JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predicting Chronic Kidney Diseases Using Machine Learning Algorithms

Samuel Mulatu,
Information Systems,
Wachemo University, Hosanna, Ethiopia

Abstract: ***A medical condition known as chronic kidney disease, or CKD, occurs when the kidneys are so severely damaged that they are unable to filter blood as effectively as they should. The objective of this study is to develop a model which predict CKD using machine learning algorithm. Classification is the most widely used machine learning task used to build classified models. The researcher will pick three classification algorithms, namely Support Vector Machine, Decision tree and K-nearest neighbors and Decision tree provide 97.85% accuracy.***
Keywords: CKD, SVM, KNN, Decision tree

Introduction

A medical condition known as chronic kidney disease, or CKD, occurs when the kidneys are so severely damaged that they are unable to filter blood as effectively as they should. Urine is created by the kidneys' primary function of eliminating waste and excess water from the circulation. The reason the illness is referred to as "chronic" is because kidney damage develops gradually over an extended period of time [1] [2] .

Among NCDs, chronic kidney disease is a significant and prevalent public health issue. Up to 10%–15% of people worldwide are impacted. Chronic kidney disease (CKD) is a serious global public health concern that might need significant financial and social resources. More over 2 million individuals worldwide now depend on dialysis or kidney transplants for their survival; of them, 20% receive care in 100 low-income nations, which account for half of the world's population. A major global public health concern, chronic kidney disease (CKD) has become much more common as the population ages and develops more chronic illnesses. According to a recent study, between 1990 and 2016, the death rate from CKD climbed by 98% and the prevalence of CKD increased by 87% globally [3] [4, 5].

Five stages of renal disease were identified by the National renal Foundation (NKF). With different tests and treatments needed at each stage, this helps physicians give the best care possible. The glomerular filtration rate, or GFR, is a mathematical formula that uses an individual's age, gender, and serum creatinine level (found through a blood test) to identify the stage of kidney disease. One of the main markers of renal function is creatinine, a waste product produced during muscle contraction. Creatinine is eliminated from the blood by the kidneys when they are functioning properly, but blood creatinine levels increase when kidney function declines. [4] [6] [7].

Table 1. risk of renal outcome according to the GFR (Filtration rate in ML/MIN/1.73 M2) and Albuminuria; Glomerular Filtration rate in ML/MIN/1.73 M2 [6].

		Albuminuria		
GFR		<30mg/g	30-300mg/g	300mg/g
Stage 1	>90	Low risk	Moderate risk	High risk
Stage 2	60-90	Low risk	Moderate risk	High risk
Stage 3A	45-59	Moderate risk	High risk	Very high risk
Stage 3B	30-44	High risk	Very high risk	Very high risk
Stage 4	15-29	Very high risk	Very high risk	Very high risk
Stage 5	<15	Very high risk	Very high risk	Very high risk

Future prediction capabilities are supported by machine learning techniques. Numerous supervised machine learning techniques exist, including Naïve Bayes, Random Forest, K-nearest neighbor, support vector machines, logistic regression, and multi-class classification. We use unlabeled data to train the dataset in the unsupervised learning technique. Using this approach, we reduce the dimensionality and split the data into two groups according to similarity. Clustering algorithms are the most often used unsupervised learning techniques. Numerous clustering algorithms exist, including K-means clustering, Hierarchical clustering, and many more. Feature extraction, variable selection, and attribute selection are other names for feature selection. It uses relevant data sets and avoids redundant and irrelevant data. Machine learning algorithms and classifiers are currently considered the most reliable technologies for diagnosing various diseases such as heart disease, diabetes, and liver disease prediction[8] [9].

Statement of the Problem

Chronic kidney disease (CKD) has emerged as one of the most prominent causes of death and suffering in the 21st century it affects nearly 10% of the general population world. A recent study reported that the global prevalence of CKD increased by 87% and the death rate from CKD rose by 98% from 1990 to 2016wide [5] [10].

More than 850 million people currently suffer from kidney disease, with a disproportionate burden being felt by people in low- and middle-income countries (LMICs), where access to healthcare is severely limited. Recent studies have shown that Africans are at higher risk of developing chronic kidney disease, that it occurs at a younger age, and that kidney failure progresses more rapidly. Data is lacking in many African countries, and the limited data available make it difficult to understand the true burden of CKD in Africa, including age-standardized rates, costs of care, and health impacts on patients, their families, and society. The epidemiological burden (epidemiological burden) is probably underestimated and therefore little known[11].

Data on the prevalence of CKD are limited. However, some studies suggest that kidney disease is a serious public health problem in Ethiopia. A cross-sectional study estimated that 12.2% of Ethiopians suffer from CKD, and the number has recently increased along with diabetes and hypertension. In Ethiopia, up to 41% of people under the age of 35 and 62% of men have CKD [3].

Hence, machine learning provides a way to get the information buried in the data. They can find patterns embedded in large and complex sets of data, where these patterns elude conventional statistical approaches to analysis [9] [2] [12].

The objective of this study is to develop a model which predict CKD using machine learning algorithm.

Methodology

A. Data Collection

In this study, we used a real-world dataset to predict CKD status in patients. The data collected is widely available data in the UCI Machine Learning repository. The available datasets are specifically used for research in chronic kidney disease. It consists of 400 records of his, each with 25 CKD-related traits. The data consisted of real numbers, decimal values, and nominal values [2] [9].

B. Modeling Techniques

Classification is the most widely used machine learning task used to build classified models. The researcher will pick three classification algorithms, namely Support Vector Machine, Decision tree and K-nearest neighbors. The researcher selects these classification algorithms because of their high fault - tolerant.

C. Evaluation Methods

The accuracy of models developed using data mining techniques is evaluated on the basis of the accuracy of the classifiers, the Precision, the Recall, the F-Measure and the True Positive rate.

Scope and Limitation of the Study

The scope of this research is limited to design and develop a productive model that can identify or predict CKD using machine learning algorithm. Since to designing and developing a predicting model for CKD using machine learning algorithm is the major goal of this research.

This study is limited to developing predictive model using classification mining techniques. However, there are other mining techniques which might show some interesting patterns or relationships in the selected dataset and also it uses.

