

**CHRONIC KIDNEY DISEASE RISK CLASSIFICATION USING MACHINE  
LEARNING**

**BY**

**Md. Takrimuzzaman**

**ID: 152-15-6130**

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

**Supervised By**

**Dr Sheak Rashed Haider Noori**  
**Professor & Associate Head**  
Department of CSE  
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**JANUARY 2023**

## APPROVAL

This Project/internship titled “Chronic Kidney Disease Risk Classification Using Machine Learning”, submitted by **Md. Takrimuzzaman, Student ID: 152-15-6130** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment 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 date.

### BOARD OF EXAMINERS

Chairman

\_\_\_\_\_  
**Dr. Touhid Bhuiyan**

**Professor and Head**

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

\_\_\_\_\_  
**Sazzadur Ahmed**

**Assistant Professor**

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

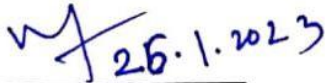
\_\_\_\_\_  
**Ms. Sharmin Akter**

**Senior Lecturer**

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



External Examiner

\_\_\_\_\_  
**Dr. Ahmed Wasif Reza**

**Associate Professor**

Department of Computer Science and Engineering

East West University

## DECLARATION

This project aims to use machine learning techniques to predict and classify chronic kidney disease (CKD) to improve early detection and treatment of the condition. CKD is a severe and often asymptomatic disease that can lead to permanent kidney damage, dialysis, or transplantation if not adequately managed. Machine learning algorithms can improve CKD diagnosis and treatment accuracy and speed by analyzing patient data and identifying key risk factors. The authors of this project have reviewed previous research on machine learning for CKD prediction and classification and propose to conduct their study using various machine-learning approaches. They hope to identify the most effective algorithms and variables for predicting and classifying CKD and contribute to the medical data analysis and disease prevention field.

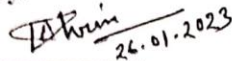
Supervised by:



---

**Dr. Sheak Rashed Haider Noori**  
Professor & Associate Head  
Department of CSE  
Daffodil International University

Submitted by:



---

**(Md. Takrimuzzaman)**  
ID: 152-15-6130  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

I want to express my sincere gratitude to all those who have supported me throughout my BSc thesis journey.

Firstly, I would like to thank my supervisor, **Dr Sheak Rashed Haider Noori**, for their invaluable guidance, support, and patience. Their expert knowledge and constructive feedback have been instrumental in completing this thesis. I am also grateful to my thesis committee members for their valuable input and insights.

I would also like to extend my gratitude to my family and friends for their unwavering support and encouragement. Their belief in me and constant motivation have kept me going through the highs and lows of this journey.

I would also like to express my appreciation to the participants of this study, who generously gave their time and energy to contribute to my research. Without their valuable input, this thesis would not have been possible.

Finally, I would like to thank the Daffodil International University community for providing a stimulating and supportive environment for research and learning. Thank you all for your invaluable support.

## **ABSTRACT**

Chronic kidney disease (CKD) is a serious and common health issue affecting millions worldwide. Early detection and treatment of CKD can prevent or delay the need for dialysis or transplantation, which can significantly improve patient outcomes. In this study, we used the UCI data repository to classify the risk of CKD using machine learning techniques. We balanced the dataset using the ADASYN technique to ensure that it was representative of the population. We then evaluated the performance of three algorithms: Random Forest, Naive Bayes, and CatBoost. Our results showed that all three algorithms had high accuracy, with Random Forest and Naive Bayes achieving 99.60% and CatBoost achieving 99.21%. Additionally, all three algorithms had a precision of 1, and the highest recall value was 99.23% for Random Forest. The F-1 score was also highest for Random Forest at 99.60%. Finally, the ROC\_AUC score was 1 for all three algorithms, indicating that they could effectively distinguish between high and low-risk individuals. These results suggest that machine learning can be a powerful tool for classifying the risk of CKD. Further research is needed to validate these findings and to develop more advanced machine-learning techniques for the early detection and treatment of CKD.

# TABLE OF CONTENTS

CHAPTER 1 .....	1
INTRODUCTION .....	1
1.1 Introduction .....	1
1.2 Motivation .....	4
1.3 Rationale of the Study .....	6
1.4 Research Questions .....	6
1.5 Expected Outcome .....	7
1.6 Report Layout.....	7
CHAPTER 2 .....	8
BACKGROUND STUDY .....	8
2.1 Related Works .....	8
CHAPTER 3 .....	21
RESEARCH METHODOLOGY .....	21
3.1 Proposed method .....	21
3.2 Data Pre-processing.....	22
3.3 Classification Methods .....	27
3.4 Classification Report Performance Measurement Criteria .....	29
CHAPTER 4 .....	30
RESULT ANALYSIS.....	30
4.1 Random Forest .....	30
4.2 Naïve Bayes.....	34

4.3 CatBoost.....	38
CHAPTER 5 .....	42
DISCUSSION.....	42
CHAPTER 6 .....	43
CONCLUSION & FUTURE WORK.....	43
REFERENCES .....	45

## LIST OF FIGURES

Figure 1: Proposed Methodology.....	22
Figure 2:ADASYN Oversampling Technique.....	25
Figure 3: Dimensionality Reduction using PCA .....	26
Figure 4: 5-fold Mean Test Accuracy for RF .....	31
Figure 5: 5-fold Mean Test Precision for RF.....	31
Figure 6: 5-fold Mean Test Recall for RF .....	32
Figure 7: 5-fold Mean Test F-1 score for RF.....	32
Figure 8: 5-fold Mean ROC-AUV Score for RF .....	33
Figure 9: 5-fold Mean Test Accuracy for NB.....	35
Figure 10: 5-fold Mean test Precision for NB .....	35
Figure 11: 5-fold Mean Test Recall for NB.....	36
Figure 12: 5-fold Mean Test f-1 Score for NB .....	36
Figure 13: 5-fold Mean ROC-AUC Score for NB.....	37
Figure 14: 5-fold Mean Test Accuracy for CatBoost .....	39
Figure 15: 5-fold Mean Test Precision for CatBoost.....	39
Figure 16: 5-fold Mean Test Recall for CatBoost .....	40
Figure 17: 5-fold Mean Test f-1 Score for CatBoost.....	40
Figure 18: 5-fold Mean ROC-AUC Scores for CatBoost.....	41
Figure 19: Performance Graph of All Methods in Each Measurement Scale.....	42



## LIST OF TABLES

Table 1: Performance Measurement Criteria.....	29
Table 2: Result Analysis for RF.....	30
Table 3: Result Analysis for NB.....	34
Table 4: Result Analysis for CatBoost.....	38
Table 5: Performance Table of All Methods in Each Measurement Scale.....	42

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

In Bangladesh, the chronic kidney disease (CKD) problem is proliferating that needs to resolve, and forties rapid growthmarks as one of the public health problems. So, it is the crying need of the country to have adequate information on the pattern of kidney and renal diseases and to take the necessary prevention program to fight against kidney diseases. Because of the increase in population and the corresponding rise in nephrologist needs, we appreciate more and more each day that we are fightinga losing battle with a mere emphasis on therapy; this is neither adequate nor progressive. Prevention, not cure, is today's challenge and, hopefully, tomorrow's achievement [1]; the principal causes of poor renal health are shared by those responsible for chronic diseases; firstof all, these factors are related to poor diet, adopting the unhealthy lifestyle, excessive use of alcohol and physical inactivity [2].

On the other hand, chronic renal disease refers to kidney damage brought on by lifestyle-related variables (CKD). Due to the hectic nature of modern lifestyles, many individuals fail to recognize the subtle changes in their health that have been linked to exposure to environmental changes [3]. Chronic renal disease has numerous risk factors, including inadequate water intake, smoking, an unhealthy diet, and insufficient rest [4]. According to the study's authors, the danger of kidney impairment is most remarkable for people with diabetes. Unlike most diseases, which are often identified early, CKD is typically discovered in its severe stages [5]. Because renal failure is the last stage of the disease, therapy is dangerous and costly. This work proposes a system for predicting the absolute risk of CKD in healthcare, considering all the symptoms and factors contributing to this condition. The signs will serve as identifiers for the different stages of renal illness. A patient's medical history may be classified according to the various stages of renal disease. Classifying patients facilitates the identification of CKD's most significant characteristics. The progression of CKD can be slowed or halted by treating its essential characteristics.

Additionally, the kidney performs a vital role in the human body through the urinary system by eliminating wastes and poisons. Chronic kidney disease (CKD) affects 1.3 billion individuals in India [6]. It causes a gradual decline in renal function and harms the kidneys. However, CKD is difficult to predict because its symptoms develop gradually and are not condition-specific, making early disease detection essential. The kidneys are responsible for filtering waste and excess fluids from the blood, with the resulting waste then being sent to the digestive tract. A few symptoms will manifest in the early stages of chronic kidney disease [7].

In doing so, it generates a unique and functioning pathway that rids the body of toxic and useless molecules while returning nutrients, amino acids, glucose, hormones, and other bioactive chemicals into the bloodstream. Things will, unfortunately, deteriorate. In some areas, recurrent nephritis is known as "chronic kidney disease" (CKD) [8]. Any deviation from normal in the structure or boundaries of the urinary organs is classified as an "infection," regardless of whether it would cause a man to feel ill or cause problems. This is a widespread issue that can harm individuals of any age. About three million people in the United Kingdom are at high risk for having chronic kidney disease. CKD is brought on by many circumstances, all of which place an extra strain on the kidneys [9].

Due to the unawareness of the causes that lead to kidney disease and not having the proper knowledge to avoid this problem, people are making wrong decisions and not taking adequate prevention to prevent this situation [10]. This project aims to predict the risk of kidney disease so that people come to know which habitant may cause chronic kidney disease. And to find these criteria, we take the help of machine learning to predict which habitant is more dangerous and which one we should emphasize most to prevent kidney disease.

Moreover, analysis of big healthcare data is necessary for the twenty-first century due to the vast amounts of data's variety, pace, integrity, and volume. This information is collected in real-time from hospitals and clinics throughout the world. Clinicians, payers, patients, and administration might benefit from applying healthcare analytics to big data. Large hospitals collect terabytes of health data via their various electronic health records, which can be used to predict whether patients are at high risk of getting a disease and to improve

the accuracy of their diagnoses. Chronic kidney disease is an issue in healthcare that the world is currently confronting. Data-driven chronic kidney disease (CKD) can be alleviated through informed decision-making throughout the healthcare industry.

Consequently, the manuscript is formatted as follows: Section II will examine a selection of research that covers similar territory. Section III discusses the suggested system for categorizing chronic renal illness (CKD). In Section IV, we describe the dataset, its characteristics, and its contents. Section V discusses the various machine learning algorithms utilized in the classification operations, their associated code, results for variable measures, and output. In addition, Section VI presents a free-flowing discussion of recent effects and viewpoints on chronic disease. In Section VII, research outcomes and attribute upgrades are discussed.

## 1.2 Motivation

In Bangladesh, the term chronic kidney disease is prevalent. Most people in Bangladesh think kidney disease has to do with the problem of fluid filtration, but the nephrological imbalance is also a part of kidney disease. Chronic kidney disease deals with the general condition of kidney failure, imbalance of liquid filtration, end-stage-renal disease, chronic glomerulonephritis, diabetes, hypertension, and associated structures [11]. At the same time, these people may develop complications like high blood pressure, anaemia, nerve damage, weak bones, and poor nutritional health [12].

In Bangladesh, it has been studied that more than 22% of the population has been affected or suffers from chronic kidney disease (CKD). If we convert it to numbers, we find that studies with a total of 225,206 participants revealed that the overall prevalence of CKD in Bangladeshi people based on meta-analysis was higher than the global prevalence of CKD [13]. Most of them suffer from end-stage renal disease, chronic glomerulonephritis (40%), diabetes (34%), hypertension (15%), etc. [14]. Surveys in a few rural, urban, and disadvantaged populations suggested that 18 million people have been suffering from CKD, as defined by Bangladesh's kidney disease outcomes quality initiative (KDOQI) [14]. About 30,000 patients reach end-stage renal failure every year in this country; they need dialysis or kidney transplantation [14]. Of 18% of kidney patients, 11% have milder to more severe kidney failure [14].

Increased CKD awareness over time in different countries and a recent increase in nephrology referrals suggested that these efforts may have some positive impact [15] [16] [17]. It has been observed that physicians other than nephrologists are less likely to recognize CKD and sometimes differ in their clinical evaluation of CKD [18]. Many CKD patients are referred to nephrologists much later than it would have been appropriate [19]. Late evaluation of CKD patients by nephrologists, especially those presenting with end-stage renal disease (ESRD), is associated with suboptimal pre-dialysis care and treatment, which ultimately increase mortality [19] [20] [21] [22].

In cities nowadays, people are growing concerned about these diseases, but in rural areas, people are comparatively ignorant regarding basic hygiene. They often consume lots of sugar, meat, and polluted water to suppress hunger and thirst. They don't even care about

their health at all. Also, there is no facility for any nephrological treatment in a rural area; people who live in a rural area must go to the nearby Upazila for health care issues. Even after that, providing conservative and prosthetic nephrological treatment to the village people is challenging, so most people can't get enough therapy after having a kidney problem.

Recent studies show that 2500 nephrologists have been registered for a population of 200 million. Therefore, it has been calculated that the ratio of dentists to population is 80000:1 [23]. So for that reason, most of the people in Bangladesh, especially in rural areas, are not getting proper treatment, so they often have to rely on traditional (unqualified) nephrologists. In conventional medicine, the most common practice is different rituals and belief in superstitions for the sudden worst health condition. If someone wants modern treatment, they need to move to larger cities, which is very difficult for people who live in rural areas. According to the World Health rankings (WHR) and WHO data published in 2020, kidney disease death in Bangladesh reached 10 841 or 1.51% of the total deaths reported. The age-adjusted death rate of 9.14 per 100,000 population ranks Bangladesh as number 143 globally [24].

On the other hand, national-level information about the distribution of oral diseases in Bangladesh kidney disease deaths almost tripled in Bangladesh last year compared to 2019, while deaths from other conditions, including brain stroke, breast cancer, and diabetes, increased significantly. The vital statistics survey of the Bangladesh Bureau of Statistics (BBS) found that 28,017 people had died from kidney disease in 2020, up from 10,622 a year before [25]. On the other hand, nephrological medicine is very costly. Hence, companies that manufacture nephrological drugs and materials need to charge each nephrologist a lot for the companies to recuperate the cost of research and development.

Therefore, we will try to anticipate if we tend to become affected by kidney disease. We have seen much research, but in terms of significance, it was not decent in this field, but some issues regarding this topic need to be fixed very soon. We will do this risk prediction using the machine learning technique in that context.

