

**CHRONIC KIDNEY DISEASE: AN ANALYTICAL PREDICTION
USING MACHINE LEARNING ON THE PATIENTS OF
DIFFERENT AGE**

BY

Md. Mosfikur Rahman
ID: 181-15-2065

Akash Ahmed
ID: 181-15-1714

AND

Mahfuja Ferdousi Mahin
ID: 181-15-1860

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By
Nazmun Nessa Moon
Associate Professor
Department of Computer Science & Engineering
Daffodil International University

Co-Supervised By
Amit Chakraborty
Lecturer
Department of Computer Science & Engineering
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2022

APPROVAL

This Project titled “Chronic Kidney Disease: An Analytical Prediction Using Machine Learning on the Patients of Different Age”, submitted by Md. Mosfikur Rahman, Akash Ahmed and Mahfuja Ferdousi Mahin to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 18.09.2021.

BOARD OF EXAMINERS



Mohammad Monirul Islam

Senior Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Mahfujur Rahman [MMR]

Senior Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Swakkhar Shatabda

Associate Professor,

Department of Computer Science and Engineering (CSE)
United International University (UIU), Dhaka, Bangladesh

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Nazmun Nessa Moon, Associate Professor Department of Computer Science & Engineering** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Nazmun Nessa Moon
Associate Professor
Department of Computer Science & Engineering
Daffodil International University

Co-Supervised by:



Amit Chakraborty
Lecturer
Department of Computer Science & Engineering
Daffodil International University

Submitted by:



Md. Mosfikur Rahman
ID: 181-15-2065
Department of Computer Science & Engineering
Daffodil International University



Akash Ahmed
ID: 181-15-1714
Department of Computer Science &
Engineering
Daffodil International University



Mahfuja Ferdousi Mahin
ID: 181-15-1860
Department of Computer Science &
Engineering
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Nazmun Nessa Moon, Associate Professor**, Department of Computer Science & Engineering, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Nazmun Nessa Moon, Amit Chakraborty and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Chronic Kidney Disease (CKD) is a term used in the medical profession to represent a range of illnesses that result in kidney damage or a decreased Glomerular Filtration Rate (GFR). In recent years, medical advances have made it feasible for doctors to treat this disease utilizing a variety of different methods, which they have done. Artificial intelligence and machine learning have gained popularity in recent years as a method of enhancing medical care and medical research in general, particularly in the field of medicine. Because Kidney Condition is a potentially fatal disease, it necessitates the application of machine learning to anticipate when it will develop in the first place. To forecast the development of "Chronic Kidney Disease," a broad variety of machine learning techniques, applications, and algorithms may be used in conjunction with one another. Using this method, a machine-learning algorithm generates a certain output, and the algorithm that outperformed all of the other algorithms is chosen as the best performance. This may make it possible for any physicians to identify the beginning of this disease as soon as the dialysis report is received. It is also possible to identify which component of the disease is the primary cause of the illness using this technique, which may be determined from the report study. It is necessary to use more complicated and dynamic algorithms in order to get the best possible outcome in this system, and these algorithms include Random Forest, Naïve Bayes, Decision Tree, K-Nearest Neighbor (KNN), XGBoost, AdaBoost, and others.

TABLE OF CONTENTS

APPROVAL	ii
DECLARATION	iii
ACKNOWLEDGEMENT	iii
ABSTRACT.....	v
LIST OF FIGURES	x
LIST OF TABLES	xi

CHAPTER 1

INTRODUCTION.....	1
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	3
1.4 Research Questions	3
1.5 Expected Outcome	4
1.6 Report Layout.....	5

CHAPTER 2

LITERATURE REVIEW	6
2.1 Introduction	6
2.2 Related Works	6
2.3 Comparative Analysis	11
2.4 Scope of the Problem	12
2.5 Challenges	13

CHAPTER 3

RESEARCH METHODOLOGY & SYSTEM ARCHITECTURE.....	14
3.1 Introduction	14
3.2 Research Subject	14
3.3 Machine Learning Techniques	14

3.3.1 Supervised Learning	15
3.4 Classification Techniques	15
3.4.1 Learning.....	15
3.4.2 Classification	16
3.5 Algorithmic Details	16
3.5.1 Logistic Regression	16
3.5.2 Support Vector Machine.....	17
3.5.3 K-Nearest Neighbors	17
3.5.4 Gaussian Naïve Bayes	17
3.5.5 Perceptron	18
3.5.6 Linear Support Vector Classifier	18
3.5.7 Stochastic Gradient Descent	19
3.5.8 Decision Tree.....	19
3.5.9 Random Forest.....	20
3.5.10 Adaptive Boosting (AdaBoost)	20
3.5.11 eXtreme Gradient Boosting (XGBoost)	21
3.6 Proposed System	22
3.6.1 Data Collection	22
3.6.2 Dataset	23
3.6.3 Data Pre-processing.....	23
3.6.4 Data Normalization.....	23
3.6.5 Data Splitting.....	23
3.6.6 Imply Algorithms	23
3.6.7 Model Analysis.....	24
3.6.8 Extract Appropriate Algorithm.....	24
3.6.9 Creating Model for Web Interface.....	24
3.6.10 Building a Web Interface.....	25
3.6.11 Execute Model.....	25
3.6.12 Input Values.....	25
3.6.13 Predictive Result.....	26
3.7 System Architecture	26

3.7.1 User Segment.....	27
3.7.2 Web Insider.....	27
3.7.3 Machine Learning Model	27

CHAPTER 4

EXPERIMENTAL RESULTS & DISCUSSION.....	28
4.1 Introduction	28
4.2 Experimental Results.....	28
4.2.1 Data Acquisition	28
4.2.2 Data Utilization.....	29
4.2.3 Feature Importance	31
4.3 Result & Discussion	32
4.3.1 Confusion Matrix.....	32
4.3.2 Classification Report	35
4.4 Result Analysis.....	36
4.4.1 Accuracy	37
4.4.2 Jaccard Score	37
4.4.3 Cross Validated Score	38
4.4.4 AUC Score.....	39
4.4.5 ROC Curve	39
4.4.6 Standard Deviation	41
4.4.7 Misclassification & Error	41
4.5 Web Implementation	42
4.5.1 Web Interface	42
4.5.2 Web Output Analysis.....	43

CHAPTER 5

IMPACT ON SOCIETY & SUSTAINABILITY.....	45
5.1 Introduction	45
5.2 Impact on Society.....	45
5.3 Ethical Aspects.....	45

5.4 Sustainability	46
 CHAPTER 6	
FUTURE SCOPE & CONCLUSION	47
6.1 Introduction	47
6.2 Implication for Further Study.....	47
6.3 Recommendations	47
6.4 Conclusion.....	48
 REFERENCES.....	 49

LIST OF FIGURES

Figure No.	Figure Name	Page No.
Figure 3.1	Proposed Method to Predict Chronic Kidney Disease	22
Figure 3.2	System Architecture	26
Figure 4.1	Accuracy Chart	37
Figure 4.2	Jaccard Score Chart	38
Figure 4.3	Cross Validated Score	38
Figure 4.4	AUC Score Chart	39
Figure 4.5	ROC Curve	40
Figure 4.6	Web Interface	43
Figure 4.7	Web Interface with Negative Output	44
Figure 4.8	Web Interface with Positive Output	44

LIST OF TABLES

Table No.	Table Name	Page No.
Table 4.1	Data Acquisition & Null Percentage	29
Table 4.2	Dataset Description	30
Table 4.3	Feature Importance	31
Table 4.4	Confusion Matrix	32
Table 4.5	Confusion Matrix for Algorithms	33
Table 4.6	Classification Report	35
Table 4.7	Accuracy, Jaccard, Cross Validated and AUC Score	40
Table 4.8	Standard Deviation	41
Table 4.9	Misclassifications & Errors	42

CHAPTER 1

INTRODUCTION

1.1 Introduction

Chronic Kidney Disease is a phrase used to describe kidney impairment that has been present for a long period of time. Getting older is typically linked to this prevalent ailment. It may affect anybody, but it is more frequent in individuals of color, especially those from the Caribbean and South Asia. CKD might worsen with time and the kidneys may potentially cease functioning entirely, but this is rare. Many persons with CKD go on to enjoy long and healthy lives despite their illness. Researchers from all over the world tried to predict the exact reason for this kidney disease.

Multiple diseases appeared in kidney that cause severe condition of any patient. Chronic Kidney Disease, Kidney Stones, Glomerulonephritis, Polycystic kidney disease and Urinary tract infections are the most lethal diseases of kidney. Among them Chronic Kidney Disease is commonly caused by high blood pressure. High blood pressure poses a threat to the kidneys because it can put an extra strain on the glomeruli, which filter out waste products. The kidneys' small blood channels known as glomeruli purify the blood. The increasing pressure affects the renal vessels over time, resulting in decreased kidney function. The kidneys' ability to do their job will ultimately degrade to the point of failure. If this were the case, a person would have to undergo dialysis. Dialysis is a process that removes waste and excess fluid from the blood. Dialysis can aid in the treatment of renal disease, but it will not cure it. Depending on patient's situation, a kidney transplant may be a possibility for him/her. Chronic kidney disease is commonly attributed to diabetes. High blood sugar is a symptom of a variety of illnesses, including diabetes. Over time, having a higher blood sugar level affects the kidneys' blood vessels. This indicates that the kidneys are unable to perform their normal function of cleaning the blood. When your body is overwhelmed with pollutants, kidney failure can develop. Researchers from multiple regions showed the actual lethality rate of Chronic Kidney Disease in their research. In variant research works, it is clear that male patients from Chronic Kidney Disease had higher mortality rate than female patients. Kidney Disease Fatalities in Bangladesh reached

16,948 in 2018 according to the most recent WHO statistics, or 2.18 percent of all deaths. Bangladesh is ranked 94th in the world for its age-adjusted Death Rate of 14.83 per 100,000 people [1]. Currently, COVID-19 is affected all over the Bangladesh. COVID-19 exposure increases the chance of developing a life-threatening illness in patients with chronic kidney disease (CKD), making access to dialysis and other healthcare treatment facilities even more challenging [2].

There are several branches of Machine Learning Algorithms. Such as, Supervised Machine Learning, Unsupervised Machine Learning, Semi-supervised Machine Learning and Reinforcement Machine Learning. Algorithms from Supervised Machine Learning Algorithms such as K-Nearest Neighbor, Decision Tree, Random Forest, Perceptron, AdaBoost, Gaussian Naïve Bayes and XGBoost etc. are the most important algorithms and these are used to predict Chronic Kidney Disease in this research work. From this system, anyone can predict their probability of Severe Kidney Disease probability. But a web implementation process can be the most useful system for pathology department and patients as well, if they want to know the exact possibilities for affected by Chronic Kidney Disease. The actual purpose of this research is to build a model trained by a relevant dataset, finding the best algorithm and then predict CKD at any stage. In this research work, after extracting the best algorithm from most important algorithms of Machine Learning, it transfers its information to a web application, where patients/doctor/pathologist can see the report and input these values to know the possibilities of kidney disease. In near future, modern era is going to be more modern and hi-tech. This research will assist individuals in better understanding kidney disease causes and taking the necessary actions to combat the condition with knowing possible report from web implementation.

1.2 Motivation

Bangladesh is home to a large number of individuals who are suffering from a broad variety of chronic illnesses, including diabetes and hypertension. Chronic illnesses are becoming more prevalent in today's society, and this tendency is expected to continue. Patients with chronic illnesses must be present at every hospital or clinic, no matter how small the number of patients present at any one moment may seem to be. Bangladesh is surrounded on all sides by rivers, which makes it a water-rich nation. As a consequence of

these characteristics, people are more susceptible to water-borne illnesses such as diarrhea, as well as the most serious illness known as Chronic Kidney Disease. And there must be a technique, based on the use of Machine Learning, for better understanding and preventing the development of this illness from spreading.

1.3 Rationale of the Study

If Chronic Kidney Disease can be diagnosed and predicted early, it can be cured by proper treatment very quickly. In the real world, people have to test many things in the lab to become sure that they have Chronic Kidney Disease or not, and it is very time-consuming. That's why, a model has been proposed and trained by a relevant dataset that can predict Chronic Kidney Disease at any stage. Main rationale focus to work on this study to make people's life easier and to save time. A web interface has been developed where patients can give data and can predict Chronic Kidney Disease at home. It will be helpful for doctors and Chronic Kidney Disease patients.