Significance of the Study

The application of machine learning (ML) in health informatics is gaining increasing attention. Timely diagnosis of kidney disease and subsequent immediate response is an example that highlights the important role of ML diagnostic algorithms. ML in kidney disease diagnosis (MLKDD) is an active research topic aimed at assisting physicians with computer-assisted systems. Various studies have attempted to test the feasibility, applicability, and mutual superiority of different ML techniques. However, the lack of a comprehensive survey of this literature has always been an obvious shortcoming. Therefore, this paper provides a comprehensive literature review on the use of ML in kidney disease diagnosis by presenting two different frameworks. One is a framework for ML that categorizes various aspects of kidney disease diagnosis, and the other is a framework for medical subdisciplines related to his MLKDD [13].

Overall, the motive of the study is to examine the applicability of specific supervised machine learning classifiers and offer their compatibility in detecting several serious diseases such as the diagnosis of CKD at an early stage [8].

Literature Review and Related Work

This area presents the review of different literatures which was carried out to create an adequate framework for the current study. Some of the latest and significant researches, relevant to the current study have been briefly reviewed in this chapter. The chapter begins with introducing CKD, Further, the machine and the data processing models adopted by this research, and the attribute selection process have been discussed.

A. Chronic Kidney Diseases

The definition and classification of Kidney Disease Outcomes Quality Initiative (K/DOQI) were approved with refinements. Chronic kidney disease is defined as kidney damage or a glomerular filtration rate (GFR) of 60 ml/min/1.73 m2 for at least three months, regardless of the cause, in other words, chronic kidney disease (CKD) means that your kidneys are damaged and cannot filter the blood as it should. This disease is called chronic because kidney damage occurs slowly over a long period of time. This damage can cause waste products in the body. Kidney disease does not develop overnight. It will happen slowly and in stages. Most people in the early stages have no symptoms. They may not know anything is wrong. But if detected and treated, kidney disease can often be slowed or stopped. [1] [2] [14].

B. Machine Learning Application

Machine learning (ML) is an umbrella term that refers to many algorithms that make intelligent predictions based on a set of data. These datasets are often large, consisting of potentially millions of unique data points. Recent advances in machine learning have achieved human-like semantic understanding and information, and sometimes the ability to perceive abstract patterns more accurately than a human expert [15]. Machine learning is defined as a field of research that gives computers the ability to learn without being specially programmed [16]. Machine learning algorithms are divided into a taxonomy based on the desired result of the algorithm. Common algorithm types include [17]

C. Supervised Learning

Supervised learning is a machine learning task that trains a function that maps input to output based on sample input-input pairs. It derives a function from identified training data consisting of a set of training examples. Supervised machine learning algorithms are those algorithms that require external help. The

input dataset is divided into train and test data. The train dataset has an output variable that must be predicted or classified. All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification. The workflow of supervised machine learning algorithms is shown in the figure below [16] [17].

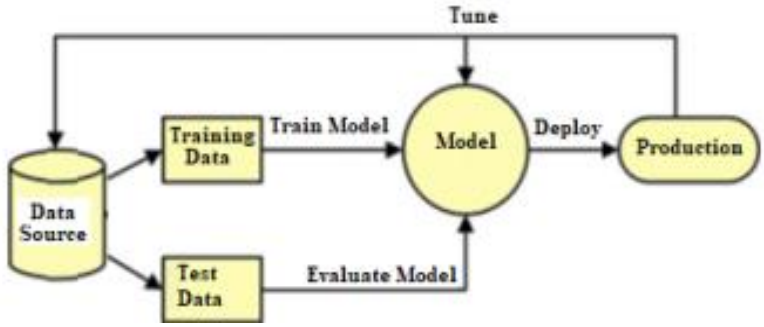

Figure1: supervised learning framework [16]

Some common algorithms of Supervised learning are [9]

Decision Tree: A decision tree classifies instances by ordering them from the root to some leaf nodes based on feature values. Each node represents some decision (test condition) in an instance attribute, while each branch represents a possible value for that attribute. Case classification starts at the root node, called the decision node. Based on the value of the node, the tree moves down along the edge corresponding to the value of the function test output. This process continues in the subtree that is controlled by the new node at the end of the previous edge. Finally, the leaf node represents the classification categories or the final decision. When using a decision tree, the focus is how to decide which attribute is the best classifier at each node level. To calculate that node value, statistical metrics such as information validation, Gini index, chi-square, and entropy are calculated for each node. Several algorithms are used to implement decision trees. The most popular are: Classification and Regression Tree (CART), Iterative Dichotomizer3 (ID3), Automatic Interaction Detection (CHAID), Chi-square C4.5 and C5.0 and M5 [16] [18].

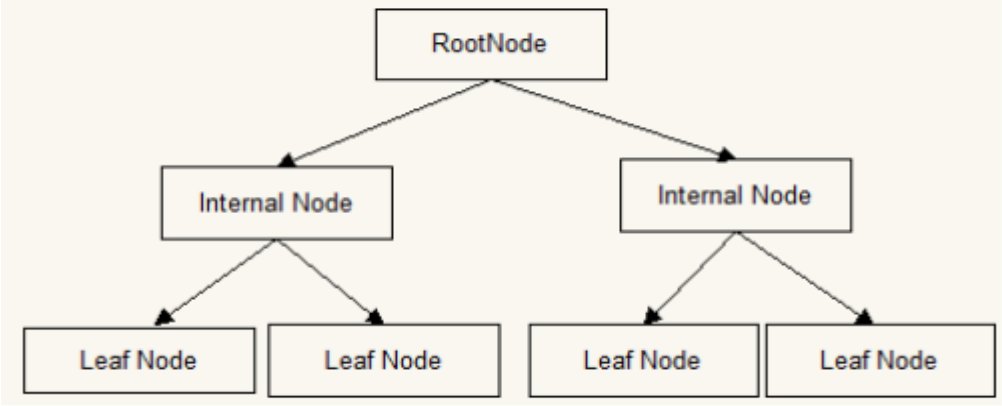

Figure 2: decision tree [19]

The main advantages of decision trees over other algorithms are that they are quick to build, efficient and easy to understand as each node is labelled in terms of the input attributes. The basic algorithm for decision tree induction is greedy algorithm that constructs decision trees in a top down recursive divide and conquer manner. The pseudo code is summarized as follows [19].

