A Deep Artificial Neural Network Model for Classification of Chronic Kidney Disease with Machine Learning Algorithms

Md. Ashik Miah* (0122310005)
Dhaka, Bangladesh
mmiah2310005@mscse.uiu.ac.bd

Dr. Riasat Azim[1]
Dhaka, Bangladesh
riasat@cse.uiu.ac.bd

Abstract— Chronic Kidney Disease (CKD) is a sort of persistent infection which implies it happens gradually throughout some stretch of time and perseveres for quite a while from that point. It is dangerous at its end stage and might be relieved by kidney substitution or normal dialysis which is a fake separating component. It is vital to distinguish CKD at the beginning phase with the goal that essential medicines can be given to forestall or fix the sickness. The principal center in this paper is around the arrangement the classification techniques, that is, Six classifier methods and Propose Artificial Neural Networks Model are used in this experiment consisting of Random Forest, Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes has been analyzed. Different measure has been used for comparison between algorithms for the dataset collected from standard UCI Machine Learning Repository. The Indians CKD dataset consists of 400 instances and 25 attributes with 2 classes collected from UCI. The Attribute of this dataset consists of two types of attributes which are numeric and nominal. Random Forest, Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, and Propose Model achieved, in that order, accuracy scores of 100%, 98.75%, 98.75%, 98.75%, 97.50%, 65% and 98.75%. With the highest accuracy, we find Random Forest (100%).

Keywords: Chronic kidney disease, machine learning, classification, prediction.

I. INTRODUCTION

Chronic Kidney disease (CKD) is a serious illness that affects millions of people worldwide and is a major cause of unfavorable health outcomes, especially in low- and middle-income nations where millions of people frequently pass away from the disease's lack of even minimal treatment [20]. Every chronic disease has stages, and the mortality rate varies according to the stage at which the sickness was not cured. A rising number of diabetic patients, hypertension, heart disease, diabetes mellitus, and a family history of kidney failure are significant risk factors for CKD [21]. Undiagnosed and consequently untreated CKD can result in hypertension and, in extreme situations, kidney failure.

A standard dataset for chronic kidney disease was obtained from UCI. If detected early and correctly, CKD can have numerous positive effects on patients. It prolongs the patient's life and raises the likelihood that the treatment will be successful. The purpose of this paper work is to predict renal illness through the application of a few chosen feature selection techniques and machine learning algorithms [11]. The aim is to gather various feature combinations and subsequently utilize them as input for machine learning algorithms. We compare the algorithms' performances once they have been developed based on specific features.

The kidney functions as a novel and practical conduit, freeing the body from waste products and poisons and reintroducing hormones, insulin, vitamins, amino acids, and other essential chemicals into the bloodstream [22]. Yet, things will unintentionally get really bad. Any form of recurrent nephritis is sometimes suggested to have CKD. Any variation from the normal structure or limits of the urinary organ is considered an "infection," regardless of the likelihood that it would make a man feel ill or cause complications [23]. It's a common problem that can affect people of any age. CKD is a worldwide health crisis. Globally, CKD directly resulted in an estimated 1.23 million deaths in 2017 with an additional 1.36 million deaths attributable to cardiovascular disease resulting from impaired kidney function (HealthData). Just over 3 million people in the UK are thought to be at risk for CKD. There were 697.5 million cases of CKD in 2017 (HealthData). Nearly one-third of those patients lived in two countries – China, with about 132 million cases, and India, with about

115 million. Further, 10 other nations the US, Indonesia, Russia, Japan, Brazil, Pakistan, Mexico, Nigeria, Bangladesh, and Vietnam each had more than 10 million cases in 2017 (HealthData). CKD is a result of a confluence of many distinct illnesses that typically train the kidneys.

The paper is organized in sections. Section II discusses research works related to this paper. The section III gives brief on the propose method followed in the research as well as describes the concepts used. Section IV describes discussion. Section V presents performance analysis and section VI concludes the research.