### **1.3 Rationale of the Study**

As mentioned earlier, less significant work has been done previously with chronic kidney disease risk prediction and classification from a Bangladeshi perspective. We are interested in working with the risk classification and prognosis of chronic kidney disease and machine learning techniques. Machine learning is a branch of artificial intelligence that employs various statistical, probabilistic, and optimization techniques that allow computers to "learn" from past examples and detect hard-to-discern patterns from large, noisy, or complex data sets. Machine learning methods are used in many applications ranging from detecting and classifying. Machine learning is used for cancer prediction [27], a systemic review of software fault prediction [26], dermatological disease detection [28], and so on. Many types of detection and risk prediction are now conducted using machine learning. Machine learning techniques may have a supplementary role in highly complex problems and provide a comparison to regression results. [29]. As machine learning has a vast field of work, we thought we should apply machine learning for our prediction work.

### **1.4 Research Questions**

- Is it possible to determine who has chronic kidney/renal disease?
- What risks, if any, contribute to kidney/renal failure?
- How closely will our present records mirror our past practices?
- Is training the system with raw data required when utilizing a machine learning model?
- How much data do we collect, and where do we obtain it?
- Do our data appear compatible with machine learning?
- When it comes to machine learning, is it preferable to apply established approaches or to experiment with something new?

## **1.5 Expected Outcome**

Our research intends to aid in the classification and diagnosis of chronic renal illness (CDK). With this procedure, the risk of developing chronic kidney disease can be identified quickly and easily. Using machine learning, and individuals can also increase their predictive abilities. Use extant or newly created machine learning algorithms to predict and categorize chronic renal illness.

When kidney disease is mentioned, most people immediately consider drinking more water than usual. Access to more drinkable water is a tremendous benefit, but that is not the only factor to consider. What we eat and how we live significantly impact kidney disease. Bengali children, in particular, have a sweet taste and are unconcerned with its repercussions, a problem that should be the parents' duty given that they are children; nonetheless, most Bengali parents are just as irresponsible as their children. People in rural areas are more likely to be addicted to unhealthy foods and to neglect the kidney-damaging effects of this addiction. In major urban centres, adolescents have an unhealthy obsession with soft drinks. Urbanization of sleeping habits is another significant concern. The fundamental objective of our technique is to identify particular instances in which individuals disregard their regular diet and lifestyle.

## **1.6 Report Layout**

The following outlines this research paper's contents: The first chapter includes an overview of the research and its context, as well as an explanation of the study's motivation, central questions, and preliminary results. The second chapter focuses on the duties at hand, a concise summary of the pertinent research, and the entire breadth of the challenges. Methods for collecting data, developing features, and doing statistical analysis are described in detail in chapter three. In Chapter 4, numerical and graphical results from the research are discussed, and their importance to the experimental evaluation. The public implications of these findings are examined in Chapter 5. Chapter 6 provides an overview of the study, including its scope, limits, and ideas for further research.



## CHAPTER 2

### BACKGROUND STUDY

#### 2.1 Related Works

Implementing machine learning to improve medical diagnoses and patient outcomes is difficult. In recent years, a broader array of industries has adopted the strategy. Various research on CKD have focused on patients at high risk for cardiovascular disease, but these studies have employed different diagnosis and evaluation criteria. In the literature review section of this work, recent efforts by other researchers in the field of CKD prognosis and risk categorization will be discussed. We've been monitoring and researching their work to appreciate better the concepts and methods it delivers. In part devoted to related work, we present a quick summary of pertinent research articles, works, underlying algorithms, classifiers, and claims of accuracy; in the section entitled "Study Summary," a concise summary of relevant literature is compiled and presented in a table for convenient reference. In this examination, we have primarily classified the efforts of numerous researchers into two categories. One is the implications of a deep learning model for risk prediction and categorization of chronic kidney disease (CKD), and the other is the representation of these processes by a machine learning model. In the "Scope of the Problem" section, we describe how our approach to the problem can be of general assistance. In the final section, "Challenges," we describe briefly the obstacles and potential threats faced during this inquiry.

At the beginning of this literature review, chronic kidney disease is permanent kidney damage caused by renal pathology or impaired kidney function. Predicting the onset of this chronic illness and delivering the proper treatments can prevent or postpone the need for dialysis or a kidney transplant. In this study, Ahmed et al. [31] assessed the effectiveness of various machine-learning approaches for the early identification of a chronic renal illness. The authors of this methodology support it with data from a predictive analytics study that examined the link between various data attributes and the desired feature for the target class. The authors introduced the optimal set of variables for machine learning training using predictive analytics and developed a suite of prediction models. The best

subset for predicting chronic renal illness was identified to consist of 30% of the original 24 indicators and the class property. In a supervised learning environment, four machine learning-based classifiers were tested, with the best producing an area under the curve (AUC) of 0.9955, a sensitivity of 0.9897, and a specificity of 1. The results of the experiment indicate that the use of predictive analytics and other forms of machine learning has facilitated the identification of intelligent solutions, demonstrating the validity of prediction outside the field of renal illness.

Additionally, Calciphylaxis is a deadly condition characterized by necrotic skin sores. Due to the rarity and intricacy of the situation, the risk factors and pathophysiology of calciphylaxis remain unknown. Ross et al. [32] focused on the application of machine learning to anticipate sickness risk and model the pertinent features extracted from an electronic medical record data set. The precise mechanisms underlying calciphylaxis are still under investigation. Four modelling approaches were applied to various groups of patients diagnosed with chronic renal illness (CKD). The author's use of random forests trained on binary feature data to model calciphylaxis risk produces robust models. The AUC-ROC for predicting the onset of calciphylaxis in patients with CKD stage 4 was 0.8718. If the prediction models presented in this paper can reliably forecast calciphylaxis, they have the potential to be utilized in the real world.

When cutting-edge machine learning algorithms are applied to colossal health screening data collections, opportunities for clinical significance in human and animal medicine are established. To do this, Richard et al. [33] used data from electronic medical records (EMRs) gathered from standard veterinary practices to develop a model that predicts the chance of cats developing chronic renal disease (CKD). In the pre-clinical years of this research, it was evaluated using an independent data set after the model was built for use in clinical diagnosis. With a high specificity (>99%) and sensitivity (63%), the algorithm reliably predicts that 93% of cats with a 15% prevalence will not be diagnosed with CKD or will acquire the correct diagnosis over the next 12 months. A fundamental strength of the current approach is incorporating medical examination data collected as part of standard veterinary practice. This idea is readily incorporated into software in veterinary hospitals and diagnostic centres.

On the other hand, chronic kidney diseases are widespread among persons with higher cardiovascular risk; examining historical patient data can aid in preventing major problems. Mohamed et al. [34] evaluated historical computerized medical data on chronic renal disease using twelve supervised machine-learning algorithms. The initial sample size was 544 outpatients; however, 48 did not cut, and 21 instances had missing information, preventing their inclusion. Based on the preliminary data and patient profiles, it was established that 88.5% of the patients had "advanced CKD," and 11.5% had "early-stage CKD." Separating the two groups accurately was critical for the classification problem and the following model analysis. As a boosted decision tree, CN2 rule induction performed poorly compared to other techniques evaluated. By utilizing logistic regression, a neural network with logistic and stochastic gradient descent, or a support vector machine with a radial basis function and polynomial, extremely high accuracy and efficiency were achieved. The polynomial support vector machine technique was the most effective, with an efficiency of 93.4% and an accuracy rate of 91.1%. The model provided 253 2-dimensional permutations of variables, with a tendency for vascular problems and smoking being the most influential. Even though machine learning techniques were successfully used to study CKD, there may be flaws in the data that require correction—which is considered a shortcoming of the study.

People are more health-conscious than ever before, but they frequently only give their health a second consideration until a problem arises. However, CKD is a disease with no indicators and, in some circumstances, no illness-specific symptoms; it is difficult to anticipate, identify, and prevent such a disease, which could result in irreparable health damage; however, machine learning may offer a solution because it excels at prediction and analysis. Siddheshwar et al. [35] examine various machine-learning techniques in this paper. The author has reviewed fourteen distinct CKD patient features and calculated the predictive potential of machine learning approaches such as decision trees and support vector machines. The results indicate that the SVM obtains an accuracy of 96.75%, while the decision tree only reaches an accuracy of 91.75%. Consider the decision tree algorithm; it generates the tree using characteristics from the entire dataset. This method requires less time to make predictions, which is one of its benefits. Helping physicians diagnose more people promptly and reliably will allow them to begin CKD treatment sooner and improve

the prognosis for those currently afflicted with the condition. Due to the limited sample size and several missing attribute values, the data in this study are less reliable than they could be. It will require millions of data with no missing values to build a machine-learning model with 99.99% accuracy for detecting chronic renal disease.

All medical diagnoses, treatments, and risk evaluations are founded on the results of meticulously conducted clinical studies. However, there is already an abundance of medical data from the actual world, and further increases in data volume would diminish completeness, uniformity, and control. The publication by Stefan et al. [36] shows a case-by-case comparison that illustrates the superior forecasting accuracy of the simple data-based technique for diabetes-related chronic kidney disease over previously reported clinical research data-based algorithms. Before these findings can be extended, additional testing on various datasets is required. Still, the current study implies that training predictive and prescriptive algorithms using RWD may achieve comparable or even improved accuracy compared to those using clinical trial data. The author hypothesizes that the diversity of RWD makes the prediction algorithm more flexible for widespread application. Insofar as they suggest that RWD-driven risk assessments may one day supplement, or in some cases even partially replace, the exploratory evaluation and interpretation of information beyond the primary objectives of fundamental, costly, long-term clinical studies on a limited number of individuals, these findings may add fuel to the essential debate on the longevity of medical evidence.

Estimating the amount of protein in the urine is essential to diagnosing and monitoring chronic renal illness (CKD). Inconveniently, the current approach to analyzing the severity of CKD necessitates monitoring 24-hour urine protein levels. Using statistical, machine learning, and neural network techniques, Jing et al. [37] developed and analyzed many prediction models to rapidly identify chronic kidney disease (CKD) severity using the more widely available demographic and biochemical blood parameters during follow-up. Blood medical and biochemical findings were collected from 551 proteinuric individuals. The nine predictive models constructed and analyzed were logistic regression, elastic net, lasso regression, ridge regression, support vector machine, random forest, XGBoost, neural network, and k-nearest neighbour. The models' analytical utility, receiver operating characteristic, recall, specificity, accuracy, log-loss, and precision were measured. The

author analyzed and ranked all estimations of variables. AUC and precision were highest for linear models such as elastic net, lasso, ridge, and logistic regression. With an area under the curve (AUC) of 0.873 and sensitivities and specificities of 0.83 and 0.82, respectively, logistic regression achieved the highest degree of accuracy. XGBoost had the highest specificity (0.95), whereas Elastic net had the most heightened sensitivity (0.85). (0.83). Effect size investigations revealed that ALB, Scr, TG, LDL, and EGFR significantly affected the accuracy of the models. CRP, HDL, and SNA were also predictive, albeit to a lesser extent. However, there are limitations to this study. Weak tuning settings and a small sample size make overfitting more likely.

In outpatient situations, non-urinary predictors such as blood tests could be used. Blood test variables, such as ALB, Scr, TG, LDL, and EGFR levels, can indicate the severity of chronic kidney disease (CKD). During clinical follow-up, the built online tool can be utilized to predict the progression of proteinuria. Artificial intelligence (AI) is meant to aid in exercising clinical judgment in the medical industry. Masaki et al. [38] use vast amounts of data, machine learning, and artificial intelligence to evaluate the viability of predicting the onset of diabetic kidney disease. Using artificial intelligence (AI), speech recognition, longitudinal data, and extensive data mining, the author constructed a new predictive model for DKD using the electronic medical records (EMR) of 64,059 diabetic patients. AI utilized a convolutional autoencoder to extract natural characteristics from the preceding six months, which served as the reference period and then used these qualities to choose 24 parameters for identifying time series patterns linked with 6-month DKD aggravation. The predictive model was constructed by an artificial intelligence system using 3,073 features, such as time series data and logistic regression analysis. 71% of the time, AI accurately predicted DKD aggravation.

In addition, the rate of hemodialysis was significantly higher in the DKD aggravation group throughout ten years (N = 2,900) than in the non-aggravation group. The new AI predictive model could detect the development of DKD, which could lead to more precise therapies that reduce the need for hemodialysis. Despite its encouraging findings, this study has considerable limitations. First, the author could not standardize the data extraction method due to the enormous variance in the information acquired from each EMR, namely the records of medical practitioners. Second, there was no set interval between testing; it

differed between patients. No other electronic medical records (EMR) from other institutions were utilized to assess the results of this investigation. Because patients without DKD are likely to be treated less vigorously, the author could not identify an association between the progression of DKD over six months and the intensity of treatment. Because of these results, the author concludes that the medication did not likely decrease the progression of DKD in this trial.

Given the high death rate during the first year following filtration initiation, a more exact estimation of post-dialysis mortality could help patients and clinicians decide whether to begin dialysis. This study by Oguz et al. [39] aimed to use machine learning to predict persons at risk for short-term mortality following dialysis by using complex data from electronic health records. Participants in this trial were a group of 27,615 male US veterans recently diagnosed with end-stage renal illness (ESRD). To predict 30-, 90-, 180-, and 365-day all-cause mortality following the initiation of dialysis, the author utilized a random forest model using 49 pre-dialysis covariates. The participants had a mean (SD) age of 68.7 years, were primarily male (98.1%), disproportionately African American (29.4%), and diabetic (71.4%). The final random forest model produced C-statistics (95% confidence intervals) of 0.7185 (0.6994-0.7377), 0.7446 (0.7346-0.7546), 0.7504 (0.7425-0.7583), and 0.7488 (0.7425-0.7583) for predicting mortality risk over four-time frames (0.7421-0.7554). Comparable to or superior to rival ML algorithms, the models demonstrated high internal validity and consistency over a broad range of patient demographic and clinical factors. Due to the study's limitations, the results of this investigation should be regarded with caution. This prediction model was developed using data from a national sample of US veterans who developed incident ESRD and subsequently began dialysis. Consequently, female patients and those with less severe stages of the renal disease will not benefit from this treatment. On the other hand, an equal number of women were included in a second predictive model that utilized the same senior populations previously validated well in a separate cohort.

Shanila et al. analyzed data from CKD patients and provided a method for predicting the likelihood of CKD [40] to assist with this problem. The primary objective of this study was to predict CKD risk using a random forest method and an artificial neural network by analyzing existing CKD data. In this study, information from 455 patients has been utilized.