Create a node N;

If samples are all of the same class, C then

Return N as a leaf node labeled with the class C;

If attribute-list is empty then

Return N as a leaf node labeled with the most common class in samples;

select test-attribute, the attribute among attributes-list with the highest information gain; label node N with test-attribute; for each known value AI of test-attribute grow a branch from node N for the condition test-attribute= AI;

let s_i be the set of samples for which test-attribute= a_i;

If s_i is empty then attach a leaf labeled with the most common class in samples;

else attach the node returned by

Generate_decision_tree(si,attribute-list_test-attribute)

Naïve Bayes: The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by calculating combinations of density and values for a given data set. The algorithm uses Bayes theorem and assumes that all attributes are independent based on the value of the class variable. This conditional independence assumption rarely holds true in real-world applications, so it is characterized as naive Bayes, but the algorithm performs well and learns quickly on a variety of supervised classification problems [15] [16] [19].

Steps to calculate Naïve Bayes formula as follows:

Step 1: Convert the dataset into a frequency table

Step 2: Create Likelihood table by finding the probabilities

Step 3: The class with the highest posterior probability is the outcome of prediction.

Use Naïve Bayesian formula to calculate the posterior probability for each class.

$P(c|x) = P(x|c)P(c)/P(x)$…… (1)

Pseudo Code of Naïve Bayes

Input:

Training dataset T, F= (f1, f2, f3, ….fn) // value of the predictor variable in testing dataset. Output: A class of testing dataset.

Steps

1. Read the training dataset T;
2. Calculate the mean and standard deviation of the predictor variables in each class;
3. Repeat Calculate the probability of fi using the gauss density equation in each class; Until the probability of all predictor variables (f1, f2, f3,.., fn) has been calculated.
4. Calculate the likelihood for each class;
5. Get the greatest likelihood

Support Vector Machine: In machine learning, another common technique that can be used for classification, regression, or other tasks is the support vector machine (SVM) [15]. In machine learning, support vector machines are supervised learning models and associated learning algorithms that analyze data for classification and regression analysis. In addition to linear classification, SVMs can efficiently perform nonlinear classification using the so-called kernel trick, implicitly mapping their input to high-dimensional feature spaces. It basically draws edges between categories. Margins are drawn so that the

distance between the margin and the classes is maximized and thus the classification error is minimized [16] [20].

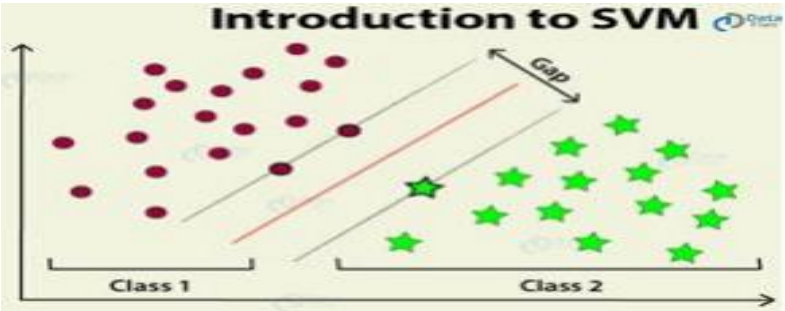

Figure 3: support vector machine [16]

Pseudo Code of Support Vector Machine
initialize Yi = YI for i ∈ I
repeat
compute svm solution vv , b for data set with imputed labels
compute outputs ii = (vv , xi) + b
for all xi in positive bags
set yi = sgn(fi) for every i e i, yi = 1
for (every positive bag bi) end
if (liei(l + yi)/2 == 0)
compute i = arg maxiei ii*
set yi = 1*
end
while (imputed labels have changed)
output (vv, b)

K-Nearest Neighbor (KNN): This algorithm is known as a non-parametric algorithm that can be used for classification. It is also known as a lazy algorithm because it does not use special training or make assumptions. In the image below, the data is classified into two categories, and the new point is selected according to the Euclidean distance of the neighbors, and the closest distance neighbors are selected for the new data point[20].

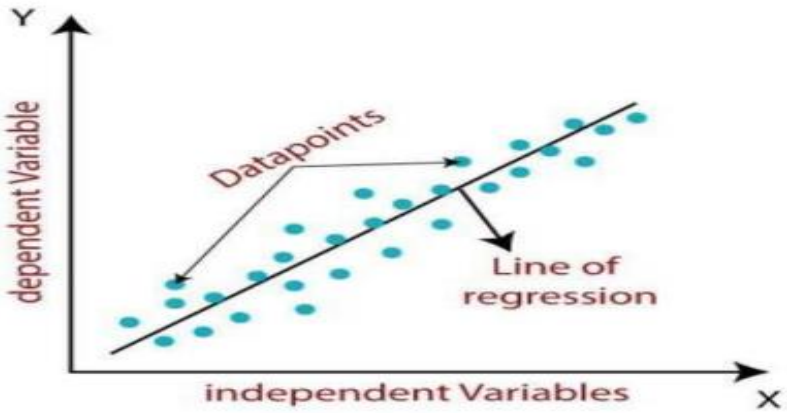

Figure 4: linear regression [20]

D. Unsupervised Learning

In unsupervised learning, a machine with only input data and the machine recognizes the pattern and learns itself without the labeled data. An unsupervised machine algorithm does not need supervision like a supervised algorithm. This algorithm aggregates the data into clusters [20]. Unsupervised learning algorithms learn few features from data. When entering new information, it uses previously learned functions to identify the class of the data. [15] [16]. Unsupervised Learning has various categories.

Cluster Analysis: Cluster analysis, also known as clustering, is an unsupervised machine learning technique that can be used to identify and group related data points in large data sets without worrying about a specific result. It groups a set of objects so that objects belonging to the same category, called a cluster, are somehow more similar than objects in other groups [15].

K-Means Clustering: This is an unverified algorithm. Here, clusters are formed without labeled data. It is almost the same as the K-nearest neighbor algorithm, but the main difference is that it used unlabeled data. K-means clustering makes neural networks easier to learn and recognize patterns. It calculates centroids and finds the best and optimal one among them and is also known as smooth clustering algorithm. [18].