II. LITERATURE REVIEW

It is not new to use machine learning algorithms to solve issues in the healthcare industry. Several investigators have attempted that. Several approaches and strategies have been employed to classify and forecast the patient's status with CKD [3]. They have developed a decision support system that uses random subspace categorization to diagnose and forecast chronic renal failure [4]. This research focused on predicting the progression of kidney disease using machine learning. It examined various predictive models, including logistic regression, K-nearest neighbors (KNN), support vector machine (SVM). The study emphasized the importance of feature selection in improving model performance [2]. Addressing the issue of imbalanced data, this study proposed techniques to improve the prediction of chronic kidney disease, with a specific focus on addressing class imbalance. It explored the use of synthetic data generation and ensemble learning methods [1]. This study concentrated on the prediction of kidney disease & differentiate the performance of various machine learning algorithms that are primarily based on its accuracy [6]. This research work has idolized Recode to compare their performance. The pivotal purpose of this study is to analyze the Chronic Kidney Disease dataset and conduct CKD and Non CKD classification cases. It utilized machine learning algorithms such as logistic regression and random forests and examined the impact of different feature sets on predictive accuracy [6]. They preprocess the dataset and then used the filter method of feature selection that is univariate selection and correlation matrix along with feature importance to find best features from the dataset and proposed algorithm that is Decision tree, Random forest and logistic regression have achieved an accuracy of 98.48, 94.16 and 99.24 respectively & Precision of 100, 95.12 and 98.82 and recall of 97.61, 96.29 and 100 [11]. In this paper they are used three machine learning models RF, SVM, DT and used two feature selection methods RFECV & UFS for build proposed models & applied to original datasets with all 19 features. SVM & RF provides highest accuracy 99.8% [13]. This research successfully created a framework for the "MedAi" system, which involved creating a machine learning model and designing a smartwatch with numerous sensors and suggesting a structure for an Android health app that forecasts twelve different types of disease [12]. The RF algorithm was used to build the model, and it produced an accuracy of 99.4% compared to other well-liked techniques like KNN (99.3%) and XGB (98.56%) [12].

III. METHODOLOGY

In our proposed method, Chronic Kidney Disease can be detected more efficiently and less costly within a short time. This project worked with preprocessed data to train and test model. In the first stage, preprocessed data are divided into two parts. Most of those are used in the training phase (80%), and the rest (20%) are used in the testing phase.

