

**CHRONIC KIDNEY DISEASE: A MACHINE LEARNING BASED
IMPROVED ANALYTICAL FORECASTING**

BY

**Md Rasel Miah
ID: 181-15-11164**

AND

**Sumaya Sarwar Prapty
ID: 181-15-11154**

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

Supervised By

Asma Mariam

Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Abu Kaisar Mohammad Masum

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY


DHAKA, BANGLADESH

JANUARY 2023

APPROVAL


The project, "**Chronic Kidney Disease: A Machine Learning Based Improved Analytical Forecasting,**" that was submitted to the Department of Computer Science and Engineering at Daffodil International University by Md. Rasel Miah and Sumaya Sarwar Prapty has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and has been approved as to its style and contents. The speech was delivered in 25 January 2023.

BOARD OF EXAMINERS



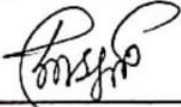
Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



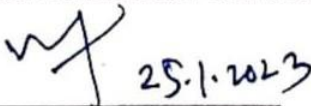
Dr. Md. Monzur Morshed
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dewan Mamun Raza
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



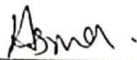
Dr. Ahmed Wasif Reza
Associate Professor
Department of Computer Science and Engineering
East West University

External Examiner

DECLARATION

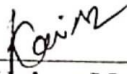
We hereby declare that, this project has been done by us under the supervision of **Asma Mariam**, Lecturer, Department of Computer Science & Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of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

Supervised by:



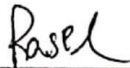
Ms. Asma Mariam
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Abu Kaisar Mohammad Masum
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Md Rasel Miah
ID: 181-15-11164
Department of CSE
Daffodil International University



Sumaya Sarwar Prapty
ID: 181-15-11154
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

We first want to give God the highest praise for His wonderful gift, which enabled us to successfully finish the senior project and internship.

We really appreciate Ms. Asma Mariam, Lecturer in the Department of Computer Science & Engineering at Daffodil International University, Dhaka, and express our sincere gratitude. Our project manager has a strong background in "Machine Learning" and a genuine interest in it. This project was made possible by her never-ending patience, academic guidance, constant encouragement, frequent and vigorous supervision, constructive criticism, helpful suggestions, reviewing numerous subpar versions and editing them at all stages.

We would like to extend our sincere appreciation to Dr. Touhid Bhuiyan, Professor and Head, Department of CSE, as well as to the other professors and employees of the CSE department of Daffodil International University, for your kind assistance in completing our project.

We'd like to thank all of our classmates at Daffodil International University who participated in this discussion while also attending class.

Finally, we must respectfully appreciate our parents' unwavering assistance and endurance.

ABSTRACT

Chronic kidney disease (CKD) refers to a variety of conditions that cause harm to the kidneys or a decrease in the Glomerular Filtration Rate (GFR). Due to recent advancements in medicine, doctors have been able to treat this issue utilising a number of various ways. A rising number of people are interested in applying artificial intelligence and machine learning, especially in the field of health, to enhance medical research and treatment. Because kidney condition can be deadly, machine learning must be used to forecast when it will first manifest. A number of machine learning approaches, applications, and algorithms may be utilised to forecast how "Chronic Disease" might progress. As a consequence, any doctor may be able to see the beginning of this condition as soon as the dialysis report is obtained. This approach may also be utilised to identify the disease component that, according to the report research, is the main contributor to the condition. To get the best results in this system, advanced and dynamic algorithms like Random Forest, Nave Bayes, Decision Tree, K-Nearest Neighbor (KNN), XGBoost, AdaBoost.

TABLE OF CONTENTS

APPROVAL	ii
DECLARATION	iii
ACKNOWLEDGEMENT	iv
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	8
2.4 Scope of the Problem	11
2.5 Challenges	12
CHAPTER 3: RESEARCH METHODOLOGY	13
3.1 Introduction	13
3.2 Research Subject	13
3.3 Machine Learning Techniques	13
	vi

3.3.1 Supervised Learning	14
3.4 Classification Techniques	14
3.4.1 Learning	14
3.4.2 Classification	15
3.5 Algorithmic Details	15
3.5.1 Logistic Regression	15
3.5.2 Support Vector Machine	16
3.5.3 K-Nearest Neighbors	16
3.5.4 Gaussian Naïve Bayes	16
3.5.5 Perceptron	17
3.5.6 Linear Support Vector Classifier	17
3.5.7 Stochastic Gradient Descent	18
3.5.8 Decision Tree	18
3.5.9 Random Forest	19
3.5.10 Adaptive Boosting (AdaBoost)	19
3.5.11 eXtreme Gradient Boosting (XGBoost)	20
3.6 Proposed System	21
3.6.1 Data Collection	21
3.6.2 Dataset	22
3.6.3 Data Pre-processing	22
3.6.4 Data Normalization	22
3.6.5 Data Splitting	22
3.6.6 Imple Algorithms	22
3.6.7 Model Analysis	23
3.6.8 Extract Appropriate Algorithm	23
3.6.9 Execute Model	23
3.6.10 User Segment	24
CHAPTER 4: EXPERIMENTAL RESULTS & DISCUSSION	25
4.1 Introduction	25
4.2 Experimental Results	25

4.2.1 Data Acquisition	25
4.2.2 Data Utilization	26
4.2.3 Feature Importance	28
4.3 Result & Discussion	29
4.3.1 Confusion Matrix	29
4.3.2 Classification Report	32
4.4 Result Analysis	33
4.4.1 Accuracy	34
4.4.2 Jaccard Score	34
4.4.3 Cross Validated Score	35
4.4.4 AUC Score	36
4.4.5 ROC Curve	36
4.4.6 Standard Deviation	38
4.4.7 Misclassification & Error	38
CHAPTER 5: IMPACT ON SOCIETY & SUSTAINABILITY	40
5.1 Introduction	40
5.2 Impact on Society	40
5.3 Ethical Aspects	40
5.4 Sustainability	41
CHAPTER 6: FUTURE SCOPE & CONCLUSION	42
6.1 Introduction	42
6.2 Implication for Further Study	42
6.3 Recommendations	42
6.4 Conclusion	43
REFERENCES	44

LIST OF FIGURES

Figure No.	Figure Name	Page No.
Figure 3.1	Proposed Method to Predict Chronic Kidney Disease	22
Figure 4.1	Accuracy Chart	35
Figure 4.2	Jaccard Score Chart	36
Figure 4.3	Cross Validated Score	36
Figure 4.4	AUC Score Chart	37
Figure 4.5	ROC Curve	38

LIST OF TABLES

Table No.	Table Name	Page No.
Table 4.1	Data Acquisition & Null Percentage	27
Table 4.2	Dataset Description	28
Table 4.3	Feature Importance	29
Table 4.4	Confusion Matrix	30
Table 4.5	Confusion Matrix for Algorithms	31
Table 4.6	Classification Report	33
Table 4.7	Accuracy, Jaccard, Cross Validated and AUC Score	38
Table 4.8	Standard Deviation	39
Table 4.9	Misclassifications & Errors	40

CHAPTER 1

INTRODUCTION

1.1 Introduction

The phrase "chronic renal disease" describes kidney impairment that has persisted for a long period. This common disease is often associated with becoming older. Although it may affect anybody, people of color—particularly those from the Caribbean and South Asia—are more likely to experience it. It is unusual for CKD to get worse over time to the point where the kidneys stop working altogether. Despite their condition, many people with CKD go on to live long, robust lives. International researchers made an effort to pinpoint the precise cause of this kidney illness. Any patient who has renal illness will have a serious condition. The most deadly kidney illnesses include urinary tract infections, chronic kidney disease, kidney stones, glomerulonephritis, polycystic kidney disease, and glomerulonephritis. High blood pressure is a frequent contributor to several of them, including chronic kidney disease. Given that the glomeruli, which filter out waste materials, might be subjected to additional stress from high blood pressure, the kidneys are at risk. Blood is purified by the glomeruli, or tiny blood vessels, in the kidneys. As the pressure builds up over time, the renal vessels are affected, which causes the kidneys to work less effectively. It will eventually get to the point where the kidneys can no longer function properly. A person would need to get dialysis if this were the situation. A procedure called dialysis. A kidney transplant may be an option for the patient, depending on their circumstances. Diabetes is frequently linked to chronic renal disease. Numerous disorders, including diabetes, can show symptoms of high blood sugar. Over time, the blood vessels in the kidneys are impacted by having increased blood sugar levels. This shows that the kidneys are unable to clear the blood, which is a typical renal function. Kidney failure can occur if your body becomes overloaded with toxins. In their investigation, scientists from many countries demonstrated the true fatality rate of chronic kidney disease. According to many research studies, male patients with chronic kidney disease died at a greater rate than female ones. According to the most recent WHO data, 16,948 deaths in Bangladesh were attributable to kidney disease in 2018. This represents 2.18 percent of all fatalities. With a death rate of 14.83 per 100,000 persons when adjusted