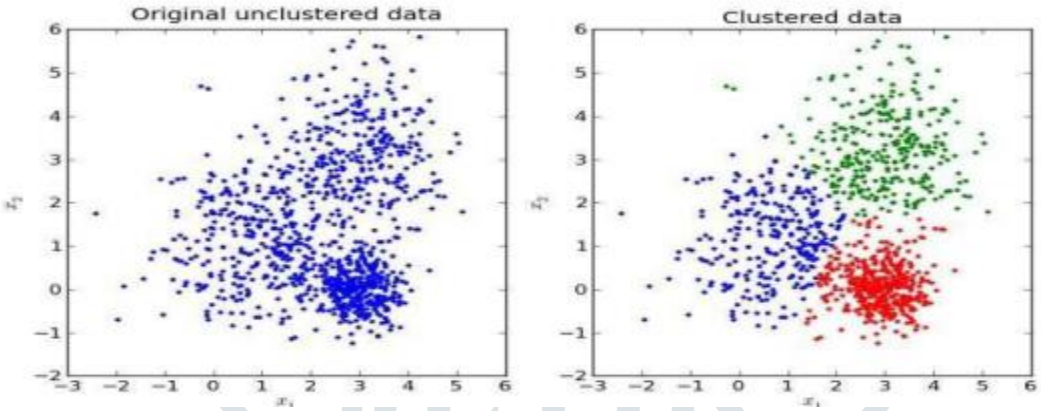

Figure 5: k-means clustering [20]

Evaluation of the Classification Model

Predictive accuracy, true positive, false positive, precision, recall, and F-measure are commonly used to measure classifier performance. The confusion matrix helps you see the distribution of classifier performance by showing how often instances of a class are misclassified into class X or misclassified into some other class, such as class Y [21].

Table 2: confusion matrix

		Predicated Class		Total instance
		+	-	
Actual Class	+	True Positive	False Negative	Positive
	-	False Negative	True Positive	Negative

Prediction Accuracy: Predictive precision measures the proportion of samples correctly classified by the classifier.

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

True Positive Rate (TP): is the proportion of samples that are classified as positive or correct compared to samples that are classified as positive or correct [19].

$$\text{True positive rate}=\frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

False Positive Rate (FP): measures the proportion of negative samples that are classified as false positives [19].

$$\text{False positive rate}=\frac{TN}{TN+FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Precision: measures the proportion of samples classified as positive and true positive [19].

$$\text{Precision}=\frac{TP}{TP+FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Recall: How many positive/negative pairs have the classifier marked as positive or negative in both correct and incorrect categories. [19].

$$\text{Recall}=\frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

F-Measure: is computed as the harmonic mean of recall and precision. [19].

$$\text{F-Measure}=\frac{2*Precision*Recall}{Precision+Recall} \quad \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Related Work

In 2018 Sujata Drall, Gurdeep Singh Drall, Sugandh Singh, Bharat Bhushan Naib studied Chronic Kidney Disease Prediction Using Machine Learning: A New Approach using python tool [9]. Their work focuses on finding best classification algorithm on basis of accuracy and execution time for prediction of Kidney Disease. They have used Naïve Bayes and KNN algorithms. The experimental result shows Naïve Bayes provide 96.25% and KNN provide 100% result.

In study which is called Detection of Chronic Kidney Disease Using Machine Learning Algorithms with Least Number of Predictors which is done by Marwa Almasoud and Tomas E Ward in 2013 [2]. They use four the machine learning algorithm, Logistic regression, SVM, Random forest and Gradient boosting ang got highest accuracy with Gradient boosting with result of 99.0%.

Dr. S. Vijayarani and Mr.S.Dhayanand [12] done their study on Data Mining Classification Algorithm for Kidney Diseases Prediction This study mainly focused on finding the optimal classification algorithm based on the performance factors of classification accuracy and execution time. This study used different data mining classification algorithms and obtained results using SVM 76.32 and Naive Bayes 70.96.This study uses experimental his MATLAB tools..

Md Nayeem Hosena , Md Ariful Islam Mozumdera , Rashedul Islam Sumona and Hee-Cheol Kim done the study with the name of Prediction of Chronic Kidney Disease Using Machine Learning [10]. They use UCI dataset repository data with an algorithm of SVM, Random Forest and ANN and got the result SVM with 95.63%, Random Forest with 95.71% and ANN with 98.61%.

Mitisha Barot [22], studied with a title Prior Stage Kidney Disease Prediction Using AI & Supervised Machine Learning Techniques with MATLAB tool. The researcher uses UCI dataset repository data with four algorithm, KNN, Naïve Bayes, Decision Tree and Supreme boosting classifier and got higher result with Supreme boosting classifier with 99%.

Table 2: summary of related works

No	Author Name	Title	Algorithm used	Performance
1.	Sujata Drall, Gurdeep Singh Drall, Sugandh Singh, Bharat Bhushan Naib (2018)	Chronic Kidney Disease Prediction Using Machine Learning: A New Approach	➢ Naïve Bayes ➢ KNN	100% (KNN)
2.	Marwa Almasoud and Tomas E Ward (2013)	Detection of Chronic Kidney Disease Using Machine Learning Algorithms with Least Number of Predictors	➢ Logistic Regression ➢ SVM ➢ Random Forest ➢ Gradient boosting	99.0% (Gradient boosting)
3.	Dr. S. Vijayarani and Mr.S.Dhayanand (2015)	Data Mining Classification Algorithm for Kidney Diseases Prediction	➢ SVM ➢ Naïve Bayes	76.32 (SVM)
4.	Md Nayeem Hosena , Md Ariful Islam Mozumdera , Rashedul Islam Sumona and Hee-Cheol Kim (2023)	Prediction of Chronic Kidney Disease Using Machine Learning	➢ SVM ➢ Random forest ➢ ANN	98.61(ANN)
5.	Mitisha Barot (2022)	Prior Stage Kidney Disease Prediction Using AI & Supervised Machine Learning Techniques	➢ KNN ➢ Naïve Bayes ➢ Decision Tree ➢ Supreme boosting classifier	99.0% (supreme boosting classifier)

Research Methodology and Proposed System Architecture

This chapter describes the methodology utilized in the study as well as the architecture of the proposed system and its components. The architecture of the proposed system is that the framework at which the general approach of the study getting to be summarized.

A. System Design and Architecture of the proposed system: The most commonly used algorithm in machine learning is classification for medical applications. The first step is the training step, where the classification algorithm builds habitats with the training set of the classifier, and the next step is the testing step, where the model is used for classification and performance analysis [19].

In this research, we use three machine learning algorithms to predict chronic kidney disease Decision tree classifier, Support Vector Machine (SVM), and KNN Algorithms.

