International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

XGBoost Learning for Detection and Forecasting of Chronic Kidney Disease (CKD)

[1]Yogesh Kale, [2]Shubhangi Rathkanthiwar, [3]P. Fulzele, [4]N. J. Bankar

Submitted: 23/12/2023 **Revised**: 29/01/2024 **Accepted**: 07/02/2024

Abstract: It's astounding that 63,538 cases have been documented. based on data from India's chronic kidney disease (CKD). Nephropathy in humans usually appears between the ages of 48 and 70. Compared to women, men are more likely to develop CKD. Regretfully, India has slipped into the top 17 countries for chronic kidney disease (CKD) since 2015. CKD is characterized by a gradual deterioration in the function of the excretory organs. Effective treatment and early illness identification may help prevent this terrible condition. Among other practical applications, machine learning is being used in fraud detection and medical research findings analysis. Chronic illness forecasting is done using a variety of machine-learning techniques. With a focus on decision trees, Adaboost, XGboost, random forests, logistic regression, support vector machines, naïve Bayes, KNN, and artificial neural networks, our primary goal is to evaluate the accuracy of various machine learning techniques. Here XGBoost ML algorithm performs well for prediction of chronic kidney disease (CKD), it provide the 99% accuracy and which is almost greater than the other ML algorithms tested. RCode has received praise from the study's performance analysis. This project's primary goal is to develop an application for sickness prediction that uses an analysis of the chronic kidney disease dataset to detect cases of chronic kidney disease (CKD) and non-CKD.

Keywords: *RCode, XGBoost technique, Classification, Accuracy, Chronic Kidney Disease (CKD).*

1 Introduction

In the current era the evolution of people's lifestyles and stress has been ingrained in people's lives, contributing to population aging and the development of numerous Chronic Kidney Diseases (CKD) affects one million people annually causing gradual loss in kidney functions [1]. Early detection of CKD is important as symptoms may not be unique to the disorder and can be difficult to detect. To identify the class of outcome for each data point, healthcare companies frequently employ machine learning classification techniques. Among these techniques are Neural Networks, K-Nearest Neighbor, Decision Trees, Support Vector Machines, and Naïve Bayes classifiers. By displaying the correlation between several risk variables, K-Nearest Neighbor may be utilized to predict CKD early on. A lot of varied data must be studied for machine learning, and the data must be analyzed and predicted using a variety of strategies, algorithms, and approaches. Many important disorders in medical research have been successfully identified and detected by machine learning. By training a predictive model using data from previous CKD patients, machine learning can predict whether a patient has CKD. The kidney plays a crucial role in removing waste and harmful substances from the body and returning essential nutrients and substances to the bloodstream. However, Chronic Kidney Disease (CKD) can develop, which is a condition that affects the structure or function of the kidneys. CKD is a common problem that can affect anyone, and it is estimated that approximately 3 million people in the UK are at risk of it. Different conditions can place a strain on the kidneys and lead to CKD. This research paper discusses the classification of CKD and is organized as follows: Section II provides a contribution of work, Section III provides related work, and in section IV provides proposed work, and in last section V result discussion and conclusion including attribute improvement.

1.1 Contribution of Work

This study report aims to predict chronic kidney diseases (CKD) at an early stage. The health care systems produce enormous amounts of data. Therefore, it is important to make good use of this data in order to assess, forecast, and manage a specific illness. Some answers are provided by a categorization model based on predetermined values. When predicting the values of their results, categorization types often anticipate either a large or little amount of input. The classification method in supervised machine learning techniques makes use of the training dataset.

[*1]*Assistant Professor Yeshwantrao Chavan College of Engineering, Wanadongri, Nagpur, India.*
[1]*ycceetyogesh@gmail.com*
[2]*Professor Yeshwantrao Chavan College of Engineering, Wanadongri, Nagpur, India.*
svr_1967@yahoo.com
[3]*Lecturer, Professor, Dept. of Pedodontics, Sharad Pawar Dental College, Deputy Director, Research, Datta Meghe Institute of Higher Education and Research, Nagpur, India.*
punitr007@gmail.com
[4]*Professor Microbiology Department Jawaharlal Nehru Medical College Sawangi Meghe Wardha, 442005, India*
drbankarnj28@gmail.com

Predicting the categorical class labels of the data is done through classification. The goal of the project is to develop a machine-learning framework for the chronic renal illness dataset's information discovery. To classify the disease at puberty, three machine learning techniques are used: K-Nearest Neighbors, Support Vector Machine, and Logistic Regression. We investigate the molarity of each algorithm. Our proposed model combines Support Vector Machine, Logistic Regression, and K-Nearest Neighbors (KNN). Complete model is comprising of these following steps:

i. Data import
ii. Data exploration includes data cleaning, descriptive statistics, data visualization, feature engineering and hypothesis testing.
iii. Data transformation includes scaling, log transformation, one-hot encoding, removing missing values, imputation.
iv. Feature engineering and preparing data for model training includes feature selection, feature scaling and normalization, creating new features.
v. Model training and evaluation.
vi. Model dumping includes deploying the model in a prediction environment, sharing the model with others, and saving checkpoints during training.
vii. Application development on stream lit.

1.2 Literature Review

The Internet of things has been used expansively in medical science to keep an eye on patient health. The works listed below are all in some way connected to this topic. IoT-based health monitoring system device was developed by Tamilselvan and colleagues on the Internet of Things web and can screen vital signs corresponding to oxygen saturation percentage, heart rate, eye movement, and body temperature in the patient. Taking apparatus comprised an eye blink sensor, oxygen saturation mainly calls it as Sp-O2 sensor, and a thermometer. Processing was done on an Arduino-UNO board in an existing IoT-based health monitoring system, but no exact performance trials exist for any of the patients [10].

