1.2	NaN	NaN	15.4	44	7800
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	18.0	0.8	NaN	NaN	11.3	38	6000
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	53.0	1.8	NaN	NaN	9.6	31	7500
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	56.0	3.8	111.0	2.5	11.2	32	6700
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	26.0	1.4	NaN	NaN	11.6	35	7300

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

IV. PROPOSED APPROACH

Dataset pre-processing, categorization, and custom evaluation are essential methods used in the suggested model. Every stage of the suggested system is crucial, significantly improves performance, and has an impact on effectiveness.
Since they are insignificant in intrusion detection, non-numeric or symbolic features must be changed or eliminated as part of the entire pre-processing process. Flag, service, and protocol are all symbolic concepts that could disappear or change.

4.1 Data Pre-Processing and Preparation

Real-world data is frequently erratic, which can affect how well models work. Preparing the data for classifiers is a crucial step in creating a machine-learning model. Similar missing values exist in the dataset used for this investigation, which must be handled correctly. Moreover, it needs to be in a modeling friendly format. As a result, pre-processing was carried out as shown in Fig. 1.

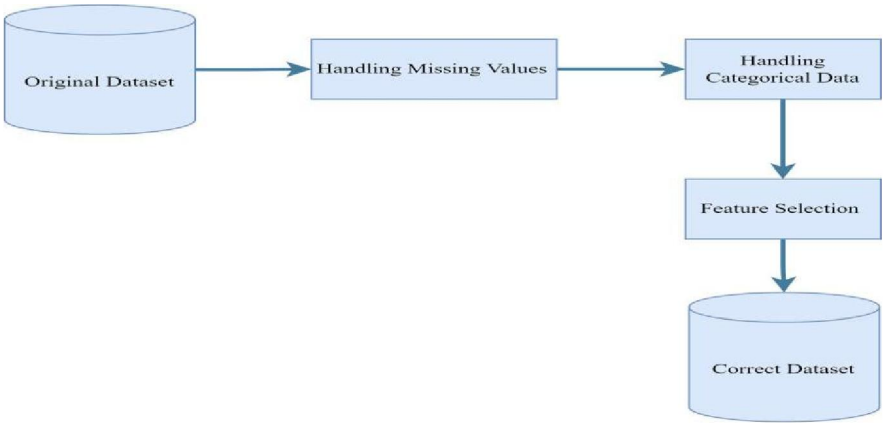

Fig. 1. Preprocessing of the dataset for chronic kidney disease

- **Cleaning Noisy Data:** A key component of preprocessing is removing outliers and smoothing noisy data. Outliers are values that fall outside the normal distribution of the other values. Outliers in clinical data can develop as a result of the data's inherent variation. The data points that fall above Q3 + 1.5(IQR) and below Q1 1.5(IQR), where Q1 is the first quartile, Q3 is the third quartile, and IQR= Q3 Q1 [8], are considered probable outliers.
- **Managing Missing Values:** Data may not always be present (or missed) as a result of equipment failure, inconsistent with other recorded data and thus deleted, not entered into the database as a result of misunderstanding, or not thought important at the time of input for some data. Diagnostic test results that would aid in predicting the likelihood of diagnoses or the efficacy of treatments are frequently missing from patient data [9]. The performance of the prediction model is impacted by the missing values. Missing values can be handled in a variety of ways, including by dropping them and filling them in. If the fraction of missing values is under 10%, they may occasionally be disregarded. However because the missing number may be a crucial component in the model's evolution, it is not thought to be healthy for the model. Sometimes zero can be used in place of the missing numbers, leaving the model unchanged. Since the missing features are numerical and mean imputation works better for numerical missing values, this study used the mean, an average of the observed characteristics, to handle the missing values.
- **Managing Categorical Data:** At this step, data has been formatted in accordance with the specifications. the nominal data transformed into 0 and 1 based numerical data. Consider the nominal value of "Gender," which can be labelled as 0-for female and 1-for male. The final CSV file contains all the integer and float values for the various CKD-related features after preprocessing the data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

4.2 System Architecture

An abstract representation of a system's structure and behaviour is called a system architecture. A system is formally represented by it. System architecture can refer to either a model used to explain the system or a process used to create the system, depending on the context. Designing a suitable system architecture aids in project analysis, particularly at the beginning. The system architecture is shown in Fig.2 and is described in the section below.