The dataset is available and obtained from Apollo Medical Research Center. First, we entered the CKD imbalance dataset, and then the dataset went through several processing steps, including pre-processing, duplicate data removal, scaling and balancing, which improves accuracy. We then used three different algorithms [2] [10]. Figure below shows the overall system design and framework of this research

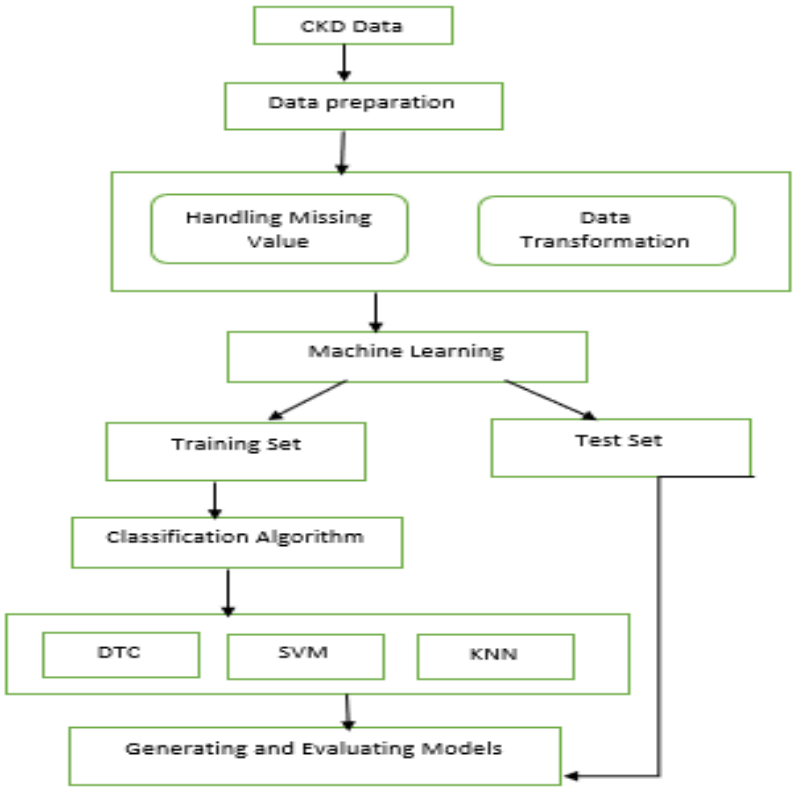

Figure 6: detail of the generic system framework

B. Dataset Description: The material supporting this study is based on CKD patients collected over a two-month period in 2015 from Apollo Hospitals, India. The data is available at the University of California, Irvine (UCI) repository called the Chronic Kidney Disease Database. This dataset of 400 observations suffers from missing and noisy values. The data includes 250 records of patients with CKD and 150 of people without CKD. Therefore, the percentage of each category is 62.5% with CKD and 37.5% without CKD. The ages of these observations vary from 2 to 90 years. Table I shows that the CKD dataset has 24 features, including 11 numerical features and 13 nominal features, and the 25th feature indicates the classification or status of CKD [2] [22].

Table 3: attributes description

Name	Description	Type: Unit/Value
Age	Age of patient	Numeric: years
Blood pressure (bp)	Patients' blood pressure	Numeric: mm/Hg
Specific gravity (sg)	Urine density ratio	Normal: 1.005, 1.010,1.015, 1.020,1.025
Albumin (albumin)	Albumin level in the blood	Nominal: 0,1,2,3,4,5
Sugar (Sugar)	Patients Sugar level	Nominal: 0,1,2,3,4,5
Red blood cells (RBC)	Patients' red blood cells count	Nominal: normal, abnormal
Pus cell (PC)	Patients pus cell number	Nominal: normal, abnormal
Pus cell clumps (PCC)	pus cell clumps present in blood	Nominal: present, not present
Bacteria (BA)	Presence of bacteria in the blood	Nominal: present, not present
Blood glucose (BGR)	blood glucose random count	Numeric: mgs/dl
Blood urea (BU)	blood urea level of the patient	Numeric: mgs/dl
Serum creatinine (SC)	Level of Serum creatinine in blood	Numeric: mgs/dl
Sodium (Sod)	Sodium level in blood	Numeric: mEq/L
Potassium (Pot)	Potassium level in blood	Numeric: mEq/L
Hemoglobin (Hemo)	hemoglobin level in the blood	Numeric: gms
Packed cell volume (PCV)	packed cell volume in the blood	Numeric
White blood cell count (WC)	white blood cell counts of the patient	Numeric: cells/cumm
Red blood cell count (RC)	red blood cell counts of the patient	Numeric millions/cmm
Hypertension (HTN)	Does the patient have hypertension on not	Nominal: yes, no
Diabetes mellitus (DM)	Does the patient have diabetes or not	Nominal: yes, no
Coronary artery disease (CAD)	Does the patient have coronary artery disease or not	Nominal: yes, no
Appetite (Appet)	Patient's appetite	Nominal: good, poor
Pedal Edema (PE)	Does patient have pedal edema or not	Nominal: yes, no
Anemia (ANE)	Does patient have anemia or no	Nominal: yes, no
Class	Does the patient have kidney disease or not	Nominal: CKD, not CKD

C. Data Preprocessing: Data preprocessing is a way to transform noisy and huge data into meaningful and clean data, because the available data is real data, so it contains inaccurate data, missing values and other noisy data, so that these conflicting data can become. data set, the proposed system must clean the raw data. This is an important part of completing the forecasting model [9].

***Outliers*:** Extreme values that are distant from the feature central tendency are called outliers. Data entry errors cause invalid outliers, which are also known as noise in the data. The extreme data points in this study that fall outside of the medically acceptable range have been regarded as missing data and subsequently altered, as the missing data section will detail. In the CKD dataset, box plots have been utilized to identify outliers; three extreme data points for potassium and sodium are considered undesirable. 7.6 mEq/L was the highest potassium level that was found. This indicates that a potassium level with 39 and 47, is not feasible and is typically the result of an error. Comparably, one extreme data point 4.5 was found for sodium, as shown in Fig. 8. In a normal range, the patient's sodium level should be between 135 and 145 mEq/L; if it is below 135, hyponatremia is the cause for this reason, a value of 4.5 is unacceptable or impossible. [2].