1.4 Research Questions

There are multiple question can be asked about this study. From different individuals, a set of question has been extracted to make this study more compact.

- ***Why Chronic Kidney Disease Prediction was the target of this study?***

Chronic Kidney Disease is one of the biggest problems in the world. It gets worse over time. It has many stages and at the last stage the kidney stops working at all. Then it leads a person to death. So, if Chronic Kidney Disease can be predicted at early stage, it can be cured with proper treatment and maintaining many rules and regulations. That's why Chronic Kidney Disease was the target of the study

- ***Why machine learning approach? Does it reliable?***

Machine learning is a most popular technique which is mainly used for any kind of prediction. Using a vast number of data, a model can train itself and can predict any outcome. Using machine learning approach in medical dataset it can be predicted Chronic Kidney Disease easily. In the current times, the world is going through a

modernization era. If a scenario can be considered about 10 years before when Artificial Intelligence or Machine Learning was not developed that much, at that time these prime points were just a name with some of mathematical logic. But now, half of the whole planet's technology depends on Artificial Intelligence. Therefore, proper practice and more precision on this field can make it more reliable though it is reliable enough now.

- ***What is the main purpose to use a web interface?***

A web interface helps a user to insert values inside a machine. It is really a generous way to general users. Using a web interface and a connection of internet one can use this project from anywhere in the world. A web interface can generate result faster too. One can get the basic idea of the condition of his or her kidney at home by using this web interface. Therefore, it is really a convenient way to predict CKD.

- ***What are the reasons to use 11 separate algorithms?***

Using 11 different algorithms, one suitable algorithm was targeted that suits the Chronic Kidney Disease dataset. Differentiating 11 algorithms and analyzing them the best suited algorithm has been obtained that will have less error rate and highest accuracy rate. If only one algorithm was selected, it is impossible to find the best algorithm as no one do not know which algorithm will suit the dataset in the most efficient and effective way.

1.5 Expected Outcome

During this study period, main theme or expected outcome has been changed a couple of times. It helps to clarify the exact outcome of this study. This research has the potential to prevent Chronic Kidney Disease at its earliest stages while also identifying the root cause of the disease's spread. It may be determined by doctors or analysts at what age individuals become afflicted with Chronic Kidney Disease (CKD). Using precise calculations and algorithms, a detailed internal examination of Chronic Kidney Disease may be disclosed for the sake of medical analysis. While this research is being carried out, a team may warn riverine communities and regions about Chronic Kidney Disease while it is still in progress.

The last method used in this research is to develop a web-based system that recommends any potential patients who should be contacted by a doctor.

1.6 Report Layout

In this report, six individual chapters are discussed to make this research report more compact and efficient for any readers or researchers.

Chapter 1 gives an important introduction about this research work. This is related to Kidney Disease and its details briefly. In this chapter, this shows research motivation, rationale of this study, relevant research questions, expected outcome and total management information along with financial aspects.

Chapter 2 gives the detailed report about background of this study. Such as, machine learning system, classification information and related work based on this research study. Comparative analysis, scope of this problem statement is also described in this chapter with perceived challenges.

Chapter 3 gives the descriptive information about methodology, proposed system and system architecture for this research study. Algorithmic details for each used algorithms is described from mathematical scratch to current condition is discussed.

Chapter 4 gives the complete result analysis for each steps result. In concludes with best accuracy score with best algorithm, jaccard score, cross validated score, confusion matrix, classification report. ROC-AUC curve is also described for each algorithms. Standard Deviation, Misclassification, Mean Absolute Error and Mean Squared Error are described in the last segment of this chapter.

Chapter 5 describes this research's impact on society with Ethical Aspects which is most important for any impactful research work. Sustainability of this research work has been described in the end segment of this chapter.

Chapter 6 shows the future scope of this research work where it is briefly described as the extension of this research study. This chapter concludes the entire research report with useful conclusion where core findings of this research is briefly discussed.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, Chronic Kidney Disease has been a grate problem for a long time. So, the background of this disease has a lot of traumatic history and many loses. So, many work has been done to prevent CKD to save lives. In this chapter the preliminaries and the background of this disease has been discussed. Some related work of this issue has been discussed in this chapter. Finally, some comparative analysis has been provided in order to demonstrate how much the current work has been influenced.

2.2 Related Works

The IoMT platform Heterogeneous Modified Artificial Neural Network (HMANN) is proposed for identifying, segmenting, and diagnosing chronic kidney disease early on. The proposed HMANN is a Multilayer Perceptron (MLP) that uses Backpropagation in addition to being a Support Vector Machine (SVM) (BP). A segmented ultrasound picture is utilized as a preprocessing step in order to determine the kidney region of interest. The suggested HMANN approach for kidney segmentation provides excellent accuracy while also drastically lowering its time to outline the contour. By evaluating candidates' facial photos, this study offers a progressively trainable network, i.e., dual-stack network (DsNet), to concurrently discriminate patients with chronic renal disease and diabetes mellitus from healthy persons. The first stack subset extracts representative high-level features efficiently from the face images. Then, during the second stack network, the extracted high-level features of the first stack subset can be further analyzed before healthy people can simultaneously classify both illnesses. When compared to existing classic noninvasive detection techniques [3].

This paper of Ali et al. focuses on the applicability issues of automated decision systems, using CKD diagnosis as a case study in developing countries. This work provides a cost-sensitive ensemble feature ranking method for group-based feature selection that is more realistic. This is one of the first studies to demonstrate that cost-sensitive ensemble

rankings for non-cutting groups may achieve the stated objectives, namely low cost and high accuracy solutions. The experiment utilizes seven well-known classification algorithms and eight techniques of comparison selection to show the technique's efficacy. It concludes that by integrating the cost factor into the objective space of solution formulations, the applicability of automated CKD systems can be improved. A cost-effective yet accurate method to be used as the CKD screening system can therefore be found [4].

Improved Teacher Learner Based Optimization (ITLBO) is the proposed research algorithm to select the best feature subject in terms of assessing the fitness function and common control parameters, such as population size and generations, based on the Chebyshev Distance Requirement to find the optimum feature set for early diagnosis of chronic conditions. Applied to Chronic Kidney Disease (CKD) data sets, the proposed function selection methodology resulted in a considerable decrease of 36% compared to the 25% feature reduction obtained with the original TLBO methodology. A valuation of the correctness of the Support Vector Machine (SVM), Convolution Neural Networks (CNN) and Gradient Boosting algorithms validate the generated optimal feature subset from TLBO algorithms and feature subsets produced from the ITLBO algorithm. Experimental results show that overall classification accuracy is improved for the three algorithms for the generated feature subset compared with the original TLBO approach in the paper of Balakrishnan et al. [5].

This research of Lambert et al. aims to forecast chronic kidney disease with solely nominal attributes that are numeric and nominal and numerical. The classification and prediction of this research contain the CFS approach used to extract and categorize key characteristics into CKD and not CKD. CFS applies to nominal, numerical, and nominal plus numerical data for selecting features. The CFS result is compared to three different ranking methods for function selection, including information gain, gain ratio, and the Relief technique. Correlation-based selection accuracy and minimum sequential optimization (CFS-SMO) resulted in a 98.5 percent accuracy for nominal, 95.25 percent accuracy for numerical, and 98.5 percent accuracy for nominal plus numerical. These experimental results show that correlation-based function selection (CFS) effectively retrieved components from the CKD dataset, and that SMO classified renal disease as the

acceptable benchmark condition. As a result, the CFS-SMO has been regarded as a hopeful tool for properly detecting renal disease, allowing medical professionals to make informed decisions [6].

This work of Nusinovici et al. assesses the achievements and the logistic regression in machine training (ML) algorithms for cardiovascular disease, chronic kidney disease (CKD), diabetes (DM), and hypertension (HTN) prediction, and in a prospective cohort study using simple clinical forecasts. The basic logistic regression model was compared to five additional machine learning models: a single hidden layer neural network, a vector support machine, a random forest, a gradient boosting machine, and k-nearest neighbor. For CKD predictions (0.905 [0.88, 0.93]) and DM (0.768 [0.73, 0.81]), logistic regression was the biggest area in the CKD and DM operating curves. The system of logistic regression, gradient boosting, and neural network was among the best models [7].

This study of Segal et al. investigated 10,000,000 health claims on a commercial health insurance database out of 550,000 patient records. The criteria comprised patients over the age of 18 who had been diagnosed with CKD Stages 1–4. 240 predictor candidates have been compiled and grouped into six feature groups. The Word2Vec algorithm is used to collect temporal information on the three principal data components, diagnostic, procedures, and drugs, via a feature embedding process. Use the gradient booster algorithm for analysis (XGBoost implementation) [8].

This study of Sealfon et al. will address how automatic approaches to the research of renal disease may be implemented, with a specific focus on how the relationship between the genotype and phenotype may be understood. An introduction to machine learning, which is broadly aimed at nephrologists and Kidney researchers, emphasizes applications of machine-learning approaches to understanding genotype-phenotype relationships. The recent generation of multidimensional, information-rich, kidney-specific data ensures that machine-teaching approaches with the potential to provide biological and translational insight to address various questions in kidney biology are increasingly successful [9].

In the study by Luo et al. of the cross-sectional associations between blood levels of four heavy metals and kidney function, data from the NHANES 2015-16 data was used for 1,435 adults over and above. eGFR and an albumin-creatinine ratio have been used to assess the kidney's estimated glomerular filtrating rate (eGFR) to detect albuminuria and

chronic kidney disease (CKD) (AKR). When it came to non-linearity, the BKMR looked at the dose-response relationships between different mixes and outcomes, as well as possible interactions between those responses. The interaction package in R detected interactions between continuous results. All four renal action measurements in dose-response patterns were strongly related to a greater metal combination. Pb had eGFR, albuminuria, and ACR as its strongest connection, and CKD was the second strongest. Pb's interaction with eGFR and interactions of Pb to Cd with ACR has also been observed [10].

The research of Sambyal et al. critically assesses existing statistical and machine-learning models for retinopathy, neuropathy, and nephropathy consequences. Inferential and descriptive analysis is made possible by statistical tests such as chi-square, odds ratio, t-tests, or the ANOVA. On the other hand, machine learning models like vector support machines, neighbour's closest to home, deep neural networks and naive mouths also provide predictive analysis. It should be pointed out that both studies could restore regulatory impairment of glucose early in their course, leading to the prevention of T2D's long-term problems. Despite changes in both analytical methodologies, their integration in medical practice may help patients adopt a healthy lifestyle effectively and reduce significant overheads for the health care of an individual and society [11].

A generic architecture for the prediction of illnesses has been presented for this work in the health sector. This system of Harimoorthy et al. has been tried with the use of reduced set-features of Chronic Kidney Disease, diabetes and heart disease data by employing an enhanced SVM-radial bias kernel approach and has been compared to others such as SVM-Linear, SVM-Polynomial and Random Forests and Decision Trees in R studio. The experiment findings showed that enhanced SVM-Radial bias kernel technology delivers accuracy in chronic kidney disease, diabetes and heart disease data sets as 98.3%, 98.7% and 89.9%, respectively [12].

In this research of Chen et al., an Adaptive hybridized Deep Convolutional Neural Network (AHDCNN) for early diagnosis of kidney disease was suggested efficiently and effectively to assess the ability of several deep-learning approaches. The function of the dataset depends on the classification efficiency of technologies. An algorithm model was constructed utilizing CNN to boost the precision of the classification system by lowering the feature dimension. These high-level features assist in creating a tissue classification

system to distinguish between two tissue kinds. The experimental Internet process of the Medical Things Platform (IoMT) leads to breakthroughs in machine training using predictive analytics to give a promising framework in which smart solutions may be recognized to demonstrate their predictive potential outside the field of retinal illness [13].

The International Federation for Diabetes (IDF) has demonstrated that there are 3,820 lakhs with diabetes patients worldwide and that this figure may quadruple over the next 15 years. The work that was done by Reddy et al. proposed is intended to design a model for predicting the highest level of accuracy of human diabetes using machine learning classifiers such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Navies Bayes (NB), and Random Forest (RF), gradient boosting (GB). Pima Indian Diabetes Database (PIDD), a dataset from the Kaggle data repository, was analyzed. Each model has its own Accuracy Score and its Receiver Operating Characteristic curve (ROC), precision, and recall, and evaluated F-measures. The performance of all six classifiers is compared [14].