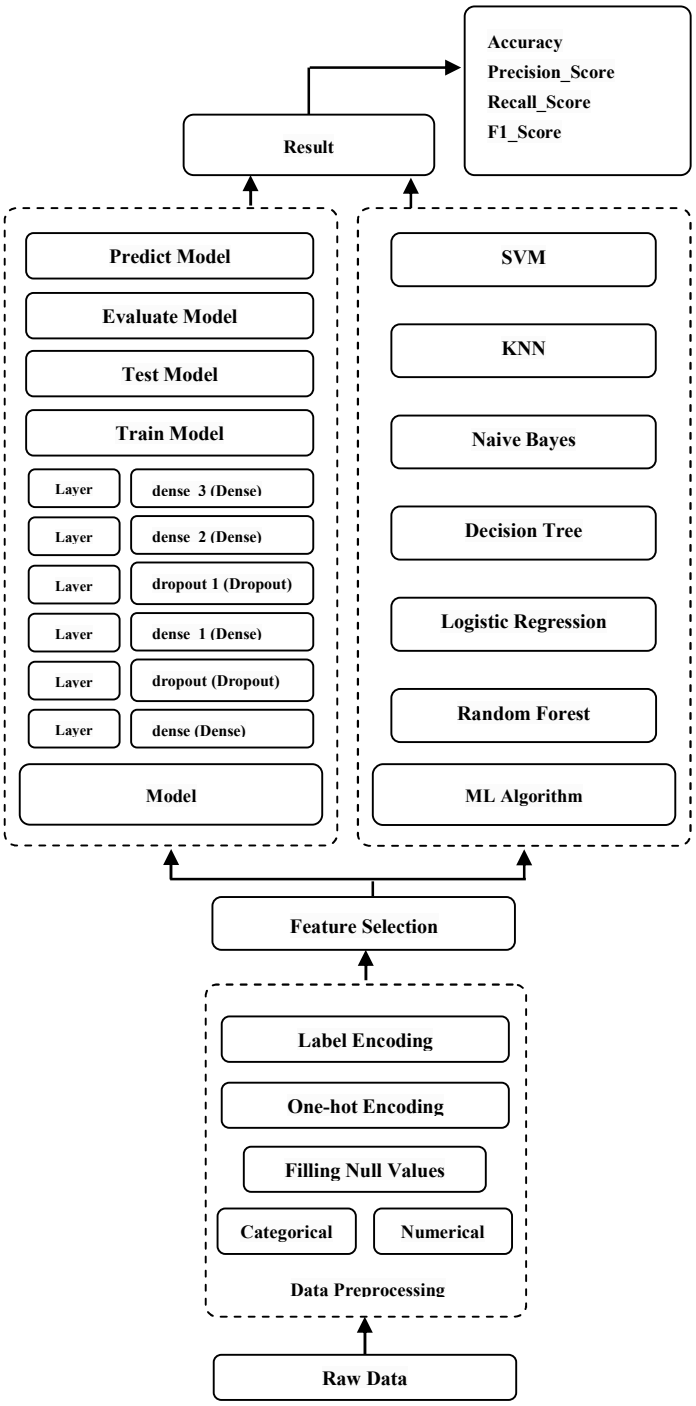

Figure 1. Proposed chronic kidney disease classification system.

The following, Figure 1 represents working procedure of our proposed CKD classification system & Table I represents in detail our input attributes.

In the training, testing phase, the proposed system has trained our dataset using machine learning algorithms like Random Forest, Logistic Regression, Support Vector Machines (SVM), k-Nearest Neighbors, Naive Bayes, Decision Tree Classifier & Propose Artificial Neural Networks Model. Using the PyCharm Community Edition platform, we have trained and finally predicted the result with testing data.

Our trained system will indicate the presence of CKD in a patient. The primary intention of this raised system is to detect kidney disease more efficiently and accurately.

A. Preprocessing

We utilized a UCI Machine Learning Repository dataset which is publicly available. This dataset contains the medical records of 400 people. There are 400 occurrences in this data set, of which 65% have CKD and 35% Non-CKD. There are a total of 25 attributes. It is necessary to classify the data into CKD and Non-CKD. There are some missing values in the dataset that require attention.

TABLE I**.** REPRESENTATION OF INPUT ATTRIBUTES.

SL	Input Attributes	Type
1	Age	numerical
2	Blood pressure	numerical
3	Specific Gravity	nominal
4	Albumin	nominal
5	Sugar	nominal
6	Red blood cells	nominal
7	Pus Cell	nominal
8	Pus Cell clumps	nominal
9	Bacteria	nominal
10	Blood glucose random	numerical
11	Blood urea	numerical
12	Serum creatinine	numerical
13	Sodium	numerical
14	Potassium	numerical
15	Hemoglobin	numerical
16	Packed cell volume	numerical
17	White blood cell count	numerical
18	Red blood cell count	numerical
19	Hypertension	nominal
20	Diabetes Mellitus	nominal
21	Coronary Artery Disease	nominal
22	Appetite	nominal
23	Pedal Edema	nominal
24	Anemia	nominal
25	Class	nominal

We have chosen the UCI dataset, whereby heart illness is caused by eleven parameters that have been allocated. While the UCI dataset is not equally divided for disease and non-disease patients, the selected UCI dataset nearly evenly contains condition and non-disease patient data. As a result, the network receives fewer inputs and learns more precisely and effectively. The system predicts the kidney condition using the following 25 appropriate attributes: age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia, class.

B. Splitting

The training and testing phases determine how well the dataset is classified. We split our entire dataset into three halves in order to obtain a better outcome: the majority of the dataset (80%) is used for training, with the remaining 20% being used for testing.

C. Classification Models

Six machine-learning algorithms Random Forest, Logistic Regression, Support Vector Machines, K-Nearest Neighbors, Naive Bayes, Decision Tree Classifier & Propose Neural Networks Model are applied to the training data. This is a detailed explanation of each algorithm below.

D. Random Forest

Random Forest is a popular machine learning algorithm used for both classification and regression tasks [19]. It belongs to the ensemble learning methods, which involve combining multiple models to improve the overall performance [19]. Random Forest is known for its robustness, scalability, and ability to handle large datasets with high dimensionality.