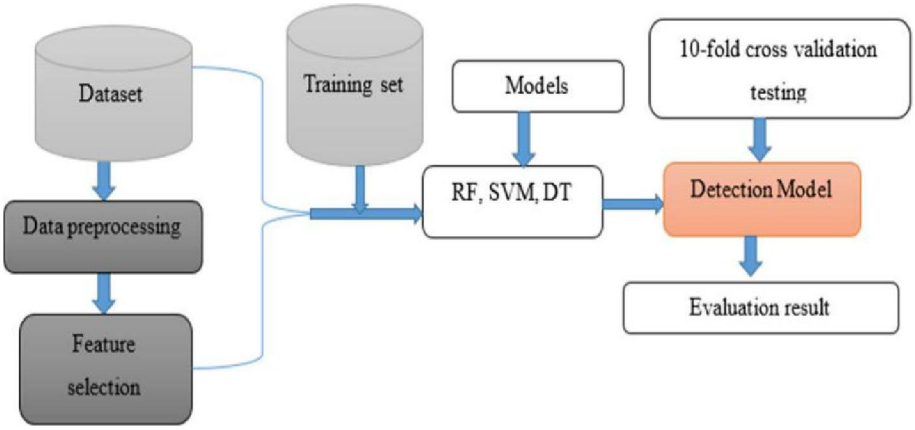

Fig. 2. Model building flow diagram for chronic kidney disease

4.3 Machine Learning Algorithms

The study's objective was to use machine learning techniques to forecast chronic renal disease. In this work, three machine learning methods were used: Random Forest, Support Vector Machine, and Decision Tree. The algorithms were chosen based on how well they classified data from other research studies [10, 11–17] and how well they were known for predicting chronic kidney disease.

- **Random Forest:** An ensemble learning technique, Random Forest is made up of several collections of decision trees. Both classification and regression use it. This model consists of a number of decision trees and produces the class target that is the target output by each tree with the highest vote result [12]. The tree is built using bagging and random feature selection in Random Forest, which results in an uncorrelated forest of trees. The collective prediction is more precise than any single tree's. Once the forest has been constructed, test cases are diffused through each tree, and the trees then make their separate predictions about the class [17].
- **Support vector machine (SVM):** One of the popular and practical supervised machine-learning algorithms, Support Vector Machine can be used for classification, instruction, and forecasting. In order to categorise each input in high dimensional data, a set of hyperplanes is constructed. In the signifier space of the training data, a discrete hyperplane is generated, and compounds are categorised according to the side of the hyperplane [14]. The decision lines that divide the data points are known as hyperplanes. The position and direction of the hyperplane are determined by support vec-tors, which are data points closest to the hyperplane. SVMs were originally designed to handle binary classification, but due to the vast amount of data that needs to be divided into more than two classes in the modern world, many academics are now attempting to use them to multiclass classification. The two most common methods used by SVM to solve multiclass issues are one-versus-rest and one-vs-one. One-versus-rest has been applied in this investigation. We employed OVR in conjunction with the SVM method for multi-classification. Each class is isolated from the other classes in the dataset using this technique. Also, one vs rest is an appropriate strategy with Linear SVC because it is employed in this study.
- **Decision Tree (DT):** One of the most well-liked supervised machine-learning methods for classification is the decision tree. By using sorted feature values to turn the data into a tree representation, Decision Tree solves the machine learning challenge. Each leaf node in a decision tree symbolises a class label that the instances belong to, and each node in the tree denotes features in an instance that need to be classed. As a predictive model that maps observations about an item to determine the goal value of instances, this model employs a tree structure

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

to divide the dataset based on the condition [18]. Fig. 3 illustrates decision-making in a binary class of chronic kidney disease.