In addition to the Machine Learning Repository data collection, a real-time dataset from Khulna City Medical College is utilized. The author was able to construct the system using Python because it is a high-level interpreted programming language. The author trained the data using a 10-fold CV, Random Forest, and ANN. Compared to ANN's 94.5% accuracy, Random Forest considerably outperforms it. The proposed approach is anticipated to aid in the early detection of CKD risk. Other techniques, such as wrapping, would have made it much easier to zero in on pertinent information than the author's sole use of the Chi-square test. It does not do so, which is one of the study's shortcomings.

Despite the greater prevalence and incidence of chronic kidney disease (CKD), which is frequently the result of delayed diagnosis, it is a severe public health concern, especially in developing nations such as Brazil. Mortality and morbidity rates increase due to CKD treatment regimens such as hemodialysis and kidney transplantation, increasing the public's healthcare expenditures. Alvaro et al. [41] evaluated the application of machine learning algorithms to the early diagnosis of chronic renal disease in situations with limited resources in this study. Weka and a CKD dataset are utilized for qualitative and quantitative comparative analysis via a systematic literature review and machine learning experimentation (precisely, the k-fold cross-validation approach). Due to the obstacles people face in emerging countries, such as a lack of access to primary health care, this study assessed the suitability of machine learning methods. It addressed how software applications can aid in the early diagnosis of chronic kidney disease (CKD). Accordingly, this study was guided by two main questions.

- MQ 1: What is the excellent machine learning technique for CKD diagnoses in developing countries?
- MQ 2: What are the adequate attributes for CKD diagnoses in developing countries?

Examining the application of machine learning algorithms for CKD risk screening in low-resource and difficult-to-access settings in developing countries, such as inadequate primary health care, was the focus of these studies. Due to its easy-to-understand classification results and 95.00% accuracy, the J48 decision tree was determined to be a suitable machine-learning strategy for such screening in developing countries. This precision is practically identical to the assessment of a seasoned nephrologist. On the other hand, random forest, naive Bayes, support vector machine, multilayer perceptron, and k-

nearest neighbour techniques, in that order, achieve 93.33%, 88.33%, 76.66%, 75.00%, and 71.67% accuracy and show at least moderate agreement with the nephrologist, albeit at the expense of more challenging interpretation of the classification results. This study's strength rests in the fact that no other research has addressed the application of machine learning approaches to the problem of CKD diagnosis, particularly in the context of the unique constraints encountered by nations with less developed healthcare infrastructures. Due to the study's limitations, the results should be regarded with considerable caution. Initially, the nephrologists were required to evaluate each participant, a time-consuming task even with a small sample size of sixty. Second, there is the possibility of bias in the upgraded data; when completing this work, great care was taken to maintain the simulated individuals in the same CKD risk category. The J48 decision tree classifier avoids making false positive or negative judgments when used in the CKD dataset.

Additionally, using machine learning techniques, Bhavya et al. [42] established a new method for predicting chronic kidney disease (CKD). The primary purpose was to assess and rank the effectiveness of several machine-learning algorithms. The programming language R is used as a comparison model for this investigation. This study's primary objective was to examine and classify the chronic renal illness dataset. K-Nearest Neighbor, Logistic Regression, and Support Vector Machines are employed to investigate chronic kidney disease (CKD). The precision of several algorithms was used to measure their efficacy. The Support Vector Machine method beat logistic regression and K-nearest neighbours to predict chronic kidney disease in this particular medical situation. In addition to identifying larger populations in less time, our technique accelerates the prediction process, allowing patients with CKD to begin treatment sooner. This analysis primarily depends on the employed data set, yet it does have certain limits. The first potential issue is that the small sample size may damage the study's credibility (400 cases). Second, problem definition is a unique data set with the same features used to evaluate the performance of the first data sets. Besides that, classification is the most extensively used machine learning technique for classifying a large population of records using a trained model and their feature values. Predicting the labels of discrete classes and developing a model for the target class are two of the most significant responsibilities of a classifier. In the emerging discipline of bioinformatics, classification techniques are frequently



employed. Predicting chronic diseases is a significant challenge for medical practitioners. Therefore, bioinformatics is vital for creating accurate disease predictions, allowing physicians to begin treatment sooner.

Lambodar et al. [43] developed a model to forecast chronic disease development using various machine-learning classification approaches. This work uses four distinct algorithms—the decision table, J48, multilayer perceptron, and Nave Bayes—to predict renal disease. The findings of their classification have been examined across six distinct criteria. In terms of performance and accuracy, the MLP classifier consistently outperforms other classifiers when predicting the risk of renal illness. The experiment's findings with the dataset indicate that the multilayer perceptron outperforms the different classifiers evaluated.

On the other hand, medical equipment creates vast amounts of data replete with information. The effectiveness of your forecast is contingent on your ability to classify this data set accurately. Chronic kidney disease (CKD) is a significant and unpredictable disorder in the medical field, making it challenging to forecast health complications. Specialists in the medical sector do not all possess the same knowledge and skills to meet the demands of their patients. The majority of diagnoses made by doctors may not be supported by evidence. Some diseases can be fatal, and sad results do indeed occur. Approximately 58 million deaths have been attributed to chronic renal disease. Therefore, Arulanthu et al. [44] conducted an exhaustive analysis of the present state of the art in chronic kidney disease (CKD) categorization and prediction. Medical records and diagnoses are classified using data mining, fuzzy and machine learning algorithms, and other innovative techniques. This work examines and summarizes all previously published classification schemes and disease diagnosis methodologies. This investigation aims to identify and resolve some of the issues and difficulties involved with present techniques. The limitations and inadequacies of current CKD categorization and disease diagnosis approaches are also reviewed.

Regarding clinical risk assessment, Arvind et al. [45] presented a genetic programming strategy for evaluating patients' clinical risk with chronic renal illness. This study utilized the clinical data repository at UCI, which has information on 400 patients. There is an

imbalance between the number of CKD samples and healthy samples in this data set. These disparities in the data have a substantial impact on the classifiers' capacity to learn. Genetic programming (GP) is an approach inspired by natural selection in machine learning (ML). The data asymmetry has a comparable effect on GP utilizing the default fitness function. To resolve this disparity, we propose a new objective function in GP based on Euclidean distance. Additional classification techniques, such as K-nearest neighbour (KNN), KNN with particle swarm optimization (PSO), and genetic programming (GP) with the baseline fitness function, are employed to assess the robustness of the presented approach. KNN obtains 83.54% accuracy with an AUC of 0.69, PSO-KNN achieves 96.79% accuracy with an AUC of 0.94, and GP, utilizing the freshly invented fitness function, achieves 99.33% accuracy with an AUC of 0.99, outperforming both KNN and PSO-KNN.

Loss of kidney function over time is a defining characteristic of chronic kidney disease (CKD), which various illnesses can cause. In most cases, early detection and treatment of CKD can only delay the onset of complete renal failure. The CKD grading system, based on a patient's estimated glomerular filtration rate, allows for more accurate risk assessment, monitoring, and treatment (eGFR). The authors Piervincenzo et al. [46] present a supervised machine-learning technique for monitoring the course of chronic renal illness. This research aimed to identify how soon a CKD patient must begin dialysis to provide tailored care and make more considered treatment decisions. A computer simulation employing supervised machine learning is developed to assess how long a CKD patient will require dialysis. Data retrieval and model training techniques are compared to determine the most effective ones. The data used to train the comparison models comes directly from the EMR system of Vimercate Hospital. The author selected an Extremely Randomized Tree classifier set as the final model, taking into account 27 factors, including creatinine clearance, urea, red blood cell count, eGFR trend (which is not the most relevant), age, and associated comorbidities. The test's accuracy, specificity, and sensitivity were 94%, 91%, and 96%, respectively, when predicting the beginning of renal failure within the following year rather than beyond. The model's performance does not decline dramatically as more and smaller time-frame intervals are selected, down to six-month granularity. The author was aware of the limitations of the created predictive models, given that the study was conducted centrally and has not been reevaluated utilizing data sources

from other institutions' EMRs. Example: the author's analysis is restricted to individuals of Caucasian descent because it is the only group from which data was collected.

Nonetheless, there is growing evidence in the scientific literature that ethnicity is crucial in the onset of chronic kidney disease (CKD). As previously stated, additional effort is required to advance the study of CKD progression, making it difficult to locate suitable public datasets. Other relevant information, such as the patient's level of proteinuria, regimen, body mass index (BMI), drug therapy, or evaluation scales created by physicians or nurses, was either unavailable for the study or had a high number of missing entries, which would have improved the overall accuracy of the model.

Elias et al. [47] developed a method employing ML techniques to produce accurate tools for predicting the development of CKD for this investigation. Class balance was utilized to equalize the proportion of instances between the two groups before performing feature ranking and analysis. Numerous ML models were ultimately trained and graded based on various performance indicators. The results demonstrate that Rotation Forest (RotF) was the most effective of the evaluated models, with an AUC of 100%, precision, recall, F-measure, and accuracy equal to 99.2%. The author should conclude by discussing the limitations of the existing study. This information is a public dataset [48] of specific attributes.

In addition, the author utilized information that did not originate from a medical facility to characterize the health status of participants using a variety of variables. This information could be time-consuming and delicate to get. The dataset's features do not include any age or gender information about the participants. Hence demographic analyses and processing cannot be conducted.

In terms of implications of deep learning techniques regarding chronic kidney disease, the diagnosis of chronic kidney disease (CKD) is frequently delayed until a more advanced stage of the disease has developed, making it a global public health concern. To effectively handle this issue, it is vital to invest in reliable forecasting. Andressa et al. [49] investigated the early diagnosis of CKD using machine-learning approaches for small and unbalanced datasets, which may aid in addressing this issue. This study intends to assist in the earlier detection of chronic kidney disease (CKD) by addressing imbalance issues and small

sample size. For this study, the author reviewed information from the medical records of Brazilians with and without a CKD diagnosis, such as blood pressure, type 2 diabetes, creatinine, urea, proteinuria, age, gender, and glomerular filtration rate. The author described an oversampling technique that involves both human and machine-driven augmentation. The author ran studies with the synthetic minority oversampling technique (SMOTE), borderline-SMOTE, and borderline support vector machine (SVM). The author then developed models using the methodologies, which included decision trees (DTs), random forests (RFs), and multi-class AdaBoosted DTs. In addition to overall local accuracy and local class accuracy, the author employed k-nearest oracles-union, k-nearest oracles-eliminate, and META-DES for dynamic ensemble selection and dynamic classifier selection, respectively. The author compared the outcomes of numerous cross-validation (CV) tests, such as hold-out validation, multiple stratified CV, and nested CV. The DT model attained an accuracy rate of 98.99% with human-assisted augmentation and SMOTE. The proposed method can contribute to developing solutions for the earlier identification of CKD using unbalanced and small data sets. A significant limitation of the study was the GridSearchCV program utilized to identify the ideal settings for each method. While searching for parameterization for each ML model, the author encountered processing limitations, particularly for ensemble models. The small sample size of manually supplemented cases may also be a limitation of this study.

More than 10% of the world's population currently has chronic kidney disease (CKD), and millions die yearly. Therefore, early detection of CKD could help patients live longer and save money on treatment. Developing such a multimedia-driven approach is vital for early and accurate disease diagnosis. Before the introduction of multimodal data-driven learning, researchers depended on many techniques connected with classic machine learning models. Using a stacked autoencoder network and visual data with a softmax classifier [50], Aditya et al. describe a novel deep-learning strategy for the classification of chronic kidney disease. First, a stacked autoencoder is used to identify significant features within the dataset, and then a softmax classifier is used to offer a classification prediction. The UCI dataset consisting of 400 patients in the early stages of CKD, 25 features, and a binary classification problem was utilized for testing. Precision, recall, specificity, and the F1 score were used to evaluate the suggested network. According to research, this multimodal

model offers a greater classification accuracy for chronic renal illness than the standard classifiers. Backpropagation can be used to fine-tune the model for classifying diseases on larger datasets.

Additionally, Due to the high cost of treatment of CKD in third-world countries, this lethal disease poses the most significant threat. Predicting who is at risk for developing chronic kidney disease (CKD) and utilizing clustering algorithms to do so is critical since it can influence the illness's progression. It is essential to recognize this "silent killer" as soon as possible to limit the danger of renal failure. Kerina et al. [51] offer a neuro-fuzzy and hierarchical clustering algorithm-based strategy for predicting the degree of CKD risk. Using a neuro-fuzzy algorithm, the risk of patients getting CKD is assessed. 97% of neuro-fuzzy systems were accurate in their predictions. By detecting risk variables using the chosen characteristics, a forecast for CKD disease is created. Patients with a high risk of renal disease are more likely to have diabetes, as established by clustering the prediction results. Cases of diabetes and chronic renal illness were divided into three groups by hierarchical cluster analysis.

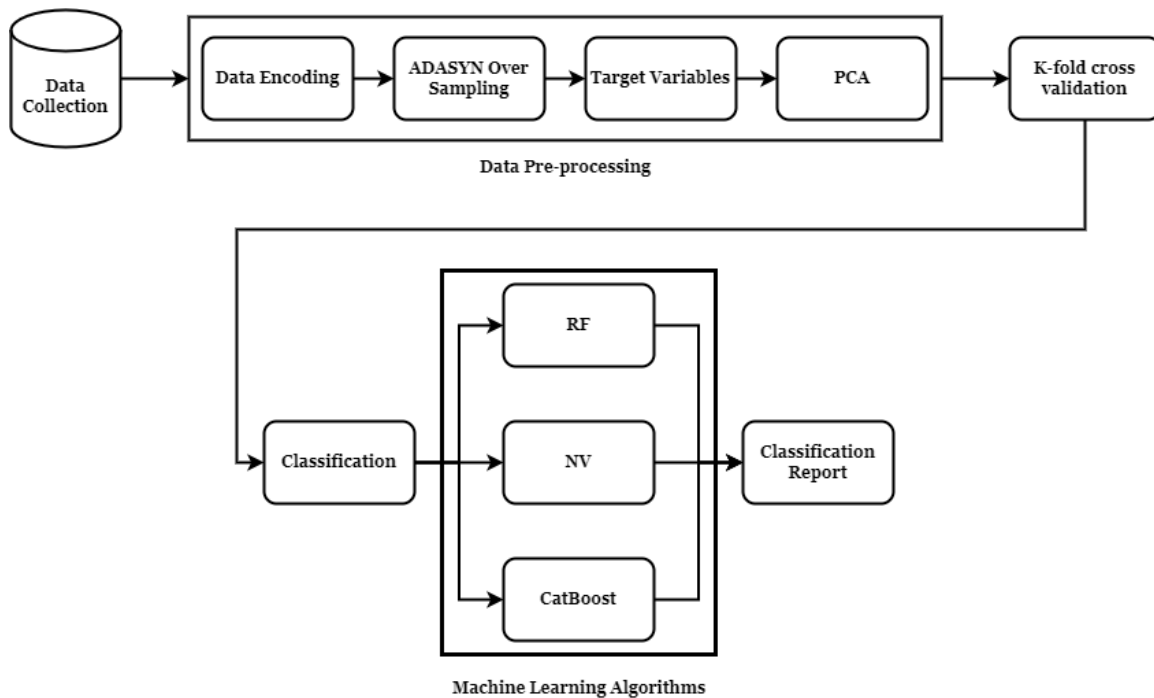
## **CHAPTER 3**

### **RESEARCH METHODOLOGY**

The techniques or tactics used in research are referred to as research methods. These methods are used to identify, acquire, organize, and manage data that comes from a vast number of sources. The term "methods" is used to refer to each of them. In this part of the article, we will talk about the technique that we have presented, a processing chain that is less complicated for an approach to risk categorization.