for age, Bangladesh is ranked 94th in the world [1]. COVID-19 is now spreading throughout Bangladesh. In individuals with chronic kidney disease (CKD), Access to dialysis and other healthcare treatment facilities is made more challenging since COVID-19 exposure raises the chance of contracting a life-threatening infection [2]. Machine learning algorithms have several sub-branches. Examples include Reinforcement Machine Learning, Semi-supervised Machine Learning, Supervised Machine Learning, and Unsupervised Machine Learning. Supervised Machine Learning algorithms In this study, To forecast Chronic Kidney Disease, the most important algorithms—K-Nearest Neighbor, Decision Tree, Random Forest, Perceptron, AdaBoost, Gaussian Nave Bayes, and XG Boost —are used. Anyone may forecast their likelihood of having severe kidney disease using this approach. However, if the pathology department and patients alike want to know the precise chances of being afflicted by chronic kidney disease, a web implementation procedure might be the most helpful approach. The true goal of this study is to create a model trained on a pertinent dataset, discovering. With a greater understanding of the causes of kidney illness and the ability to take the appropriate action to treat it, people will benefit from this research.

1.2 Motivation

Many people in Bangladesh are afflicted with a wide range of chronic diseases, such as diabetes and high blood pressure. In today's culture, chronic diseases are more common, and this trend is predicted to continue. No matter how few patients appear to be present at any one time, people with chronic diseases must be present in every hospital or clinic. Bangladesh is a country rich in water since rivers round it on all sides. Due to these traits, people are more prone to both the most serious ailment, chronic kidney disease, as well as water-borne disorders like diarrhea Additionally, a machine learning-based approach is required for better understanding and halting the spread of this illness.

1.3 Rationale of the Study

If chronic kidney disease is diagnosed and anticipated in time, it may be treated promptly and effectively to the point of curing the disease. In practice, it takes a long time and multiple blood tests to establish if a person has chronic renal disease. Thus, a model that can predict any stage of chronic kidney disease has been introduced and trained with appropriate data. The fundamental goals of this research were to simplify people's lives and save them time. To better prepare for chronic renal disease at home, patients can contribute data using a web interface. It will be useful for both healthcare providers and patients suffering from chronic renal disease.

1.4 Research Questions

There is a wide variety in the kinds of questions that can be posed about this research. Multiple people were asked the same set of questions to streamline the data collection process for this study.

Why was the prediction of chronic kidney disease the focus of this study?

Chronic kidney disease is a worldwide epidemic. Eventually, it will get much worse. Ultimately, the kidney stops working altogether. When that point is reached, it's fatal. Chronic Kidney Disease can be prevented and treated effectively if its presence is detected at an early stage. This is why CKI was selected as the primary outcome measure for the study.

How come you're using a machine learning strategy? Is it a dependable source?

Machine learning is the gold standard for producing predictions of any kind. With enough information, a model can educate itself and make any prediction. Using machine learning on medical datasets may allow for more accurate diagnosis of chronic kidney disease. At present, the world is through a period of rapid modernization. Imagine a time about ten years ago, when machine learning and artificial intelligence were in their infancy and these prime points were nothing more than names with a touch of mathematics. But today, AI is

what drives 50% of all technology in the world. Therefore, even if it is reliable enough now, with the right training and improved precision, it can be even more so.

Why are there 11 different algorithms used?

One acceptable method that fits the Chronic Kidney Disease dataset was selected using 11 different algorithms. The optimal algorithm, with the lowest mistake rate and maximum accuracy rate, has been found after differentiating and evaluating 11 algorithms. If only one method was chosen, it would be difficult to determine which approach would fit the dataset the most effectively and efficiently.

1.4 Expected Outcome

Over the course of this study, there have been some adjustments made to the central theme or anticipated outcome. It sheds light on the exact findings of this study. Chronic kidney disease could be halted with the help of this study if the fundamental reasons of the illness's progression could be identified and treated. Doctors and statisticians can evaluate the age at which a person develops chronic renal disease (CKD). Precise calculations and algorithms may enable the revelation of a thorough internal study of Chronic Kidney Disease for the purpose of medical analysis. While this study is ongoing, a notice about Chronic Kidney Disease may be sent to riverside communities and places.

1.5 Report Layout

This study report is broken up into six separate sections to make it easier to navigate and more useful for readers and researchers.

Chapter 1 provides a crucial introduction to this study project. This is a brief explanation of kidney disease connected to it. This chapter explains the purpose of the study, the pertinent research questions, the anticipated results, the overall management information, and the financial issues.

Chapter 2 presents a thorough account of the study's history. Including, but not limited to, categorization data, machine learning systems, and other relevant work based on this research study. This chapter also describes the extent of the problem statement and the apparent difficulties with comparative analysis.

Chapter 3 provides a description of the research study's suggested system and methods. From the very beginning of mathematics through the discussion of the current state, each employed algorithm's algorithmic specifics are given.

Chapter 4 provides a thorough analysis of each step's results. The best algorithm, Jaccard score, cross-validated score, confusion matrix, and classification report are used to conclude it. Each algorithm's ROC-AUC curve is also explained. The final section of this chapter describes Standard Deviation, Misclassification, Mean Absolute Error, and Mean Squared Error.

Chapter 5 discusses the ethical factors that are crucial for any study that will have a significant influence on society. The sustainability of this study effort is covered in the chapter's concluding part.

Chapter 6 where it is succinctly characterized as the expansion of this research study, demonstrates the future extent of this research activity. The full study report is concluded in this chapter with a helpful summary that quickly discusses the main results

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This problem of chronic renal disease that we cover here has been around for quite some time. The history of this disease is therefore fraught with suffering and death. That's why there's been a lot of effort put into finding a cure for CKD; it could save a lot of people's lives. This chapter serves as an introduction to the condition and its background. This chapter offers discussion of some current studies on the subject. As a last step, some comparison analysis has been offered to show how much the present position

2.2 Related Works

For CKI detection, segmentation, and diagnosis, we suggest a heterogeneously modified artificial neural network operating on the IoMT platform. Using Backpropagation, the Multilayer Perceptron (MLP) is what gives the proposed HMANN its power. The first stage involves selecting the kidney region of interest from a segmented ultrasound image. The HMANN technique is recommended for kidney segmentation since it is both accurate and fast. This study proposes using a dual-stack network (DsNet), a network that can be trained in stages to differentiate between patients with diabetes and chronic kidney disease and healthy people. the goal of this network is to improve patient care. From the photographs of different people's faces, we were able to successfully elicit generalizable intermediate features in the first stack subset. A second stack network can be used to examine the first stack's high-level properties before healthy people are able to simultaneously categorize both illnesses. compared to the traditional, noninvasive techniques that are now the industry standard [3]

In this article, Ali et al. use CKD diagnosis in developing countries as a case study to address challenges associated with using automated decision support tools. This research improves upon previous methods of group-based feature selection by providing a cost-sensitive ensemble feature ranking methodology. According to our knowledge, \sthis is one of the first studies to demonstrate that cost-sensitive ensemble rankings for\snon-cutting

groups may yield both low-cost and high-accuracy answers. This was one of the first studies to establish this concept. The experiment demonstrates how effective the strategy is by making use of eight different comparison selection procedures and seven well-known classification algorithms. Incorporating the cost aspect into the objective space of potential solution formulations has been found to have the potential to increase the utility of automated CKD systems. Therefore, it is possible to develop a method that is reliable and cost-effective for diagnosing chronic kidney disease [4]

The study method known as Improved Teacher-Learner Based Optimization, or ITLBO for short, was created to find the best feature subset by analyzing the fitness function and common optimization parameters such as population size and thresholds throughout generations. An effective method for focusing on the optimal set of criteria for the early detection of chronic illnesses is the Chebyshev Distance Requirement. This can be accomplished by reducing the number of possible combinations of features. In comparison to the feature reduction obtained using the original TLBO approach, which was 25%, the suggested function selection strategy produced a notable drop of 36% when applied to data sets for Chronic Kidney Disease (CKD). With the aid of the Support Vector Machine (SVM), Convolution Neural Network (CNN), and ITLBO methodology, the accuracy of the optimal feature subsets created by the TLBO techniques and the feature subsets produced by the ITLBO methodology are tested. Boosting algorithms for gradients. This evaluation is carried out in order to determine which methodology yields the most accurate optimal feature subsets. These algorithms are what are utilized to determine which feature subsets are the best. According to the outcomes of the studies, all three strategies for the newly introduced feature subset perform significantly better than the first TLBO method that was applied by Balakrishnan et al. [5]