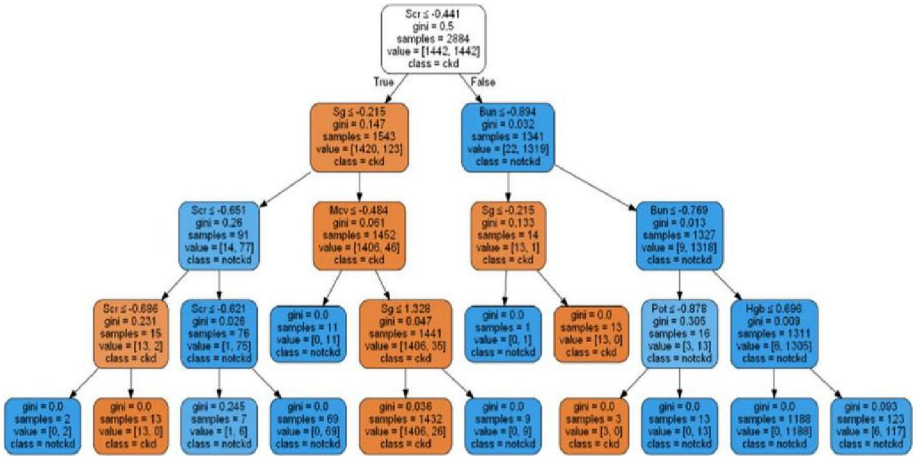

Fig. 3. Decision making in binary class of chronic kidney disease

V. EVALUATION OF A PREDCTION MODEL

The crucial process of creating an accurate machine-learning model is performance evaluation. To make sure the model fits the dataset and performs well on unobserved data, the prediction model must be assessed. The goal of the performance evaluation is to determine how well a model generalises on unobserved or out-of-sample data. One performance evaluation technique that compares and rates models by partitioning data is called cross-validation (CV). The original dataset was divided into k folds of equal size, of which nine were used to train a model and one to test or validate it. The average performance will be determined once this process has been repeated k times. In this investigation, tenfold cross-validation was performed. Several performance evaluation metrics have been computed, including accuracy, precision, recall, f1-score, sensitivity, and specificity.

- When both the actual value and the anticipated value are positive, this is referred to as a true positive (TP).
- When both the actual and anticipated values of a data point are negative, the condition is known as a true negative (TN).
- False positive (FP) situations occur when the anticipated value is positive even if the actual value of the data point was negative.
- When a data point's real value is positive and the predicted value is negative, this is known as a false negative (FN).

VI. PERFORMANCE ANALYSIS

We have employed five well-known algorithms for this project: KNN, Decision Tree, Neural Network, and Logistic Regression. The foundation of every algorithm is supervised learning. We are choosing the optimum approach by taking into account four criteria: specificity, sensitivity, log loss, and accuracy. The best algorithm to utilise to detect chronic kidney disease was determined to be logistic regression when all the algorithms were placed on a graph.

Accuracy: The proportion of accurately predicted data points among all the data points is known as accuracy. The number of true positives and true negatives divided by the total number of true positives, true negatives, false positives, and false negatives is how it is more precisely described.

Accuracy is (TP + TN)/(TP + TN + FP + FN)

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

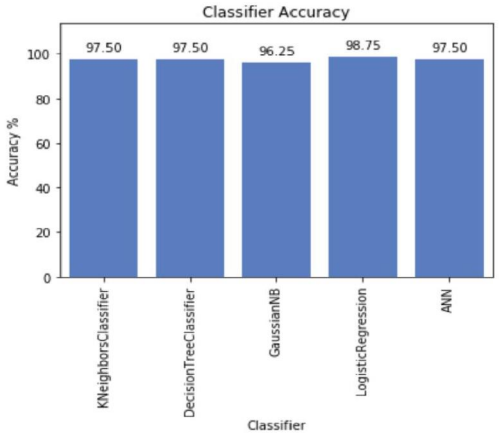

Fig. 4. Accuracy comparison

Specificity: The percentage of actual negatives that were anticipated as negatives is known as specificity (or true negative)

Specificity is TN/(TN + FP)

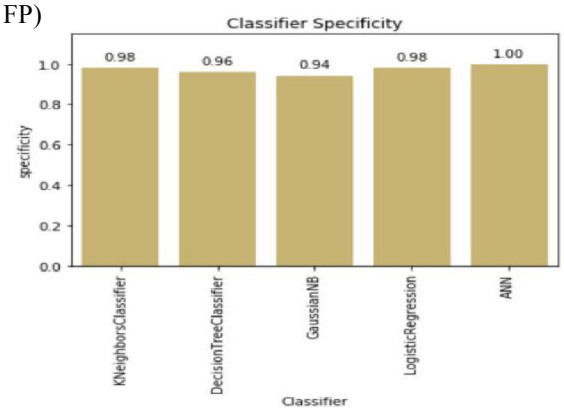

Fig. 5. Specificity comparison

Sensitivity: Sensitivity is a measure of the percentage of cases that were actually positive but were misclassified as positive (or true positive).