#### **3.1 Proposed method**

Data collection is the first step in our fundamental methodology. Afterwards, the gathered information is processed in various methods to prepare for classification. After completing the preliminary stages of the classification process, we will move on to accurate categorization. Figure 1. It should be no surprise that the first step included data collection, followed by encoding the data to render it usable for the other stage. The ADASYN method is used to complete the following degree, called data balancing and is an essential part of the research on machine learning. Then, to have a clearer understanding of the data, we separated the variables involved in our feature from those involved in our objective. The principal component analysis (PCA) method is then used for the higher dimension analysis, which enables the study to zero in on the data's most important aspects. The K-fold cross-variation approach is the next step, which allows us to train five distinct models and may also be used to assess a model's performance when presented with new information. We have arrived at the classification step, during which we prepare separate reports for each of the three classification strategies we use (Random Forest, Naive Bayes, and Cat Boost). In the following sections, you'll find condensed explanations of each topic.



**Figure 1: Proposed Methodology**

### 3.2 Data Pre-processing

#### Data Encoding

Any system that employs machine learning must first have the capacity to grasp the data before it can be put into operation. For instance, it is essential to convert qualitative descriptors like "Age," "Number of sexual partners," and "First sexual intercourse" into their equivalent numerical values. The problem can be fixed by changing them, for instance, into numeric labels with the values "1" for age, "2" for the number of sexual partners, and "3" for First sexual intercourse. This will remove the issue. This will eliminate the complication.

We also use it to deal with situations in which data is lacking. Dealing with missing data in the dataset used to build the predictive model may be accomplished with the help of the Simple Imputer class included in scikit-learn. When you input new values, those values will replace those now set to the NaN method. Data specific to the dataset will be used in place of NaNs.

## ADASYN Over Sampling

Machine learning becomes challenging when one form of data has a much smaller total number than another, which is known as a data imbalance issue. The bulk of machine learning algorithms works optimally when the number of samples in each class is virtually the same. When there are more instances of one class than the other, difficulties may arise. An example of an unbalanced classification problem is one in which the distribution of samples across recognized classes is biased or skewed. There may be only one occurrence of the minority class for every hundred, thousands, or millions of cases of the dominating class or classes.

Unbalanced classifications pose a barrier to predictive modelling because the bulk of machine learning algorithms used for classification was developed, assuming each class would include an equal number of samples. As a result, models ultimately produce unsatisfactory prediction results, especially for excluded populations. Given that the majority of the time, the minority group, this is a hurdle.

There is a wide variety of methods available to lessen the impact of this class imbalance problem; however, in this instance, we use ADASYN to address the class imbalance problem. The ADASYN algorithm computes the minority-to-majority ratio using a formula such as  $d = \frac{m_s}{m_l}$ , where  $m_s$  and  $m_l$  stand for the number of cases in the minority and majority classes, respectively. The algorithm will not be initialized if  $d$  exceeds a predetermined limit.

Calculate the total number of synthetic minority data to generate.  $G = (m_l - m_s)\beta$  Here,  $G$  is the total number of minority data to generate.  $\beta$  is the ratio of minority: Majority data desired after ADASYN.  $\beta = 1$  means a perfectly balanced data set after ADASYN.

Find the  $k$ -Nearest Neighbors of each minority example and calculate the  $r_i$  value. After this step, each minority example should be associated with a different neighbourhood.

$r_i = \frac{\# \text{majority}}{k}$  The  $r_i$  value indicates the dominance of the majority class in each community. Higher  $r_i$  neighbourhoods contain more majority-class examples and are more challenging to learn. See below for a visualization of this step. In the example,  $K = 5$ , look for the 5 nearest neighbours. Normalize the  $r_i$  values so that the sum of all  $r_i$  values equals 1.



$$\hat{r}_i = \frac{r_i}{\sum r_i}$$

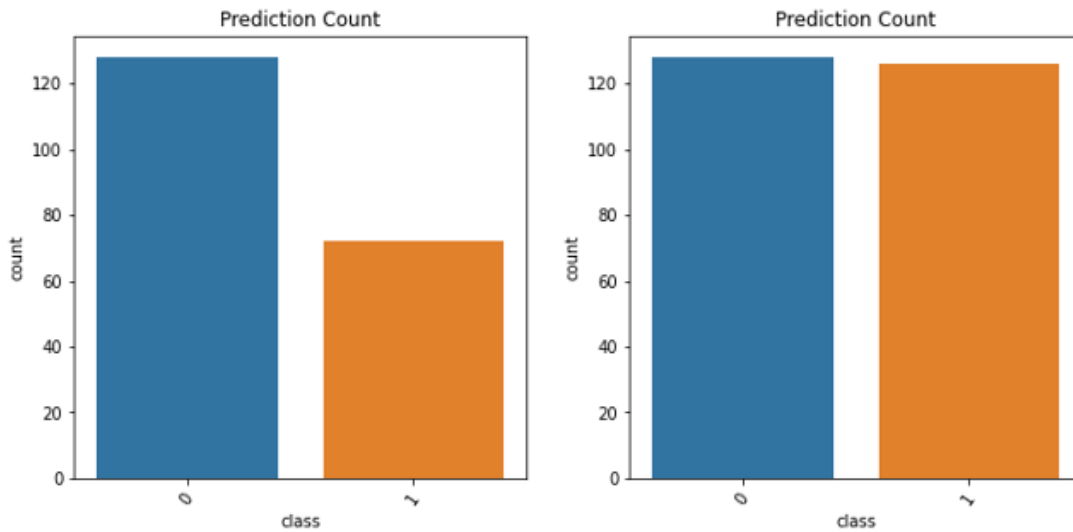
$$\sum \hat{r}_i = 1$$

Calculate the number of synthetic examples to generate per neighbourhood.  $G_i = G\hat{r}_i$ . Because  $r_i$  is higher for areas dominated by majority class examples, more synthetic minority class examples will be developed for those neighbourhoods. Hence, this gives the ADASYN algorithm its adaptive nature; more data is generated for "harder-to-learn" areas.

Generate  $G_i$  data for each neighbourhood. First, take the minority example for the areas  $x_i$ . Then, randomly select another minority example within that neighbourhood,  $x_{z_i}$ .

The new synthetic example can be calculated using:  $s_i = x_i + (x_{z_i} - x_i)\lambda$

In the above equation,  $\lambda$  is a random number between 0–1,  $s_i$  is the new synthetic example, and  $x_i$  and  $x_{z_i}$  are two minority examples within the same neighbourhood. A visualization of this step is provided below. Intuitively, synthetic measures are created based on a linear combination of  $x_i$  and  $x_{z_i}$ . White noise can be added to the synthetic examples to make the new data more realistic. Also, planes can be drawn between 3 minority examples instead of linear interpolation, and points can be generated on the plane instead. With the above steps, any imbalanced data set can now be fixed, and the models built using the new data set should be much more effective.



**Figure 2: ADASYN Oversampling Technique**

### **Target Variables**

The term "target variable" refers to the variable inside a dataset on which we focus our efforts to achieve a greater level of comprehension. It is the variable the user wants to create a hypothesis to explain the rest of the information included in the dataset. To extract the aim variable from it, supervised machine learning approaches are used most of the time. This sort of algorithm looks at data from the past to identify patterns and find connections between different aspects of your dataset and the goal. Target variables are subject to change depending on the intended objective and the facts at hand. In the absence of a clearly stated purpose, it is mathematically impossible for supervised machine learning algorithms to convert the data they have access to into results. It is possible that until adequate training data is acquired, it will be necessary to use a system that is less than ideal for locating the variable of interest 3.3.4 principal component analysis

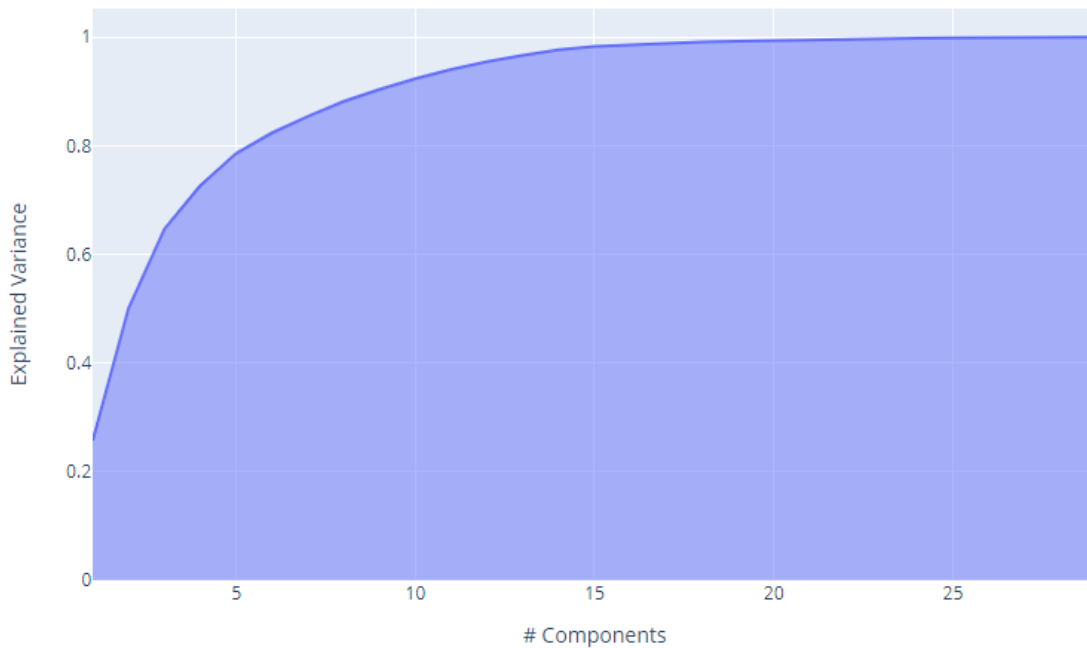
### **Principal Component Analysis**

Principal component analysis (PCA) is a dimensionality reduction technique that takes a dataset's set of features and reduces them to a smaller group of features (called principal

components) while attempting to retain as much information as possible. Consider the following details to comprehend better why we implement PCA: It liberates us from the qualities generally associated. The principal component analysis is a method for removing multi-collinearity and associated characteristics (PCA). Finding traits that are interconnected requires time, especially when there are a large number of elements. Increases the efficiency of the machine learning algorithms. As a direct result of PCA's capacity to lower the total number of features, the time required to train your model is drastically decreased. Avoid snug clothing. PCA prevents overfitting by eliminating redundant data points from the analysis.

Here we use PCA and divide 37 data features into 11 without losing any feature information; look at figure 3 to understand the PCA use.

Number of features before PCA: (254, 28)  
Number of features after PCA: (254, 18)



**Figure 3: Dimensionality Reduction using PCA**

### **K-fold Cross Validation**

The core technique is straightforward, and the following steps may be employed: The dataset must be randomly divided into k subsamples, and each subsample must undergo a

random selection procedure. (It is necessary to verify that the subsample sizes are comparable and that  $k$  is less than or equal to the total number of items in the dataset.) The first subset is used as the test data for the first cycle, and the following subsets are considered training data. After "training" the model using the training data, you may assess it using the test data. You can eliminate the model but must maintain its rating or error rate. Now, for the following iteration, choose a new subset to serve as the test data set, and include the test data set from the previous iteration as part of the training data for the remaining data sets. Instead of maintaining the model, you should retrain it using the new test data set before putting it to the test. Continue by repeatedly carrying out steps 1-4k. A random sample of the data will be selected for each iteration's study, and this procedure will continue until the whole dataset has been evaluated. You will get a total of  $k$  distinct assessment score results. The overall error rate is derived by averaging these results.

It is crucial to choose an appropriate value for  $k$ . The model's performance will probably be underestimated if  $k$  is not chosen correctly. In other words, it may result in overestimating the model's ability with high bias or substantially fluctuating assessed power across training datasets. These results are both undesired and able high variances. The following possibilities for picking the letter  $k$ : It is feasible that  $k$  will have a value of 5 or 10. Experimentation shows us that for findings to be judged acceptable,  $k$  must have a value between 5 and 10. Assume that the number of dataset records,  $n$ , equals  $k$ . This guarantees that the validation data accurately represents every sample. In addition, we have the option of choosing  $k$  so that each subsample of the data is big enough to be statistically significant in proportion to the whole set. For this study purpose, we set the  $K$  value to 5.

### **3.3 Classification Methods**

The dataset was subjected to five (3) classification algorithms to determine the most effective performing method prediction accuracy and other statistical properties. Random Forest (RF), Naïve Bayes (NB), and CatBoost were always the techniques used.

#### **Random Forest (RF)**

RF is a data classification approach achieved by the proposed learning and DT [22]. While in the training stage, it generates a vast number of trees as well as a forest of decision trees [23]. Every tree in the forest predicts the class label for every event during the testing period. When each tree indicates a class label, the final selection for each test data is made via a majority vote [24]. The class label that receives the most votes is considered the most appropriate label for the test data. This cycle is repeated for each piece of data in the collection. The best fit randomize responsible for this experiment was 123, which offered the best effectiveness for the presented collection.

### **Naïve Bayes (NV)**

The classification process is carried out with a probabilistic machine learning model known as a Naive Bayes classifier. The Bayes theorem forms the backbone of the classifier's  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$  Given that event B has taken place, we may use Bayes' theorem to calculate the likelihood that event A will also occur. In this case, Hypothesis A is being tested against Evidence B. In this case, it is assumed that the predictors and features can be considered separate entities. That is to say, the presence of one characteristic does not affect the other. Because of this, we refer to it as naive.

### **CatBoost**

CatBoost, a machine learning algorithm developed by Yandex, was recently made available to the open-source community. It is straightforward to include in other deep learning frameworks, such as Google's TensorFlow or Apple's Core ML. It is beneficial for tackling a wide range of difficulties that current businesses face due to its flexibility to a large variety of data formats, making it suitable for usage in modern companies. In addition to this, its accuracy is unparalleled.