Table 1 Differentiable Analysis of Previous Techniques

Ref. No.	**Dataset/ Algorithm**	**Findings**	**Limitation**
[1]	KNN, Naïve Bayes	Researchers have utilized machine learning algorithms to solve health sector problems, using techniques like random subspace classification and K-nearest neighbor (KNN) to classify and predict chronic renal failure (CRF) status in patients. This decision support system aids in diagnosing and predicting CRF presence.	-Lack of comprehensive comparison. -Limited Algorithm diversity
[2]	Ensemble model, Weka tool	Improving results is the study's main goal. Machine learning methods are used to predict chronic kidney disease utilizing both ensemble and individual learners with the WEKA tool.	-Lack of explanation for ensemble learner selection -Limited evaluation metrics -Dataset limitations -Limited discussion of feature selection.
[3]	KNN, LR, ANN	Three continuous, categorical, and simple prediction algorithms were studied by the researchers; the best fitted models included 10 predictors, while the simplified model had eight.	- The limited scope of algorithms -Less number of parameters used for analysis
[4]	Radial Basis Function Network, Multi-layer perceptron	The research uses five feature selection strategies to examine the accuracy, sensitivity, and specificity of two classification algorithms: radial-basis	-Less number of evaluation parameters - Minimal

		function networks (RBFN) and multilayer preceptor network (MLPN).	explanation Of the feature selection
[5]	Random forest, Decision tree	Researchers utilized the random forest algorithm to predict chronic kidney disease in candidates, using original samples for drawing and tree bootstrap, and creating an unpruned classification tree.	-Limited scope of algorithms -Fewer evaluation metrics -Limited discussion of interpretability
[6]	Perceptron Network, Sampling algorithm	P. Yildrim utilized a multilayer perceptron network and a range of sample approaches to assess and compare the accuracy, recall, and f-measure of the sampling algorithm for chronic renal illness prediction.	-Absence of external validation -Limited explanation of sampling theorem.
[7]	SVM stands for stochastic neural network; radial basis function	Probabilistic Neural Networks (PNN), Multilayer Perceptrons (MLP), Radial Basis Function (RBF), and Support Vector Machine (SVM) were among the methods whose performance was compared in this study.	-The study's main limitation is its small dataset, which may not accurately represent disease categories' diversity, and the proposed method's effectiveness in certain exceptional conditions.

Internet of Things healthcare monitoring kit was developed by Acharya. The intended system supervised several essential health metrics, including respiration, body temperature, ECG, and heartbeat. In this research, the researcher used an ECG sensor on the Raspberry Pi as well as blood and heart rate monitoring. Whatever the clinical data measured, first it will be processed in a Raspberry processor and then transferred to the Internet of Things network. The researcher concludes one thing that insufficient interfaces for data visualization are the system's most serious problem [11]. The researcher Gregoski first proposed the idea of using a smartphone to show one's heart rate (HR). A mobile brightness and camera were used to trail portion blood drift, and the cardiac signal was considered based on that. The outcome of this latest proposed technology, users can square up their heart rate (HR) just by observing their phones instead of using their hands, which was earlier required. The researcher concludes, if you need to continuously monitor your heartbeat throughout the day without disturbing the patient's mobility, this proposed technology doesn't work [12]. If the user doesn't mind the cost and time, then the same technique is implemented by using cell phones. This technique is used to detect cardiovascular illness suggested by Oresko [13], and it also concludes that only coronary rhythm was observed in real-time by the proposed prototype, not heart rate over time and hence it's difficult to detect cardiovascular disease. Now the latest and more popular healthcare technology which comprises (a smartphone and Arduino-based system) was proposed by Trivedi [14]. Using this technology medical professionals can monitor and analyze clinical data remotely and prescribe things remotely. Similarly using IoT, Kumar [15] established a safety nursing device that may be personalized. The background is planned into three segments: control, device, and convince. The most popular temperature sensor IC-DS18B20 and a pulse sensor were used to monitor the patient's temperature and pulse rate. Measured clinical data was transferred from Arduino to the cloud, this data was retrieved by the medical professionals by using a laptop, desktop, or mobile phone and prescribed things accordingly. Researcher Desai [16][17] used a wireless sensor network (WSN), to provide monitoring of the heartbeats of patients to a medical professional. This application makes use of Spartan3 to process data in parallel on an FPGA. Mainly three limitations were found in that related work, the first limitation is the availability of data not on a cloud platform [10], the second limitation is no timer facility [24] and the third limitation is more power consumption in wireless sensor network [16][17].

2 Proposed Methodology

Through the use of machine learning algorithms, as seen in figure 1, this research focuses more on chronic kidney illnesses, and the description of this in further sections.

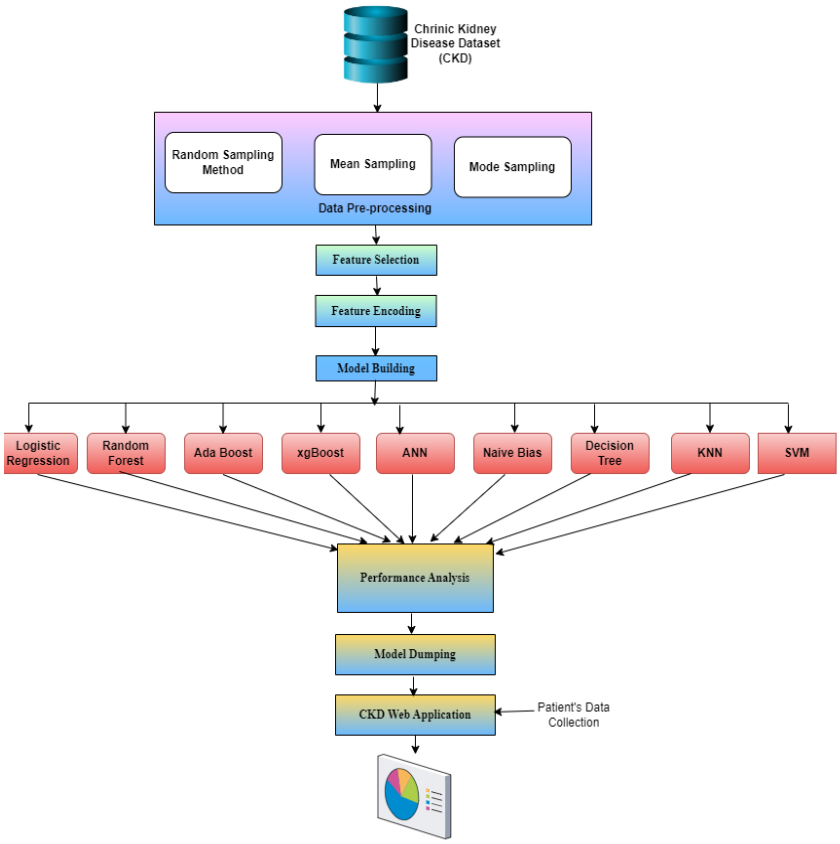

Fig 1 Chronic Kidney Disease (CKD) prediction Workflow

2.1 Dataset

The suggested approach makes use of the Chronic Kidney Disease (CKD) dataset from the UCI Machine Learning Repository, which comprises a total of 25 characteristics—11 of which are numerical and 14 of which are nominal. Using all 400 dataset instances, machine learning algorithms are trained. A total of 400 patients were diagnosed with either Non-Chronic Kidney Disease (NCKD) or Chronic Kidney Disease (CKD) in 250 instances. The attributes included in the data set are: pus cell, anemia, pedal edema, age, hemoglobin, diabetes mellitus, classification, appetite, coronary artery disease, blood pressure, specific gravity, packed cell volume, pus cell chumps, specific gravity, white blood cell count, hypertension, red blood cell count, potassium, and blood glucose random. Two groups are created from the dataset: one is used to test the samples, and the other is used to train them. The proportions of training to testing data are 70% and 30%, respectively [18]. Table 2 below lists the data characteristics that were used.