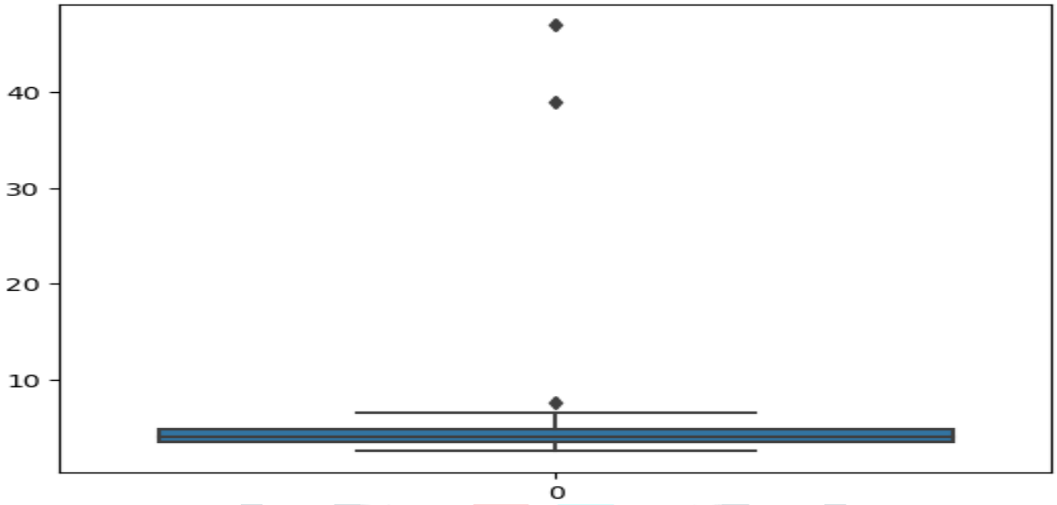

Figure 7: boxplot for potassium with outliers

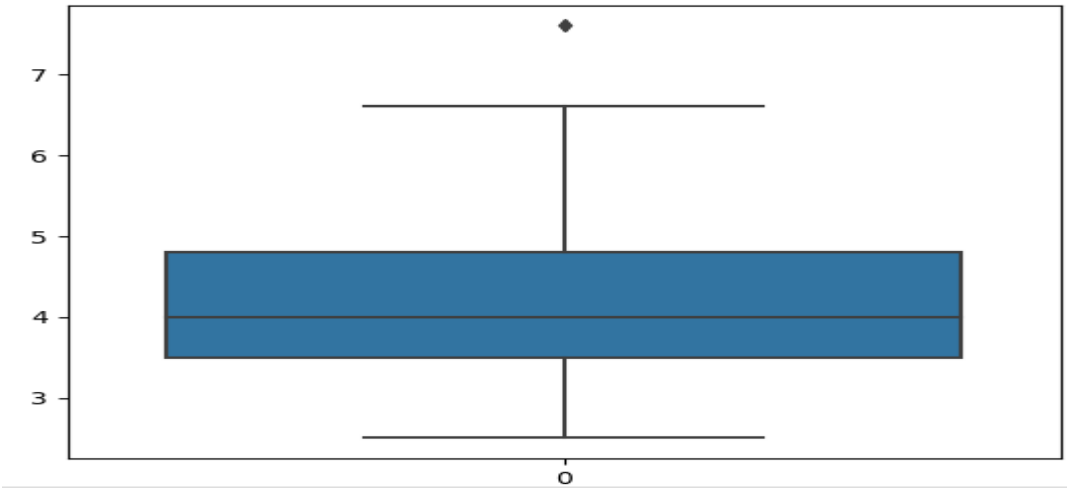

Figure:8: boxplot for potassium after removing outliers

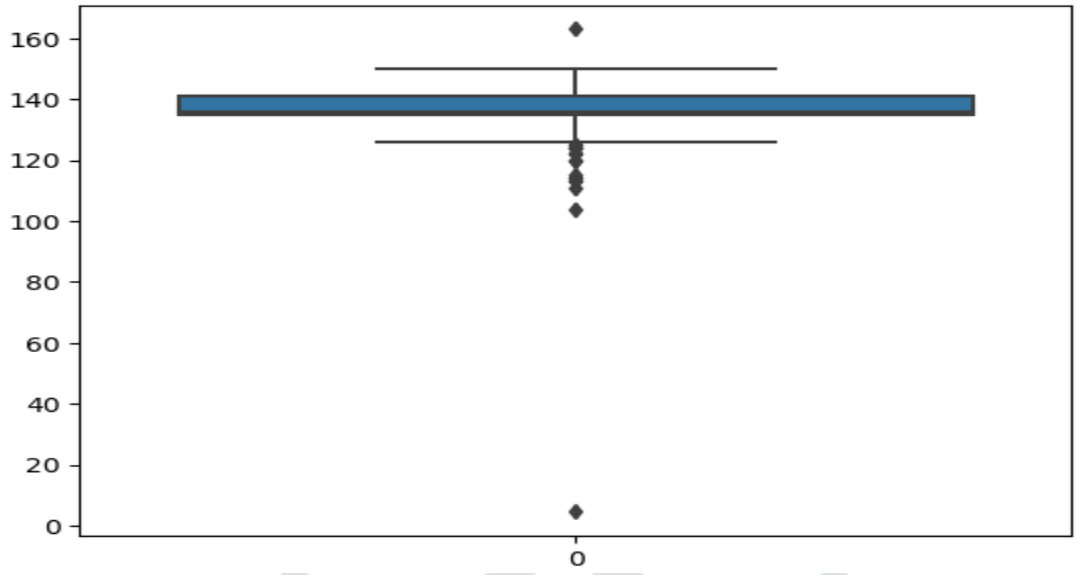

Figure 9: boxplot for sodium with outliers

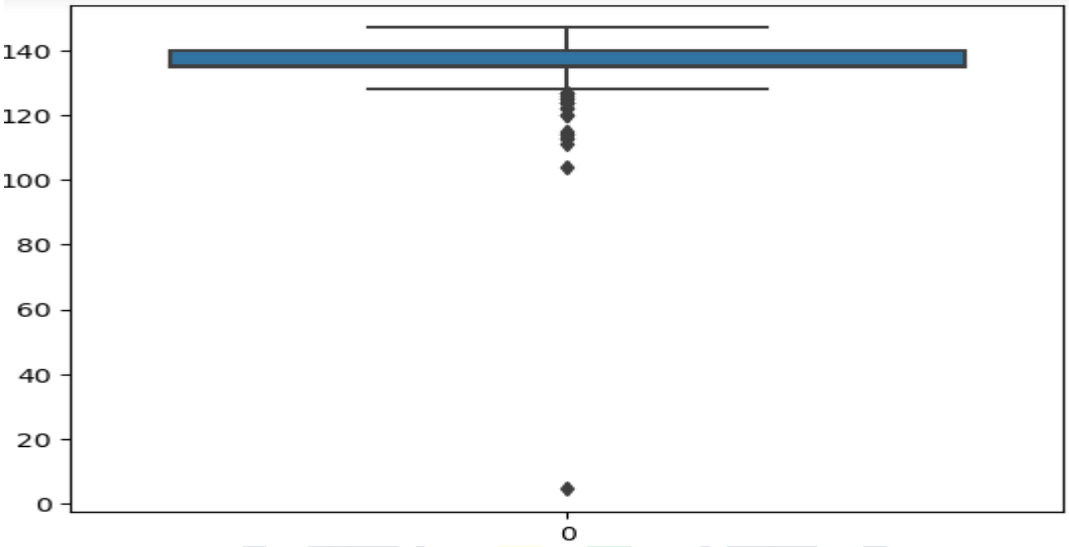

Figure 10: boxplot for sodium after removing outliers

***Handling Missing Values*:** CKD dataset has instances with total records of 10,000 and 25 attributes including class label, among these records 1013 (10.13%) are missing values. We use Jupyter Notebook as a tool to handle missing values by using mean and mode average techniques to fill the missing values.

Data Transformation: Min-max normalization has been used on various kinds of numerical data in this investigation. For variables that are categorical, another data transformation has been performed. This is a result of categorical variables being incompatible with some ML methods. Consequently, SVM classifiers dummied categorical variables having n values by turning each one of them into n-1 dummy variables.

EXPERMENTAL RESULT AND ANALYSIS

In this place, the researcher describes the techniques adopted for developing the model to predict the status of CKD using Machine Learning Approach. Machine Learning classification methods were chosen to develop the predictive model. Three experiments were done using three algorithms: Decision Tree Classifier, Support Vector Machine (SVM) and K-nearest neighbor (KNN). The result described below

Table 4: predictive performance of ML model result

Classifier	Precision	Recall	F1-Score	Specificity	Accuracy	Correctly classify instance	Incorrectly classify instance
Decision Tree	97.5%	98%	97.5%	98%	97.85%	64.3%	35.7%
SVM	91%	90.5%	91%	87.03%	91.43%	61.43%	38.57%
KNN	96.5%	96%	96%	93.8%	96.42%	65%	35%