Especially useful in two different respects: It produces state-of-the-art solutions without the considerable data training that is sometimes required by other machine learning techniques and provides solid out-of-the-box support for the more descriptive data formats that accompany many business concerns. This strategy for improving overall performance is referred to as "CatBoost," which comes from the combination of the words "Category"

and "Boosting." As mentioned before, the library is suitable for use with a broad range of data types. The term "boost" was chosen for this library because its creators were motivated by the gradient-boosting machine learning approach. Gradient boosting is a powerful method of machine learning that has found broad usage in the solution of a wide range of business issues, such as the detection of fraud, the making of item suggestions, and the making of forecasts. It can deliver highly accurate findings with considerably fewer data than is required by DL models, which need a large quantity of data to train from.

### 3.4 Classification Report Performance Measurement Criteria

We build the classification report for each technique by applying these evaluation criteria. We will test the efficiency of each classifier by utilizing precision, recall, and f-1 score. In addition, we will employ confusion matrices and ROC-AUC curves to comprehend each classifier's overall performance.

**Table 1: Performance Measurement Criteria**

Metrics	Explanation	Formula
Accuracy	The proportion of observations that are correctly categorized [25].	$A = \frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	It is a breakdown of the actual positive vs all the expected positives [26].	$S_n = \frac{TP}{TP + FN}$
Specificity	This calculates the percentage of real negatives compared to all expected negatives [27].	$S_p = 1 - \left( \frac{FP}{FP + TN} \right)$
Precision	Precision is the ratio of accurately predicted positives to all predicted positives. [28].	$P = \frac{TP}{TP + FP}$
Recall	A machine learning model's prediction of the system is characterized by True Positives. [29].	$R = \frac{TP}{TP + FN}$
AU ROC	A ROC is a simple medical test screening test that is created by comparing actual performance against the false positive rate at estimation circumstances [31].	$TPR = \frac{TP}{TP + FN}$ $FPR = \frac{FP}{FP + TN}$

## CHAPTER 4

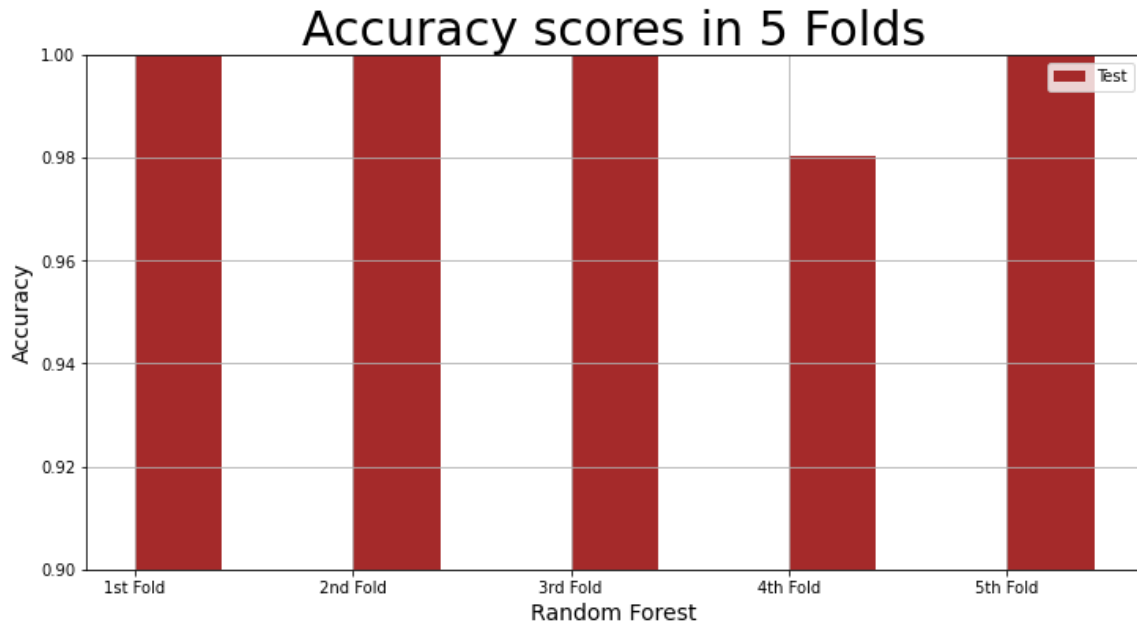
### RESULT ANALYSIS

#### 4.1 Random Forest

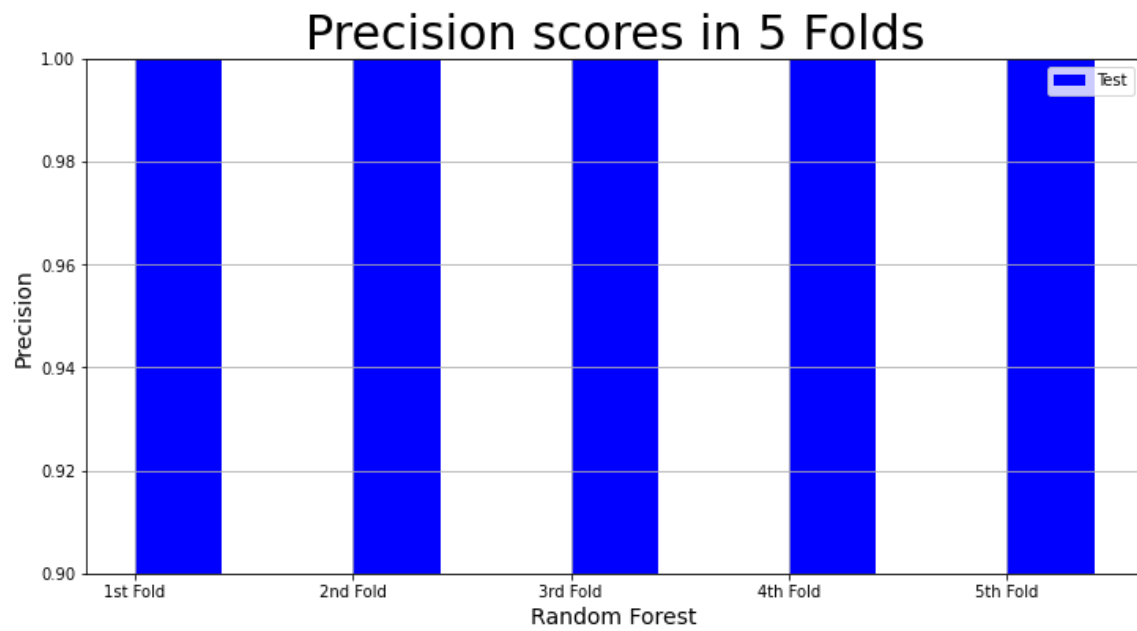
We employed the Random Forest classifier in combination with 5-fold cross-validation. Consequently, we took each fold's accuracy, precision, and recall, as well as the f-1 score and ROC-AUC. If we look at Table 3, we can see that for folds 1, 2, 3, and 5, we received a score of one in each of the measurement categories. But in the instance of fold 4, we obtain an accuracy score of 0.9803, a precision score of 1, a recall score of 0.9615, an F-1 score of 0.9803, and a ROC-AUC score of 1.

**Table 2: Result Analysis for RF**

<b>k-fold</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F-1 score</b>	<b>ROC-AUC</b>
1	1	1	1	1	1
2	1	1	1	1	1
3	1	1	1	1	1
4	0.9803	1	0.9615	0.9803	1
5	1	1	1	1	1