E. Logistic Regression

Logistic Regression is a statistical method used for binary classification tasks, where the target variable has two possible outcomes [6]. Despite its name, it is a classification algorithm rather than a regression algorithm. It estimates the probability that a given input belongs to a particular class [6].

F. Support Vector Machines

Support Vector Machine is one of the most widely used supervised learning algorithms for classification and regression issues shown in Figure 2.

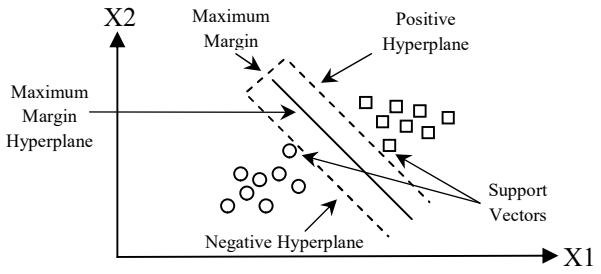

Figure 2. Support vector machine.

However, it is mostly utilized in machine learning for classification difficulties [14]. It is widely used for classification objectives as it produces significant accuracy with reduced computation power. By altering the distance between the data points and the hyperplane, this classifier seeks to construct a hyperplane that can divide the classes as much as feasible. It creates a hyperplane for each attribute that has a coordinate in the dataset. It is simple to create a linear hyper-plane between these two classes in the SVM classifier [14]. Another pressing concern is whether we need to manually implement this functionality in order to have a hyper-plane.

Support vector machines can manipulate multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate diverse classes. SVM iteratively generates an optimal hyperplane, used to minimize an error. The main idea of SVM is to find a maximum marginal hyperplane (MMH) that best divides the dataset into classes.

$$(x) = wT + b \quad (1)$$

Where w is a dimensional vector and b is an offset. A subsequent optimization problem can solve that.

G. K-Nearest Neighbors

K-nearest neighbors algorithm is a type of supervised ML algorithm that can be used for both classification as well as regression predictive problems [15]. However, it is mostly employed in industry for categorization and prediction challenges. KNN is well defined by the following two characteristics: KNN is a lazy learning algorithm since it does not have a dedicated training phase and instead uses all of the data for training and classification [15]. KNN is also a non-parametric learning method since it makes no assumptions about the underlying data. A KNN categorizes an example to the class that appears the most frequently among its k nearest neighbors. The restriction k is used to fine-tune the categorization job.

H. Naive Bayes

Bayesian networks are one of the most effective operational classifiers [4]. Naive Bayes classifiers are probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between features. A Naive Bayesian model is simple to construct and does not require iterative parameter estimate, making it very useful in medical science for identifying heart patients. Despite its simplicity, the Naive Bayesian classifier often works admirably and is commonly used because it outperforms more complex classification algorithms. From P(c), P(x), and P(x|c), the Bayes theorem can be used to get the posterior probability, P(c|x) [4]. The effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors, according to the Naive Bayes classifier. This is known as class conditional independence. The Bayesian Classification is both a supervised learning method and a statistical classification method. It assumes an underlying probabilistic model and allows us to convey uncertainty about the model in a logical manner by calculating probabilities of outcomes. It has the ability to tackle diagnostic and predictive issues. The Bayesian Theorem is the foundation of the Naive Bayes method.

I. Decision Tree Classifier

The decision tree is the most powerful and widely used tool for classification and prediction. A decision tree is a flowchart-like tree structure in which each internal node represents a test on an attribute, each branch represents a test outcome, and each leaf node (terminal node) stores a class label [16]. Decision trees categorize instances by sorting them along the tree from the root to some leaf node that offers the instance's categorization.

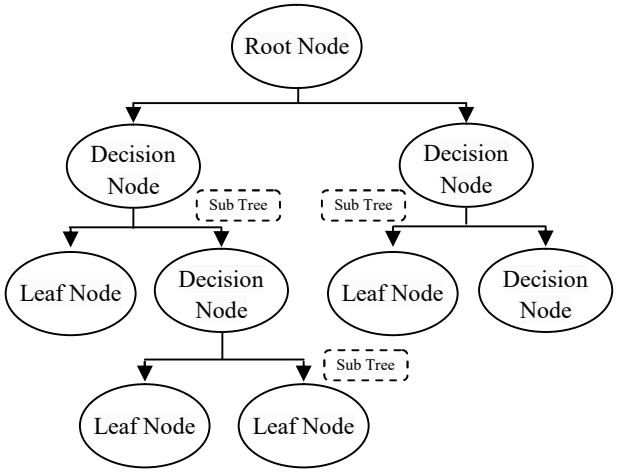

Figure 3. Decision tree.