This paper by Fauvel et al. intends to compare each PH subgroup's prognostic importance in a reference center cohort and identify phenotypically distinct high-risk PH patients using machine learning. From 2002 to 2019, PH patients were included and followed up routinely. Hemodynamic, clinical, laboratory, imaging and variables have been collected for this research. Each subgroup then compared the 4-year survival rate. Furthermore, phenotypical domains with 5 watching real for missing values were imputed and filtered when the Pearson correlation coefficient was >0.6 . Agglomerative hierarchical clustering subsequently carried out the grouping of phenotypic variables and patients: a heat map was generated with participation separation by the Model-Based Clustering Penalized. Significant $P < 0.05$ was considered [15].

CKD is a grave cause of death and incapacity. In 1990, this was the 27th focus reason and in 2010 became 18th. In 2013, about 1 million individuals died from CKD. In this study, a suggested methodology was introduced by analyzing the data of CKD patients, through which the risk from CKD may be predicted. An online data set was gathered from the UCI Machine Learning Repository, and real-time datasets from the Khulna City Medical College are used for this research. Two Machine Learning Algorithms, Random

Forest and ANN, were used in this paper of Yashfi et al. with the accuracy of 97.12% and 94.5% [16].

2.3 Comparative Analysis

A method proposed in the previous study with deep learning technique and a novel detection methodology to detect Chronic Kidney Disease from saliva samples. Convolutional Neural Network (CNN) and Support Vector Machine (SVM) was used to overcome the challenges faced by the data. It provided an average 97.67% accuracy for the CNN-SVM network and a sensitivity and specificity of 97.5% and 97.83%, respectively. The standard CNN model had a 96.51 % average accuracy. The performance of this proposed system was compared to other existing algorithms, and it was discovered that it outperformed different well-known data classification approaches. On the other hand, the authors of this paper got high accuracy of 98.55% in XGBoost Classifier with Jaccard Score of 96.14% and Cross Validated Score of 98.97% and AUC score of 98.21% which is better than previous studies [17].

The data of this paper was collected from the University of California Irvine (UCI) machine learning repository that contained a lot of missing values. This problem was solved using KNN imputation. In this paper, six machine learning algorithms were used to establish models. Random forest outperformed better than the other machine learning models, with a diagnostic accuracy of 99.75%. After examining the errors produced by the existing models, an integrated model was presented that integrates logistic regression and random forest with perceptron, achieving an average accuracy of 99.83% after ten simulations. In this study of the authors, the paper contains UCI data and extra data that was collected from different hospitals in Bangladesh. About 10321 data was collected. In this paper, the authors used 11 algorithms and have tried to find the best suitable algorithm that can give the best accuracy for the data set [18].

The prediction of CKD is used as an example of health care services in the cloud-computing environment in this paper. Using two intelligent techniques, linear regression (LR) and neural network (NN), this study provide a hybrid smart model for predicting CKD based cloud-IoT. The results reveal that the brilliant hybrid model is 97.8% accurate in predicting CKD. The proposed study by the authors has about 11 algorithms. Calculating

the accuracy of the 11 algorithms the best algorithm was proposed. A model was build using the best algorithm and using the model a user-friendly interface was made where one can enter the necessary attributes and can predict the outcome [19].

A method proposed in the previous study was to focus on reliable diagnosis rather than finding the optimum remedy. The primary goal is to determine which of the two algorithms is the best for predicting chronic kidney disease (CKD). Two data mining techniques, the Random Forest algorithm and the Back Propagation Neural Network were used. The Random Forest Algorithm obtains 88.7% certainty and Back Propagation algorithm gives 98.40% certainty. In the study of the authors have implemented 11 algorithms. Among them, four algorithms obtained accuracy over 95% (XGBoost Classifier with 98.55%, Decision Tree and AdaBoost Classifier with 98.11% and Random Forest with 97.09% accuracy) [20].

Previously a strategy for predicting CKD status from clinical data, including data preparation, a missing value management approach, collaborative filtering, and attribute selection was presented. Out of the 11 machine learning algorithms studied, the extra tree classifier and random forest classifier are demonstrated to produce the highest accuracy (100.00%). The study also considers the practical issues of data gathering and emphasizes the need to apply domain knowledge when utilizing machine learning to predict CKD status. In this paper, the authors have chosen the best model for predicting CKD at any stage. The authors have also created a user-friendly interface where people can enter necessary data for predicting CKD and get the result. The interface that was shown in the paper can also be used for different hospitals and can be converted into a website so that one can use it through the internet in the future [21].

2.4 Scope of the Problem

Bangladesh is known as the "River Kingdom". Bangladesh has about 700 rivers that run through it. As a result, all of the settlements are established around waterways. However, as a result of the Industrial Revolution, the water has become more contaminated. As a result, various kinds of diseases affect the citizens of this nation. Chronic Kidney Disease is a significant public health issue in the United States. According to the most recent WHO statistics, the number of people who died from kidney disease in

Bangladesh reached 16,948 in 2018, accounting for 2.18 percent of all fatalities. It may result in death or permanent disability. As a result, it requires immediate treatment. It is possible to anticipate Chronic Kidney Disease in its early stages by monitoring aberrant health conditions and health reports. It is possible to prevent chronic renal disease from progressing further than it already has by adopting the appropriate precautions.

2.5 Challenges

Before the kidneys can generate urine, they must go through a series of complicated excretion and reabsorption processes, which may take many hours. This mechanism must be correctly functioning in order for the body's chemical equilibrium to stay constant. Chronic Kidney Disease, on the other hand, causes this process to be either partly or totally slowed down in those who have it. When you don't know how much fluid someone consumes, it may be difficult to establish whether or not they have a kidney problem. The most challenging part of this research is determining the appropriate variables that must be considered in order to establish whether or not someone has chronic renal disease. A challenge is eliminating all of the null values from the dataset that has been collected. This may be a time-consuming process. It was also challenging to develop an algorithm that was suitable for the situation. The dataset was trained using a number of algorithms, and the scientists then chose the algorithms that had the greatest accuracy for identifying chronic renal disease from the training data (CKD). For the team, one of the most challenging tasks was the creation of an intuitive user interface for this system, which enables anybody to input data and predict CKD at any point.

CHAPTER 3

RESEARCH METHODOLOGY & SYSTEM ARCHITECTURE

3.1 Introduction

In order to start research, the research approach is very important. First of all, a problem is needed to be found then the methods are needed to solve that particular problem. In this chapter the research subject is explained. Then the algorithms that were used to solve were explained then the methodology was described with visualized diagram for better understanding. Even a system architecture was also shown for better visualization.

3.2 Research Subject

Main purpose of a research is to find a problem as a topic and then find out the solution of that current problem. Chronic Kidney Disease is a common disease that's intrinsically intertwined to getting older. It is a long-term condition which can get worse over time. Every year many people die from Kidney Disease. If anyone progresses to the last stage of CKD, his/her kidney works less than 15%. If any types of Kidney Disease can be predicted at an early stage, it can be cured by proper treatment. That's why it is important to work on Kidney Diseases and build a model which can predict CKD at different age. A web interface which will also be user-friendly and people will be able to get their report at home.

3.3 Machine Learning Techniques

It is possible to utilize categorization together with supervised machine learning to categorize objects. For purposes of making decisions, a machine learning system is defined as one that is self-contained and capable of gathering and integrating data on a continuous basis. In order to create a system that is always improving, it is possible to learn from past experiences, make analytical observations, and use other methods. Machine learning methods come in a variety of shapes and sizes.

3.3.1 Supervised Learning

In this work, supervised machine learning techniques are widely used. To make predictions about the future, supervised machine learning techniques use labeled samples of past events. In order to generate predictions about the output values, the learning technique creates an inferred function from an analysis of a well-known training dataset. There are no limits to how much training the system can take after that. If the learning algorithm compares its output to what was intended, it may spot any errors and adjust the model as needed. This thesis deals with Chronic Kidney Disease classification at first stage with multiple techniques of supervised learning.

3.4 Classification Techniques

Classification is a form of data analysis that extracts models describing important data classes. It is the most popular and widely used machine learning technique. Such models which are known as classifier can predict categorical class labels under supervision learning. The predictions are discrete and unordered. No intermediate value can be obtained from classifier. For example, a classifier is built to detect whether an image contains the picture of a dog or a cat. The prediction will be either “dog” or “cat”. No intermediate value can be obtained from classifier. Classification learning technique can be used on labeled data. There are two types of data in classification learning. One type is called training data, another is called test data. Training data are used to build the model and test data are used to validate the model. The classification process can be divided into two steps.

3.4.1 Learning

During the learning phase, a suitable technique and training data are utilized to construct a classifier, which is then evaluated against the real world. When a classification algorithm and training data are combined, the result is a classifier, which is essentially a set of rules that can be applied to a variety of different scenarios in various situations.

3.4.2 Classification

As a result of the learning phase, it is possible to predict which class of unknown data will be predicted using the classifier or model that was developed during the prediction phase. This section makes use of the test data in order to evaluate whether or not a model's predictions are true or not.

3.5 Algorithmic Details

To complete this research work, 11 best algorithms of Supervised Machine Learning algorithms are used. Generally, an algorithm may be defined as an organized collection of instructions that directs computer software on how to transform an input set of data into usable data. Statistics are facts, and any knowledge that is useful to humans, machines or algorithms is valuable information. Just like this, machine learning algorithms work in a flow and some of mathematics. In general, mathematical transformation for all Machine Learning Algorithms are not same. However, this research study contains most important machine learning algorithms which contains relevant algorithmic process in entire system architecture.

3.5.1 Logistic Regression

Logistic Regression is a classification algorithm that can predict a binary outcome 0 or 1. Since this system is predicting a binary form of Chronic Kidney Disease detection as CKD or Not Chronic Kidney Disease as NotCKD, this algorithm can be used to understand and predict the outcome easily. These models may address problems that are far more difficult by combining a variety of features rather than just one. As the Y-axis goes from 0 to 1. This is because the sigmoid function always uses these two values as maximum and lowest, which is ideal for the purpose of categorizing data into two groups. This system acquires a probability (between 0 and 1 obviously) of an observation belonging to one of the two categories by computing the sigmoid function of X. The formula from equation (i) shows the sigmoid function.

The sigmoid function has the following formula:

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \dots \dots (i)$$

Therefore, for Logistic Regression the equation becomes equation (ii).

$$P = \frac{1}{1 + e^{-(b' + b''x)}} \dots \dots (ii)$$

3.5.2 Support Vector Machine

Because the Support Vector Machine (SVM) is a machine learning method, it is essential to closely monitor its performance. It may be used to handle problems involving classification and regression, respectively. Despite the fact that it is used in a range of activities, the majority of the time it is used for categorization. It was discovered that the inventors of this method attempted to plot data items in n-dimensional space using a graphing calculator but were unsuccessful. Data charting in two dimensions was investigated in this study, with the researchers deciding on either the CKD or the NotCKD techniques to use for their analysis.

3.5.3 K-Nearest Neighbors

It is also possible to utilize the k-nearest neighbors' method (KNN) to tackle both classification and regression problems by using a straightforward supervised machine learning approach. The understanding and Implementation of KNN is simple. Euclidean distance is the basic theorem of KNN. KNN is employed in this classification since the dataset is separated into two classes. Equation (iii) derives the actual formula for K-Nearest Neighbor algorithm.

$$D(x_i, x_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \dots \dots (iii)$$

3.5.4 Gaussian Naïve Bayes

Gaussian Naïve Bayes is a Naïve Bayes variation that allows Gaussian normal distribution and ongoing data. Naive Bayes is a collection of supervised algorithms based on the Bayes theorem for classification machine. The categorization approach is basic, yet

very useful. When using continuous data, it is frequently assumed that the values associated with each class are divided by a normal (or Gaussian) distribution. In the equation (iv), complete formulization method is derived for Gaussian Naïve Bayes algorithm.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \dots \dots (iv)$$

3.5.5 Perceptron

In the context of a binary classification machine, the perceptron algorithm is a kind of neural network model that is implemented as an algorithm. Essentially, it is made up of a single node or neuron that receives input from a row of data and predicts the classification of the data into categories. Predictions are produced utilizing a predictor function and a feature vector, both of which are linear classification methods, in this technique of making predictions. In this section, how to train a threshold function that translates an input x to an output $f(x)$ in the form of a binary value is discussed. The binary value of the input x is mapped to the binary value of the output $f(x)$. In the equation (v), complete functionalized method is derived for Perceptron algorithm.