2.3 Comparative Analysis

Researchers from a previous study advocated the use of deep learning in conjunction with an innovative detection approach in order to diagnose CKD through the use of mouth swabs. In order to resolve the problems with the data, Convolutional Neural Networks (CNN) and Support Vector Machines (SVM) were utilized (SVM). The CNN-SVM network has an accuracy rating of 97.67% overall, with a sensitivity of 97.5% and a specificity of 97.83%, respectively. The CNN model had an accuracy rating of 96.51% on average. When it comes to the task of data classification, our proposed method performed significantly better than the status quo when compared to other methods In contrast to earlier research [17].Using a Jaccard Score of 96.14%, a Cross Validated Score of 98.97%, and an AUC Score of 98.21%, the authors of this study achieved a high accuracy of 98.55% in the XGBoost Classifier. Cross-validated score is 98.97%, while the Jaccard score is 96.14%] The machine learning repository at UCI was the source of the vast bulk of the investigation's data, the majority of which contained missing values. Imputation using KNN was helpful with resolving this issue. Six distinct approaches to machine learning were utilized throughout the development of the models.Random Forest outperformed the other machine-learning models with a diagnosis accuracy of 99.75%. 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 Following an analysis of the errors generated by the currently available models, we developed a new one that incorporates logistic regression, random forest, and perceptron. The new model achieved an accuracy of 99.83% on average following 10 simulations. The researchers who were responsible for this study combined the data that they obtained from a number of hospitals in Bangladesh with the data that they obtained from the University of California, Irvine. We tallied a total of 10321 separate data points in our investigation. The authors of this study conducted a series of experiments using 11 different algorithms [18] to discover which one would yield the most accurate results based on the data that was provided. In this investigation, we apply cloud computing to the process of providing medical treatment, specifically to the forecasting of patients who have chronic renal disease (CKD). This study creates a hybrid smart model for CKD prediction in the cloud-based Internet of Things by combining two intelligent methods—a neural network (NN) and linear regression. Both of these methods are intelligent in their own right (LR). According to the findings, the innovative hybrid model had a degree of accuracy of 97.8 percent when it came to predicting CKD. In this paper, approximately eleven different algorithms are suggested for further investigation. Following an analysis of the effectiveness of eleven distinct algorithms, a strategy that is suggested was formulated. We utilized the best method to develop a model, and then we used that model to design a user-friendly front end for entering the relevant characteristics and generating a prediction [19]. This front end was based on the model that we constructed using the best method. Considering the results of the previous investigation, it was indicated that obtaining an accurate diagnosis is more crucial than locating the treatment that is going to be the most beneficial. The major objective is to compare the two approaches in terms of their ability to accurately forecast chronic renal disease (CKD). In this study, data mining techniques like the Random Forest and the Back Propagation Neural Network were both employed. In contrast to the accuracy of the Random Forest Algorithm, which is just 87%, the Back Propagation Algorithm has an accuracy of 98.40%. The analysis carried out by the authors made use of eleven different algorithms in all. Four of these algorithms, including XGBoost (98.55%), Decision Tree and AdaBoost (98.11%), and Random Forest (97.09%), each obtained an accuracy level more than 95%. [20] A method for predicting

the presence of CKD from clinical data has been published in previous research. The phases of this technique were attribute selection, collaborative filtering, a missing value management system, and data preparation. The random forest classifier and extra tree classifier were found to have the highest accuracy. Out of the eleven different machine learning techniques that were examined, (100%) were successful. The study demonstrates the need of incorporating domain expertise when using machine learning to predict CKD status and takes into consideration the practical limitations of data collecting. The authors of this study used the most accurate model for predicting all phases of chronic kidney disease (CKD). [21] Size of the Issue Bangladesh is referred to as the "River Kingdom" frequently. More than 700 rivers traverse Bangladesh's landscape. As a result, every town is situated beside a river. But since the advent of industry, water quality has deteriorated dramatically contaminated. This makes people in the country vulnerable to a wide variety of illnesses. In the United States, chronic kidney disease is a major health concern. For 2018, kidney disease was responsible for 2,18% of all deaths in Bangladesh, or 16,948 deaths according to the latest data from the WHO. Death or permanent disability might result. Therefore, immediate treatment is required. In order to recognize Chronic Kidney Disease in its early stages, it is important to keep an eye out for any abnormalities in health data and to keep track of any uncommon symptoms. When preventative actions are taken, chronic kidney disease can be halted before it reaches the end-stage renal failure that it is currently in.

2.4 Scope of the Problem

The "River Kingdom" is the nickname for Bangladesh. Bangladesh is traversed by over 700 rivers. All of the communities are built around rivers as a consequence. The water has gotten increasingly polluted as a consequence of the Industrial Revolution, however. The people of this country are thus susceptible to a variety of ailments. In the United States, chronic kidney disease is a serious public health concern. According to the most recent WHO data, 16,948 deaths in Bangladesh were kidney-related in 2018, or 2.18 percent of all fatalities. It could lead to demise or long-term incapacity. Therefore, it has to be treated right away. Monitoring unusual health symptoms and health data may help detect Chronic Kidney Disease in its early stages. By taking the necessary measures, chronic renal disease may be stopped before it reaches its current stage.

2.5 Challenges

It may take the kidneys many hours to make urine due to the extensive excretion and reabsorption processes they must through. Maintaining stability in the body depends on a healthy chemical balance. However, this process is slowed or halted altogether in those with chronic kidney disease. The hardest element of this research was pinpointing what exactly makes a difference in assessing whether or not a person has chronic renal illness. Eliminating all blanks from a data set is a challenging task. We may be in for a long haul here. Conceiving of an appropriate algorithm was difficult. Scientists employed a variety of algorithms to train the dataset, and then selected the most promising methods for detecting chronic renal illness (CKD). The most challenging part of the project was developing an easy-to-use interface for the system so that data could be entered and a CKD prognosis could be generated at any time.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Before starting, it is essential to develop a research approach. It is necessary to identify the issue at hand before looking for a solution in order to handle it. An explanation of the study's subject is given in this chapter. After discussing the techniques used to answer the issues, the methodology was further explained using a graphical illustration to facilitate comprehension.

3.2 Research Subject

Finding a problem to explore and then determining a solution is the basic goal of research. A common condition that has a direct correlation to becoming older is chronic renal disease. It is a chronic condition that might get worse. Numerous individuals lose their lives to kidney illness each year. When CKD reaches its most advanced state, kidney function drops to less than 15%. Any kind of kidney illness may be cured if it can be detected early and treated well. Focusing on renal conditions is essential, as is creating a model that can predict CKD at different ages.

3.3 Machine Learning Techniques

Both supervised machine learning and categorization may be used to classify objects. A self-contained machine learning system is one that can continually gather and integrate data and can do so with the goal of making judgements. It is possible to design a system that is always getting better by reflecting on past mistakes, making analytical observations, and utilizing other techniques. Different techniques to machine learning exist in different forms and sizes.

3.3.1 Supervised Learning

Here, we make extensive use of supervised machine learning techniques. In order to forecast the future, supervised machine learning algorithms use labelled samples of the past. A learning technique is used to anticipate the output values; it builds an inferred function based on examination of a well-known training dataset. The system may continue to train indefinitely after that. The learning process may identify any differences and make the necessary corrections if the output of the model is compared to the predicted one. Using a variety of supervised learning approaches, we first categorise Chronic Kidney Disease in this dissertation.

3.4 Classification Techniques

To develop models that reflect pertinent data categories, classification-based data analysis is employed. In terms of machine learning, it is now considered the norm. In supervised learning, classifiers are the models that are used to generate predictions about predefined classes. These projections are scattered and distinct. The classifier does not provide a value between the two extremes. A classifier may be developed, for instance, to detect whether an image depicts a dog or a cat. It's expected to be either a dog day or a cat day. The classifier does not provide an in-between value. Data with labels may be subjected to a classification learning approach. Data may be divided in two different ways when used to categorization learning. Data may be divided into two categories: training data and test data. The model is initially built with training data, and it is then verified with test data. To categorize something, there are two steps required.