Table 2 data is imported from Kaggle into the notebook, and the dataset used is the UCI Irvine dataset. Then data exploration is an essential step in any data analysis or machine learning project. It involves analyzing and summarizing the main characteristics of a dataset to gain insights and identify patterns, trends, and relationships between variables. It includes the following steps, Data cleaning, descriptive statistics, data visualization, feature engineering, hypothesis testing, data transformation and Handling missing values, Scaling, log transformation, one-hot encoding, removing missing values, imputation, feature engineering, and preparing data for model training some common techniques used in feature engineering include Feature selection, feature scaling, and normalization, creating new features, train-test split, encoding categorical variables, handling missing values. The accuracy and performance of the model may be greatly impacted by feature engineering and data preparation for model training, which are crucial phases in machine learning. These steps require careful consideration and analysis of the data and should be done with the ultimate goal of improving the predictive power of the model.

2.2 Data Pre-processing

Pre-processing is a data mining technique that converts unprocessed data into a comprehensible format. It removes hash tags, null values, prepositions, links, and abbreviations. Once data is acquired, it must be pre-processed using various techniques to remove unnecessary and variable data. The preprocessing stage includes removing null values, punctuation, stop words, white spaces, and changing uppercase characters to

lowercase. This process is essential for achieving data mining goals.

2.2.1 Random Sampling Methods

This method is used for filling higher null values. The probability of selecting a sample just once is given by, where P is the probability, n is the sample size, and N is the population. Every item in the population has an equal chance of being selected for the sample taken to the random sampling process. It is sometimes referred to as the chance sampling technique denoted by equation (1) to (4).

$$\boldsymbol{P = 1 - \left(\frac{N-1}{N}\right).\left(\frac{N-2}{N-1}\right) \ldots \left(\frac{N-n}{N(n-1)}\right)} \quad (1)$$

$$\boldsymbol{Cancelling = 1 - \left(\frac{N-n}{n}\right)} \quad (2)$$

$$\boldsymbol{P = \frac{n}{N}} \quad (3)$$

Table 2 Attribute table of CKD

Sr. no.	Attribute	Attributes Description	Sr. no.	Attribute	Attributes Description
1	Age	Age of Persons in years	13	Sodium	sod in mEq/L
2	BP	mm/hg	14	Potassium	Pot in mEq/L
3	Specific Gravity	Hg (1.005/1.010/1.015/1.020/1.025)	15	Haemoglobin	Hemo in grams
4	Albumin	0/1/2/3/5	16	Packed Cell Volume	Packed Cell Volume
5	Sugar	su – (0/1/2/3/4/5)	17	White blood Cell Count	Wc in cells/cumm
6	Red Blood Cells	RBC – (normal/ abnormal)	18	Red Blood Cell Count	Rc in cells/cumm
7	Pus cells	PCC – (present / not present)	19	Hypertension	htn – (yes/no)
8	Pus cells clumps	PC – (normal / abnormal)	20	Diabetes Mellitus	dm – (yes / no)
9	Bacteria	ba – (present / not present	21	Coronary Artery Disease	CAD – (yes / no)
10	Blood Glucose Random	BGR in mgs/dl	22	Appetite	appet – (good / poor)
11	Blood Urea	Bu in mgs/dl	23	Peda edema	pe – (yes / no)
12	Serum Createnine	Sc in mgs/dl	24	Aanemia	ane – (yes / no)

The probability of having a sample chosen more than once is provided by.

$$\boldsymbol{P = 1 - \left(1 - \left(\frac{1}{N}\right)\right).n} \quad (4)$$

2.2.2 Mean/Mode Sampling

Lower null values are filled using this procedure. The arithmetic mean of all the terms is the most widely used formula for the means of a statistical distribution containing a discrete random variable. To compute it, take the total number of phrases and divide it by the sum of the values of each phrase. One may get the mean, or expected value, of a continuous random variable in a statistical distribution by multiplying the variable by the probability that the distribution specifies. The Greek character mu (μ) in lowercase indicates the expected value. It's computed by taking the total number of values in the data collection and dividing it by the total value sum. The mean of the data is as follows when a collection of data with n values (x1, x2, x3,..., xn) is present in equation (5) and (6):

$$\overline{x} = \frac{x_1+x_2+x_3+\cdots\ldots+x_n}{n} \quad (5)$$

It can also be denoted as:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6)$$

The most often occurring word's value is the mode of a distribution including a discrete random variable noted in equation (7). A distribution containing a discrete random variable often includes a large number of unique modes, particularly when the distribution has few terms. When two or more words appear more frequently than any other phrase and equally frequently, this occurs. The mode for grouped data may be found using the following formula:

$$\boldsymbol{Mode = l + (\frac{f_1-f_0}{2f_1-f_0-f_2}) \times h} \quad (7)$$

Bimodal distributions are those that have two modes [16]. Trimodal distributions are those that have three modes. A distribution with a continuous random variable has a mode that is its greatest value. Several modes may exist, much like discrete distributions.

2.3 Feature Encoding

Since each categorized column has two categories, a label encoder may be used. Pre-processing of the input data is done using the following encoding methods:

- *One-hot encoding:* Each category is given a vector. The vector indicates the presence or absence of the related characteristic (1), (0) [18].
- *Target-mean encoding:* This method substitutes the target variable's mean value for categorical values.
- *Frequency encoding:* Considers the number of times a specific category value appears in connection with a feature.

2.4 Model Training and Evaluation

Important phases in machine learning are model assessment and training [19]-[23], which entail using training data to train the model and testing data to assess its performance. Model training involves using the training data to train the machine learning algorithm to make predictions. This process involves selecting the appropriate algorithm and hyperparameters, fitting the model to the training data, and validating the model to ensure that it is not overfitting or underfitting the data. Some common techniques used in model training include:

2.5 Method selection

This involves selecting the appropriate algorithm that is best suited for the type of problem and the characteristics of the data.