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0, \\ 0 & \text{otherwise} \end{cases} \dots \dots (v)$$

3.5.6 Linear Support Vector Classifier

A Linear SVC (Support Vector Classifier) is a machine learning method that is designed to accommodate the data provided, resulting in a hyperplane that is the "best fit" that divides or categorizes the data according to its classification. Following the creation of the hyperplane, it may be required to provide specific features into the classifier in order to get the "predicted" classification result. This makes this algorithm particularly well suited for a certain objective; nevertheless, it may be used to a wide range of circumstances. When training samples are given to SVC as input, the classifier takes two arrays of shape as input: an array X of shape containing the training samples, and an array y of shape class labels, which are also both collections of form.

3.5.7 Stochastic Gradient Descent

The Stochastic Gradient Descent (SGD) method is a simple yet effective method for learning linear and regressive classifiers with convex loss functions such as (linear) Support Vector Machines and Logistic Regression. Despite the fact that SGD has been available for a long time in the machine learning field, it has only lately garnered a lot of attention in the context of large-scale learning. The class SGD Classifier provides a simple stochastic gradient descent learning procedure that supports a variety of classification loss functions and penalties. An SGD Classifier trained with the hinge loss, which is equal to a linear SVM.

The mathematical principles of the SGD process are described here. Here is given a set of retraining examples like $(x_1, y_1) \dots (x_n, y_n)$ where $x_i \in \mathbf{R}^m$ and $y_i \in \mathbf{R}$ ($y_i \in -1, 1$ for classification), this system have to learn a linear scoring function $f(x) = w^T x + b$ with model parameters $w \in \mathbf{R}^m$ and intercept $b \in \mathbf{R}$. For making predictions for binary classification, the system must consider at the sign of $f(x)$. For finding model parameters, the normalized training error given by equation (vi).

$$E(w, b) = \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) + \alpha R(w) \dots \dots \dots (vi)$$

where L is a loss function that evaluates model fit and R is a normalization term that prohibits complexity of the model; $\alpha > 0$ is a non-negative linear combination that controls the normalization strength.

3.5.8 Decision Tree

In supervised learning, decision trees may be used to address classification and regression problems. However, they are most frequently used to solve classification problems. Throughout this tree-structured classifier, internal nodes reflect dataset properties, branches represent decision rules, and each leaf node represents the conclusion. In order to do this, the Decision Tree must be used to build a training model that employs basic decision rules in order to predict the class or value of the input variables based on the training dataset.

$$E(S) = \sum_{i=1}^c -p_i - \log_2 p_i \dots \dots (vii)$$

In the equation (vii), S means the current state and p_i is the likelihood of an event, i in state S or the percentage of class i in the node state S.

3.5.9 Random Forest

Random forests or random decision-making forests constitute an ensemble learning process in order to classify, regress and carry out other tasks by building many decision-making trees. The random forest output is the class chosen by the majority of trees for classification tasks. The mean or average forecast of the individual trees is returned for regression tasks.

$$\sigma = \sqrt{\frac{\sum_{b=1}^B (f_b(x') - \hat{f})^2}{B-1}} \dots \dots (viii)$$

In the equation (viii), the sample number, B, is a parameter free of charge. Usual use depending upon the size and nature of the workshop are some hundreds to thousands of trees. Cross validation or observation of the out-of-bag error can help to find an optimum number of B trees: the average prediction error for each sample of training X' , which is just the trees with no X' in their bootstrap sample. After some trees are fit, the training and test errors begin to deteriorate.

3.5.10 Adaptive Boosting (AdaBoost)

Adaptive classification boosting is a group learning approach or simply known as AdaBoost classification. It uses a group of weak classifiers to build a strong classifier. A poor classifier learns from its previous classifier mistakes in this method. Consider the n-sample dataset. Each sample is initially allocated $1/n$ weight. And a weak classifier is built with this dataset. This classifier calculates the overall ε error. And the impact of such a classificatory α is measured by this total inaccuracy in the classification of the data samples.

$$\alpha = \frac{1}{2} \ln \left(\frac{1 - \epsilon}{\epsilon} \right) \dots \dots \dots (ix)$$

In the equation (ix), the α is here utilized to change the weight of the dataset samples so that a new dataset is created.

3.5.11 eXtreme Gradient Boosting (XGBoost)

XGBoost is a Machine Learning technique for decision-taking, using a gradient boosting framework. It applies initially a series of classified and regressed (also known as CART) trees as weak learners and then increases trees' performance by generating a set of trees that reduce a regularized target function to a minimum. The algorithm included principles like split-wisdom finding in each tree, cache friendly approximate algorithms for determining divisions, and effective out-of-core calculation of gradient boosting methods for creation of an algorithm at high computer speed with a good capacity to forecast.

For data set $D = \{(x_i, y_i)\} (x_i \in R^m, y_i \in R, i = 1, 2, \dots, n)$ containing n m dimensions, the XGBoost model can be expressed as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in F (i = 1, 2, \dots, n)$$

Where in the CART decision tree structure set, $F = \{f(x) = w_{q(x)}\} (q: R^m \rightarrow \{1, 2, \dots, T\}, w \in R^T)$ represents the set of CART decision tree structures, q represents the tree structure of the sample map to the leaf nodes, T represents the number of leaf nodes, and w represents the actual score of the leaf nodes.

3.6 Proposed System

After consider each and every algorithms that described above now the desired system can be proposed. A system diagram is more suitable which helps to understand the exact procedure of the system which is shown in Figure 3.1.

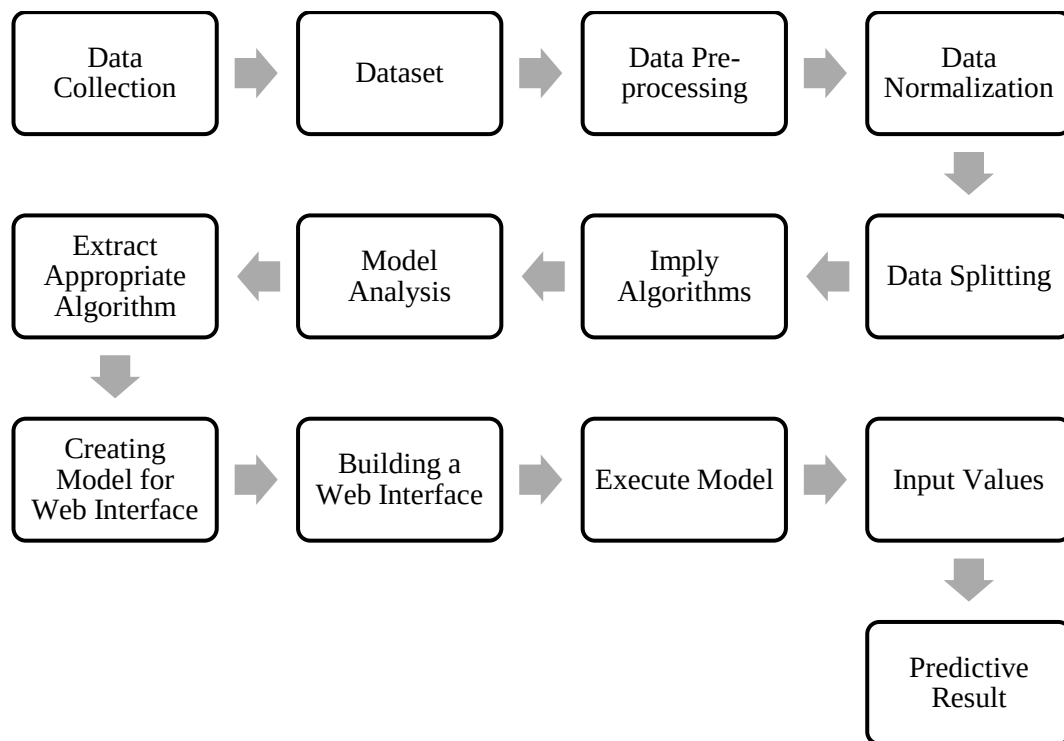


Figure 3.1: Proposed Method to Predict Chronic Kidney Disease

3.6.1 Data Collection

To analyze CKD, the system needed real-life data that must be required. For this study, the data was collected from the University of California Irvine (UCI), different Hospitals and Kidney Foundation in Bangladesh. Collecting all the necessary data from various hospitals is just the first part of this study. Next, all the data was converted to a single CSV file for analysis and understanding.

3.6.2 Dataset

Merging all the data set into a single CSV file, the number of rows became perfect for implementing different Machine Learning Algorithms. Machine learning algorithms need a vast number of data to predict something. The raw dataset contains 10321 rows and 25 columns with the presence of some missing values.

3.6.3 Data Pre-processing

The dataset contains a lot of missing values and qualitative data that was required to be converted. First of all, the qualitative data was converted into quantitative data. Then comes dealing with the missing values. All the missing values were replaced with the mean value. It was divided into dependent variables and independent variables X and Y.

3.6.4 Data Normalization

Normalization is converting numeric columns to a standard scale while preserving ranges of values. The independent variable (X) was then normalized for better accuracy.

3.6.5 Data Splitting

Before applying any machine learning algorithm, the data set needs to be divided into two parts: train and test. 20% data was taken for testing the model and 80% for training. Using the Training part of the dataset, the specific model can be trained to predict something, and the Testing part can be used to see how accurately the data is being predicted.

3.6.6 Imple Algorithms

Eleven different algorithms were implemented to find the best accuracy and select the best algorithm. The names of the 11 algorithms are: Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Decision Tree, Random forests or random decision-making forest, Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost),

Perceptron Algorithm, Gaussian Naive Bayes, K-nearest Neighbors (KNN), Logistic Regression, Linear Support Vector Classification (Linear SVC). Using all these algorithms, different analytical results were found.

3.6.7 Model Analysis

Measuring the Confusion Matrix, Accuracy Score, Jaccard Score, Cross Validated Score, AUC Score, Misclassification, Mean Absolute Error and Mean Squared Error for all the algorithms, the data was converted into tables. The confusion matrix gives a summary of how accurately the data is being predicted. The accuracy score, Jaccard Score, Cross Validated Score, AUC Score provide a percentage of accuracy of the predicted data. Misclassification, Mean Absolute Error and Mean Squared Error gives an error rate of the algorithms.

3.6.8 Extract Appropriate Algorithm

The best algorithm was extracted by measuring and observing all the necessary results from the tables. The extracted algorithm contains the best accuracy rate and minimum error rate in that specific dataset.

To make proper use of the dataset, first and foremost, an appropriate algorithm must be developed. In this case, it is preferable to utilize many algorithms as models and then choose the most appropriate one. Various analytical criteria, such as accuracy score, Jaccard score, cross validation score, AUC score, and so on, were utilized to determine the most effective method in this study. The XGBoost Classifier is the best method that has been shown to be appropriate for the CKD dataset in this study. It received the highest marks in all of the criteria listed above. Following the selection of an algorithm, the process proceeds to the next stage.

3.6.9 Creating Model for Web Interface

After extracting the best algorithm, the authors have built a web-based interface. To connect the interface with the best algorithm, a “pickle” was used. Here Serializing objects in python is accomplished through the use of the Pickle library. The pickling technique can

be used to serialize machine learning algorithms and store them in a file. In this case, the authors serialized the selected model into a “.sav” file or model. To create a model first of all an object of that particular algorithm is needed. Then the trained dataset is used to train the model using .fit() function of that model. After training the model with the appropriate algorithm, the model is ready to be used. The preceding step saves the model to a file and then loads it as a new object named pickled_model. The loaded model is then used to compute the accuracy score and make predictions on previously unknown (test) data.

3.6.10 Building a Web Interface

To build a web interface, the authors took the help of the “flask” module in python. Basic HTML and CSS were also needed to create a user-friendly interface. In the backend of the website, the pickle file was connected with the website.

3.6.11 Execute Model

After creating the model, the model is needed to be saved in a folder. Here comes the use of pickle. Pickle is a Python module for serializing and decoding Python object structures. As a result, Python objects are converted to byte streams, which may be stored in files or databases, used to preserve program state between sessions, and even used to transmit data over the internet. Then the model is needed to be dump in python using pickle. For that pickle.dump() is used. Inside the function the model’s name is denoted and then a file is opened in write binary mode and a location is given. This is how the model is created and opened in a folder.