3.3.1 Learning

During the learning phase, a classifier is constructed using a suitable methodology and training data. Next, the classifier is evaluated against the real world. A classification method and training data are combined to build a classifier. A classifier is just a group of rules that may be used in various situations.

3.4.2 Classification

It is now possible to forecast which category of unknown data will be anticipated as a result of the learning phase using the classifier or model that was generated during the prediction phase.

3.5 Algorithmic Details

Eleven of the top Supervised Machine Learning methods in total are used in this research. Computer programmes called algorithms are designed to take raw data and turn it into something more useful. Any information that may benefit people, machines, or algorithms is important because knowledge is power. Algorithms for machine learning also make use of mathematics in this manner. Undoubtedly, not every machine learning algorithm is mathematically transformed the same way.

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 acquire 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:

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

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

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

3.5.2 Support Vector Machine

Constant assessment is necessary for the Support Vector Machine (SVM) to be effective as a machine learning tool. It might be applied to classification and regression issues. Although it has other uses as well, the categorization process is where it is most often used. This method's developers attempted to use a graphing calculator to plot data points in n-dimensional space, but ultimately failed. The analysis of two-dimensional data charting was looked at, and either the CKD or the NotCKD methodologies were used.

3.5.3 K-Nearest Neighbors

To deal with classification and regression problems, one simple supervised machine learning method called k-nearest neighbours (KNN) may be used. KNN is simple to use and comprehend. The Euclidean distance serves as the main guiding concept of KNN. Due to the dichotomous nature of the dataset, KNN is used to categorise the data. Equation provides the true formula for the K-Nearest Neighbor method (iii).

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

3.5.4 Gaussian Naïve Bayes

A variation of Naive Bayes that accepts continuous data and the Gaussian normal distribution is called Gaussian Naive Bayes. The Bayes theorem serves as the foundation for the Naive Bayes family of supervised algorithms. Despite being straightforward, the classification process functions pretty well.

When dealing with continuous data, it is often assumed that it follows a normal (or Gaussian) distribution. Equation gives the Gaussian Naive Bayes method's complete formulation (iv).

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

3.5.5 Perceptron

The perceptron algorithm is one particular type of neural network model that is used as an algorithm in a binary classification machine. One node, or neuron, processes data from a row of observations and produces a projected category. Using a predictor function and a feature vector, two linear classification techniques, this method produces predictions. In this section, we outline the procedures needed to train a threshold function that accepts a single integer input (x) and outputs a single binary result (f(x)). At the output, the binary value of x is changed into the binary value of f. (x). Equation gives us the whole functionalized Perceptron algorithm approach (v)

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

3.5.6 Linear Support Vector Classifier

A Support Vector Classifier (SVC) uses a hyperplane that offers the "best fit" for the data to classify data according to its attributes. To get the "predicted" classification result after the hyperplane has been formed, it could be required to provide the classifier extra data. As a result, the algorithm excels at a certain job while still being applicable in other situations. Two arrays of shapes—one holding the training samples (X) and another (Y) containing the shape class labels—are supplied to SVC when it is given training examples. The feature of being a collection of forms applies to both two lists.

3.5.7 Stochastic Gradient Descent

To train a classifier, data scientists feed it training examples using SVC. The Stochastic Gradient Descent (SGD) technique is a straightforward yet powerful approach to learning linear and regressive classifiers with convex loss functions, such as (linear) Support Vector Machines and Logistic Regression. SGD has been used in the area of machine learning for some time, but it has only lately gained general acceptance in the context of extensive education. The class SGD Classifier provides a useful stochastic gradient descent learning method for data classification that works with a variety of loss functions and penalties. a comparable linear SVM classifier that uses the hinge loss training to SGD. In this article, we will discuss the mathematical underpinnings of the SGD procedure. Examples of retraining are supplied, such as (x_1, y_1) . This system requires training on a linear scoring function, $f(x) = w^T x + b$, where $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$ are model parameters and $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$ are inputs, respectively ($y_i \in \{-1, 1\}$ for classification). For binary classification, the system needs to account for the sign of $f(x)$. To find the parameters of a model, we first normalize the training error of the corresponding 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 assesses model fit and R is a normalisation term that controls the strength of the normalisation; a combination that is > 0 is said to be non-negative linear.

3.5.8 Decision Tree

Both classification and regression problems in supervised learning may be solved using decision trees. They are often used to categorise problems. Classification outcomes are represented by leaf nodes in this tree-like classifier, while inner nodes represent attributes of the dataset and outside branches represent decision criteria. Using the training dataset, the Decision Tree may be used to create a model that predicts the class or value of the input variables based on the basic rules of decision-making

c

$$E(S) = \sum - p_i - \log_2 p_i \dots \dots \dots (vii)$$

Pi is the probability that an event, I, will occur in state S or the percentage of the class I in the node state S in equation (vii), where S is the current state.

3.5.9 Random Forest

Random forests, also known as random decision-making forests, are an ensemble learning method that may be applied to classification, regression, and other tasks by generating a number of decision-making trees. When conducting a classification task, the majority of trees select the category indicated by the random forest's output. The mean or average forecast of each individual tree is provided for regression tasks.

$$\sigma = \sqrt{\frac{\sum^B (f(x') - \bar{f})^2}{B - 1}} \dots \dots \dots (viii)$$

B, the sample size, is the only free variable in Eq. (viii). Trees can number in the hundreds to the thousands, depending on the size and purpose of the workshop. By computing the average prediction error for each training X' sample—the collection of trees lacking X' in their bootstrap sample—either by cross-validation or by keeping track of the out-of-bag error, it is possible to identify the ideal number of B trees. The training and test error rates increase when just a few trees are fitted.

3.5.10 Adaptive Boosting (AdaBoost)

Adaptive classification boosting, sometimes referred to as AdaBoost classification, is a group learning strategy. To create a powerful classifier, it combines a number of weak classifiers. This strategy allows a bad classifier to learn from its prior classifier errors. Think about the dataset n_{sample}. A weight of 1/n is originally assigned to each sample. A weak classifier is created as well. use this dataset. The total error is calculated by this classifier. This complete error in the classification of the data samples serves as a proxy for the impact of such a classificatory method.

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

The is used in equation (ix) to modify the weight of the dataset samples and produce a new dataset.

3.5.11 eXtreme Gradient Boosting (XGBoost)

A gradient boosting framework is used by the machine learning approach known as XGBoost to produce choices. It builds multiple trees that minimise a regularised goal function after first using numerous categorised and regressed (CART) trees as weak learners to enhance tree performance. In order to create a quick-running algorithm with strong predictive power, the approach depended on split-wisdom discovery in every tree, cache-friendly approximation methods for identifying splits, and effective out-of-core gradient boosting techniques..

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:

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

$F = f(x) = w_q(x)$ The collection of CART decision tree structures is represented as $(q: R^m \rightarrow R, T, w)$, where q refers to the leaf nodes of the sample map's tree structure, T determines the number of leaf nodes, and w provides the actual score of the leaf nodes.

3.6 Proposed System

Now that all of the algorithms discussed above have been considered, the desired system may be suggested. It is better to grasp the precise operation of the system by using a system diagram, as illustrated in Figure 3.1.

