Diagnosis of Chronic Kidney and Liver Disease using Machine Learning with Hybrid Feature Selection

This project is submitted to the Department of Computer Science and Engineering, Dhaka International University, in partial fulfillment to the requirements of Bachelor of Science (B. Sc.) in Computer Science and Engineering (CSE).

Submitted by

NAME	REG. NO	ROLL NO
LUTHFURZZAMAN BABU	CS-D-58-19-111668	06
FORHADUL ISLAM	CS-D-58-19-111848	18
TANIA SULTANA	CS-D-58-19-111887	22
GOLAM MASUM	CS-D-58-19-111889	23
ABID BIN AHOSAN	CS-D-58-19-111993	43

Project Work, CSE-425

Batch: 58th, Session (2019-20)

Supervised by

Khandaker Mohammad Mohi Uddin
Assistant Professor

Knowledge is Power
DIU
Estd. 7th April, 1995

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
FACULTY OF SCIENCE AND ENGINEERING
DHAKA INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
FEBRUARY-2024

SUPERVISOR'S STATEMENT

This is to certify that the project paper entitled as "**Diagnosis of Chronic Kidney Disease and Liver Disease using Machine Learning with Hybrid Feature Selection**" submitted by LUTHFURZZAMAN BABU, Roll No:- 06; FORHADUL ISLAM, Roll No:- 18; TANIA SULTANA, Roll No:- 22; GOLAM MASUM, Roll No:- 23; ABID BIN AHOSAN, Roll No:- 43; has been carried out under my supervision. This project has been prepared in partial fulfillment of the requirement for the Degree of B.Sc. in Computer Science & Engineering, Department of Computer Science & Engineering, Dhaka International University, Dhaka, Bangladesh.

Supervisor's Signature

Date: …………………

...........…...…………………………………

Khandaker Mohammad Mohi Uddin
Assistant Professor
Dept. of Computer Science & Engineering
Dhaka International University

APPROVAL

The project report as **“Diagnosis of Chronic Kidney Disease and Liver Disease using Machine Learning with Hybrid Feature Selection”** submitted by LUTHFURZZAMAN BABU, FORHADUL ISLAM, TANIA SULTANA, GOLAM MASUM, ABID BIN AHOSAN to the Department of Computer Science & Engineering, Dhaka International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents.

<u>Board of Honorable Examiners</u>

1. Chairman

…………………………………………

Prof. Dr. A. T. M. Mahbubur Rahman
Dean
Faculty of Science and Engineering and
Chairman
Dept. of Computer Science and Engineering
Dhaka International University

2. Internal Member

....

Associate Prof. Md. Abdul Based
Professor,
Dept. of Computer Science and Engineering,
Dhaka International University

3. Internal Member

....

Mst. Jahanara Akhtar
Professor,
Dept. of Computer Science and Engineering,
Dhaka International University

4. Supervisor and Member

....

Khandaker Mohammad Mohi Uddin
Assistant Professor,
Dept. of Computer Science and Engineering
Dhaka International University

5. External Member

…...

Dr. Md. Manowarul Islam
Associate Professor
Dept. of Computer Science and Engineering,
Jagannath University

DECLARATION

We hereby declare that; this project has been carried out by us and it has been submitted for the award of the B.Sc. degree. We also certify that this project was prepared by us for the purpose of fulfillment of the requirements for the Bachelor of Science (B.Sc.) in Computer Science & Engineering.

Authors Signature

....
LUTHFURZZAMAN BABU
B.Sc. In CSE, Roll No: 06
Reg. No: CS-D-58-19-111668
Batch: 58th, Session: 2019-20
Dhaka International University

....
FORHADUL ISLAM
B.Sc. In CSE, Roll No: 18
Reg. No: CS-D-58-19-111848
Batch: 58th, Session: 2019-20
Dhaka International University

....
TANIA SULTANA
B.Sc. In CSE, Roll No: 22
Reg. No: CS-D-58-19-111887
Batch: 58th, Session: 2019-20
Dhaka International University

....
GOLAM MASUM
B.Sc. In CSE, Roll No: 23
Reg. No: CS-D-58-19-111889
Batch: 58th, Session: 2019-20
Dhaka International University

....
ABID BIN AHOSAN
B.Sc. In CSE, Roll No: 43
Reg. No: CS-D-58-19-111993
Batch: 58th, Session: 2019-20
Dhaka International University