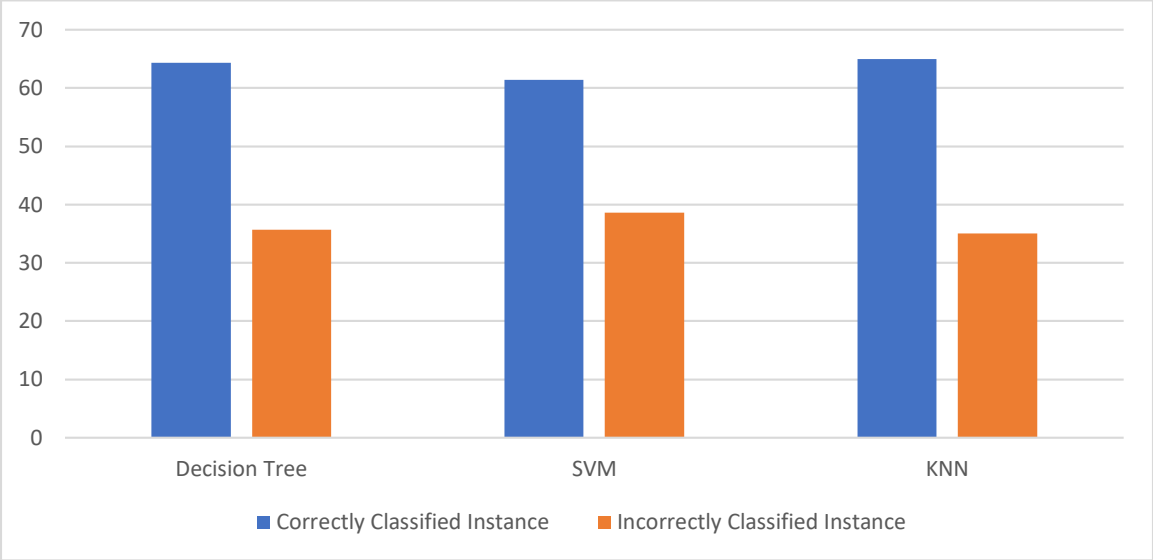

Figure 11: correctly and incorrectly classified instances

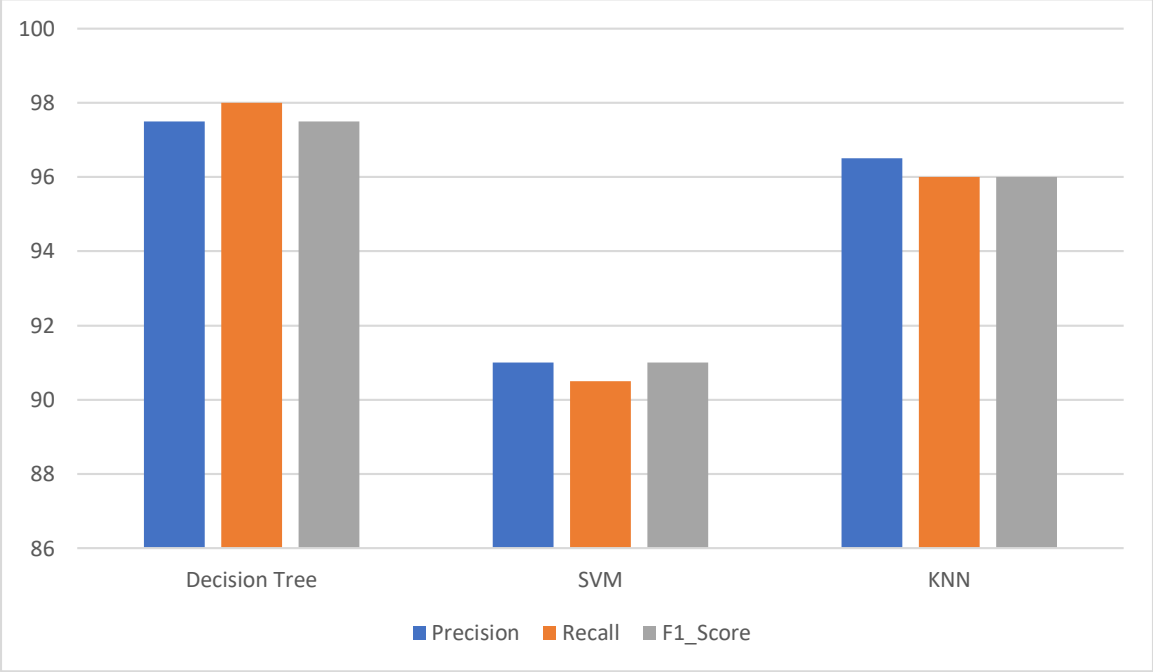

Figure 12: performance measure for classification algorithms

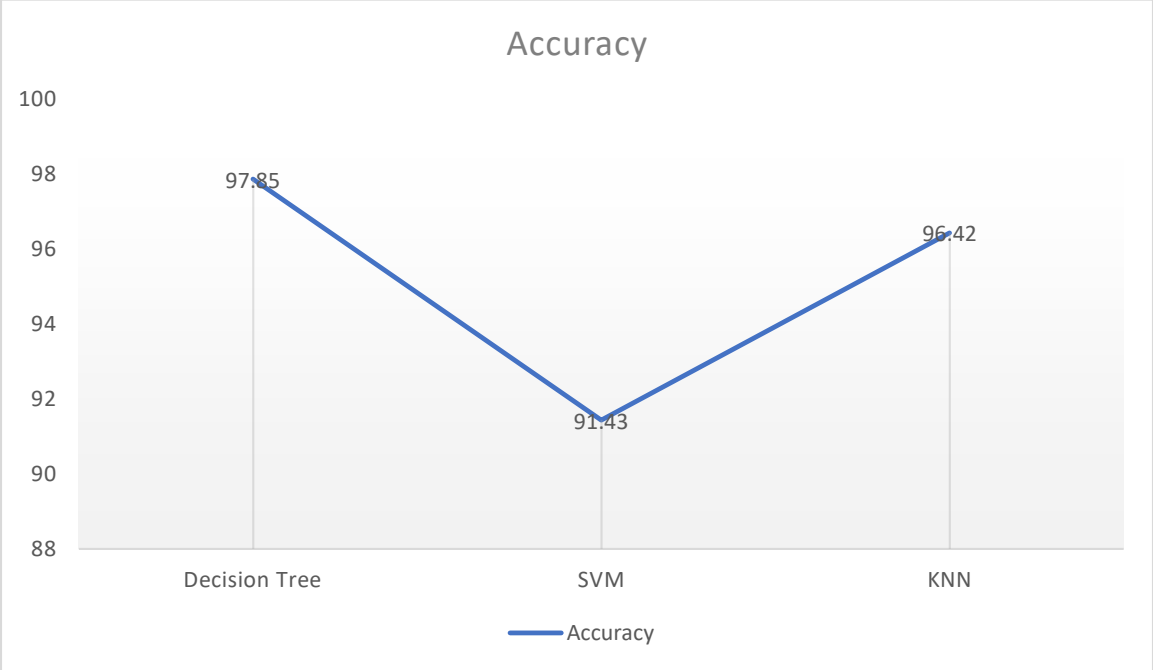

Figure 13: accuracy measure for classification algorithms

Conclusion

This paper tries to classify chronic kidney diseases datasets using different classification algorithm and based on that Decision tree classifier provide highest accuracy compared to the rest of two algorithm SVM and KNN.

References

[1] Md. Ariful Islam, "Chronic kidney disease prediction based on machine learning algorithms," 23 01 2023.

[2] M. Almasoud, "Detection of Chronic Kidney Disease Using Machine," *core.ac.uk*, vol. XXX, p. 9, 2013.

[3] W. Ahmed Ali, "Prevalence of chronic kidney disease and associated factors among patients with underlying chronic disease at Dessie Referral Hospital, East Amhara Region, Ethiopia," *Front. Epidemiol*, vol. 3, 2023.

[4] M. Robert Thomas, "Chronic Kidney Disease," *Prim Care Clin Office Pract*, p. 16, 2008.

[5] K. M. Kim, "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6727899/," 30 9 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6727899/. [Accessed 25 11 2023].

[6] A. L. Ammirati, "Chronic Kidney Disease," *REV ASSOC MED BRAS 2020;*, p. 7.

[7] "https://www.davita.com/education/kidney-disease/stages," [Online]. Available: https://www.davita.com/education/kidney-disease/stages.

[8] H. Khalid, "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease," *computational intellegence and neuroscience*, 2023.

[9] S. drall, "Chronic Kidney Disease Prediction Using Machine Learning:," *International Journal of Management, Technology And Engineering*, vol. 8, no. v, p. 10, 2018.