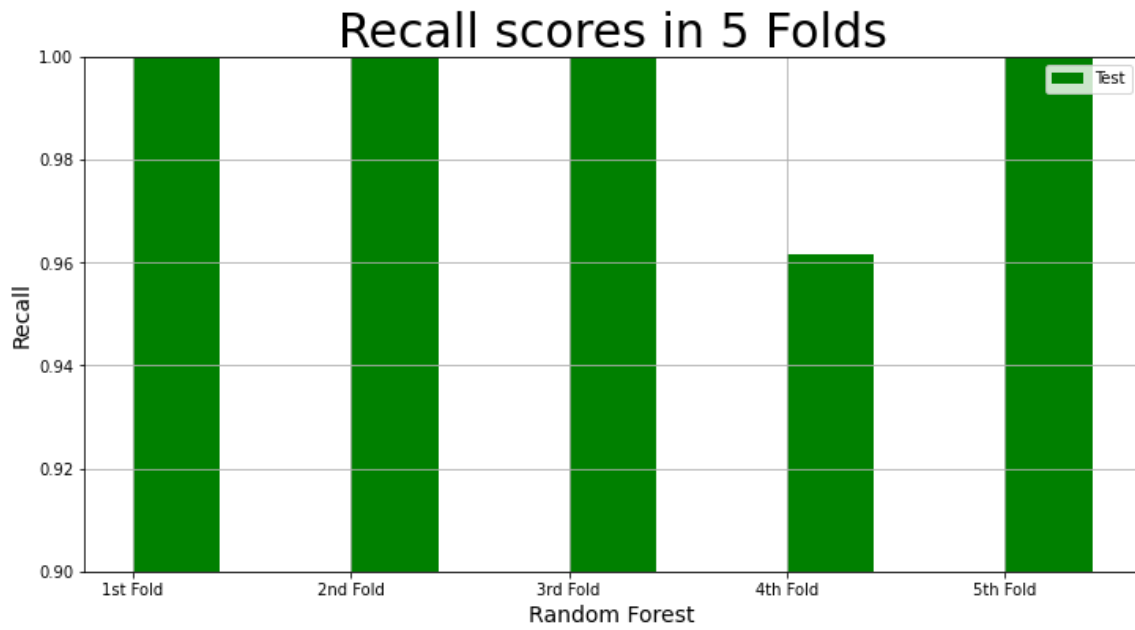


**Figure 4: 5-fold Mean Test Accuracy for RF**

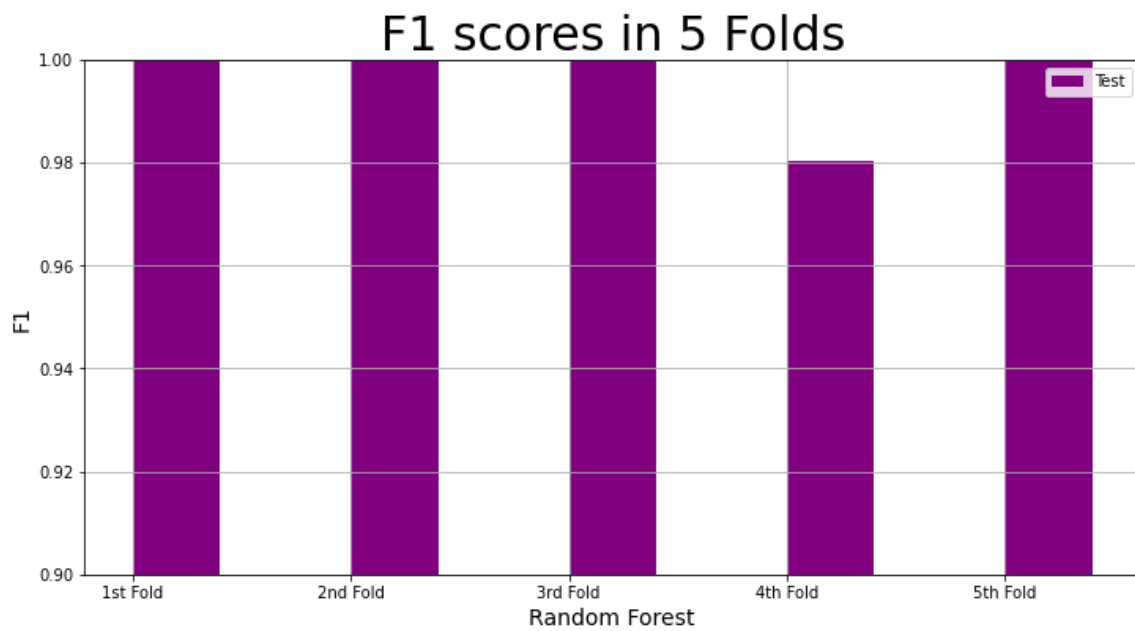


**Figure 5: 5-fold Mean Test Precision for RF**

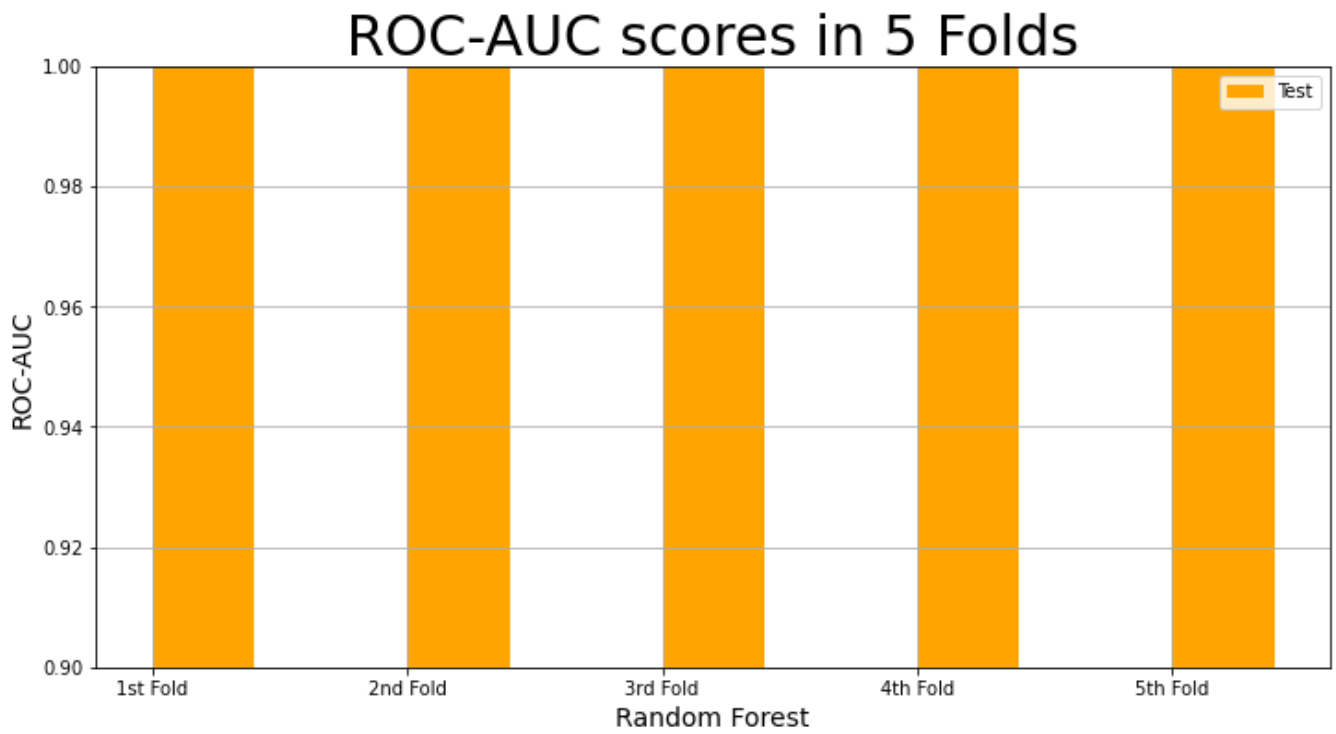




**Figure 6: 5-fold Mean Test Recall for RF**



**Figure 7: 5-fold Mean Test F-1 score for RF**



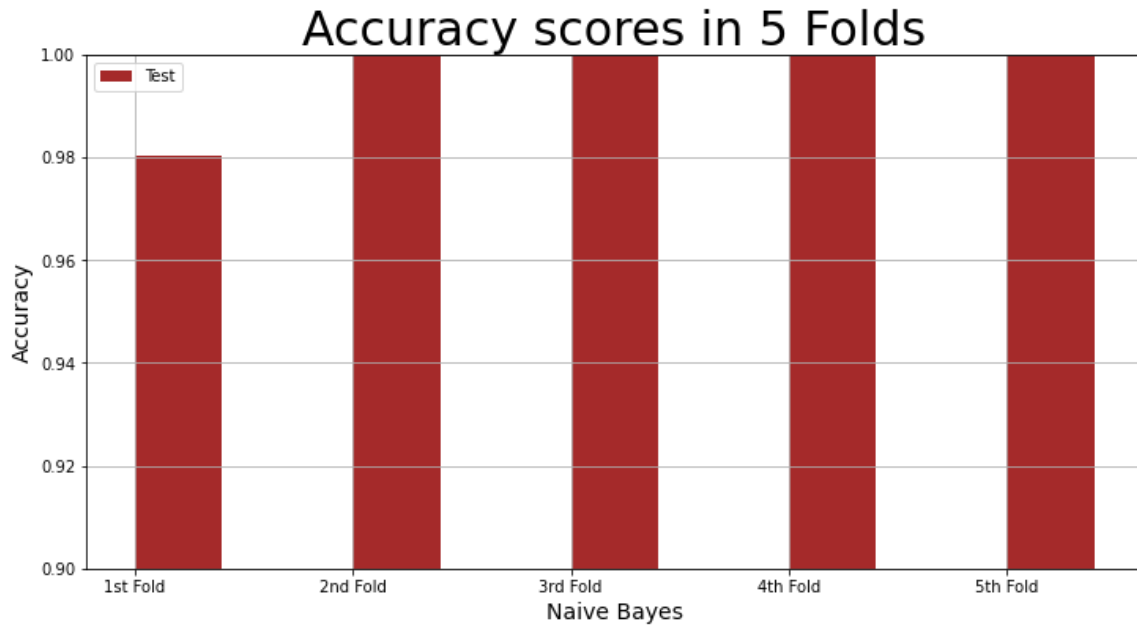
**Figure 8: 5-fold Mean ROC-AUV Score for RF**

## 4.2 Naïve Bayes

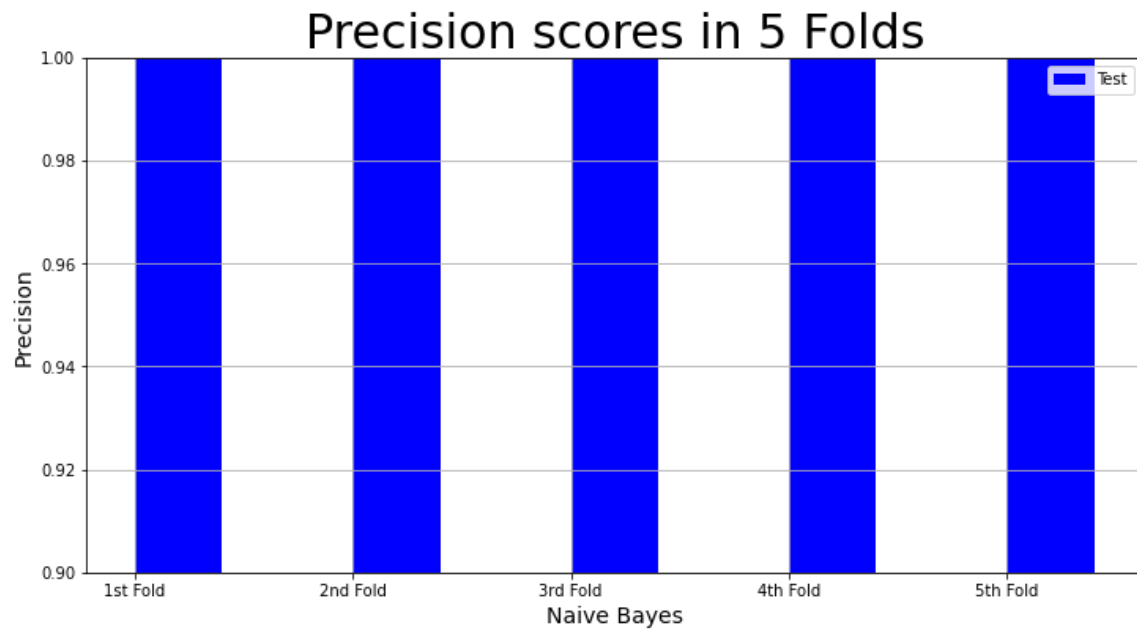
We employed the Naïve Bayes classifier in combination with 5-fold cross-validation. Consequently, we took each fold's accuracy, precision, and recall, as well as the f-1 score and ROC-AUC. If we look at Table 3, we can see that for folds 2, 3, 4 and 5, we received a score of one in each of the measurement categories. But in the instance of fold 1, we obtain an accuracy score of 0.9803, a precision score of 1, a recall score of 0.96, an F-1 score of 0.9795, and a ROC-AUC score of 1.

**Table 3: Result Analysis for NB**

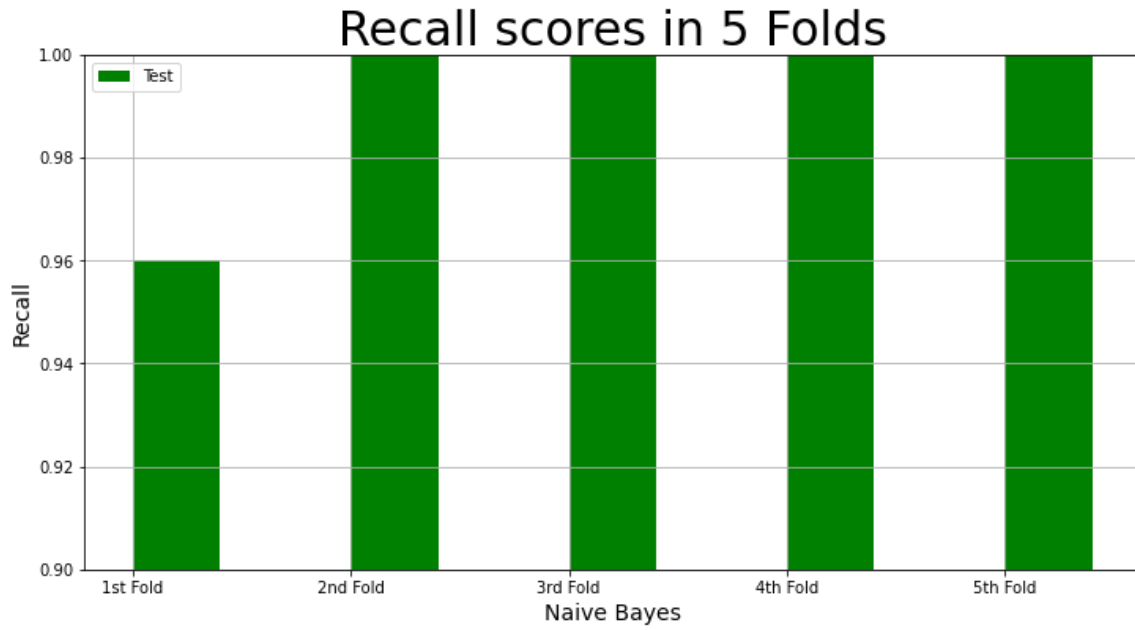
<b>k-fold</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F-1 score</b>	<b>ROC-AUC</b>
1	0.9803	1	0.96	0.9795	1
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	1	1
5	1	1	1	1	1



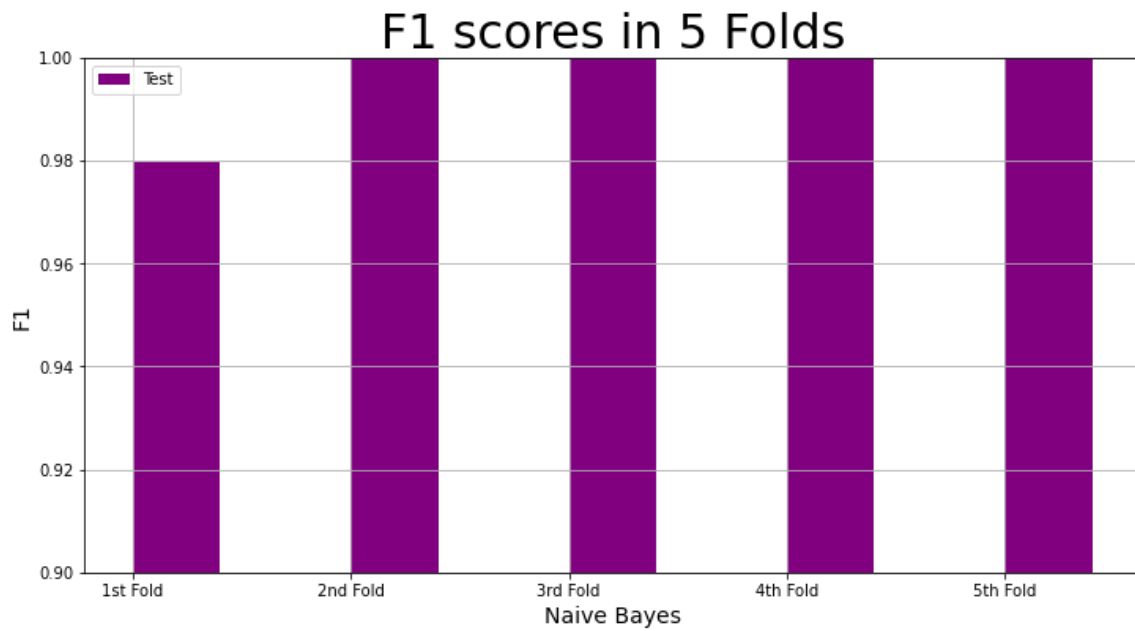
**Figure 9: 5-fold Mean Test Accuracy for NB**



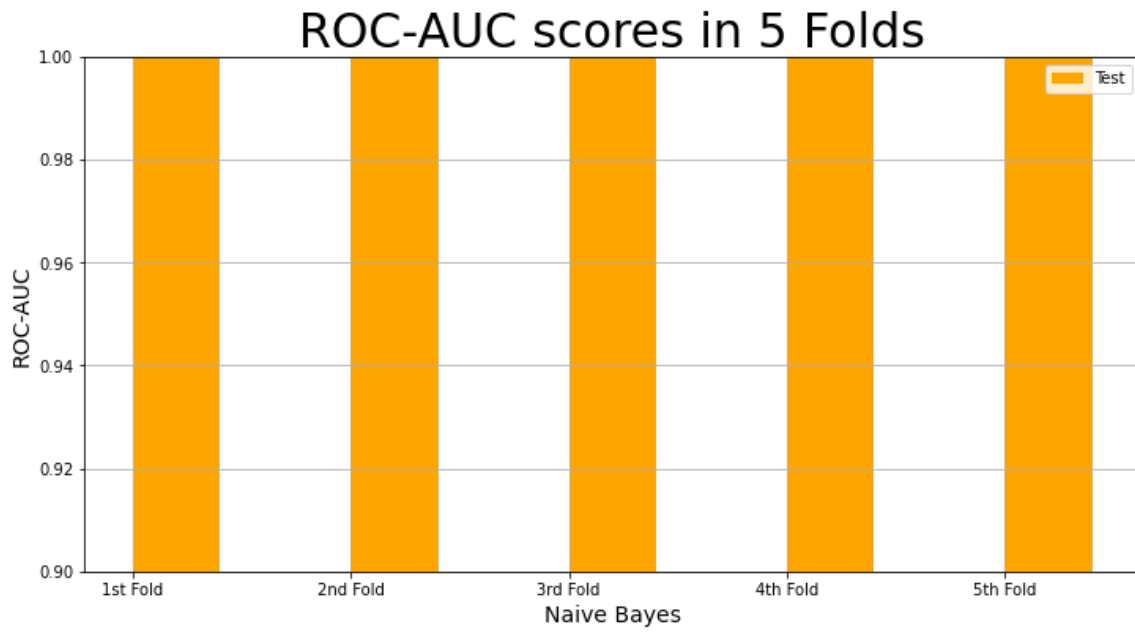
**Figure 10: 5-fold Mean test Precision for NB**



**Figure 11: 5-fold Mean Test Recall for NB**



**Figure 12: 5-fold Mean Test f-1 Score for NB**



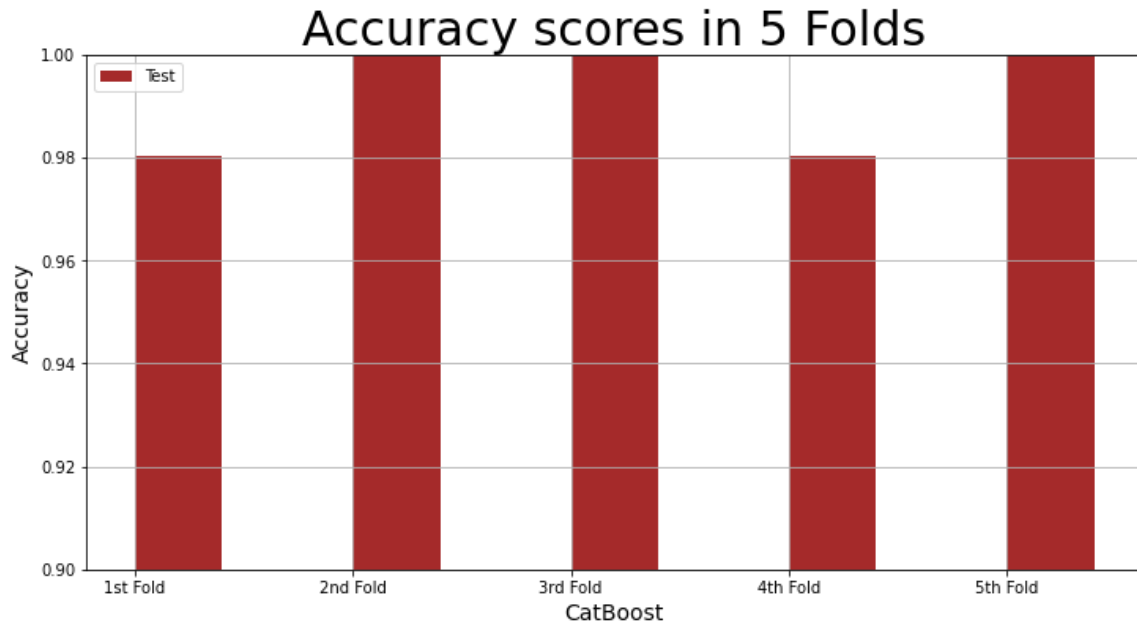
**Figure 13: 5-fold Mean ROC-AUC Score for NB**