Supervisor's Signature

Date: ……………

....
Khandaker Mohammad Mohi Uddin
Assistant Professor
Dept. of Computer Science and Engineering
Dhaka International University

ABSTRACT

Among the most critical illnesses in the current scenario, Chronic Kidney Disease (CKD) and Liver Disease (LD) stand out, and a prompt diagnosis is essential to effectively address them. Early detection of these diseases is crucial for preventing millions of deaths. In the field of medical treatment, machine learning techniques have become reliable. To ensure timely detection, doctors are availing assistance from various Machine Learning (ML) classifiers. This study focuses on implementing a diagnostic tool to identify CKD and LD using different ML algorithms with a fused feature selection approach. Various methods of data preparation and the Synthetic Minority Oversampling Technique (SMOTE) have been employed to handle data imbalance. Ten ML classifiers are utilized in this study. To eliminate redundant features, a fused strategy-based feature selection approach, including Pearson correlation matrix, Mutual Information (MI), and Chi-squared test (Chi2), is applied. Out of the ten ML models, the AdaBoost classifier achieved the highest accuracy of 99.19% in diagnosing CKD, while the Light Gradient Boosting Machine (LGBM) classifier attained the highest accuracy of 83.74% in diagnosing LD. The second-highest accuracy for LD, 83.25%, was obtained from the voting classifier, which was merged with the four classifiers with the highest accuracy.

ACKNOWLEDGEMENTS

We would like to pay our gratitude to the Almighty Allah who created us with not only the ability to design and program this system but also the power of practice.

We would also like to express our sincere thanks to our respected supervisor **Khandaker Mohammad Mohi Uddin, Assistant Professor**, Department of CSE, Dhaka International University for his continuous encouragement, motivation and professional guidance during the work of this project.

We would like to thank all the faculty members for their valuable time spent on requirements analysis and evaluation of the project work.

We would like to express our sincere and cordial gratitude to the people those who have supported us directly, provided mental encouragement. We are also thankful to our family and friends who have contributed directly or indirectly the development word and its associated activities.

We warmly thank **Prof. Dr. A. T. M. Mahbubur Rahman,** Dean, Faculty of Science and Engineering and Chairman, Department of Computer Science and Engineering. His extensive discussion around work and interesting exploration in operations has been very helpful for this study.

Finally, we would like to dedicate this project to our teachers for their love, encouragement and professional guidance throughout the project.

DEDICATION

Dedicated to

Our Parents

&

Teachers

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 INTRODUCTION

1.1 OVERVIEW

This research submission presents a comprehensive exploration into the development and implementation of Machine Learning(ML) models for the early detection and prediction of Chronic Kidney Disease(CKD) and Liver Disease(LD). The integration of diverse ML techniques showcases a holistic approach towards improving diagnostic accuracy and facilitating timely interventions for these critical medical conditions. This research submission provides a detailed account of our exploration into the integration of feature selection techniques within machine learning models for the early detection and prediction of CKD and LD. By incorporating advanced feature selection methodologies, we aim to enhance the efficiency and interpretability of our models, facilitating more accurate predictions and furthering the understanding of key factors influencing disease outcomes.

1.2 PROBLEM DEFINITION

A decline in kidney function or kidney damage due to any cause that lasts longer than three months is the hallmark of CKD, an irreversible disorder that progresses over time. There are two kidneys in a human body. These are essential organs required for the correct functioning of the peritoneal cavity, and they are situated near its rear. The kidneys are bean-shaped organs that are situated one on each side of the spine, beneath the rib cage. [1]. Every half an hour, they filter all of the blood in the body. Toxins, waste materials, and extra fluid are all eliminated from the body via the kidneys. Additionally, they maintain proper levels of electrolytes, control blood pressure, promote the development of red blood cells, maintain the integrity of bones, and control blood molecules that are necessary for life. Blood waste and excess fluid remain in the body and can cause heart disease and stroke, among other health problems, if the kidneys are damaged and cannot properly filter blood.