- **Hyperparameter tuning:** This involves selecting the optimal hyperparameters that can improve the accuracy and performance of the model. Bayesian optimization, random search, and grid search are a few methods that may be used for this.
- **Cross-validation:** This involves validating the model using cross-validation techniques to ensure that it is not overfitting or underfitting the data.

To assess the correctness and performance of the trained model [24]-[26], testing must be done using the testing data. ROC-AUC score, accuracy, precision, recall, and F1-score are just a few of the performance metrics that are used in this process to evaluate the model's performance. Common methods for evaluating models includes, A table known as the confusion matrix displays the model's predicted true positives, true negatives, false positives, and false negatives. The ROC-AUC curve shows how the true positive rate and false positive rate predictions provided by the model are traded off. Cross-validation: Cross-validation techniques are used to test the model's performance to ensure that it is neither over- nor under fitting the data. Overall, model training and evaluation are important steps in machine learning that require careful consideration and analysis of the data. These steps can significantly impact the accuracy and performance of the model and should be done with the ultimate goal of improving the predictive power of the model.

2.6 Model Dumping

Model dumping, also known as model persistence, is the process of saving a trained machine-learning model to a file so that it can be used later for prediction or further training. This is an important step in machine learning as it allows you to reuse the trained model without having to

retrain it every time you need to make a prediction or make further improvements. The process of model dumping involves using a serialization library, such as pickle or joblib, to save the trained model to a file. The saved file contains all the information needed to recreate the trained model, including the trained parameters, hyper parameters, and other metadata. Once the model has been dumped, it can be loaded back into memory using the deserialization function provided by the serialization library. As a result, you may continue train the model with more data or use it to generate predictions on fresh data. Some common use cases for model dumping include: Deploying the model in a production environment: By dumping the trained model to a file, you can easily deploy it in a production environment without having to retrain it every time you need to make a prediction. Sharing the model with others: By dumping the trained model to a file, you can share it with others who can use it to make predictions on their own data or to further improve the model. Saving checkpoints during training: By dumping the trained model at regular intervals during training, you can save the progress of the training process and In the event of an interruption or failure, pick up where you left off with your training. Overall, model dumping is an important step in machine learning that allows you to save and reuse trained models. This can save you time and effort by avoiding the need to retrain the model every time you need to make a prediction or make further improvements.

2.7 Application development on streamlit

An open-source framework called Streamlit is used to create interactive web apps for data science and machine learning initiatives. It makes it simple for developers to create and implement data-driven apps with little to no coding knowledge. For the purpose of generating interactive visualizations, data exploration tools, and machine learning models, Streamlit offers an easy-to-use interface. Because it enables developers to rapidly prototype and refine their concepts, it's a perfect tool for developing proof-of-concept apps or launching machine learning models. Streamlit makes it simple to add machine learning models into your application by integrating with well-known machine learning libraries like Tensor Flow, Scikit-learn, and PyTorch. To develop an application with Streamlit, you first define the UI elements such as sliders, dropdowns, and text inputs that allow users to interact with the application. Then, you define the logic that responds to user input, such as loading data, performing calculations, and displaying the results. Streamlit also provides a number of built-in components for data visualization, including line charts, scatter plots, and heat maps. These components can be customized and configured to display data in a variety of formats. Once the application is built, it can be deployed to a variety of hosting services, including Heroku, Google Cloud, and AWS. Streamlit provides a simple CLI command for deploying the application, making it easy to get started with deployment. Overall, Streamlit is a powerful and easy-to-use tool for building interactive web applications for machine learning and data science. It provides a simple and intuitive interface for building data-driven applications and integrates seamlessly with popular machine learning libraries. Table 3 shows the optimal features for Chronic Kidney Disease.

Table 3 Optimal features for Chronic Kidney Disease

Sr. No.	Features	Scores
1	White_Blood_Cell_Count	9701.050391
2	Blood_Urea	2343.097145
3	Blood_Glucose_Random	2241.651289
4	Serum_creatinine	357.792101
5	Packed_cell_Volume	308.18145
6	Albumin	216
7	Haemoglobin	123.856342
8	Age	115.85994
9	Sugar	94.8
10	Hypertension	88.2

The ideal characteristics chosen according to an individual's age are displayed in table 3 above. But as people age, CKD becomes increasingly prevalent. Kidney filtration starts to decline by around 1% a year after the age of 40. People in their later years are more prone to have diseases like diabetes, high blood pressure, and heart disease, which can damage the kidneys, in addition to the normal aging process.

3 Description of Performance Analysis of Machin Learning Algorithm

The Parameter for the performance analysis for the algorithm is accuracy based on accuracy the random forest yielded the best result.

3.1 Logistic Regression

The supervised learning algorithm used mainly as an approach for single-class classification, it works on categorical values rather than numerical values [27]. In this research work the use of the logistic regression with two categories 0 denote having chronic kidney disease and 1 denote having no kidney disease.

3.2 Support Vector Machine (SVM)

This well-liked supervised machine learning approach is applied to regression analysis and classification [28]. SVM locates the hyperplane in a high-dimensional space that best divides the various classes in classification tasks. Within the framework of this study project: The dataset was divided into two groups using Support Vector Machines (SVM): diseases and non-diseases. Regression analysis and categorization are performed using a machine learning technique known as K-nearest neighbors (KNN). The KNN algorithm locates the k examples that are closest to a new example that hasn't been seen before and assigns the new example the class label or value that appears the most frequently among the k examples. KNN is a simple yet effective algorithm for classification and regression tasks, However, the choice of neighbor count (k) and the distance metric employed to compute the sample distance determine the accuracy and performance of the system. In the context of this research paper, the model shown in Figure 1 takes the input features for the new example and calculates the distance between the new example and all the stored labeled examples. Then, it chooses the k labeled instances that are closest to the new example and gives it the class label that appears the most frequently among the k examples. The decision tree technique may be applied to two different kinds of problems: regression analysis and categorization. Because the data provided for this research work are categorical, it was used for categorization. Random Forest is the name of the supervised machine learning example that was utilized for the regression and classification tasks. It is an ensemble learning approach that creates better results by combining many decision trees.