3.6.12 Input Values

When the model is created, flask can be used to make an interface to predict the CKD easily. Flask is a Python programming interface that enables us to create web-based applications. In addition to being clearer than Django's framework, Flask's framework is also less difficult to understand since it has less basic code to build a simple web-Application.

3.6.13 Predictive Result

The interface helps any user to insert necessary data into the website to predict Chronic Kidney Disease. If a patient had a CKD, the website will show a positive result and suggest meeting a doctor. And if the patient has an expected outcome, the website will display a negative result and tells him that everything is alright.

3.7 System Architecture

To imply the entire project system, a system architecture formation is badly needed to simplify the actual formation in machine learning method and web implementation. In the Figure 3.2, a basic system architecture is given that is a larger view of proposed system.

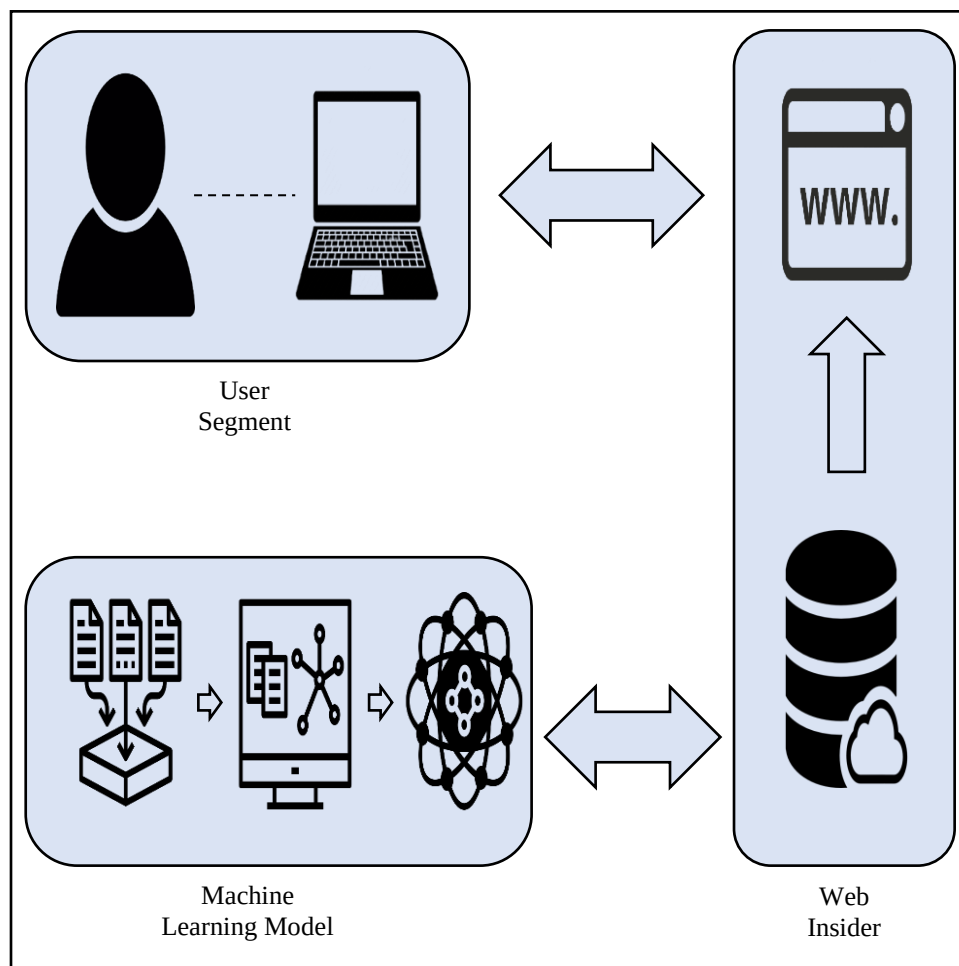


Figure 3.2: System Architecture

3.7.1 User Segment

Here in user segment, it shows how a user uses machine to see his or her kidney condition. Here a user will enter the attribute values of diagnosis report using a device like laptop. Then the system will give the user a basic condition of his or her kidney. The user segment has been designed very simply. The interface is very user friendly so, anyone can use the user segment. Just entering the value and pressing the submit value is that is needed and the machine will predict the current condition.

3.7.2 Web Insider

To build a web interface, the authors took the help of the “flask” module in python. Basic HTML and CSS were also needed to create a user-friendly interface. In the backend of the website, the pickle file was connected with the website. The interface helps any user to insert necessary data into the website to predict Chronic Kidney Disease. If a patient had any Kidney Disease, the website will show a positive result and suggest meeting a doctor. And if the patient has an expected positive outcome, the website will display a negative result and tells him that everything is alright.

3.7.3 Machine Learning Model

The best algorithm was extracted by measuring and observing all the necessary results from the tables. The extracted algorithm contains the best accuracy rate and minimum error rate in that specific dataset. After extracting the best algorithm, it was necessary to build a web-based interface. To connect the interface with the best algorithm, a “pickle” was used. Here serializing objects in python is accomplished through the use of the Pickle library. The pickling technique can be used to serialize machine learning algorithms and store them in a file. In this case, it was serialized the selected model into a “.sav” file or model.

CHAPTER 4

EXPERIMENTAL RESULTS & DISCUSSION

4.1 Introduction

For any kind of research and project the result is very important. As the result contains the outcome of any project. In this chapter all result has been given shown with table. The details of the data acquisition, data utilization, feature importance has been shown in this chapter for understanding the dataset for the CKD. Then the confusion matrix table has been generated to show result for different algorithms. With the help of a classification report the precision, recall, F1-Score also has been shown in a table. The accuracy, jaccard score, cross validated score, AUC score, ROC curve has been showed in different graphs. The details are also has been given in a table for better understanding. The standard deviation also given in a form of a table. Finally, the misclassification & error was also shown in a form of a table. Since this project has a web implementation, the output display was given in this chapter as screenshot.

4.2 Experimental Results

After successful implementation of Machine Learning Model Creation, each algorithms showed their own accuracy and scores that is needed to extract the best algorithm to predict Chronic Kidney Disease. Therefore, Experimentation results is an analytical segment, where each and every possible scores for all algorithmic application and procedure can be examine.

4.2.1 Data Acquisition

The dataset that was used to train the model is collected from University of California Irvine (UCI), different Hospitals and Kidney Foundation in Bangladesh. About 10321 samples of data were used to train the model. Each sample in this CKD dataset includes 24 significant factors or features (13 categorical (nominal) variables and 11 numerical variables) plus a target variable (class). Each class has two nominal values; CKD (sample

with CKD) and notCKD (sample without CKD). The data has a significant number of missing values. The Table 4.1 gives a short description of the dataset.

Table 4.1: Data Acquisition & Null Percentage

Attribute	Scale	Data Type	Missing Values (%)
Age	age in years	Numerical	2.27
Blood Pressure	in mm/Hg	Numerical	0
Specific Gravity	(1.005,1.010,1.015,1.020,1.025)	Nominal	0
Albumin	(0,1,2,3,4,5)	Nominal	0
Sugar	(0,1,2,3,4,5)	Nominal	0
Red Blood Cells	(normal, abnormal)	Nominal	0
Pus Cell	(normal, abnormal)	Nominal	16.2
Pus Cell Clumps	(present, notpresent)	Nominal	0.97
Bacteria	(present, notpresent)	Nominal	0.97
Blood Glucose Random	in mgs/dl	Numerical	11.06
Blood Urea	in mgs/dl	Numerical	0
Serum Creatinine	in mgs/dl	Numerical	0
Sodium	in mEq/L	Numerical	0.72
Potassium	in mEq/L	Numerical	0
Hemoglobin	in gms	Numerical	0
Packed Cell Volume	-	Numerical	17.88
White Blood Cell Count	in cells/cumm	Numerical	0
Red Blood Cell Count	in millions/cmm	Numerical	0
Hypertension	(yes, no)	Nominal	0
Diabetes Mellitus	(yes, no)	Nominal	0
Coronary Artery Disease	(yes, no)	Nominal	0.5
Appetite	(good, poor)	Nominal	0.25
Pedal Edema	(yes, no)	Nominal	0.25
Anemia	(yes, no)	Nominal	0.25
Class	(ckd, notckd)	Nominal	0

4.2.2 Data Utilization

A computer system's data was made simpler to manage by coding each category (nominal) variable separately. When rbc and pc had normal or abnormal values, they were recorded as 1 and 0, respectively. It was decided that the presence or absence of pcc and ba would be classified as 1 or 0, respectively. As a result, the yes/no answers were coded as 1 and 0, with 1 being yes, and 0 denoting no. To calculate appet's value, good appet was assigned a value of 1 and poor appet a value of 0. It is important to note, however, that although though the three variables sg, al, and su were originally described as categorical types, their values were determined by their relationship to numbers. All of the category

variables were converted using a factorization method. Each sample was given a random number between 1 and 10321 in order to differentiate them. In this dataset, there were a lot of blank spaces. Patients may overlook important measures before a diagnosis is made for a variety of reasons. As a consequence, missing values will show in the data if the diagnostic categories of the samples are not known and an appropriate imputation technique is required. The missing values of the main CKD dataset were initially processed and filled after the categorical variables had been encoded. Then the data description for the 25 attributes were extracted for a better understanding of the dataset. The Count, Mean, Std, Min, 25%, 50%, 75%, Max was extracted from the dataset and shown in the Table 4.2.

Table 4.2: Dataset Description

	Count	Mean	Std	Min	25%	50%	75%	Max
Age	10321	51.51	16.95	2	42	54	64	90
Blood Pressure	10321	79.62	70.39	0	70	76	80	1400
Specific Gravity	10321	1.02	0.01	1.01	1.02	1.02	1.02	1.03
Albumin	10321	1.02	1.27	0	0	1	2	5
Sugar	10321	0.40	1.03	0	0	0	0	5
Red Blood Cells	10321	0.88	0.32	0	1	1	1	1
Pus Cell	10321	0.76	0.39	0	0.76	1	1	1
Pus Cell Clumps	10321	0.12	0.32	0	0	0	0	1
Bacteria	10321	0.06	0.23	0	0	0	0	1
Blood Glucose Random	10321	148.40	74.87	22	101	127	150	490
Blood Urea	10321	57.73	49.63	1.5	27	44	64	391
Serum Creatinine	10321	3.04	5.31	0.4	0.9	1.4	3.07	76
Sodium	10321	144.03	87.07	104	135	137.53	141	1436
Potassium	10321	4.43	0.73	1.4	3.9	4.63	4.8	7.6
Hemoglobin	10321	12.46	2.83	3.1	10.8	12.53	14.6	17.8
Packed Cell Volume	10321	38.75	8.09	9	34	38.75	44	54
White Blood Cell Count	10321	8403.41	2534.28	2200	7000	8406	9400	26400
Red Blood Cell Count	10321	4.85	2.81	2.1	4.5	4.71	5.1	58
Hypertension	10321	0.37	0.48	0	0	0	1	1
Diabetes Mellitus	10321	0.35	0.48	0	0	0	1	1
Coronary Artery Disease	10321	0.09	0.28	0	0	0	0	1
Appetite	10321	0.79	0.40	0	1	1	1	1
Pedal Edema	10321	0.19	0.39	0	0	0	0	1
Anemia	10321	0.15	0.36	0	0	0	0	1
Class	10321	0.62	0.48	0	0	1	1	1

4.2.3 Feature Importance

The term "feature importance" refers to methods that give a value to input characteristics based on their predictive power for a target variable. Feature importance is a term that refers to a group of methods for giving scores to the input features of a predictive model, indicating the relative significance of each feature when generating a prediction. The score of feature importance can provide insight into the dataset, the model, and can be used to improve a predictive model. From the Table 4.3, the feature importance of different attributes for different algorithms is shown. In the Table 4.3, it is clear that Hemoglobin has the highest value. So, the feature importance of Hemoglobin is much important than other attributes.