Among the several causes of CKD include heart disease, diabetes, and high blood pressure. High blood sugar levels are a hallmark of diabetes, which harms the heart, blood vessels, kidneys, and eyes. Additionally, insufficient blood pressure management can significantly increase the risk of CKD, heart attack, and stroke. Glomerulonephritis, genetic disorders, dysplasia, kidney stones, tumors, recurrent UTIs, metabolic diseases, obesity, and ageing are other issues that can affect the kidneys [2,3]. Figure 1.1 illustrates a few of the influences that affect CKD. In stages 1 through 5, the initial phases of CKD, there are frequently no external symptoms. As CKD advances, you may have symptoms including nausea and vomiting, cramping in your muscles, edema in your feet and ankles, dry, itchy skin, dyspnea, and excessive or insufficient urination. Generally, the first stages of the disease are referred to as stages 1 through 3. However, these are typically in the advanced phases. When the kidney is damaged badly only then CKD does give signs otherwise not. CKD also depends on age and gender.

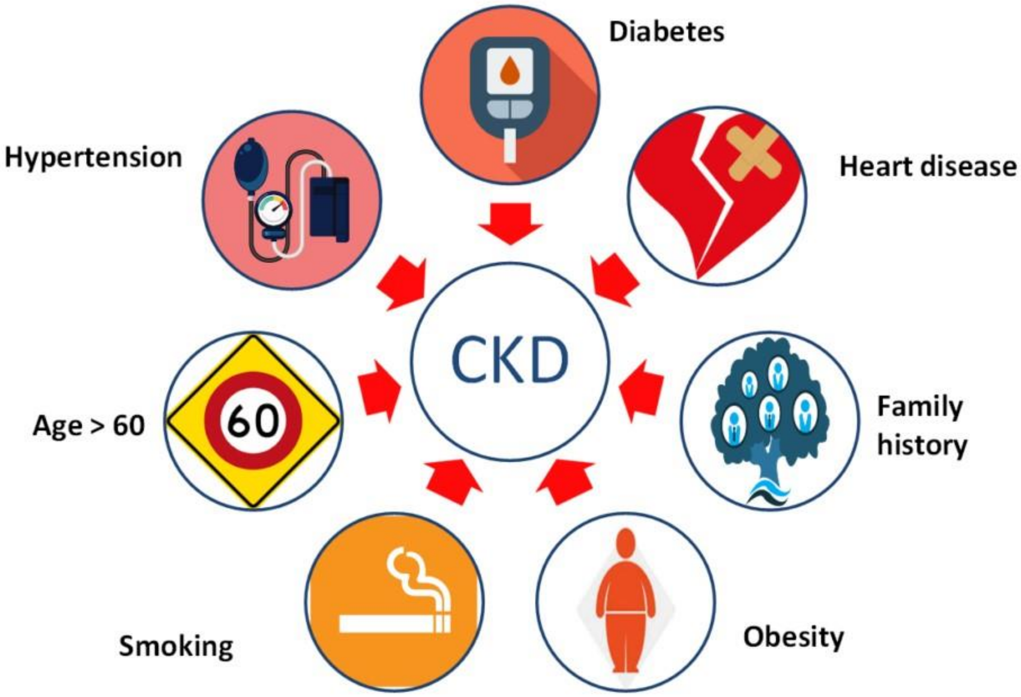


Figure 1.1 Factors influencing CKD.

Based on the rate of glomerular filtration, it is split into phases. It reflects how well the kidneys are working. It displays the pace at which filtrate moves from the nephron's glomerulus right inside Bowman's capsule and then into the tubules of Renal. The phases are marked one to five with two groups in stage three. The GFR must be below 60 milliliters per minute to be considered decreased, however in groups one and two, the GFR may be around 60 despite the presence of kidney injury. A GFR of less than 15 or the requirement for treatment of renal replacement, such as kidney transplantation or dialysis, are considered signs of end-stage kidney illness. Signs of kidney problems can be identified through imaging or examination, as an elevated albumin to ratio of creatinine. Which in conjunction with GFR, measures the quantity of protein emitted in urine. Considering the ratio of albumin to creatinine, which shows greater values signifying bigger harm and a higher chance of growth, these individuals are further separated into three categories. Doctors utilize the kidney function indicator known as the (GFR) to identify renal diseases. The patient's gender, age, blood test results, and other relevant factors are used to compute GFR. Based on the GFR number, doctors can divide CKD into five phases [4]. The phases of renal disease as determined by GFR are shown in Table 1.1. Kidney failure can be avoided if CKD is identified and treated in the early phases.

Table 1.1 The CKD development phases (Anon, 2015)

Phases of CKD	The rate of Globular Filtration(ml/min/1.73m^2)	Description
1	>=90	Impairment of the kidneys with adequate renal function proteinuria is present.
2	60-89	Kidney damage and a little drop in GFR.
3	30-59	Damage in the kidneys and a mild decline in GFR.
4	15-29	Damage to the kidneys and a significant drop in GFR.
5	<=15	End-stage renal disease and kidney failure.