3.3 Extreme Gradient Boost

The XGboost is the ensemble model that combines various weak models for achieving better performance in terms of results [29]. The next phase in designing a neural network is to define its architecture, which includes how many layers there are, how many nodes are in each layer, what kind of output layer it has, and how the activation functions operate. Train the model using all available machine learning methods. For example, an optimization algorithm like Adam or stochastic gradient descent is used to train a neural network on a dataset. The network's weights are adjusted during training to reduce the discrepancy between the expected and actual labels. Model assessment is the following step, when the performance of the neural network model is evaluated on an alternative validation dataset using measures such as accuracy, precision, recall, F1-score, and confusion matrix. The following stage involves fine-tuning the neural network's hyperparameters, including the learning rate, number of epochs, regularization strength, and number of hidden nodes, in order to enhance the model's performance. The study article concludes with a prediction section. Following training and validation, predictions on previously unidentified data may be made using the neural network.

4 Result Analysis

The correlation matrix of the CKD 14-attributes is displayed in Figure 2. Based on this matrix, it can be concluded that the RBC count has a positive correlation with specific gravity, hemoglobin, and packed cell volume, and a negative correlation with albumin, blood urea, and packed cell volume. Additionally, there is a strong positive correlation between packed cell volume and hemoglobin, and a negative correlation between packed cell volume and blood urea, as well as a negative correlation between hemoglobin and albumin.

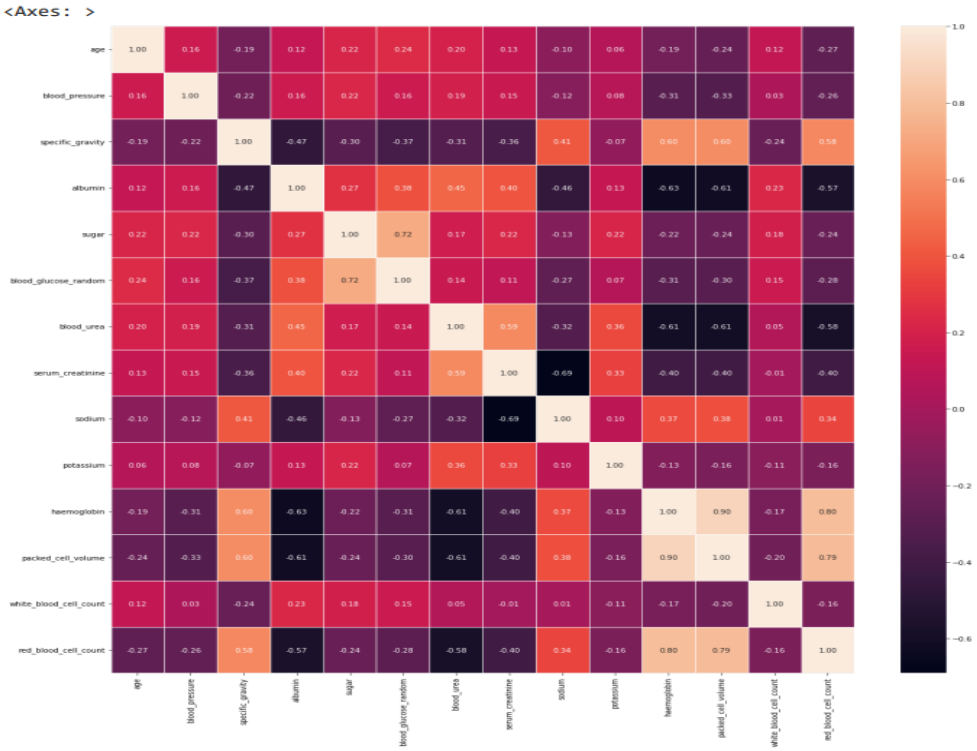

Fig 2 Correlation matrix of CKD 14-attributes

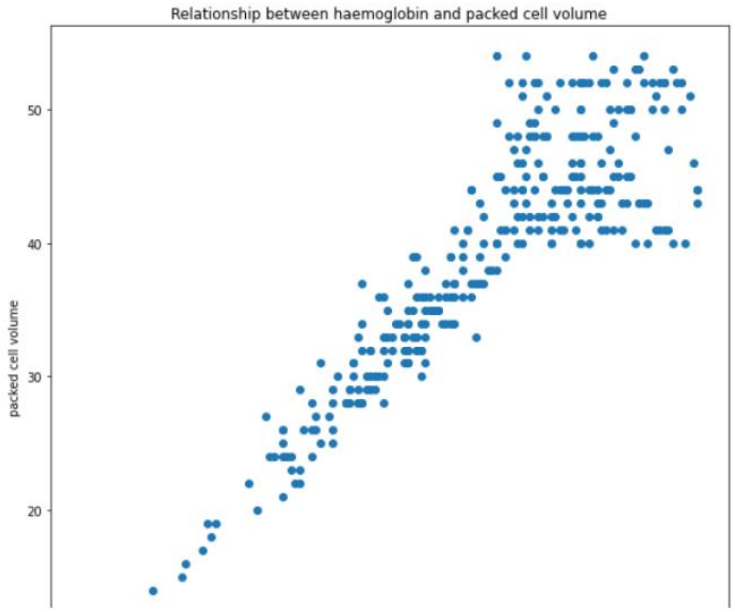

Fig 3 Haemoglobin and packed cell volume relationship

The linear relationship between haemoglobin and packed cell volume is shown by the graphical datapoints in Figure 3, which depicts the relationship between the two variables. Whenever haemoglobin is below 13-14 then the person is positive for chronic disease & whenever haemoglobin is near 18 then the person is negative for the disease. The link between haemoglobin and red blood cell count is seen in figure 4, where data points illustrate the roughly linear relationship between the two parameters. The kidney disease that is chronic is represented by blue data points in the above points, whereas non-chronic kidney disease and one outlier with CKD are represented by green data points.