### 4.3 CatBoost

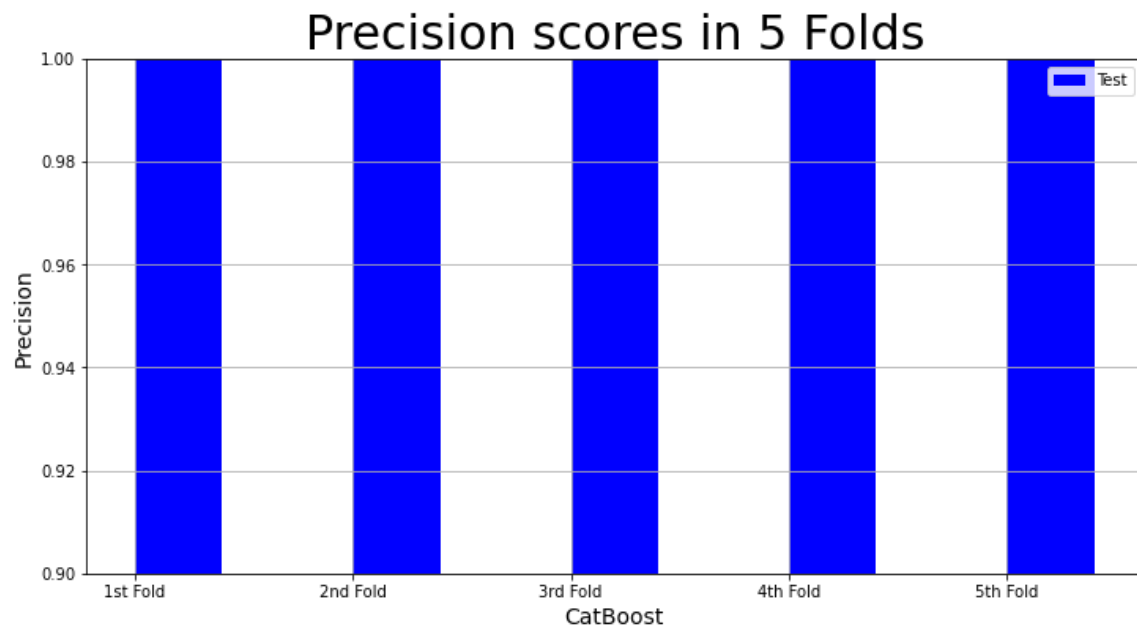
We used the Nave Bayes classifier in conjunction with 5-fold cross-validation. As a result, we calculated each fold's accuracy, precision, and recall, in addition to the f-1 score and ROC-AUC. Examining Table 3 reveals that for folds 2, 3, and 5, we earned a score of one in each measurement category. In folds 1 and 5, however, we obtained inconsistent results: an accuracy score of 0.9803, a precision score of 0.9615, a recall score of 1, an F-1 score of 0.9803, and a ROC-AUC score of 1. The fold 1 accuracy score is 0.9803, the precision score is 1, the recall score is 0.9615, the F-1 score is 0.9803, and the ROC-AUC score is 1.

**Table 4: Result Analysis for CatBoost**

<b>k-fold</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F-1 score</b>	<b>ROC-AUC</b>
1	0.9803	0.9615	1	0.9803	1
2	1	1	1	1	1
3	1	1	1	1	1
4	0.9803	1	0.9615	0.9803	1
5	1	1	1	1	1

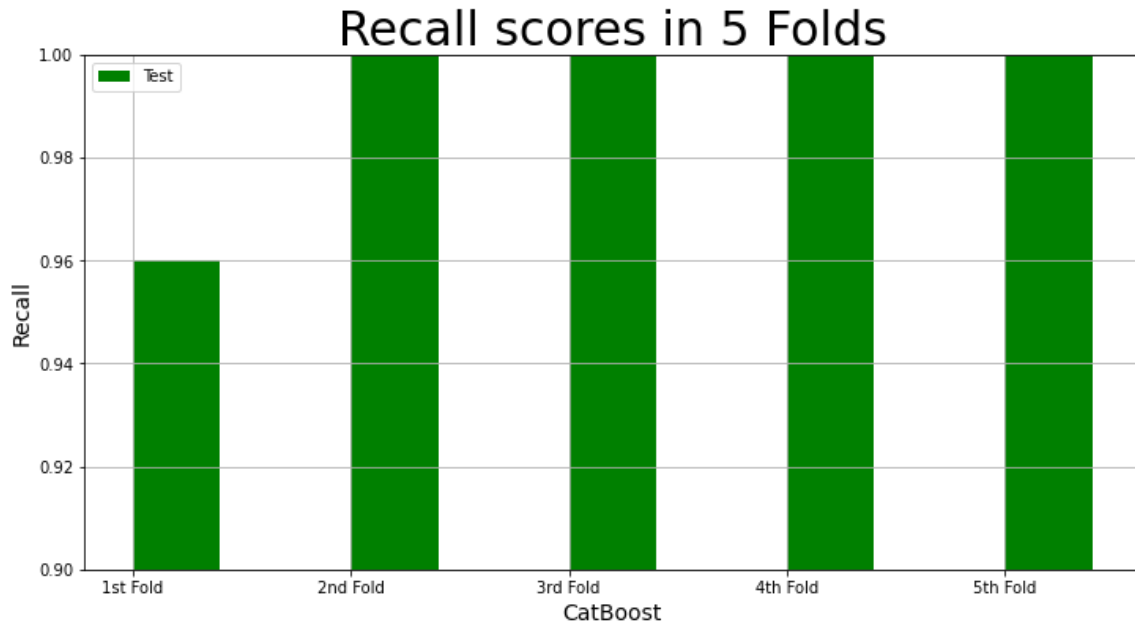


**Figure 14: 5-fold Mean Test Accuracy for CatBoost**

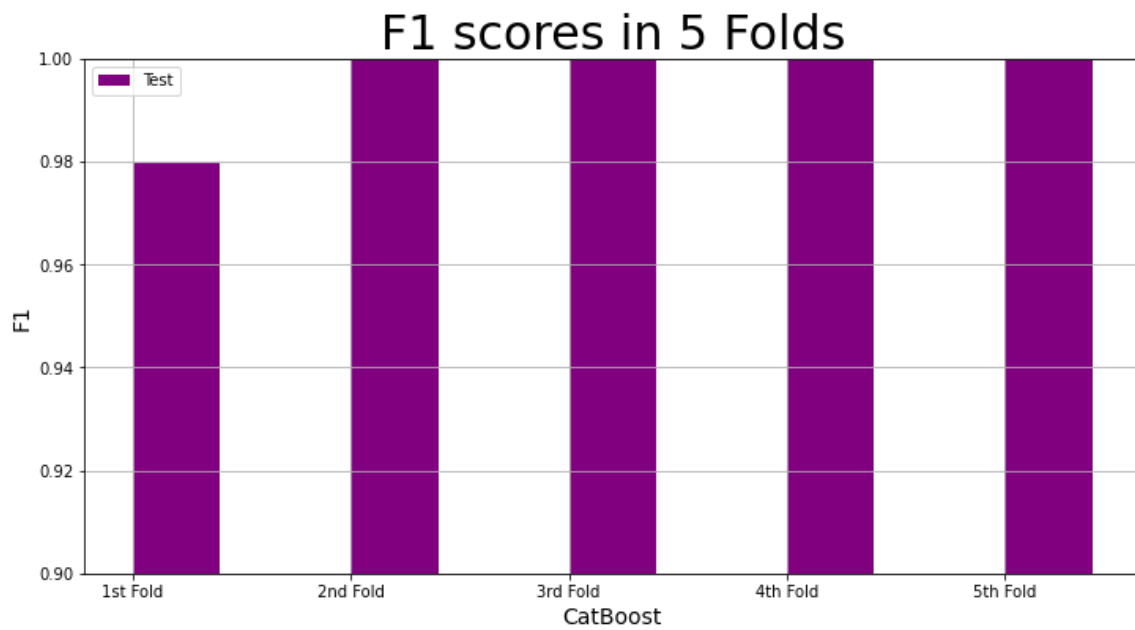


**Figure 15: 5-fold Mean Test Precision for CatBoost**

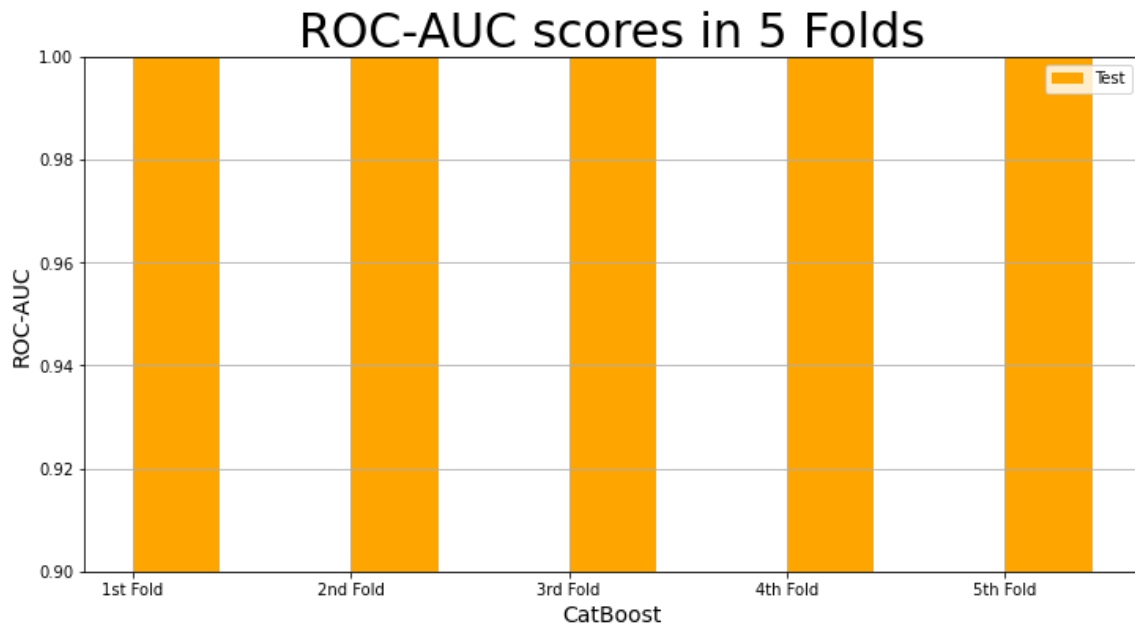




**Figure 16: 5-fold Mean Test Recall for CatBoost**



**Figure 17: 5-fold Mean Test f-1 Score for CatBoost**



**Figure 18: 5-fold Mean ROC-AUC Scores for CatBoost**

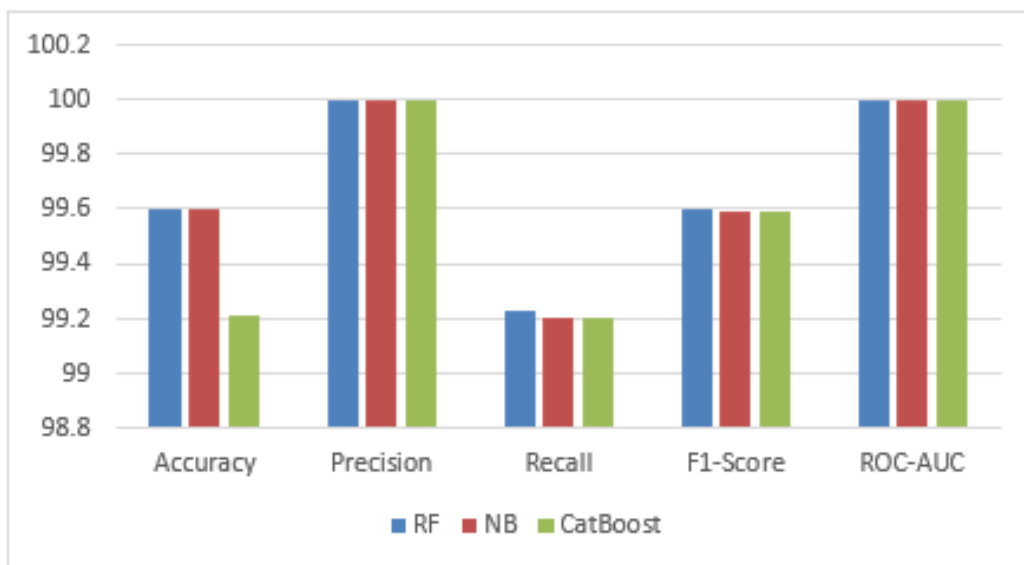
## CHAPTER 5

### DISCUSSION

Table 1 and figure 1 reveal that for this chronic kidney disease risk classification investigation, we employed a total of three ML approaches (RF, NB, and CatBoost) to choose the optimal one. In this instance, nearly every method did well on every measuring scale, but RF fared the best. The approaches produce identical results in terms of accuracy, precision, F1-score, and ROC-AUC curve; however, in terms of recall, RF fared better than NB and CatBoost, whose recall scores are 99.20 and 99.23, respectively.