CKD = Chronic kidney disease, GFR = Glomerular filtration rate.

On the other hand, Liver is the largest and powerful organ that performs hundreds of essential functions in our body. It helps in the removal of toxins, energy storage, and the digestion of food. It is a major multitasker. It performs about to 500 function by itself and in connection with another body system. Right upper quadrant of the abdominal cavity, directly below the diaphragm, is home to the liver, a triangle organ and it has reddish brown color and averaging 1.6 kg (3.5 pounds) in weight. The liver's greater mass, the right lobe, is located on the right side of the body and descends downward towards the right kidney, whereas the smaller left lobe is located on the stomach. The kidney receives its abundant blood supply primarily from the hepatic portal vein and the hepatic artery. Nutrient-rich blood from the spleen and the digestive tract (stomach, intestine, and color) is supplied by the hepatic portal veins. The heart pumps blood enriched with oxygen through the hepatic artery. It supports the body's immune system, metabolism, digestion, and storage of vital nutrients. The liver is a vital component of the body because without it, the body's tissues run the risk of dying from a lack of nutrients and energy. An assortment of ailments affecting the liver might be referred to as hepatic or liver disorders. It is also known as chronic liver disease if it persists for a long time. LD frequently share characteristics in common, despite their detailed differences.

It is also hard for a person to exist without a functional liver because of this function. Compared to other parts of the body, the liver has a significant volume of blood running through it. The liver's primary function is to filter blood arriving from the digestive tract before it is sent to the other parts of the body, using about 13% of the blood flowing through the body at any given time. A pair of lobes form the liver – the right and the left lobe made up of 8 segments each. There are thousands of lobules in each segment.

Disease has the ability to disrupt these processes, and in severe circumstances, completely stop them. Liver has two unique qualities that impact the progression of disease. Initially, a large portion of the harm we inflicted on the liver can be repaired and regeneration occurs naturally. Secondly, the liver lacks nerve endings that would indicate the presence of illness. Consequently, the majority of individuals are ignorant about the onset and progression of disease.

Damage to the liver might become more irreversible as it starts to outweigh its natural ability to repair itself. A common misunderstanding is that excessive alcohol uses is the only causes of LD. Although, alcohol is a common cooperate it is only one of many. Other cause, include for diet, and obesity, viruses such as hepatitis A, B, C, D or E [5]. An individual may be considered at-risk if they have a family history of LD, excessive use of prescription and over-the-counter medications, excessive use of illicit drugs, or an autoimmune condition that prompted the immune system to attack the liver. Examples include fatty LD and cirrhosis [6]. Disorders passed down via families, as Wilson disease and hemochromatosis [7].

One of the primary LDs, Non-Alcoholic Fatty Liver Disease (NAFLD), is characterized by a buildup of lipids in the liver. It's known as "non-alcoholic steatohepatitis" if there is liver cell injury and inflammation [8]. Cirrhosis is also one of the most dangerous LDs. Scar tissue replaces healthy tissue as a result of this condition. As a result, the liver sustains irreversible damage and becomes dysfunctional. Liver cirrhosis is mostly caused by drinking, NAFLD, chronic hepatitis B, and chronic hepatitis C [9].

Hepatitis or alcoholism are typically linked to LD, but obesity and diabetes are also increasingly linked to possibly catastrophic liver damage [10]. A person with advanced fatty LD has a nearly seven-fold increased risk of dying. It is a quiet "killer," and things become worse quickly if indications of fatty LD start to show [11,12].

While LD symptoms can vary, they frequently include jaundice, or yellowing of the skin and eyes, bruising easily, changes in the color of your urine and stool, and swelling of the abdomen and legs. There may not always be any symptoms. Tests that measure liver damage and aid in the diagnosis of liver illnesses include imaging tests and liver function tests. It is recommended that patients avoid heavy alcohol consumption in order to prevent liver damage.

On the other hand, for those with alcoholic hepatitis, hepatitis B or C, etc, the simple recommendation is to fully abstain from alcohol [13]. Other preventative methods include using a condom during sexual activity, avoiding sharing syringes or needles, receiving the hepatitis A and B vaccine, and protecting the skin from dangerous substances. In the end, maintaining a healthy weight, eating a balanced food, and exercising all support the liver's optimal function [14].