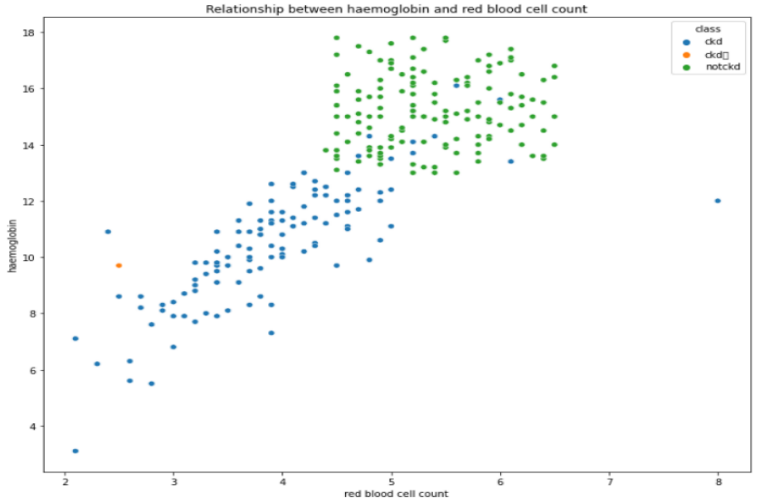

Fig 4 Relationship between haemoglobin and red blood cell count

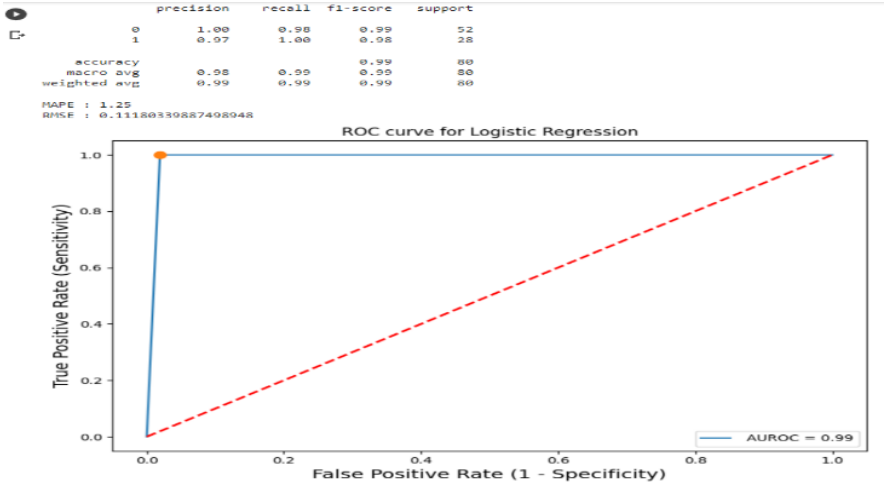

Fig 5 Logistic regression's ROC curve

The true positive rate (TPR) and false positive rate (FPR) have a linear relationship of logistic regression is seen in figure 5, which displays the ROC curve for linear regression. As the range near 1 is consider as good region of curve (ROC).

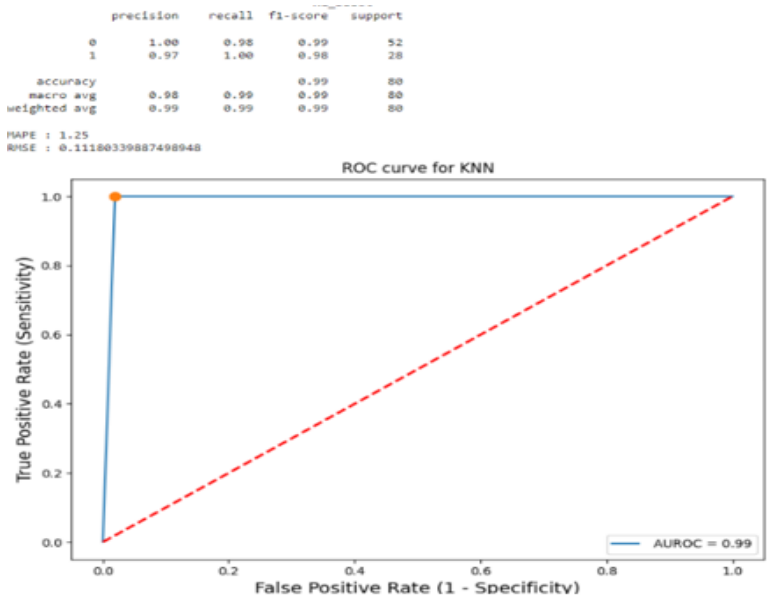

Fig 6 KNN ROC curve

The ROC curve for KNN is displayed in figure 6 and illustrates the linear relationship between the sensitivity and specificity true positive rates (TPR and FPR) in K-Nearest Neighbor (KNN). The curve displays a 0.99 best outcome.

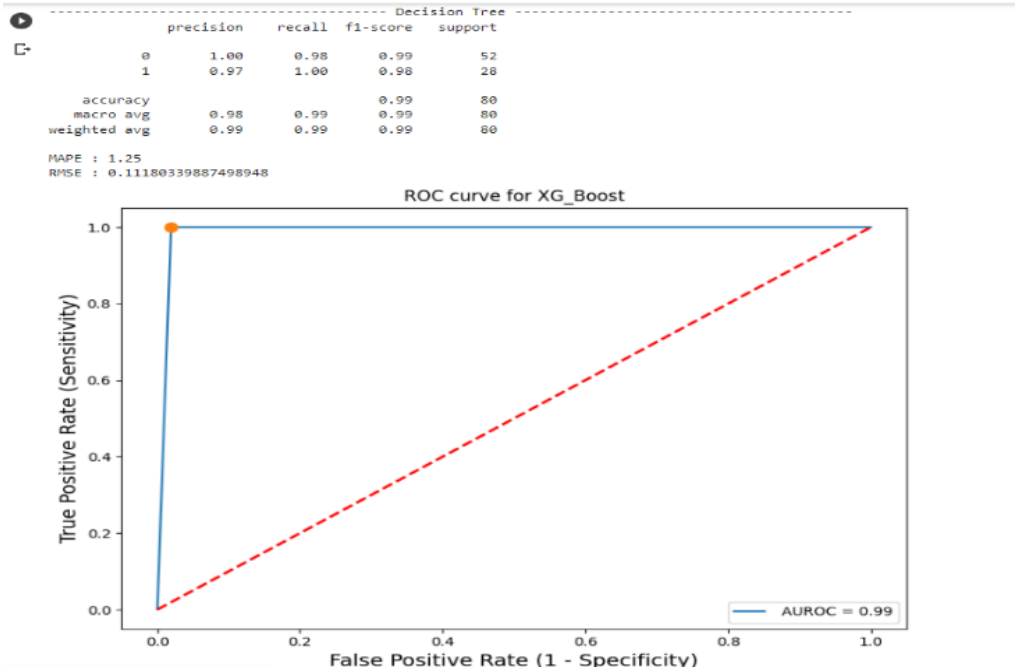

Fig 7 ROC Curve of XG_Boost

According to figure 7, it illustrates the linear relationship between the XG-Boost's True positive rate (TPR) of sensitivity and false positive rate (FPR) of specificity can be found. Following figure 8 shows that the ROC curve for the decision tree and figure 9 shows the ROC curve for Ada Boost.