[10] Md Nayeem Hosena, "Prediction of Chronic Kidney Disease Using Machine Learning," *The 20th World Congress of the International Fuzzy Systems Association*, p. 5, 2023.

[11] K. Cindy George1, "The Chronic Kidney Disease in Africa (CKD-Africa) collaboration: lessons from a new pan-African network," *BMJ journals*, vol. 6, no. 8, 2022.

[12] M. Dr. S. Vijayarani, "DATA MINING CLASSIFICATION ALGORITHMS FOR," *International Journal on Cybernetics & Informatics*, vol. 4, no. 4, p. 13, 2015.

[13] S. B. Jaber Qezelbash-Chamak, "A survey of machine learning in kidney disease diagnosis," *Science Direct*, p. 17, 2022.

[14] T. ANDREW S. LEVEY, "Definition and classification of chronic kidney disease: A," *International Society of Nephrology*, vol. 67, p. 12, 2005.

[15] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research," *SN Computer Science*, vol. 2, p. 21, 2021.

[16] B. Mahesh, "Machine Learning Algorithms - A Review," *International Journal of Science and Research*, vol. 9, no. 1, p. 7, 2020.

[17] t. Ayodele, "Types of Machine Learning Algorithms," *ResearchGate*, p. 31, 2014.

[18] K. Jafar Alzubi, "Machine Learning from Theory to Algorithms: An Overview," in *Journal of Physics: Conference Series*, 2018.

[19] T. B. GASHAWTENA, "Integrating Data Mining Results with Knowledge Based System for Diagnosis and Treatment of Cattle Diseases," *Debrebirhan University*, p. 118, 2018.

[20] S. Neelam Dahiya, "A Review paper on Machine Learning Applications, Advantages and Techniques," *The Electrochemical Society*, vol. 107, p. 15, 2022.

[21] B.Max, "princiles of Data Mining," *Portsmouth Springer*, 2007.

[22] M. Barot, "Prior Stage Kidney Disease Prediction Using AI &," *Journal of Emerging Technology and Innovative Research*, vol. 9, no. 4, p. 4, 2022.