Sensitivity is TP /(TP + FN)

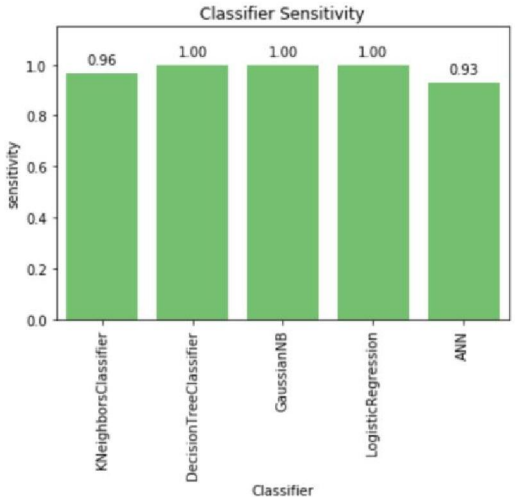

Fig. 6. Sensitivity comparison

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

VII. CONCLUSION

The best prediction method for early-stage CKD prediction was presented by this system. The models are trained and verified using the input parameters that were obtained from the CKD patients in the dataset. To perform the CKD diagnosis, learning models for K-Nearest Neighbors Classifier, Decision Tree Classifier, GaussianNB, Logical Regression, and Artificial Neural Networks are created. The effectiveness of the models is assessed using a number of comparative measures, including Accuracy, Specificity, Sensitivity, and Log Loss. The study's findings demonstrated that, when all measures are taken into account, the logical regression model predicts CKD better than the other models. A person's likelihood of developing CKD later in life could be determined with the aid of this approach. This technique would make it possible to predict a person's likelihood of developing CKD later in life, which would be extremely beneficial and affordable for most people. If a person is at risk, this model might be coupled with standard blood report creation to instantly flag that person. People wouldn't need to visit a doctor until the algorithms flagged them. For the current busy person, it would be less expensive and simpler as a result.

VIII. FUTURE SCOPE

This would make it possible to predict a person's likelihood of developing CKD later in life, which would be extremely beneficial and economical. If a person is at risk, this model might be coupled with standard blood report creation to instantly flag that person. People wouldn't need to visit a doctor until the algorithms flagged them. For the current busy person, it would be less expensive and simpler as a result.

REFERENCES

[1]. Rady EA, Anwar AS. Informatics in Medicine Unlocked Prediction of kidney disease stages using data mining algorithms. Informatics Med. 2019;15(2018):100178.

[2]. Almasoud M, Ward TE. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int J Adv Computer. 2019;10(8):89–96.

[3]. Salekin A, Stankovic J. Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. In: Proc. - 2016 IEEE Int. Conf. Healthc. Informatics, ICHI 2016, pp. 262–270, 2016.

[4]. Tekale S, Shingavi P, Wandhekar S, Chatorikar A. Prediction of chronic kidney disease using machine learning algorithm. Disease. 2018;7(10):92–6.

[5]. Kumar V. Evaluation of computationally intelligent techniques for breast cancer diagnosis. Neural Comput Appl. 2021;33(8):3195–208.

[6]. Xiao J, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med. 2019;17(1):1–13.

[7]. Alsuhibany SA, et al. Ensemble of deep learning based clinical decision support system for chronic kidney disease diagnosis in medical internet of things environment. Comput Intell Neurosci. 2021;3:2021.

[8]. Jasim A, Kaky M. Iintelligent systems approach for classification and management of by. 2017.

[9]. Saar-tsechansky M, ProvostF. Handling Missing Values when Applying Classi … cation Models. vol. 1, 2007.

[10]. Priyanka K, Science BC. Chronic kidney disease prediction based on naive Bayes technique. 2019. p. 1653–9.

[11]. Aqlan F, Markle R, Shamsan A. Data mining for chronic kidney disease prediction. 67th Annu Conf Expo Inst Ind Eng. 2017;2017:1789–94.

[12]. Subas A, Alickovic E, Kevric J. Diagnosis of chronic kidney disease by using random forest. IFMBE Proc. 2017;62(1):589–94.

[13]. Kapoor S, Verma R, Panda SN. Detecting kidney disease using Naïve bayes and decision tree in machine learning. Int J Innov Technol Explor Eng. 2019;9(1):498–501.

[14]. Vijayarani S, Dhayanand S. Data Mining Classification Algorithms for Kidney Disease Prediction. Int J Cybern Informatics. 2015;4(4):13–25.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

[15]. Drall S, Drall GS, Singh S. Chronic kidney disease prediction using machine learning : a new approach bharat Bhushan Naib. Learn. 2014;8(278):278–87.

[16]. KadamVinay R, Soujanya KLS, Singh P. Disease prediction by using deep learning based on patient treatment history. Int J Recent Technol Eng. 2019;7(6):745–54.

[17]. Ramya S, Radha N. Diagnosis of Chronic Kidney Disease Using. pp. 812–820, 2016.

[18]. Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J. Supervised machine learning algorithms: classification and comparison. Int J Comput Trends Technol. 2017;48(3):128–38.