Table 4.3: Feature Importance

Attribute	Algorithms				
	Logistic Regression	Random Forest	AdaBoost Classifier	Decision Tree	XGBoost Classifier
Hemoglobin	-3.32391	0.4042	0.36	0.7749	0.36743
Red Blood Cell Count	-4.10634	0.26318	0.27	0.07877	0.07197
Hypertension	2.98574	0.12988	0.01	0.07549	0.38551
White Blood Cell Count	0.05221	0.07531	0.28	0.03789	0.05247
Age	0.07557	0.01393	0.01	0.00529	0.00524
Packed Cell Volume	-0.08325	0.01401	0	0.00494	0.00684
Potassium	0.02034	0.01041	0	0.00482	0.00671
Blood Urea	-0.07015	0.01284	0.01	0.00445	0.00649
Sodium	-0.00714	0.0105	0.01	0.00334	0.01034
Blood Glucose Random	-0.04264	0.01468	0.02	0.00316	0.00663
Blood Pressure	0.04297	0.00714	0	0.00177	0.00507
Albumin	0.03239	0.00416	0	0.00133	0.00522
Serum Creatinine	-0.01995	0.01145	0.02	0.00079	0.00713
Sugar	0.00864	0.00292	0	0.00071	0.01002
Bacteria	0.03948	0.00152	0	0.00047	0
Red Blood Cells	-0.02106	0.00104	0	0.00044	0.00626
Pus Cell Clumps	-0.00889	0.00194	0	0.00041	0.00907
Pus Cell	0.0285	0.00369	0	0.00034	0.00649
Anemia	0.0338	0.00221	0	0.00034	0.0043
Appetite	-0.0092	0.00182	0	0.00033	0.00483
Specific Gravity	-0.01694	0.00528	0	0	0.00804
Diabetes Mellitus	0.15624	0.00402	0.01	0	0.01044

Coronary Artery Disease	0.07053	0.00159	0	0	0
Pedal Edema	-0.06433	0.00229	0	0	0.00352

4.3 Result & Discussion

The value of CKD was made positive whereas the value of notckd was created negative in this research. The confusion matrix has been used to show specific results and assess the effectiveness of machine learning algorithms. Table-4 shows the template for the confusion matrix for various algorithm types.

4.3.1 Confusion Matrix

It is important to generate confusion matrix to validate result from implementation aspect. When assessing a classification model's performance, a Confusion matrix is an N x N matrix used with N target classes. The matrix evaluates the accuracy of the machine learning model by comparing the actual target values to the predicted ones. In this way, it is visible how effectively algorithmic model is working and what mistakes it is making. From binary classification, it will be easier to identify the Precision, Recall, and Accuracy with some of mathematical equation. Furthermore, for multi-class classification it is must to have to go through these average value with micro average or macro average. Before discussing about these, it is important to know about four building blocks that are used in computing various evaluation measures. And these are True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Table 4.4 shows the exact formation of Confusion Matrix containing True Positive, False Positive, False Negative and True Negative. Table 4.5 contains the confusion matrix for each algorithms.

Table 4.4: Confusion Matrix

Confusion Matrix		
Predicted Class	Actual Class	
	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

Table 4.5: Confusion Matrix for Algorithms

Algorithm	Confusion Matrix			Confusion Matrix Percentage	
		CKD	Not CKD	CKD	Not CKD
SVM	CKD	1228	65	59%	3%
	Not CKD	70	702	3%	34%
KNN	CKD	1153	140	56%	7%
	Not CKD	95	677	5%	33%
Logistic	CKD	1237	56	60%	3%
	Not CKD	67	705	3%	34%
Random Forest	CKD	1266	27	61%	1%
	Not CKD	28	744	1%	36%
Naive Bayes	CKD	1051	242	51%	12%
	Not CKD	47	725	2%	35%
Perceptron	CKD	1184	109	57%	5%
	Not CKD	117	655	6%	32%
SGD	CKD	1236	57	60%	3%
	Not CKD	52	720	3%	35%
Linear SVC	CKD	1238	55	60%	3%
	Not CKD	75	697	4%	34%
Decision Tree	CKD	1273	20	62%	1%
	Not CKD	17	755	1%	37%
AdaBoost	CKD	1286	7	62%	0%
	Not CKD	32	740	2%	36%
XGBoost	CKD	1287	6	62%	0%
	Not CKD	24	748	1%	36%

A. True Positive (TP)

Those that were properly classified by the classifier are referred to as positive tuples. It is indicated by the letter TP in the acronym. Here in XGBoost, AdaBoost, Decision Tree 62% of the value were True Positive (TP) value. Followed by Random Forest and Logistic Regression with 61% and 60%.

B. True Negative (TN)

Negative tuples are positive tuples that the classifier incorrectly categorized. These occurrences may be denoted by the letter TN. Here in Decision Tree the value is 37% followed by XGBoost, AdaBoost and Random Forest with 36% True Negative (TN) values. And Logistic regression with 34%.

C. False Positive (FP)

It is these negatively labeled tuples that have been erroneously classified as positive by the classifier that is interested in today. FP may be used to indicate this type of relationship. In this part of the analysis the Random Forest, Decision Tree and XGBoos produced less False Positive (FP) values with only 1%. Followed by Naive Bayes and AdaBoost with 2% and Logistic Regression generated 3% of the False Positive (FP) values.

D. False Negative (FN)

The classifier misclassified these positive tuples as negative. It is denoted by the letter FN. The False Negative (FN) values the XGBoost and AdaBoost had 0% outcome. Then comes Decision Tree and Random Forest with 1% and Logistic Regression with 3% False Negative (FN) values.

E. Precision

Precision may be seen as a metric for determining how accurate something is (i.e. what percentage of tuples labeled as positive are actually such). In other words, it's the percentage of retrieved occurrences that are really relevant. Equation (x) shows the mathematical formula to measure Precision.

$$Precision = \frac{TP}{TP + FP} \dots \dots \dots (x)$$

F. Recall

It is a measure of completeness (how many positive tuples are identified as such) that is used in machine learning. A relevant instance is defined as the proportion of relevant examples that have been retrieved in comparison to the entire number of relevant instances. Equation (xi) shows the mathematical formula to measure Recall.

$$Recall = Sensitivity = \frac{TP}{TP + FN} \dots \dots \dots (xi)$$

G. F1-Measure

An evaluation of a test's precision and recall is known as the weighted harmonic mean, and this measurement is known as the F measure. Equation (xii) shows the mathematical formula to measure F1-Measure.

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \dots \dots \dots (xii)$$

H. Accuracy

Accuracy of a classifier on a given test set is the proportion of test set tuples that are properly categorized by the classifier on that test set. From equation (xiii), it can be clearer to measure accuracy.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \dots \dots \dots (xiii)$$

4.3.2 Classification Report

A classification report is a statistic used in machine learning for evaluating the system's performance. It's used to demonstrate trained classification model's accuracy, recall, F1 Score, and support. It is a classification-based machine learning model's performance statistic. The accuracy, recall, F1 score, and support of model are all shown. It gives a clearer picture of trained model's overall performance. Classification reports from machine learning models need knowledge of all metrics shown in the research study in order to be understood. Table 4.6 shows the classification report for each algorithms which contains Precision, Recall and F1-Score and Accuracy percentage value.

Table 4.6: Classification Report

Algorithm	Class	Precision	Recall	F1-Score	Accuracy (%)
XGBoost	Not CKD	0.99	0.97	0.98	98.55
	CKD	0.98	1	0.99	
	Macro Avg.	0.99	0.98	0.98	
	Weighted Avg.	0.99	0.99	0.99	
Decision Tree	Not CKD	0.97	0.98	0.97	98.11
	CKD	0.99	0.98	0.98	
	Macro Avg.	0.98	0.98	0.98	
	Weighted Avg.	0.98	0.98	0.98	
AdaBoost	Not CKD	0.99	0.96	0.97	98.11

	CKD	0.98	0.99	0.99	
	Macro Avg.	0.98	0.98	0.98	
	Weighted Avg.	0.98	0.98	0.98	
Random Forest	Not CKD	0.96	0.96	0.96	97.09
	CKD	0.98	0.98	0.98	
	Macro Avg.	0.97	0.97	0.97	
	Weighted Avg.	0.97	0.97	0.97	
Logistic	Not CKD	0.93	0.91	0.92	94.04
	CKD	0.95	0.96	0.95	
	Macro Avg.	0.94	0.93	0.94	
	Weighted Avg.	0.94	0.94	0.94	
SGD	Not CKD	0.92	0.92	0.92	93.95
	CKD	0.95	0.95	0.95	
	Macro Avg.	0.94	0.94	0.94	
	Weighted Avg.	0.94	0.94	0.94	
Linear SVC	Not CKD	0.93	0.9	0.91	93.7
	CKD	0.94	0.96	0.95	
	Macro Avg.	0.93	0.93	0.93	
	Weighted Avg.	0.94	0.94	0.94	
SVM	Not CKD	0.92	0.91	0.91	93.46
	CKD	0.95	0.95	0.95	
	Macro Avg.	0.93	0.93	0.93	
	Weighted Avg.	0.93	0.93	0.93	
Perceptron	Not CKD	0.86	0.85	0.85	89.06
	CKD	0.91	0.92	0.91	
	Macro Avg.	0.88	0.88	0.88	
	Weighted Avg.	0.89	0.89	0.89	
KNN	Not CKD	0.83	0.88	0.85	88.62
	CKD	0.92	0.89	0.91	
	Macro Avg.	0.88	0.88	0.88	
	Weighted Avg.	0.89	0.89	0.89	
Naive Bayes	Not CKD	0.75	0.94	0.83	86
	CKD	0.96	0.81	0.88	
	Macro Avg.	0.85	0.88	0.86	
	Weighted Avg.	0.88	0.86	0.86	

4.4 Result Analysis

After calculation of all possible aspects such as Precision, Recall, F1-Measure, Accuracy etc. it is time to go through result analysis. In this analysis, it will be figured out which algorithm is performing best among all algorithms and which algorithm is performing less comparing with others.

4.4.1 Accuracy

The accuracy measures the best performance of an algorithm. How good it performs based on the data that is provided to it. Accuracy can measure the performance using probabilistic method. From these algorithms, best accurate algorithm is eXtreme Gradient Boosting and lower accurate algorithm is Gaussian Naïve Bayes comparing to others. eXtreme Gradient Boosting algorithm have shown that it is very powerful and scalable implementations of gradient boosting machines, and it have shown that they are capable of pushing the limits of computing power for boosted trees algorithms. It was designed and developed solely for the purpose of enhancing the performance of models and increasing the speed at which computers could process data. Figure 4.1 and Table 4.7 shows the accuracy chart and percentage of each algorithms used to predict in this model.

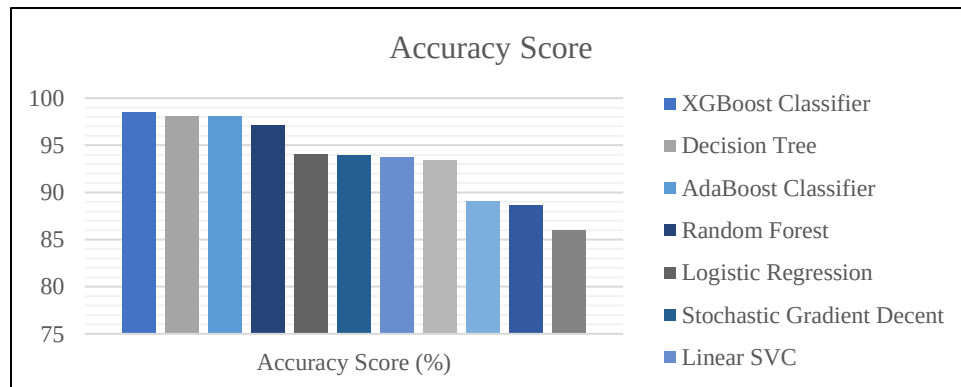


Figure 4.1: Accuracy Chart

4.4.2 Jaccard Score

The Jaccard score is a statistic for calculating sample set similarity and diversity. In terms of the ratio of Intersection to Union, they are equivalent. There is a way to compare the similarity of two finite sample sets mathematically by using the Jaccard coefficient, which is defined as the intersection size divided by the union size. Equation (xiv), Figure

4.2 and Table 4.7 shows the accuracy chart and percentage of each algorithms used to predict in this model.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \dots \dots (xiv)$$

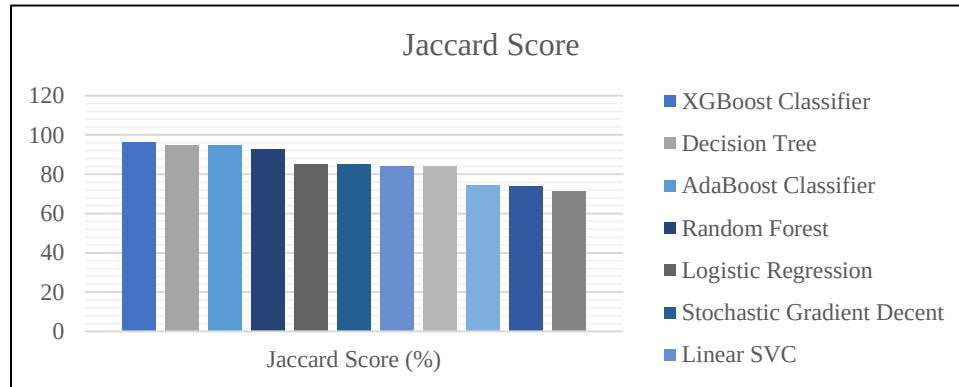


Figure 4.2: Jaccard Score Chart

4.4.3 Cross Validated Score

The statistical method of cross-validation is used to evaluate the skill of machine learning models. Cross Validation begins by shuffling the data and dividing it into k folds. Then k models are fitted to $\frac{k-1}{k}$ of the data and $\frac{1}{k}$ of the data is evaluated. By averaging the outcomes of each evaluation, the final score is calculated, and the resulting model is then fitted to the complete dataset for implementation. Figure 4.3 and Table 4.7 shows the cross validated score chart and percentage of each algorithms used to predict in this model.