As indicated in the diagram above, an instance is categorized by starting at the root node of the tree, checking the attribute given by this node, and then going down the tree branch according to the value of the attribute [15]. This method is then repeated for the new node-rooted sub tree. The decision tree in the above graphic categorizes a certain morning based on whether it is suitable for playing tennis and returns the classification linked with the specific leaf (In this situation, either Yes or No). For example, the instance would be sorted down the decision tree's leftmost branch and so classed as a negative instance. In other words, a decision tree represents a disjunction of conjunctions of restrictions on instance attribute values.

IV. DISCUSSION

Compared to other systems, the suggested working model is a useful and time-saving method. By delivering an initial diagnosis in a timely manner, the method reduces treatment expenses. Patients can use our computer-aided technique to detect chronic kidney disease. Anyone can check for CKD on a daily basis using the proposed approach, saving a significant amount of money and time. A thorough explanation of our model architecture can be found in Section C.

A. Confusion Matrix

Figure 6 shows that are often used to describe the performance of a classification model. It allows visualization of the performance of an algorithm by comparing actual and predicted classes. The confusion matrix is particularly useful for evaluating the performance of machine learning algorithms in binary classification tasks, although it can also be extended to multi-class classification.

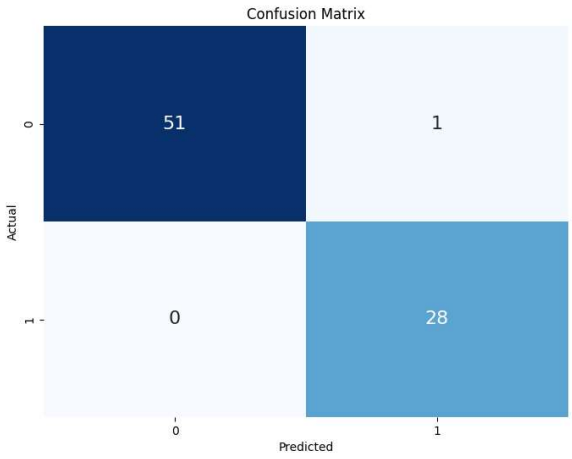

Figure 4. Confusion matrix.

B. Performance Metrics

Performance metrics in machine learning (ML) are used to evaluate the performance of a predictive model or algorithm. These metrics help in assessing how well the model is performing in terms of its ability to make accurate predictions on unseen data.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

C. Propose Deep Artificial Neural Model Architecture

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 32)	800
dropout (Dropout)	(None, 32)	0
dense_1 (Dense)	None, 64)	2112
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 128)	8320
dense_3 (Dense)	(None, 1)	129

Total params: 11361 (44.38 KB)
Trainable params: 11361 (44.38 KB)
Non-trainable params: 0 (0.00 Byte)

Figure 7. Propose model architecture

V. PERFORMANCE ANALYSIS

A. Experimental Results

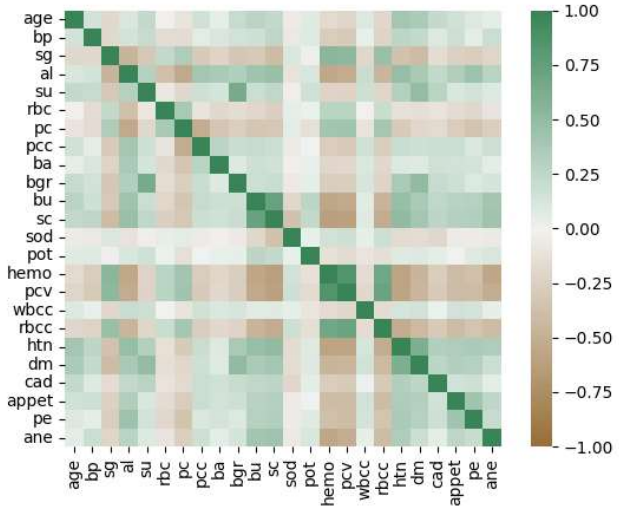

Figure 6. Heatmap.