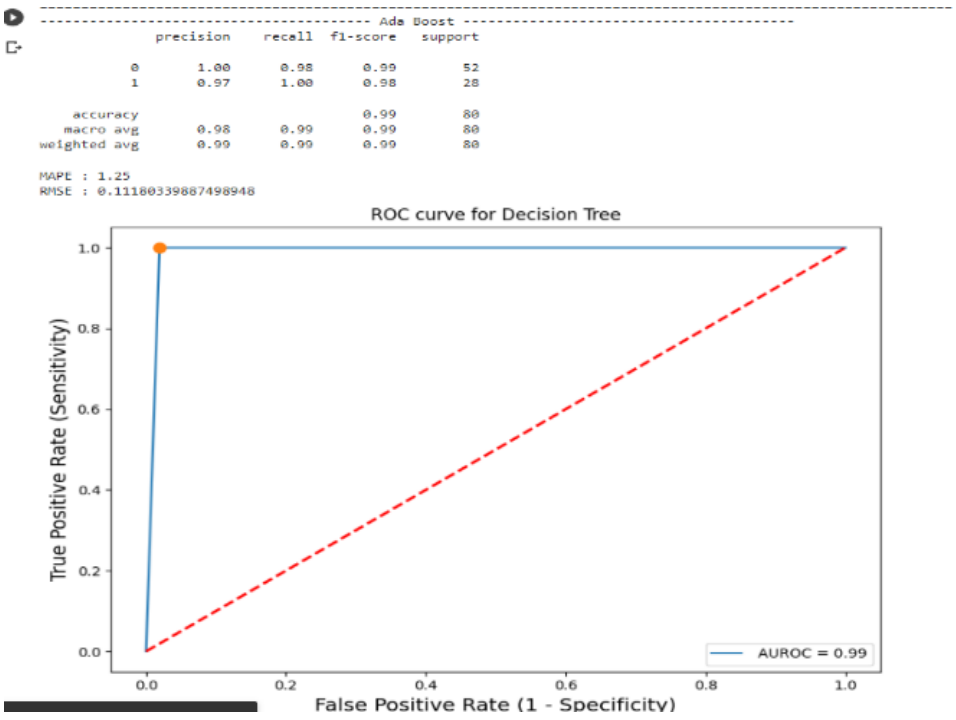

Fig 8 ROC of Decision Tree

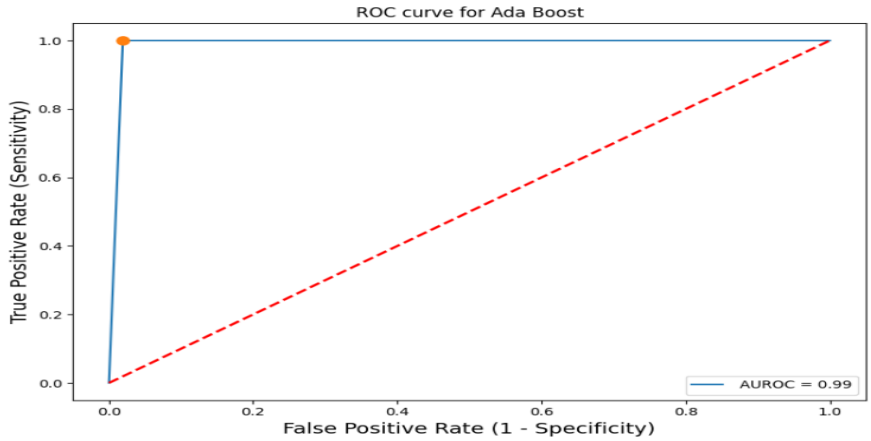

Fig 9 ROC for Ada boost

4 Classification Report:

Performance metrics such as accuracy, F1 score, recall, precision, AUC, and ROC are used to assess the presence of a proposed model. These metrics include condition positive (P), condition negative (N), the total number of truly negative data points, true positive instances, precise negative test results (TN), true positive (TP), positive error (FP), and false negatives (FN). Positive errors indicate the absence of a disease or trait, while false negatives falsely suggest the absence of a specific ailment or characteristic.

Table 4 Analysing several machine learning methods for renal disease in chronic patients

Classifier	F1-Score	Accuracy	Precision	Recall
Logistic Regression	0.99	0.98	1.00	0.98
K-NN	0.99	0.99	1.00	0.98
XG_Boost	0.99	0.99	1.00	0.98
Decision Tree	0.99	0.98	1.00	0.98

From the above table 4, it's confirmed that the K-NN and XG_Boost algorithms are performed better in all aspects. Hence XG-Boost is the best classifier to check the person with chronic kidney disease or not.

5 Conclusion and Future Scope

Finally, this paper investigated the efficiency of different machine learning algorithms for detection and forecasting of chronic heart disease. Logistic regression, KNN, XG Boost, and decision tree were thoroughly assessed for discrete proficiencies and contribution to the challenge. The outcome of this paper was encouraging, with all algorithms displaying excellent performance. Logistic regression and decision tree demonstrated 98% accuracy, where XG Boost and KNN demonstrated 99% accuracy, these models show a lot of potential for making detection with 99% accuracy in real-world applications.

In conclusion, the thorough study of machine learning algorithms in chronic heart disease diagnosis adds new insights to the field of healthcare. Medical professionals can make informed judgments and deliver individualized treatment strategies by using the power of these applied algorithms, ultimately leading to enhanced patient outcomes and overall kidney health issues as early as possible. The search for early detection of kidney disease continues, and machine learning is a valuable tool in that pursuit.

References

[1] M. Irmansyah, E. Madona, A. Nasution, and R. Putra, Low-Cost Heart Rate Portable Device for Risk Patients with IoT and Warning System, *International Conference on Applied Information Technology and Innovation, Padang, Indonesia, 2018, pp. 46–49.*

[2] S. Cherrier, V. Deshpande., From BPM to IoT, *International Conference on Business Process Management, Barcelona, Spain, 2017, pp. 310-318.*

[3] S. Abba, A. M. Garba, An IoT-Based Smart Framework for a Human Heartbeat Rate Monitoring

and Control System, Proceedings, *Vol. 42, No. 1, pp. 1-8, November 2019.*

[4] M. N. I. Khan, D. F. Noor, M. J. H. Pantho, T. S. Abtahi, F. Parveen, M. I. H. Bhuiyan, A low-cost optical sensor based heart rate monitoring system, *International Conference on Informatics, Electronics and Vision, Dhaka, Bangladesh, 2013, pp. 1-5.*

[5] K. N. Sree, G. M. Rao, IoT Based Low Cost and Ageing Healthcare Monitoring System, International Journal of Engineering and Advanced Technology, *Vol. 9, No. 6, pp. 6–14, August 2020.*