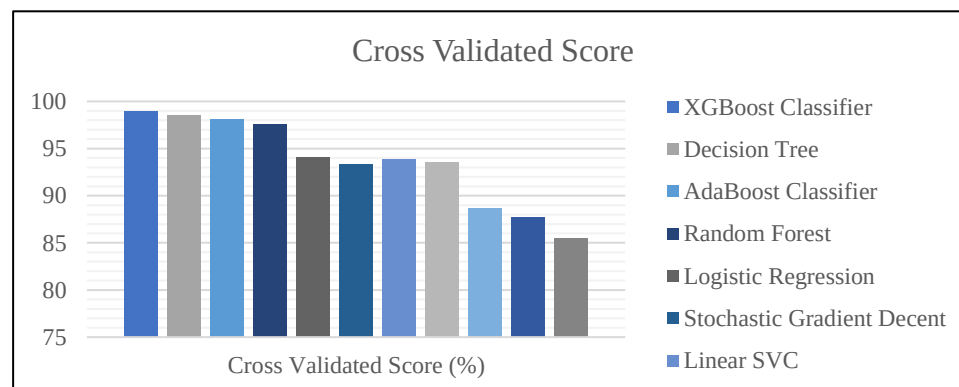


Figure 4.3: Cross Validated Score

4.4.4 AUC Score

Using this performance measure in combination with machine learning applications, a system's prospective categorization levels may be determined, which can be useful in many situations. When comparing a percentile of random positive cases in which the model performs considerably better to a percentile of random negative examples in which the model performs significantly worse, the AUC may be computed. It is conceivable for this number to have four different values, with the greatest potential value being one. The values are in the range of 0 to 1, with 0 being the lowest possible value. As seen in the Figure 4.4 and Table 4.7, models that make 100 percent incorrect predictions have an accuracy of zero, while models that make 100 percent correct predictions have an accuracy of one.

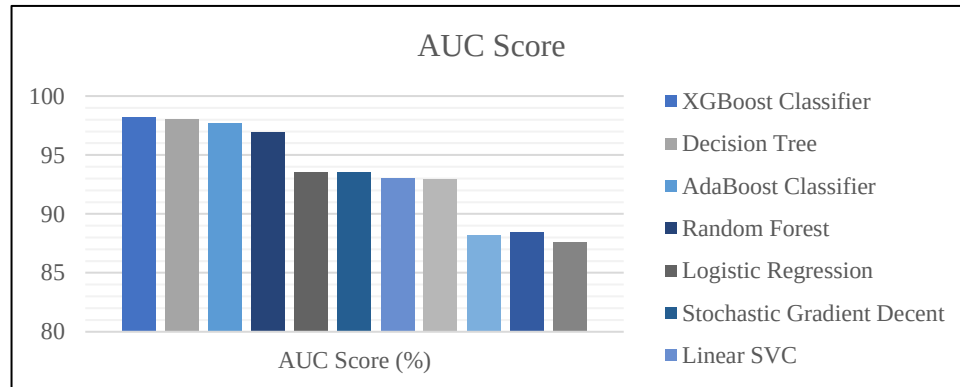


Figure 4.4: AUC Score Chart

4.4.5 ROC Curve

ROC analysis is a very important method for evaluating diagnostic test performance and, more extensively, the accuracy of a statistical model that categorizes people into one of two categories: diseased or non-diseased. Probably one of the best applications of ROC

curve analysis is as a simple graphical tool for illustrating the accuracy of a medical diagnostic test. In the Figure 4.5, this important curve score is shown.

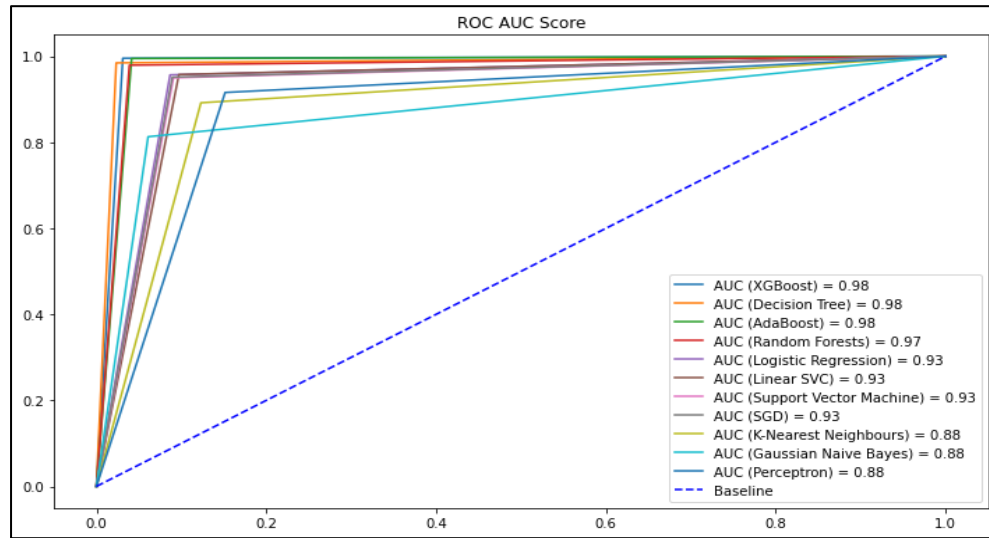


Figure 4.5: ROC Curve

After completing overall analysis, the best accuracy that was achieved from the model is XGBoost Classifier that had an accuracy score of 98.55%, Jaccard Score of 96.14%, Cross Validated Score of 98.97% and AUC Score of 98.21%. The Table 4.7 gives a short description about the accuracy.

Table 4.7: Accuracy, Jaccard, Cross Validated and AUC Score

Algorithm Name	Accuracy Score (%)	Jaccard Score (%)	Cross Validated Score (%)	AUC Score (%)
XGBoost Classifier	98.55	96.14	98.97	98.21
Decision Tree	98.11	95.08	98.54	98
AdaBoost Classifier	98.11	94.99	98.12	97.66
Random Forest	97.09	92.53	97.61	96.92
Logistic Regression	94.04	85.14	94.1	93.5
Stochastic Gradient Decent	93.95	85.01	93.34	93.52
Linear SVC	93.7	84.28	93.81	93.02
Support Vector Machines	93.46	83.87	93.53	92.95
Perceptron	89.06	74.35	88.65	88.21
KNN	88.62	74.23	87.71	88.43
Naive Bayes	86	71.5	85.45	87.6

4.4.6 Standard Deviation

From this study, it can also be computed the standard deviation. A standard deviation quantifies a dataset's dispersion from its mean. The standard deviation is computed by taking the square root of the variance of each data point. The standard deviation increases as the data points move away from the mean. Here the standard deviation for the top five algorithms is shown in the Table 4.8.

Table 4.8: Standard Deviation

Algorithm Name	Standard Deviation
Random Forest	0.09
XGBoost Classifier	0.1
AdaBoost Classifier	0.1
Decision Tree	0.15
Logistic Regression	1.23

4.4.7 Misclassification & Error

While finding the accuracy in any algorithm error becomes a problem. After misclassification means absolute error and mean square error becomes a part of the accuracy of a machine-learning model. Misclassification can occur when an unsuitable attribute is chosen. Misclassification occurs when all classes, groups, or categories of a variable have the same error rate.

The amount of inaccuracy in a measurement is referred to as the absolute error. Mean Absolute Error (MAE) is the average of all absolute errors in measurement. Equation (xv) represents the formula for Mean Absolute Error.

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - x| \dots \dots \dots (xv)$$

How close a regression line is to a set of points is indicated by the Mean Squared Error (MSE). Equation (xvi) represents the formulized calculation for Mean Squared Error.

$$MSE = \frac{1}{n} \sum_{i=1}^n |y_i - y|^2 \dots \dots \dots (xvi)$$

The Table 4.9 shows the misclassification, mean absolute error and mean square error in the algorithms. The Error rate for the XGBoost was also less for Misclassification with 1.45%, Mean Absolute Error with 1.45% and Mean Squared Error 1.45%.

Table 4.9: Misclassifications & Errors

Algorithm Name	Misclassification (%)	Mean Absolute Error (%)	Mean Squared Error (%)
XGBoost Classifier	1.45	1.45	1.45
Decision Tree	1.89	1.89	1.89
AdaBoost Classifier	1.89	1.89	1.89
Random Forest	2.91	2.91	2.91
Logistic Regression	5.96	5.96	5.96
Stochastic Gradient Decent	6.05	6.05	6.05
Linear SVC	6.3	6.3	6.3
Support Vector Machines	6.54	6.54	6.54
Perceptron	10.94	10.94	10.94
KNN	11.38	11.38	11.38
Naive Bayes	14	14	14

4.5 Web Implementation

The machine learning model development implementation has been completed successfully, and the appropriate technique has been identified and implemented. It is now time to show this model implementation on the Web Interface, as described in the System Architecture section of the previous chapter.

4.5.1 Web Interface

Chronic Kidney Disease will be predicted on a website using a technique known as "Flask," as previously described in Chapter 3. A web interface based on the "Flask" approach will be created, as previously stated in Chapter 3. To ensure that this job is completed successfully, a highly effective and functional website has been created for this purpose. That Web Interface is shown in the Figure 4.6.

Chronic Kidney Disease

Age*	Blood Pressure*	Specific Gravity*	Albumin*
Sugar*	Red Blood Cells*	Pus Cells*	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/swell*	Anemia*

[Submit](#)

Figure 4.6: Web Interface

4.5.2 Web Output Analysis

This research method and collected data only allow for two classification options: CKD and NotCKD, which stand for Chronic Kidney Disease and Not Chronic Kidney Disease, respectively. Two independent output analyses will be required to complete this section in its entirety. In the Figure 4.7, all data has been input randomly and got the required output as previously trained model to predict Chronic Kidney Disease. As well as, in the Figure 4.8, all data has been input randomly from test case and got the required output as previously trained model to predict Chronic Kidney Disease. After completion of these process, it can be clearly said that this applied model can predict Chronic Kidney Disease successfully as trained.