The following Figure 6 shows the heatmap to visualize the correlation between attributes. A heatmap is a graphical representation of data where values in a matrix are represented as colors. It's particularly useful for visualizing two-dimensional data, such as a correlation matrix, confusion matrix, or any other type of matrix where the values can be represented using a color gradient.

TABLE II. ACCURACY RESULT.

ML Algorithms	Accuracy	Model	Accuracy
Random Forest	100%	Propose Model	98.75%
Logistic Regression	98.75%		
Decision Tree	98.75%		
Naive Bayes	98.75%		
KNN	97.50%		
SVM	65%		

Best Accuracy Result is Random Forest “100%” & Our Propose Model Accuracy “98.75%”.

TABLE III. CLASSIFICATION REPORT

ML Algorithms\Model	Precision	Recall	F1-Score
Random Forest	1.0	1.0	1.0
Logistic Regression	0.97	1.0	0.98
Decision Tree	0.97	1.0	0.98
Naive Bayes	0.97	1.0	0.98
KNN	0.93	1.0	0.97
SVM	-	-	-
Propose Model	0.96	1.0	0.98

B. Training and Validation Loss, Accuracy

The following Figure 8 shows the training and validation loss, accuracy. The training loss is a measure of how well the model is performing during the training phase. It represents the error between the model's predictions and the actual target values on the training dataset.

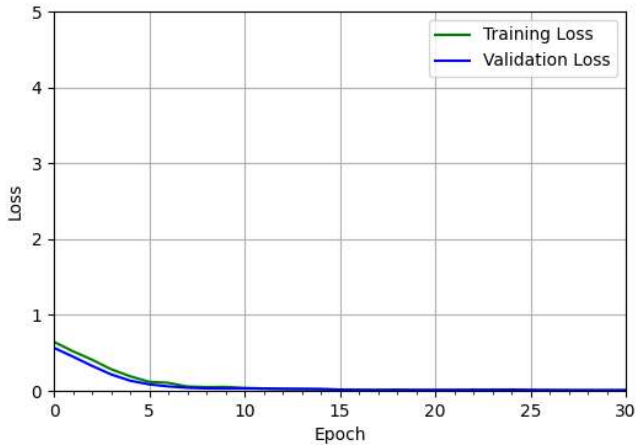

Figure 8. Training and validation loss

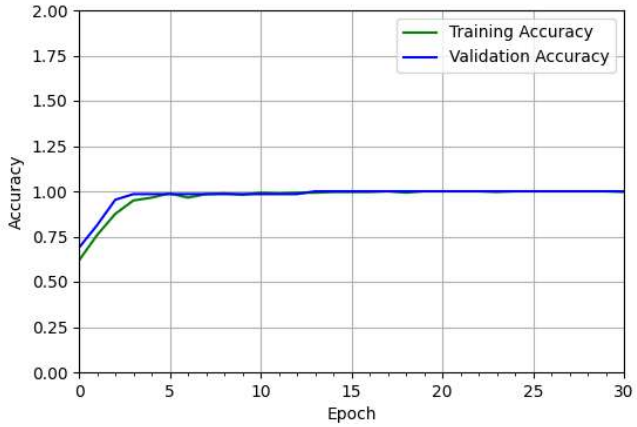

Figure 9. Training and validation accuracy

The validation loss is similar to the training loss, but it is computed on a separate validation dataset that the model has not seen during training. This helps assess whether the model is overfitting. Accuracy gives a measure of the model's correctness in classification tasks

VI. CONCLUSION

In conclusion, using the Chronic Kidney Disease data set that we extracted from UCI, we were able to assess how well various machine learning algorithms performed. To identify the best features from the dataset, we preprocessed it before using the filter method of feature selection, which combines feature importance, correlation matrix, and univariate selection. The suggested algorithms Random Forest, Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, and Propose Model achieved, in that order, accuracy scores of 100%, 98.75%, 98.75%, 98.75%, 97.50%, 65% and 98.75%. With the highest accuracy, we find Random Forest (100%).