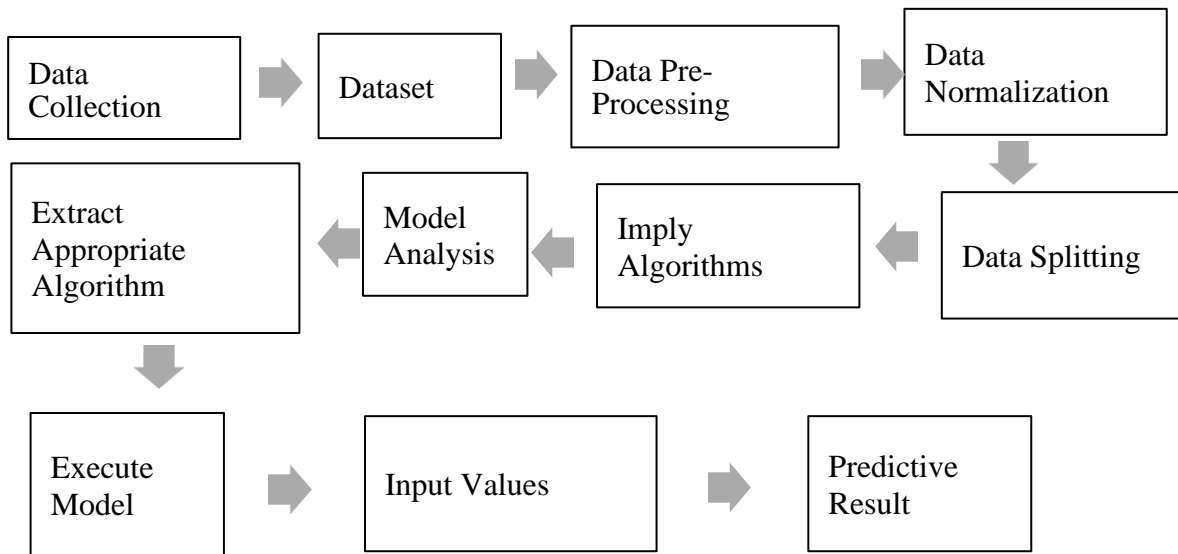


Figure 3.1: Proposed Method to Predict Chronic Kidney Disease

3.6.1 Data Collection

Using real-world data is essential for the system's CKD evaluation. The Kidney Foundation of Bangladesh, a number of hospitals, and the University of California, Irvine all contributed data to this study (UCI). In the first stage of this study, we compiled all of the necessary data from various medical facilities. Data was then merged into a single comma-separated values (CSV) file for ease of study and comprehension.

3.6.2 Dataset

A CSV file was created once all the data had been combined, allowing various machine learning techniques to be used on the data. Machine learning algorithms require massive amounts of data in order to produce a prediction. The raw data collection has 25 columns and 10321 rows, some of which are blank.

3.6.3 Data Pre-processing

The dataset had several missing values and has to be converted qualitative data. The qualitative input was first transformed into quantitative data. Dealing with the missing values follows. The mean was used to replace each missing value.

value. The independent variables X and Y were separated from the dependent variables.

3.6.4 Data Normalization

The process of normalizing involves keeping the ranges of values while translating numerical columns to a similar scale. For greater accuracy, the independent variable (X) was afterwards normalized.

3.6.5 Data Splitting

Any machine learning method must divide the dataset into a train set and a test set before it can be used. 80% of the data were used to train the model, while 20% were utilized for testing. A model can be taught to predict anything using the dataset's training component, and its performance can be assessed using the dataset's testing component.

3.6.6 Impley Algorithms

To determine which algorithm produced the best results and then choose that one, eleven alternative ones were used. In total, 11 algorithms can be used: the Support Vector Machine (SVM), the Stochastic Gradient Descent (SGD), the Decision Tree, the Random Forests

(RF), the Adaptive Boosting (AdaBoost), the eXtreme Gradient Boosting (XGBoost), the Gaussian Naive Bayes, Perceptron Algorithm, K-Nearest Neighbors (KNN), Logistic Regression, and Linear Each of these methods produced distinctive analytical results.

3.6.7 Model Analysis

Area Under the Curve, Accuracy Score, Jaccard Score, Cross Validated Score, and Confusion Matrix (AUC), Misclassification, mean absolute error, and mean squared error were computed for each technique using the tabulated data. The accuracy of the data prediction can be evaluated using the confusion matrix. There are a variety of metrics that can be used to estimate how accurate the data will be, including the area under the curve (AUC), jaccard score, cross validation score, and accuracy score. We may evaluate the algorithms' accuracy using measures like misclassification, mean absolute error, and mean squared error.

3.6.8 Extract Appropriate Algorithm

The best method was selected after meticulously measuring and analyzing all essential data from the tables. The extracted approach has the best accuracy and lowest error rate in the dataset. First, an efficient algorithm needs to be developed so that the dataset may be used to its greatest potential. It is recommended to employ multiple algorithms as models in order to find the most effective one. The optimal method in this research was determined using the accuracy score, Jaccard score, cross validation score, area under the curve (AUC), and other analytical criteria. The CKD dataset was utilised in this study, and the XGBoost Classifier delivered the best results. Since every requirement was satisfied, the grade was a perfect 10. The procedure will go on to the next stage after the algorithm decision.

3.6.9 Execute Model

After finishing the model, it must be saved to a specific location. Pickles are employed at this juncture. The Pickle module in Python provides a mechanism for encoding and decoding object structures. By converting Python objects to bytes, data can be saved in

files or databases, saved between sessions of the same program, or transferred over the internet. Next, the Python library pickle must be used to dump the model. For this purpose, we use `Pickle.dump()`. After getting the model's name back from the function, you open a file in write binary mode and choose a location. The model can then be accessed from within that folder.

3.6.10 User Segment

This user guide shows how a person could check their kidney function. Attribute values for the diagnostic report will be entered by a user on a computer. The user will then get a brief evaluation of their kidney health via the device. The layout of the user interface is simple and basic. The user interface is so simple and straightforward that anyone can use it. To have the computer make a prediction, all you have to do is enter a value and hit the submit value button.

CHAPTER 4

EXPERIMENTAL RESULTS & DISCUSSION

4.1 Introduction

The results matter greatly in any investigation or undertaking. For the simple reason that the end result is the sum total of all project outcomes. This chapter presents all of its findings in tabular form. This chapter has demonstrated the particulars of data collection, data utilization, and feature importance for comprehending the CKD dataset. In order to compare and contrast the performance of different algorithms, a confusion matrix table was constructed. The accuracy, recall, and F1-Score may be shown in a table using a classification report. The accuracies, Jaccard scores, cross-validation scores, area under the curve (AUC), and receiver operating characteristic (ROC) curve have all been shown graphically. Details are presented in a table format for ease of reading. The standard deviation is also tabulated for your convenience. Finally, the incorrect categorization and error were laid up in a table.

4.2 Experimental Results

Following the creation of a machine learning model, several algorithms were assessed in terms of how well they could anticipate the beginning of chronic renal illness. Therefore, it is possible to analyze all possible scores for each algorithmic application and procedure, as shown in the results of the experiments.

4.2.1 Data Acquisition

The Bangladeshi Kidney Foundation, many hospitals, and the University of California, Irvine all contributed data that was utilized to teach the algorithm (UCI). The model was trained using a total of 10321 data samples. There are a total of 24 distinguishing features or components in this CKD data set, one for each sample: a "target" variable, plus 13 category (nominal) variables, and 11 numerical variables (class). Each category can take on one of two possible nominal values: CKD (sample containing CKD) or notCKD (sample without CKD). The data is missing key crucial values. Table 4.1 provides a high-level

summary of the data set. .

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

To facilitate the organization of the data in a database, each category (nominal) variable was given its own unique set of codes. RBC and PC counts were noted as normal and abnormal, respectively, as 1 and 0. . The group settled on a classification system assigning a 1 to pcc and a 0 to ba. Thus, 1 meant yes and 0 meant no for the yes/no responses. Good appets were given a value of 1, while bad ones were given a value of 0. Remember that while sg, al, and su were classified as categorical variables originally, their real values were determined by their numerical connections. A factorization process was used to modify all of the category variables. An individual number between one and 10321 was assigned to each sample. This dataset has a large number of blank cells. There are several reasons why

patients could omit crucial steps in the diagnostic procedure. When sample diagnostic categories are unclear, missing data might be found and a proper imputation approach must be implemented. After determining where the gaps were in the core CKD dataset, we processed and populated the categorical variables. The data description for each of the 25 features was then extracted to give a more thorough overview of the entire dataset. The number of samples, mean, standard deviation, and In Table 4.2, the minimum, 25%, and maximum are displayed.

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

Methods that assess an input feature's significance based on its capacity to predict an outcome variable are referred to as having "feature significance." When constructing a prediction for the input characteristics of a predictive model, the term "feature importance" can be used to refer to a variety of methods for ranking the importance of each feature. Predictive models can be improved by using the feature significance score, which can also shed light on the dataset and the model. Table 4.3 displays the feature relevance of various attributes for various algorithms. Table 4.3 displays the feature relevance of various attributes for various algorithms. It is obvious that haemoglobin is the most important component based on the information in Table 4.3. This means that the qualities of hemoglobin are far more important than any others.

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 results of this study showed a positive value for people with CKD but a negative value for those without CKD. Using the confusion matrix, the efficacy of various machine learning algorithms has been measured, and specific results have been demonstrated. In Table-4, you'll find a sample confusion matrix that can be used with several distinct kinds of algorithms.