**Table 5: Performance Table of All Methods in Each Measurement Scale**

Method	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF	99.60	1	0.9923	0.9960	1
NB	99.60	1	0.9920	0.9959	1
CatBoost	99.21	1	0.9920	0.9959	1



**Figure 19: Performance Graph of All Methods in Each Measurement Scale**

## CHAPTER 6

### CONCLUSION & FUTURE WORK

Even though it is not a discovery that existing machine learning performs well in risk classification of renal disease and that some results are comparable to ours, in this particular case, we received nearly the same impact on all measurement scales, indicating that we did make a modest contribution to this field. Knowing that in the machine learning field, data ambalancemen is a significant challenge to train the model and perform well in prediction makes it somewhat more challenging to uncover the secret of this constant contribution across all measurement scales for all existing approaches. The problems we encountered are sufficiently pervasive that other researchers have addressed them in their experiments; however, none of these other researchers has adopted the data ambalancemen technique or ADASYN. By eliminating data ambalancemen problems with the ADASYN technique, we achieved a continuous phase for all the approaches we examined. This is a new contribution to our study and has made this possible. On the other hand, RF improved its F1-score performance compared to the different corporations. Since the proposed pipeline worked so well for risk classification in our future plans, we want to apply it to more datasets and for clinical detection and classification.

In future work, we plan to apply the data ambalancemen technique and ADASYN to other datasets for clinical detection and classification. This will allow us to evaluate the effectiveness of our approach on a wider range of data, and determine if it can be used to improve the performance of machine learning models in other areas of healthcare. Additionally, we plan to investigate other machine learning algorithms that could benefit from the use of ADASYN and the data ambalancemen technique. For example, we will investigate the use of deep learning algorithms, such as convolutional neural networks, to see if they can achieve even better performance when combined with these techniques. Furthermore, we will also study the use of other techniques like feature selection, feature extraction and ensemble learning to enhance the performance of our model. As we move forward, it will be important to continue to engage with the clinical community to ensure that our research is aligned with their needs, and that our results are translated into practical

tools that can be used to improve patient outcomes. Ultimately, our goal is to contribute to the development of more accurate and reliable diagnostic tools that can help to improve the diagnosis and treatment of renal disease, and other diseases.

## REFERENCES

1. Anwarul A, Bachchu MA, 1989. "Prevention of Dental caries and the science of nutrition." Bangladesh Dental Journal, JULY 2020.
2. Poul Erik Petersen, Denis Bourgeois, Hiroshi Ogawa, Saskia Estupinan-Day and Charlotte Ndiaye, The global burden of oral diseases and risks to oral health", Bull World Health Organ, JULY 2020.
3. Sebastià, C., Páez-Carpio, A., Guillen, E., Paño, B., Arnaiz, J. A., De Francisco, A. L., ... & Oleaga, L. (2022). Oral hydration as a safe prophylactic measure to prevent post-contrast acute kidney injury in oncologic patients with chronic kidney disease (IIIb) referred for contrast-enhanced computed tomography: subanalysis of the oncological group of the NICIR study. *Supportive Care in Cancer*, 30(2), 1879-1887.
4. Oweira, H., Ramouz, A., Ghamarnejad, O., Khajeh, E., Ali-Hasan-Al-Saegh, S., Nikbakhsh, R., ... & Sadeghi, M. (2022). Risk factors of rejection in renal transplant recipients: a narrative review. *Journal of Clinical Medicine*, 11(5), 1392.
5. Aparicio-Trejo, O. E., Aranda-Rivera, A. K., Osorio-Alonso, H., Martínez-Klimova, E., Sánchez-Lozada, L. G., Pedraza-Chaverri, J., & Tapia, E. (2022). Extracellular Vesicles in Redox Signaling and Metabolic Regulation in Chronic Kidney Disease. *Antioxidants*, 11(2), 356.
6. Sindhu, C., Prasad, P., Elumalai, R., & Matcha, J. (2022). Clinical profile and outcomes of COVID-19 patients with acute kidney injury: a tertiary centre experience from South India. *Clinical and Experimental Nephrology*, 26(1), 36-44.
7. Shazzad-"F"-DIU: Jamaludin, T. S. S., Nurumal, M. S., Hasan, M. K. C., Rizuan, S. H. S., & Faizal, N. F. F. M. (2022). Prevention and Early Detection of Acute Kidney Injury in Intensive Care Unit: A Systematic Review. *INTERNATIONAL JOURNAL OF CARE SCHOLARS*, 5(1), 72-84.
8. Shazzad-"F"-DIU: Cody, E., & Hooper, D. K. (2022). Kidney transplantation in pediatric patients with rheumatologic disorders. *Current Opinion in Pediatrics*, 34(2), 234-240.
9. Shazzad-"F"-DIU: Geng, T., Li, X., Ma, H., Heianza, Y., & Qi, L. (2022, January). Adherence to a healthy sleep pattern and risk of chronic kidney disease: the UK Biobank Study. In *Mayo Clinic Proceedings* (Vol. 97, No. 1, pp. 68-77). Elsevier.
10. Shazzad-"F"-DIU: Rossing, P. (2022). Clinical perspective—evolving evidence of mineralocorticoid receptor antagonists in patients with chronic kidney disease and type 2 diabetes. *Kidney International Supplements*, 12(1), 27-35.
11. Asri, H., Mousannif, H., Moatassime, H. and Noel, T., 2016. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, 83, pp.1064-1069.
12. <https://www.kidney.org/atoz/content/about-chronic-kidney-disease> [Accessed November 22nd, 2022]

13. Kashem, T. S., Begum, N. A. S., Nobil, F., Arefin, S. Z., & Rashid, H. U. (2021). POS-529 COVID 19 INFECTION AMONG HAEMODIALYSIS PATIENTS OF A SPECIALIZED KIDNEY HOSPITAL IN BANGLADESH. *Kidney International Reports*, 6(4), S230-S231.
14. Jahan, F., Rahman, A. S., Mahbub, T., Noman, M. U., Akter, Y., Rahaman, M. M., ... & Chowdhury, M. J. (2019). Awareness of chronic kidney disease among patients attending tertiary care hospital in Bangladesh. *Journal of Biosciences and Medicines*, 7(8), 106-118.
15. Rashid, H.U. (2011) Reports of President, 9th Conference of NUTS of SAARC Country.
16. Plantinga, L. C., Boulware, L. E., Coresh, J., Stevens, L. A., Miller, E. R., Saran, R., ... & Powe, N. R. (2008). Patient awareness of chronic kidney disease: trends and predictors. *Archives of internal medicine*, 168(20), 2268-2275.
17. Hemmelgarn, B. R., Zhang, J., Manns, B. J., James, M. T., Quinn, R. R., Ravani, P., ... & Alberta Kidney Disease Network. (2010). Nephrology visits and health care resource use before and after reporting estimated glomerular filtration rate. *Jama*, 303(12), 1151-1158.
18. Shah, S. N., Abramowitz, M., Hostetter, T. H., & Melamed, M. L. (2009). American journal of kidney diseases: the official journal of the National Kidney Foundation. *Am J Kidney Dis*, 54(2), 270-277.
19. Boulware, L. E., Troll, M. U., Jaar, B. G., Myers, D. I., & Powe, N. R. (2006). Identification and referral of patients with progressive CKD: a national study. *American journal of kidney diseases*, 48(2), 192-204.
20. Gøransson, L. G., & Bergrem, H. (2001). Consequences of late referral of patients with end-stage renal disease. *Journal of internal medicine*, 250(2), 154-159.
21. Kinchen, K. S., Sadler, J., Fink, N., Brookmeyer, R., Klag, M. J., Levey, A. S., & Powe, N. R. (2002). The timing of specialist evaluation in chronic kidney disease and mortality. *Annals of internal medicine*, 137(6), 479-486.
22. Stack, A. G. (2003). Impact of timing of nephrology referral and pre-ESRD care on mortality risk among new ESRD patients in the United States. *American journal of kidney diseases*, 41(2), 310-318.
23. Chan, M. R., Dall, A. T., Fletcher, K. E., Lu, N., & Trivedi, H. (2007). Outcomes in patients with chronic kidney disease referred late to nephrologists: a meta-analysis. *The American journal of medicine*, 120(12), 1063-1070.
24. Rashid, H. U., Alam, M. R., Khanam, A., Rahman, M. M., Ahmed, S., Mostafi, M., ... & Islam, N. (2021). Nephrology in Bangladesh. In *Nephrology Worldwide* (pp. 221-238). Springer, Cham.
25. <https://www.worldlifeexpectancy.com/bangladesh-kidney-disease> [Accessed November 22nd, 2022]
26. <https://www.tbsnews.net/bangladesh/health/kidney-disease-deaths-tripled-2020-bbs-214390> [Accessed November 22nd, 2022]
27. Cruz, Joseph A., and David S. Wishart. "Applications of Machine Learning in Cancer Prediction and Prognosis.", *Cancer Informatics*, JULY 2020
28. Cagatay Catal, Banu Diri, "A systematic review of software fault prediction studies", *Expert Systems with Applications*, Volume 36, Issue 4, 2009, Pages 7346-7354, ISSN 0957-4174. JULY 2020

29. V. B. Kumar, S. S. Kumar and V. Saboo, "Dermatological disease detection using image processing and machine learning," 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, pp. 1-6, JULY 2020.
30. Ewout W Steyerberg, Tjeerd van der Ploeg and Ben Van Calster, "Risk prediction with machine learning and regression methods", *Biometrical Journal*, 56: 601-606, JULY 2020.
31. Kleiman, R. S., LaRose, E. R., Badger, J. C., Page, D., Caldwell, M. D., Clay, J. A., & Peissig, P. L. (2018). Using machine learning algorithms to predict risk for development of calciphylaxis in patients with chronic kidney disease. *AMIA Summits on translational science proceedings, 2018*, 139.
32. Tekale, S., Shingavi, P., Wandhekar, S., & Chatorikar, A. (2018). Prediction of chronic kidney disease using machine learning algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 7(10), 92-96.
33. Aljaaf, A. J., Al-Jumeily, D., Haglan, H. M., Alloghani, M., Baker, T., Hussain, A. J., & Mustafina, J. (2018, July). Early prediction of chronic kidney disease using machine learning supported by predictive analytics. In *2018 IEEE congress on evolutionary computation (CEC)* (pp. 1-9). IEEE.
34. Ravizza, S., Huschto, T., Adamov, A., Böhm, L., Büsser, A., Flöther, F. F., ... & Petrich, W. (2019). Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nature medicine*, 25(1), 57-59.
35. Bradley, R., Tagkopoulos, I., Kim, M., Kokkinos, Y., Panagiotakos, T., Kennedy, J., ... & Elliott, J. (2019). Predicting early risk of chronic kidney disease in cats using routine clinical laboratory tests and machine learning. *Journal of veterinary internal medicine*, 33(6), 2644-2656.
36. Alloghani, M., Al-Jumeily, D., Hussain, A., Liatsis, P., & Aljaaf, A. J. (2020). Performance-based prediction of chronic kidney disease using machine learning for high-risk cardiovascular disease patients. In *Nature-inspired computation in data mining and machine learning* (pp. 187-206). Springer, Cham.
37. Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., ... & Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine*, 17(1), 1-13.
38. Makino, M., Yoshimoto, R., Ono, M., Itoko, T., Katsuki, T., Koseki, A., ... & Suzuki, A. (2019). Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Scientific reports*, 9(1), 1-9.
39. Akbilgic, O., Obi, Y., Potukuchi, P. K., Karabayir, I., Nguyen, D. V., Soohoo, M., ... & Kovesdy, C. P. (2019). Machine learning to identify dialysis patients at high death risk. *Kidney international reports*, 4(9), 1219-1229.
40. 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.



41. Sobrinho, A., Queiroz, A. C. D. S., Da Silva, L. D., Costa, E. D. B., Pinheiro, M. E., & Perkusich, A. (2020). Computer-aided diagnosis of chronic kidney disease in developing countries: A comparative analysis of machine learning techniques. *IEEE Access*, 8, 25407-25419.
42. Gudeti, B., Mishra, S., Malik, S., Fernandez, T. F., Tyagi, A. K., & Kumari, S. (2020, November). A novel approach to predict chronic kidney disease using machine learning algorithms. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1630-1635). IEEE.
43. Wang, W., Chakraborty, G., & Chakraborty, B. (2020). Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm. *Applied Sciences*, 11(1), 202.
44. Jena, L., Patra, B., Nayak, S., Mishra, S., & Tripathy, S. (2021). Risk prediction of kidney disease using machine learning strategies. In *Intelligent and cloud computing* (pp. 485-494). Springer, Singapore.
45. Arulanthu, P., & Perumal, E. (2021). Risk factor identification, classification and prediction summary of chronic kidney disease. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, 14(8), 2551-2562.
46. Kumar, A., Sinha, N., Bhardwaj, A., & Goel, S. (2021). Clinical risk assessment of chronic kidney disease patients using genetic programming. *Computer Methods in Biomechanics and Biomedical Engineering*, 1-9.
47. Ventrella, P., Delgrossi, G., Ferrario, G., Righetti, M., & Masseroli, M. (2021). Supervised machine learning for the assessment of chronic kidney disease advancement. *Computer Methods and Programs in Biomedicine*, 209, 106329.
48. Dritsas, E., & Trigka, M. (2022). Machine Learning Techniques for Chronic Kidney Disease Risk Prediction. *Big Data and Cognitive Computing*, 6(3), 98.
49. Silveira, A. C. D., Sobrinho, Á., Silva, L. D. D., Costa, E. D. B., Pinheiro, M. E., & Perkusich, A. (2022). Exploring Early Prediction of Chronic Kidney Disease Using Machine Learning Algorithms for Small and Imbalanced Datasets. *Applied Sciences*, 12(7), 3673.
50. Khamparia, A., Saini, G., Pandey, B., Tiwari, S., Gupta, D., & Khanna, A. (2020). KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network. *Multimedia Tools and Applications*, 79(47), 35425-35440.
51. Chimwayi, K. B., Haris, N., Caytiles, R. D., & Iyengar, N. C. S. (2017). Risk level prediction of chronic kidney disease using neuro-fuzzy and hierarchical clustering algorithm (s).

# Chronic\_Kidney\_Disease\_Minor\_Revision\_1.pdf

## ORIGINALITY REPORT

29%

SIMILARITY INDEX

24%

INTERNET SOURCES

17%

PUBLICATIONS

9%

STUDENT PAPERS

## PRIMARY SOURCES

1	<a href="https://dspace.daffodilvarsity.edu.bd:8080">dspace.daffodilvarsity.edu.bd:8080</a> Internet Source	8%
2	Submitted to Daffodil International University Student Paper	3%
3	<a href="https://medium.com">medium.com</a> Internet Source	3%
4	<a href="https://www.researchgate.net">www.researchgate.net</a> Internet Source	2%
5	<a href="https://www.scirp.org">www.scirp.org</a> Internet Source	2%
6	Submitted to Berlin School of Business and Innovation Student Paper	1%
7	<a href="https://mdpi-res.com">mdpi-res.com</a> Internet Source	1%
8	<a href="https://etd.astu.edu.et">etd.astu.edu.et</a> Internet Source	1%
9	<a href="https://doaj.org">doaj.org</a> Internet Source	1%