The advantage of this strategy is that the classification process is much faster, allowing physicians to treat CKD patients as soon as possible and classify a larger patient population in a shorter amount of time. Due to the small size of the 400 items dataset utilized in this paper, it has been recommended to work with larger datasets in the future or compare the dataset's results with those of another dataset that contains the same. Aims have also been made to use relevant datasets to predict the likelihood that an individual with this syndrome will experience chronic risk factors like diabetes, hypertension, and a family history of kidney failure, in an effort to reduce the incidence of CKD.

REFERENCES

[1] Pinar Yildirim, “Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron,” IEEE 41st Annual Computer Software and Applications Conference – 2017.

[2] A. Charleonnan et al., "Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques," MITiCON – 2016.

[3] Q. Zhang and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: systematic review," BMC Public Health – 2008.

[4] R Ani, Greeshma Sasi, U Resmi Sankar, O. S. Deepa, “Decision Support system for diagnosis and prediction of Chronic Renal Failure using Random Subspace Classification,” IEEE – 2016.

[5] J. Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Translational Medicine – 2019.

[6] Bhavya Gudeti et al., “A Novel Approach to Predict Chronic Kidney Disease using Machine Learning Algorithms,” IEEE – 2020.

[7] W.H.S.D Gunarathne, K.D.M Perera, K.A.D.C.P Kahandawaarachchi, “Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)”, IEEE 17thInternational Conference on Bioinformatics and Bioengineering – 2017.

[8] Dr. Vijayaprabakaran.K, Pratheek Reddy.P, Puthin Kumar Reddy.T, Munnaf.K, Reddi Prasad.G, "Chronic Kidney Disease Diagnosis using Machine Learning," IRJET – 2021.

[9] S. A. Shinde and P. R. Rajeswari, “Intelligent health risk prediction systems using machine learning: a review,” IJET – 2018.

[10] P. Yildirim, “Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction,” in Proceedings - International Computer Software and Applications Conference – 2017.

[11] Rahul Gupta , Nidhi Koli, Niharika Mahor, N Tejashri, “Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease”, INCET – 2020.

[12] Shinthi Tasnim Himi, Natasha Tanzila Monalisa, Md Whaiduzzaman, Alistair Barros, Mohammad Shorif Uddin, “MedAi: A Smartwatch-Based Application Framework for the Prediction of Common Diseases Using Machine Learning”, IEEE – 2023.

[13] Dibaba Adeba Debal, Tilahun Melak Sitote, “Chronic kidney disease prediction using machine learning techniques”, Journal of Big Data – 2022.
[14] Sourish Ghosh, Anasuya Dasgupta, Aleena Swetapadma, “A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification”, ICISS – 2019.
[15] Kashvi Taunk, Sanjukta De, Srishti Verma, Aleena Swetapadma, “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification”, ICICCS – 2019.
[16] Bahzad Taha Jijo, Adnan Mohsin Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning”, JASTT – 2021.
[17] Sujan Banik, Antara Ghosh, “Prevalence of chronic kidney disease in Bangladesh: a systematic review and meta-analysis”, International Urology and Nephrology – 2020.
[18] Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Selectiing Important Predictive Attributes", International Conference on Healthcare Informatics – 2016.
[19] Dilip Singh Sisodia, Akanksha Verma, “Prediction Performance of Individual and Ensemble learners for Chronic Kidney Disease”, International Conference on Inventive Computing and Informatics – 2017.
[20] Debal, D.A., Sitote, T.M., “Chronic kidney disease prediction using machine learning techniques," Journal of Big Data – 2022.
[21] AS Levey et al., “Chronic kidney disease as a global public health problem: Approaches and initiatives – a position statement from Kidney Disease Improving Global Outcomes,” Kidney International – 2007.
[22] B. Khan, R. Naseem, F. Muhammad, G. Abbas and S. Kim, “An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy,” IEEE – 2020.
[23] Chen TK, Knicely DH, Grams ME, “Chronic Kidney Disease Diagnosis and Management: A Review,” JAMA – 2019.