4.3.1 Confusion Matrix

A confusion matrix has to be created so that the implementation's outcomes may be evaluated. The performance of a classification model is evaluated with the use of a confusion matrix, which takes the form of an N-by-N matrix and a set of N target classes. A machine learning model's performance can be evaluated by comparing its projected values to the actual target values using the matrix. This exposes the shortcomings and successes of the algorithmic model. The metrics of Precision, Recall, and Accuracy in binary classification can be calculated with the use of some simple equations. These average values must also be processed using either a micro average or a macro average for multiclass labelling. Because of their widespread use in the computation of various assessment metrics, it is important to first familiarize oneself with four foundational elements. False Positives (FP), True Negatives (TN), False Positives (FN), and True Positives (TP) are additional terms (FN). The Confusion Matrix, which comprises True Positive, False Positive, False Negative, and True Negative values, is really formed as shown in Table 4.4. The confusion matrices for each of the various techniques are shown in Table 4.5.

Table 4.4: Confusion Matrix

Confusion Matrix		
	Actual Class	
Predicted 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%

True Positive (TP)

Positive tuples are ones that the classifier properly classified as falling within a certain category. The acronym's TP letter serves as a clue. Here, 62% of the values in XGBoost, Decision Tree and AdaBoost were True Positive (TP) values. With 61% and 60%, respectively, Random Forest and Logistic Regression finished in second and third.

True Negative (TN)

Positive tuples that the classifier misclassified are known as negative tuples. These instances may be recognised by the letter TN. Here, the True Negative (TN) value for Decision Tree is 37%, followed by 36% TN values for XGBoost, AdaBoost, and Random Forest. and 34% using logistic regression.

False Positive (FP)

The classifier that is of relevance today has mistakenly classed these negatively labeled tuples as positive. FP may be used to denote this kind of connection. In this part of the investigation, the False Positive (FP) scores for the Random Forest, Decision Tree, and XGBoos were lower, at only 1%. Naive Bayes and AdaBoost came in second and third, each producing 2% of the False Positive (FP) results, while Logistic Regression produced 3%.

False Negative (FN)

The classifier misclassified these positive tuples as negative. The letter FN is used to identify it. The results of the XGBoost and AdaBoost's False Negative (FN) values were both zero. After that, Logistic Regression has 3% False Negative (FN) values, followed by Decision Tree and Random Forest with 1% each.

Precision

One criteria for judging how accurate something is may be its precision (i.e. what percentage of tuples labelled as positive are actually such). In other words, it calculates the proportion of events that were really relevant when they were retrieved. Equation (x) displays the formula for calculating Precision.

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

Recall

In machine learning, It serves as a gauge of thoroughness (how many positive tuples are recognised as such). The percentage of pertinent cases among all accessible pertinent examples is what is referred to as a relevant instance. The mathematical formula to calculate Recall is shown in Equation (xi).

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

F1-Measure

The weighted harmonic mean, often known as the F measure, is a tool used to assess a test's recall and accuracy. The mathematical formula to calculate F1-Measure is shown in Equation (xii).

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

Accuracy

The percentage of test set tuples that the classifier correctly categorizes on a given test set serves as a gauge of its accuracy. Understanding how to determine accuracy from an equation is simpler (xiii).

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

4.3.2 Classification Report

A classification report is a measure used to assess the effectiveness of the system in machine learning. It demonstrates the reliability of a trained classification model by measuring its accuracy, recall, F1 Score, and other metrics. The exam assesses how well a classification-based machine learning model performs. The model's efficacy is displayed together with its accuracy, recall, F1 score, and support metrics. It gives more information about the trained model's overall performance. One has to be acquainted with all of the metrics offered in the research in order to understand the categorization results from machine learning models. For each approach, Table 4.6 displays the classification report, which includes Precision, Recall, F1-Score, and Accuracy%.

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	

4.1 Result Analysis

It is time to do the result analysis after calculating every conceivable factor, including Precision, Recall, F1-Measure, Accuracy, etc. In this analysis, it will be determined which algorithm performs the best overall and which algorithm performs the worst when compared to others.

4.4.1 Accuracy

The accuracy gauges an algorithm's optimal performance. Based on the facts given to it, how well it operates. The performance may be measured accurately using a probabilistic approach. Extreme Gradient Boosting is the most accurate of all techniques, whereas Gaussian Naive Bayes is the least accurate of them all. Extreme Gradient Boosting has shown to be a very effective and scalable implementation of gradient boosting machines that can push the computing capacity of boosted tree algorithms to their physical limits. It was designed and constructed primarily to improve model performance and speed up computer data analysis. Figure 4.1 and Table 4.7 provide the accuracy graph and percentage for each forecasting technique employed in this model.

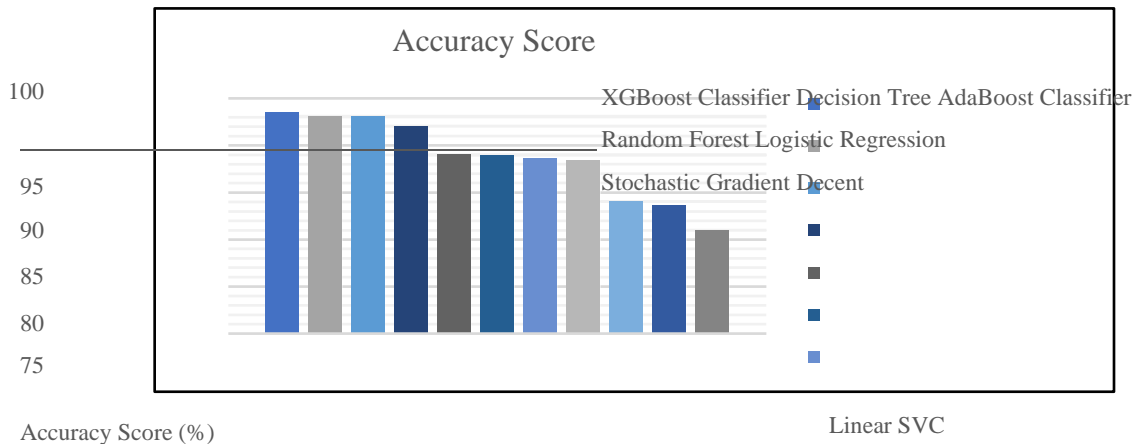


Figure 4.1: Accuracy Chart

4.4.2 Jaccard Score

A metric called the Jaccard score is used to assess how similar and diverse a sample is. In terms of the ratio of intersections to unions, they are on par with one another. The Jaccard coefficient, where is the size of the intersection and is the size of the union, may be used to compare the similarity of two finite sample sets quantitatively. The accuracy charts and percentages for the various strategies used to produce predictions in this model are shown in Table 4.7, Figure 4.2, and Equation (xiv)

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

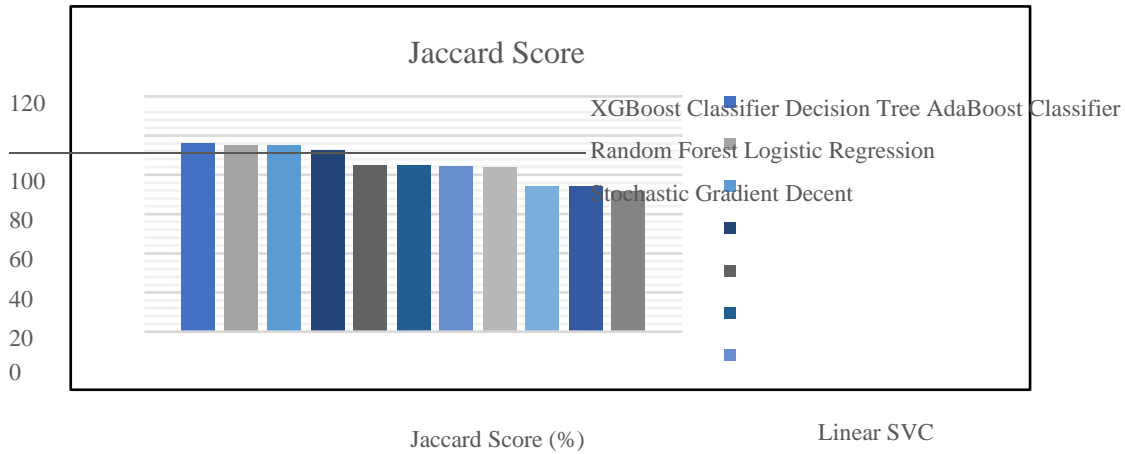
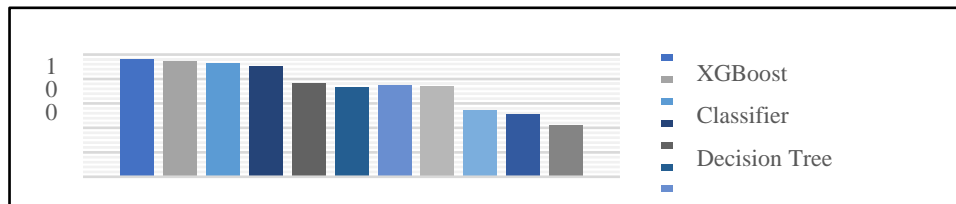


Figure 4.2: Jaccard Score Chart

4.4.3 Cross Validated Score

When paired with machine learning technologies, this efficiency metric can be used to estimate the breadth of a system's possible classifications. The AUC may be determined by comparing the number of instances where the model performs better than its peers with the number of instances where it performs worse. One of the four possible values for this integer is the highest possible value. The numbers lie on a scale from 0 (the lowest possible value) to 1 (the highest possible value). As illustrated in Figure 4.4 and Table 4.7, models with no predictions have an accuracy of zero, whereas models with flawless predictions have an accuracy of one..