[6] M. Murali, M. Bhargava, G. Snehaa, A. Anand, M. A. Haque, V. Sarobin, Data analytics on IoT-based health monitoring system, *International Journal of Recent Technology and Engineering, Vol. 8, No. 1, pp. 220–223, May 2019.*

[7] M. A. Al-Sheikh, I. A. Ameen, Design of Mobile Healthcare Monitoring System Using IoT Technology and Cloud Computing, *3rd International Conference on Sustainable Engineering Techniques, Baghdad, Iraq, 2020, pp. 1-17.*

[8] P. S. Banerjee, A. Karmakar, M. Dhara, K. Ganguly, S. Sarkar, A novel method for predicting bradycardia and atrial fibrillation using fuzzy logic and Arduino supported IoT sensors, Medicine in Novel Technology and Devices, *Vol. 10, p. 1-11, January 2021.*

[9] M. M. Islam, A. Rahaman, M. R. Islam, Development of Smart Healthcare Monitoring System in IoT Environment, *SN Computer Science, Vol. 1, No. 3, pp. 1-11, 2020.*

[10] V. Tamilselvi, S. Sribalaji, P. Vigneshwaran, P. Vinu, J. Geetha Ramani, IoT-based health monitoring system, *6th International Conference on Advanced Computing and Communication Systems, Coimbatore, India, 2020, p. 386–389.*

[11] D. Acharya, S. N. Patil, IoT-based Health Care Monitoring Kit, 4th International Conference on Computing Methodologies and Communication, Erode, India, 2020, pp. 363–368.

[12] M. J. Gregoski, M. Mueller, A. Vertegel, A. Shaporev, B. B. Jackson, R. M. Frenzel, S. M. Sprehn, F. A. Treiber, development, and validation of a smartphone heart rate acquisition application for health promotion and wellness telehealth applications, *International Journal of Telemedicine and Applications, Vol. 2012, No.1, pp. 1-11, January 2012.*

[13] J. J. Oresko, Z. Jin, J. Cheng, S. Huang, Y. Sun, H. Duschl, A. C. Cheng, A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing, *IEEE Transactions on Information Technology in Biomedicine, Vol. 14, No. 3, pp. 734–740, April 2010.*

[14] S. Trivedi, A. N. Cheeran, Android-based health parameter monitoring, *International Conference on Intelligent Computing and Control Systems, Madurai, India, 2017, pp. 1145–1149.*

[15] S. P. Kumar, V. R. R. Samson, U. B. Sai, P. L. S. D. M. Rao, K. K. Eswar, Smart health monitoring system of a patient through IoT, *International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud), Palladam, India, 2017, pp. 551–556.*

[16] M. R. Desai, S. Toravi, A Smart Sensor Interface for Smart Homes and Heartbeat Monitoring using WSN in IoT, *International Conference on Current Trends in Computer, Electrical, Electronics and Communication, Mysore, India, 2017, pp. 74–77.*

[17] Wesam S. Bhaya, Review of Data Preprocessing Techniques in Data Mining, Journal of Engineering and Applied Sciences *12(16):4102-4107.*

[18] Puneet Mishra, Alessandra Biancolillo, Jean Michel Roger, Federico Marini, Douglas N. Rutledge, New data preprocessing trends based on an ensemble of multiple preprocessing techniques, TrAC Trends in Analytical Chemistry, *Volume 132, 2020,116045, ISSN 0165-9936,https://doi.org/10.1016/j.trac.2020.116045*

[19] Waskom, M. L., (2021). seaborn: statistical data visualization. *Journal of Open-Source Software, 6(60), 3021.*

[20] Ms. Deepali Bhende, Gopal Sakarkar, Review of Machine Learning Techniques for Analysis of Medical Data Sets, *Springer's, 2nd FICR International Conference on Rising Threats in Expert Applications and Solutions, (FICR- TEAS-22), Jaipur.*

[21] S. Trivedi, A. N. Cheeran, Android-based health parameter monitoring, *International Conference on Intelligent Computing and Control Systems, Madurai, India, 2017, pp. 1145–1149.*

[22] J. J. Oresko, Z. Jin, J. Cheng, S. Huang, Y. Sun, H. Duschl, A. C. Cheng, (201)}. A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing, *IEEE Transactions*

on Information Technology in Biomedicine, Vol. 14, No. 3, pp. 734–740, April 2010.

[23] M.Jaya lakshmi, C.Sadia Sameen, D.Maneesha, G.Dharani, K.Farhat Mubeena "Smart Home using Blynk App Based On IOT" *International Journal of Creative Research Thoughts (IJCRT), Volume 10, Issue 5 May 2022 | ISSN: 2320-2882.*

[24] Thomas G. Pickering, Daichi Shimbo, and Donald Haas, "Ambulatory Blood-Pressure Monitoring", *N Engl J Med 2006; 354:2368-2374 DOI: 10.1056/NEJMra060433.*

[25] Kale, Yogesh and Rathkanthiwar, Shubhangi V and Rajurkar, Sarvadnya and Parate, Himanshu and Ninawe, Anshul and Bharti, Aditya, "Analysis for Determining Best Machine Learning Algorithm for Classification of Heart Diseases", *2023 IEEE 8th International Conference for Convergence in Technology (I2CT), page 1-6, April 2023, IEEE.*

[26] Albur M, Hamilton F, MacGowan AP. Early warning score: a dynamic marker of severity and prognosis in patients with Gram-negative bacteraemia and sepsis. Ann Clin Microbiol Antimicrob. *2016 Apr 12;15:23. doi: 10.1186/s12941-016-0139-z. PMID: 27071911; PMCID: PMC4830018.*

[27] Ng R, Sutradhar R, Kornas K, et al. Development and Validation of the Chronic Disease Population Risk Tool (CDPoRT) to Predict Incidence of Adult Chronic Disease. *JAMA Netw Open. 2020;3(6):e204669. doi:10.1001/jamanetworkopen.2020.4669.*

[28] Siddiqui E, Jokhio AA, Raheem A, Waheed S, Hashmatullah S. The Utility of Early Warning Score in Adults Presenting With Sepsis in the Emergency Department of a Low Resource Setting. *Cureus. 2020 Jul 6;12(7):e9030. doi: 10.7759/cureus.9030. PMID: 32775109; PMCID: PMC7406184.*

[29] Khekare, Ganesh, Pushpneel Verma, Urvashi Dhanre, Seema Raut, and Ganesh Yenurkar. "Analysis of Internet of Things based on characteristics, functionalities, and challenges." *International Journal of Hyperconnectivity and the Internet of Things (IJHIoT) 5, no. 1 (2021): 44-62.*