Figure 4.7: Web Interface with Negative Output

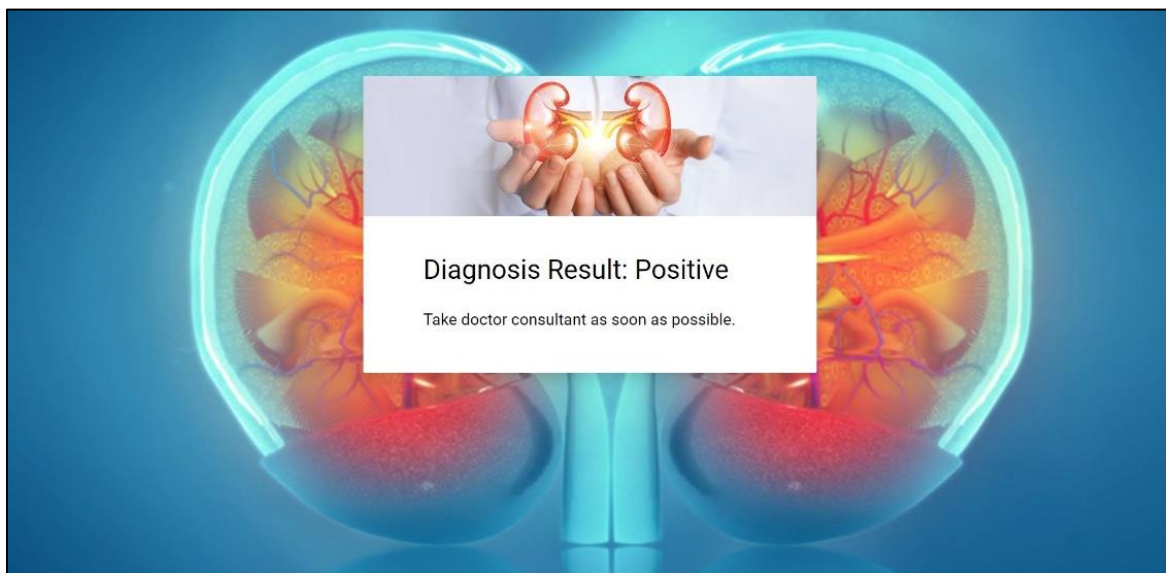


Figure 4.8: Web Interface with Positive Output

CHAPTER 5

IMPACT ON SOCIETY & SUSTAINABILITY

5.1 Introduction

If a project has been made, the impact of that project in the society need to be discussed and analyzed. Here in this chapter the impact of the Chronic Kidney Disease project has been discussed in three parts, in the impact on society is discussed how this project will impact the society in a positive way. Then comes ethical aspects. Here the ethical aspect was discussed properly for understanding how this project will help the patients. And in the end the project sustainability was discussed. There how this project can grow in the future and help more people is discussed.

5.2 Impact on Society

This research has a significant social impact. Because people nowadays all have busy life, going to the hospital to ensure that people are not infected with any sickness takes a long time. An interface has been created where people can enter the data they need to predict Kidney Disease and get a quick result. It is risky to go to the hospital for tests now as people all over the world are in a risk COVID-19. It would be quite helpful for us if anyone could analysis their report at home. A survey on patients and doctors is ongoing about this research work and its effectiveness. And it is quite sure that that survey report will show a positive outcome on the impact on society and patients-doctors as well.

5.3 Ethical Aspects

People don't need to go to any diagnosis hospital for any kind of test. They can get a basic idea of Chronic Kidney Disease at home. They can predict by themselves. As a web interface has been built, all data will be saved in the database and the model will be stronger day by day. As it is a machine learning based project, that's why it is quite in doubt to be fully dependent on this model right now. A machine can't predict anything accurately. It takes time. In future when the amount of data in the database will be million, the model will be stronger and it can be hoped that time it will predict accurately. In future if it is

possible to connect Artificial Intelligence and Internet of Things with database, then it will be unnecessary to go diagnosis hospital. People will be able to predict Chronic Kidney Disease at home and then they will consult a doctor on their status. And if it is possible to invent such type of machine which can test blood and urine and will be portable, then people will be able to test CKD at home more precisely

5.4 Sustainability

This research has a great sustainability because of its unique approach since one can use a website to check his or her Chronic Kidney Disease condition. In this current condition the website only uses machine learning techniques to predict Kidney Disease. But in future there are many scopes with deep leaning, Artificial Intelligence and Internet of Things that can furnish this project with more accurate results. Therefore, in case of sustainability in the future many works can be done using this project. Even a mobile app can be created. Here in this paper only Chronic Kidney Disease is predicted. But using this machine learning project and the web approach other disease about Kidney and other organs can also be predicted. Even using a web interface uploading an image of skin, skin disease can also be detected in the future.

CHAPTER 6

FUTURE SCOPE & CONCLUSION

6.1 Introduction

In this chapter the future scope was discussed. What approaches may be used to make a better machine in the future and how this project can aid in the efficient growth of the company in the future. This chapter has come to a neat and tidy conclusion, which has been provided at the end of this chapter. Finally, at the end of this chapter, a list of the references is given.

6.2 Implication for Further Study

Regardless of how difficult it may be, the concept may be used as the basis for a website for any Kidney Disease Hospital, regardless of location. For this project, the authors have developed an interface that is straightforward and straightforward to use. It is conceivable that in the future, a website based on the Internet of Things will be developed and made available to the public through the internet. The simple act of anticipating Chronic Kidney Disease at home can be accomplished by entering all of the necessary information into the website that is needed to identify CKD. The individual will then be able to decide whether or not they have CKD. Because it is an Internet of Things-based website, and because users will enter new data each time, the website will store the information. In the next steps, the model will learn from each new piece of data, enabling it to grow more accurate over time. A Deep Learning approach with Neural Network algorithms it can be more effective by time and Artificial Intelligence can be involved as well.

6.3 Recommendations

For any diagnostic part there is a normal level and according to the normal level if any abnormal stage found, this system can aware of that particular condition. Here in Chronic Kidney Disease the main attributes that takes a vest action on predicting the Chronic Kidney Disease are Serum Creatinine, Creatinine Clearance, Urine Albumin, Hemoglobin

(Hgb) etc. For Serum Creatinine the normal range is 0.6–1.1 mg/dL in women and 0.7–1.3 mg/dL in men. For Creatinine Clearance normal range is 88–128 mL/min for healthy women and 97–137 mL/min for healthy men. A rise in the amount of protein in the urine may indicate a kidney issue. It's critical to know whether or not your urine contains protein. A normal level is from 0 to about 8 mg/dL. A rise in the amount of protein in the urine may indicate a kidney issue. It's critical to know whether or not your urine contains protein. For healthy women, the normal Hgb level is 12–16 g/dL, while for healthy men, the normal Hgb level is 14–18 g/dL. If the ranges change and vary in some cases then it can come to an indication of chronic kidney disease early stage. Therefore, some precautions can be taken in the early stage. Like change food habit, exercise, drinking more water etc. Taking doctor's recommendation is mandatory at that time.

6.4 Conclusion

Assuming chronic kidney disease (CKD) is detected and treated early enough, it has the potential to be successfully and efficiently managed, with the patient perhaps being cured within a short period of time. There are a variety of laboratory tests that people must go through to identify whether or not they have chronic kidney disease (CKD). In the actual world, it is possible that these exams will take a considerable amount of time. According to the researchers that conducted this study, a model trained on a relevant dataset may predict the development of chronic kidney disease at any stage. An interface that the authors have developed for this project allows users to get their CKD report by simply entering the necessary information into a form that they have provided. In order to find the best suitable model for the dataset, eleven machine learning algorithms are trained on the dataset and evaluated based on their accuracy, Jaccard score, Cross Validated score, and AUC score. There are three different scores computed: the accuracy score, the Jaccard score, and the Cross Validated score. XGBoost outperforms the competition in terms of performance when compared to the other classifiers. One hundred and eighty-seven percent of the time, it is the most appropriate option to choose. It will be made accessible to any Kidney Disease Hospital after a website for this project has been completed. People will not have to leave their homes in order to receive CKD reports since they will be accessible online and available 24 hours a day, seven days a week.

REFERENCES

- [1] Kidney Disease in Bangladesh. (n.d.). World Life Expectancy. Retrieved September 19, 2021, from <https://www.worldlifeexpectancy.com/bangladesh-kidney-disease>
- [2] COVID-19 and Your Health. (2020, February 11). Centers for Disease Control and Prevention. <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html>
- [3] Ma, F., Sun, T., Liu, L., & Jing, H. (2020). Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. *Future Generation Computer Systems*, 111, 17-26.
- [4] Ali, S. I., Bilal, H. S. M., Hussain, M., Hussain, J., Satti, F. A., Hussain, M., & Lee, S. (2020). Ensemble feature ranking for cost-based non-overlapping groups: A case study of chronic kidney disease diagnosis in developing countries. *IEEE Access*, 8, 215623-215648.
- [5] Balakrishnan, S. (2020). Feature Selection Using Improved Teaching Learning Based Algorithm on Chronic Kidney Disease Dataset. *Procedia Computer Science*, 171, 1660-1669.
- [6] Lambert, J. R., Arulanthu, P., & Perumal, E. (2020, July). Identification of Nominal Attributes for Intelligent Classification of Chronic Kidney Disease using Optimization Algorithm. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 0119-0125). IEEE.
- [7] Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, 122, 56-69.
- [8] Segal, Z., Kalifa, D., Radinsky, K., Ehrenberg, B., Elad, G., Maor, G., & Koren, G. (2020). Machine learning algorithm for early detection of end-stage renal disease. *BMC nephrology*, 21(1), 1-10.
- [9] Sealfon, R. S., Mariani, L. H., Kretzler, M., & Troyanskaya, O. G. (2020). Machine learning, the kidney, and genotype–phenotype analysis. *Kidney international*, 97(6), 1141-1149.
- [10] Luo, J., & Hendryx, M. (2020). Metal mixtures and kidney function: An application of machine learning to NHANES data. *Environmental Research*, 191, 110126.
- [11] Sambyal, N., Saini, P., & Syal, R. (2020). Microvascular Complications in Type-2 Diabetes: A Review of Statistical Techniques and Machine Learning Models. *Wireless Personal Communications*, 115(1), 1-26.
- [12] Harimoorthy, K., & Thangavelu, M. (2021). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *Journal of Ambient Intelligence and Humanized Computing*, 12(3), 3715-3723.
- [13] Chen, G., Ding, C., Li, Y., Hu, X., Li, X., Ren, L., & Xue, W. (2020). Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform. *IEEE Access*, 8, 100497-100508.
- [14] Reddy, D. J., Mounika, B., Sindhu, S., Reddy, T. P., Reddy, N. S., Sri, G. J., & Kora, P. (2020). Predictive machine learning model for early detection and analysis of diabetes. *Materials Today: Proceedings*.

- [15]** Fauvel, C., Raitière, O., Belkacem, N. S., Dominique, S., Artaud-Macari, E., Viacroze, C., & Bauer, F. (2020). Prognostic importance of Kidney, Heart and Interstitial lung diseases (KHI triad) in PH: A machine learning study. *Archives of Cardiovascular Diseases*, 113(10), 630-641.
- [16]** Yashfi, S. Y., Islam, M. A., Sakib, N., Islam, T., Shahbaaz, M., & Pantho, S. S. (2020, July). Risk Prediction of Chronic Kidney Disease Using Machine Learning Algorithms. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.
- [17]** Navaneeth, B., & Suchetha, M. (2020). A dynamic pooling based convolutional neural network approach to detect chronic kidney disease. *Biomedical Signal Processing and Control*, 62, 102068.
- [18]** Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access*, 8, 20991-21002.
- [19]** Abdelaziz, A., Salama, A. S., Riad, A. M., & Mahmoud, A. N. (2019). A machine learning model for predicting of chronic kidney disease based internet of things and cloud computing in smart cities. In *Security in Smart Cities: Models, Applications, and Challenges* (pp. 93-114). Springer, Cham.
- [20]** Snegha, J., Tharani, V., Preetha, S. D., Charanya, R., & Bhavani, S. (2020, February). Chronic Kidney Disease Prediction Using Data Mining. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-5). IEEE.
- [21]** Ekanayake, I. U., & Herath, D. (2020, July). Chronic Kidney Disease Prediction Using Machine Learning Methods. In 2020 Moratuwa Engineering Research Conference (MERCon) (pp. 260-265). IEEE.

Chronic Kidney Disease An Analytical Prediction Using Machine Learning on the Patients of Different Age

ORIGINALITY REPORT

15%	10%	10%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ijesi.org Internet Source	2%
2	doctorpenguin.com Internet Source	1%
3	www.ripublication.com Internet Source	1%
4	Jiongming Qin, Lin Chen, Yuhua Liu, Chuanjun Liu, Changhao Feng, Bin Chen. "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease", IEEE Access, 2020 Publication	1%
5	lifeoptions.org Internet Source	1%
6	www.scribd.com Internet Source	1%
7	Manonmani. M, Sarojini Balakrishnan. "Feature Selection Using Improved Teaching Learning Based Algorithm on Chronic Kidney	1%