Figure 4.3: Cross Validated Score



4.4.4 AUC Score

It is feasible to ascertain the potential categorization levels for a system using this performance metric and machine learning approaches, which may be useful in a variety of circumstances. AUC is calculated by contrasting a percentile of randomly selected positive instances, where the model performs notably better, compared to a percentile of randomly selected negative examples, where the model performs noticeably worse. Up to four alternative values for this integer are possible, with one being the most likely. The potential values fall within the range of 0 to 1, with 0 being the lowest. According to Figure 4.4 and Table 4.7, models with a total error in their predictions have an accuracy of zero, while models with a total success rate in doing so have a value of one.

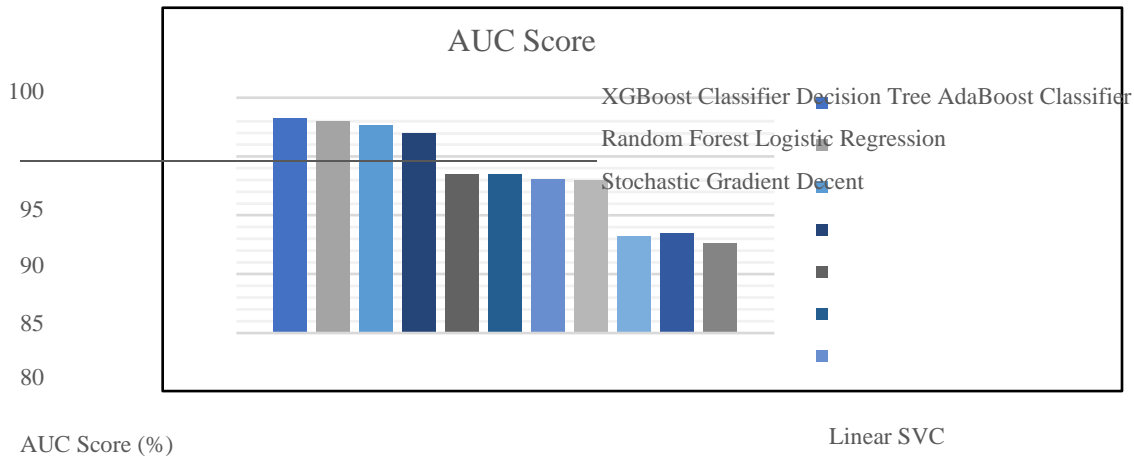


Figure 4.4: AUC Score Chart

4.4.5 ROC Curve

Important for determining how well diagnostic tests perform, ROC analysis is also a statistical model for classifying individuals into healthy and diseased categories. As a simple graphical aid, ROC curve analysis is excellent for showing the reliability of a medical diagnostic procedure. Figure 4.5 displays the final score determined by the key curve. shown.

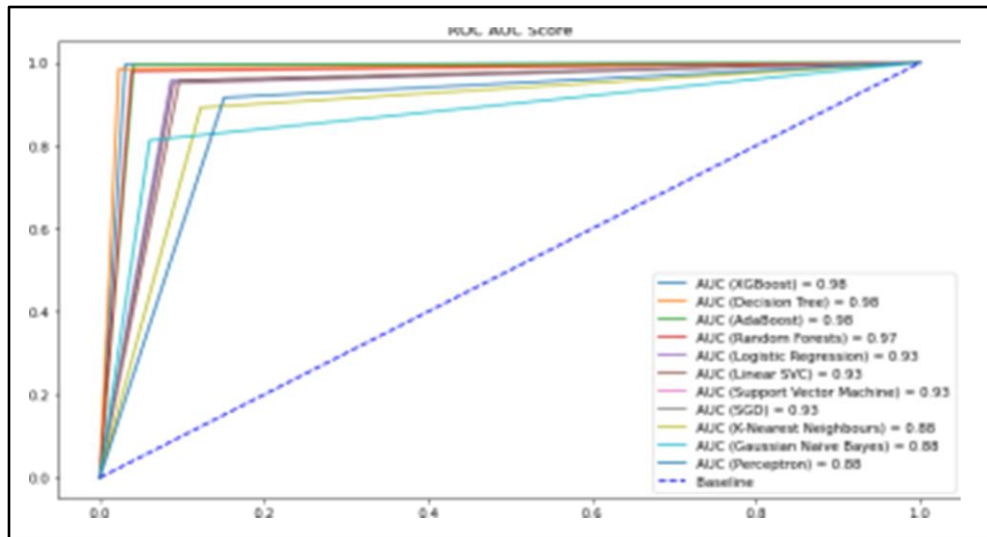


Figure 4.5: ROC Curve

After doing a comprehensive examination, the model's XGBoost Classifier had the highest accuracy, with scores of 98.55% accuracy, 96.14% Jaccard, 98.97% crossvalidation, and 98.21% AUC. Table 4.7 provides a brief explanation of 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

The standard deviation may also be calculated using the results of this investigation. The variation of a dataset from its mean is measured by a standard deviation. The square root of each data point's fluctuation is used to compute the standard deviation. The standard deviation increases as the data points vary from the mean. The standard deviations for the top five techniques are shown in 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

Error is a barrier when judging whether an algorithm is right. A machine-learning model's accuracy is determined by mean absolute error and mean square error following misclassification. The selection of an inappropriate characteristic might lead to misclassification. Misclassification occurs when all classes, groups, or categories of a variable have the same error rate. Absolute error is the term used to describe the degree of measurement error. The average of all measurement's absolute errors is known as the mean absolute error (MAE). The Mean Absolute Error formula is given in Equation (xv).

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

A regression line's mean squared error indicates how closely a line is fitting a group of points (MSE). Equation displays the Mean Squared Error formalised calculation (xvi)

Table 4.9 shows the algorithms' misclassification, mean absolute error, and mean square error. With 1.45%, 1.45%, and 1.45% respectively, the XGBoost had lower error rates for misclassification, mean absolute error, and the error's mean square

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

CHAPTER 5

IMPACT ON SOCIETY & SUSTAINABILITY

5.1 Introduction

After a project has begun, its effects on the public should be carefully evaluated. This section focuses on three areas where the Chronic Kidney Disease project has had an impact. The positive effects of the program are outlined in the following section. Then there are moral factors to think about. In this case, we looked closely at the ethical implications to see how the patients could benefit from this project. After the project was finished, its long-term viability was analyzed as well. Plans for the project's future growth and its capacity to assist additional individuals are currently being addressed.

5.2 Impact on Society

The societal significance of this research is substantial . Today, everyone has a busy schedule, so checking in at the hospital to make sure no one is ill takes a lot of time. Anyone may input the necessary details to foresee renal illness and obtain a timely result thanks to the development of a user interface Nowadays, going to the hospital for testing is perilous since COVID-19 is a risk that affects individuals worldwide. If anyone could review their report at home, it would be really beneficial to us. About this research's efficacy, clinicians and patients are now being surveyed. And it is almost certain that the survey's results will reflect favorably on the influence on society. And it is almost certain that the survey's findings will have a favorable effect both on society and on patients and physicians.

5.3 Ethical Aspects

People can get the testing they need without having to go to a specialized facility. They might read up on the basics of renal illness at home. They can foresee their own futures. The model's accuracy will rise over time if a web interface is created, where all data will

be saved One is a machine learning experiment, therefore it's not wise to put all your eggs in this particular basket just yet. No machine can make an accurate prediction.

Putting up the effort will take some time. In the future, when there are millions of data points in the database, the model will be more robust and, maybe, more accurate. When that time comes, if ever, Visits to hospitals for diagnostics won't be essential if artificial intelligence and the internet of things can be connected with databases. One can predict chronic kidney illness at home, saving time and money on doctor visits. A portable device that can test blood and urine would allow for more precise at-home testing of CKD. Sustainability Due to its new methodology, this study will be useful for a very long time because anyone can use a website to evaluate their health for Chronic Kidney Disease. The results of this investigation could be improved greatly with the help of future developments in deep learning, AI, and the Internet of Things. In light of these findings, numerous further sustainability-related studies may be conducted in the near future. It's also doable to make a website or a mobile app. The only kind of kidney illness predicted in this article is chronic. However, this machine learning study may be used to foresee a variety of illnesses that impact the kidneys and other organs. Even while submitting a picture of the skin via a web interface, skin diseases can be diagnosed. In the future, sending a snapshot of the skin using a web interface may be able to identify skin diseases..

5.4 Sustainability

Since of its novel methodology, this study has a very long shelf life because anybody may use a website to assess their level of Chronic Kidney Disease. Future opportunities with deep learning, AI, and the Internet of Things may provide this study endeavour with more precise outcomes. As a result, various activities may be done utilising this work in the case of sustainability in the future. Even a website or a mobile app may be made. Only chronic kidney disease is anticipated in this article. However, additional diseases affecting the kidney and other organs may also be predicted using this machine learning approach. Skin illness may potentially be diagnosed in the future by submitting a photograph of the skin using a web interface..

CHAPTER 6

FUTURE SCOPE & CONCLUSION

6.1 Introduction

In this chapter, the problem's scope was covered. Explain how this work can pave the way for the organization's continued and fruitful expansion in the future. Possible techniques of the future that could be used to build a superior device. At the chapter's end, you'll find the neat and tidy summary that you've been looking for. A bibliography is located at the end of this chapter.

6.2 Implication for Further Study

Regardless of location or degree of difficulty, the concept may serve as the foundation for a website for any hospital treating renal illness. Internet of Things-based websites could be developed and made available to the public at some point in the near or distant future. As long as you have access to the information required to diagnose CKD, you can easily make an at-home prediction of the disease. Once this is done, the individual can know for sure if they have CKD. Since the platform is based on the IoT and users will regularly upload new data, the website will retain the information. Over time, the model will get increasingly precise as it absorbs and applies new information. Some research suggests that enhancing a Deep Learning approach by using AI and Neural Network algorithms can improve its performance over time

6.3 Recommendations

Every diagnostic component has a normal range, and the system will know to look into it further if an aberrant value is discovered. Several variables, including serum creatinine, creatinine clearance, urine albumin, hemoglobin (Hgb), and others, play an important role in predicting chronic kidney disease. For males, the normal range of serum creatinine is 0.7–1.3 mg/dL and for women, 0.6–1.1 mg/dL. In healthy men and females, respectively, creatinine clearance ranges from 97 to 137 millilitres per minute. . An abnormally high protein level in the urine may point to a renal disorder. Determining whether or not protein

can be found in your urine is crucial. Normal readings vary from below 0 mg/dL to just above 8 mg/dL. An increase in urine protein could be a sign of renal trouble. It's crucial to know if your urine contains any protein. Women should have a Hgb between 12 and 16 g/dL, while men's should range from 14 to 18 g/dL. The onset of chronic renal illness may be indicated by a change or widening of the ranges. Therefore, preventative measures can be taken before any damage occurs. in the form of making adjustments to one's diet, physical activity routine, water intake, etc. It is imperative to listen to the physician's orders in such a situation.

6.4 Conclusion

Chronic kidney disease (CKD) has the potential to be effectively and efficiently controlled, and the patient may be cured in a short amount of time if it is identified and treated early enough. A series of laboratory tests is used to determine the presence of chronic renal disease (CKD). Theoretically, such verifications may take a long time to complete in practise. According to the study's authors, a model that has been trained on a relevant dataset may be able to properly forecast every stage of the development of chronic renal disease. Users may obtain their CKD report by filling out a form using an interface developed by the authors for this project and entering the required data. In order to select the best model that matches the dataset, eleven machine learning algorithms are trained on the data and evaluated based on their accuracy, Jaccard score, Cross Validated score, and AUC score. The three distinct scores that are computed are Accuracy, Jaccard, and Cross Validated. Overall performance-wise, XGBoost outperforms other classifiers by a wide margin. The majority of the time, this is the best course of action. After a website for this project is ready, it will be available to any renal disease hospitals. Since CKD results will be stored digitally and made available online around the clock, seven days a week, getting a copy won't even necessitate leaving the house

REFERENCES

- [1] Bangladesh's renal disease (n.d.). Life expectancy worldwide. September 19, 2021, retrieved, from <https://www.worldlifeexpectancy.com/bangladesh-kidney-disease>
- [2] You and COVID-19's Health (2020, February 11). Department of Health and Human Services. <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html>
- [3] Sun, T., Liu, L., Ma, F., and Jing, H. (2020). employing a heterogeneous adjusted artificial neural network based on deep learning to detect and diagnose chronic kidney disease. *Future Generation Computer Systems* 111, pgs. 17–26.
- [4] Hussain, M., Hussain, J., Satti, F. A., Ali, S. I., Bilal, H. S. M., Hussain, M., and Lee, S. (2020). An example of diagnosing chronic kidney illness in a developing nation using ensemble feature ranking for cost-based non-overlapping groupings. *IEEE Access*, 8(215622–215648).
- [5] S. Balakrishnan (2020). Feature Selection on the Chronic Kidney Disease Dataset Using Improved Teaching Learning Based Algorithm. 1660–1669 in *Procedia Computer Science*.
- [6] Arulanthu, P., E. Perumal, and J. R. Lambert (2020, July). Finding Nominal Attributes for Intelligent Classification of Chronic Kidney Disease Using Optimization Algorithm. 2020 will see the ICCSP (International Conference on Communication and Signal Processing) (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., and M. Hendryx (2020). Kidney function and metal mixtures: A machine learning application to NHANES data. 110126 *Environmental Research*, 191.
- [11] N. Sambyal, P. Saini, & R. Syal (2020). Type-2 Diabetes Microvascular Complications: A Review of Statistical Methods and Machine Learning Models. 115(1), 1–26. *Wireless Personal Communications*.
- [12] K. Harimoorthy, M. Thangavelu, and others (2021). In a healthcare monitoring system, a multi-disease prediction model employing enhanced SVM-radial bias approach. *The Ambient*

Intelligence and Humanized Computing Journal 12(3), 3715–3723. Computing.

[13] G. Chen, C. Ding, Y. Li, X. Hu, X. Li, X. Ren, & W. Xue (2020). Using an adaptive hybridized deep convolutional neural network on the Internet of Medical Things platform, predict chronic kidney disease. *IEEE Access* 8, 100497-100508.

[14] Among the authors are Reddy, D. J., Mounika, B., Sindhu, S., Reddy, T. P., Reddy, N. S., Sri, G. J., and Kora (2020). An early diabetes analysis and diagnosis algorithm using predictive machine learning. *Resources for 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] Shahbaaz, M., Islam, T., Sakib, N., Yashfi, S. Y., Islam, M. A., and Pantho, S. S. (2020, July). Machine Learning Algorithms for Risk Prediction of Chronic Kidney Disease. 2020 will be the eleventh iteration of the International Conference on Computing, Communication, and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[17] & Suchetha, M. Navaneeth & (2020). a convolutional neural network method based on dynamic pooling for the detection of chronic renal disease. 62, 102068; *Biomedical Signal Processing and Control*

[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] The authors are Abdelaziz, Salama, Riad, and Mahmoud (2019). An internet of things- and cloud-based machine learning approach for chronic kidney disease prediction in smart cities. *Models, Applications, and Challenges of Security in Smart Cities* (pp. 93-114). Cham Springer.

[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.

final ckd

ORIGINALITY REPORT

29%	29%	7%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	26%
2	dokumen.pub Internet Source	<1%
3	tudr.thapar.edu:8080 Internet Source	<1%
4	www.um.edu.mt Internet Source	<1%
5	Tanuja Sudhakar, Marina Gavrilova. "Deep Learning for Multi-instance Biometric Privacy", ACM Transactions on Management Information Systems, 2020 Publication	<1%
6	Juhua Luo, Michael Hendryx. "Metal Mixtures and Kidney Function: An Application of Machine Learning to NHANES Data", Environmental Research, 2020 Publication	<1%
7	core.ac.uk Internet Source	<1%