1.3 MOTIVATION FOR FINDING MOTIF

The motivation for identifying motifs in the context of CKD and LD prediction stems from a deep-seated commitment to enhancing the precision and interpretability of predictive models. Several key factors drive the pursuit of motifs in this project:

Complex Multifactorial Nature of Diseases: CKD and LD are inherently complex, often influenced by a multitude of genetic, environmental, and lifestyle factors. Identifying motifs, i.e., recurring patterns or combinations of features, allows us to capture the intricate interplay of variables contributing to disease manifestation. Understanding these motifs can unravel hidden relationships and provide insights into the nuanced nature of these conditions.

Personalized and Precision Medicine: The era of personalized medicine emphasizes the need to tailor healthcare interventions to individual patient characteristics. Motif discovery enables the identification of patient-specific patterns, facilitating the development of personalized treatment plans. By discerning motifs related to disease progression or susceptibility, we aim to contribute to a more nuanced and individualized approach to patient care.

Interpretability and Trust in Models: The incorporation of motifs enhances the interpretability of ML models. Instead of viewing predictions as black-box outputs, understanding the motifs allows healthcare professionals to grasp the underlying rationale behind predictions. This transparency builds trust in the models, fostering greater acceptance and adoption in clinical settings.

Early Detection and Intervention: Early detection is critical in managing CKD and LD effectively. Motif-based analysis offers a pathway to identify subtle yet indicative patterns that precede overt symptoms. By capturing these early motifs, our models aim to facilitate timely intervention strategies, potentially preventing disease progression and improving overall patient outcomes.

Optimizing Feature Selection: Motif discovery aligns with our feature selection approach, aiming to identify compact subsets of features that contribute significantly to predictive accuracy. By prioritizing motifs, we can streamline the feature set, leading to more efficient models that are less prone to overfitting and more adaptable to different patient populations.

Facilitating Biomarker Discovery: Motif analysis holds promise for uncovering potential biomarkers associated with CKD and LD. Identifying recurring motifs may highlight specific combinations of clinical, genetic, or demographic features that serve as reliable indicators of disease risk. These discovered motifs may pave the way for the development of novel diagnostic and prognostic biomarkers.

Advancing Scientific Understanding: Beyond immediate clinical applications, the pursuit of motifs contributes to the broader scientific understanding of the intricate mechanisms underlying CKD and LDs. Unraveling recurrent motifs provides valuable insights into the complex interdependencies of variables, potentially opening avenues for further research and therapeutic innovation.

1.4 CASE HISTORY

CKD caused the death of huge amount of people globally. According to a recent medical study, CKD affects 324 million people worldwide [15], The estimated prevalence in the US is 14%. An estimated one in seven adult Americans in the United States has CKD. It is considered as a most common cause of death in the US [16], While CKD is considered to be 13.4% common worldwide [17] it's claimed to be 13.9% common in the African Sahara [18,19]. The highest frequency of CKD in Africa, according to a different study, was found in West Africa at 16% [20]. According to numerous research studies, CKD is increasingly common for the nations that are developing [21]. Significantly, it has been asserted that 1 in 10 residents of South Asia, which include Bangladesh, Bhutan, India, Pakistan, and Nepal, suffer from CKD [22]. It is flattering a common disease worldwide. With an estimated 1.2 million fatalities each year, Prof. Harun-Ur-Rashid noted in his presentation that CKD ranks as the eleventh leading cause of death worldwide [23]. The Renal Foundation estimates that every year 35,000–40,000 of Bangladesh's 18 million CKD patients experience renal failure. According to the National Institute of Health, adult Bangladeshis had a 17.3% prevalence of CKD, which ranged from 12.8% to 26.0% [24]. CKD is increasing rapidly.

Every year, almost 2 million people worldwide pass away from LD. Of these deaths, a million are due to complications from cirrhosis, and a further million are from hepatocellular carcinoma and viral hepatitis. Globally, cirrhosis ranks 11th, and liver cancer ranks 16th respectively, contributing to 3.5% of all fatalities. At 1.6% and 2.1% of the global burden, respectively, cirrhosis is among the top 20 causes of years of life lost and disability-adjusted life years. Worldwide, there are about 2 billion alcohol drinkers, of whom up to 75 million people have been diagnosed with problems related to alcohol use and who are susceptible to LD associated with alcohol use. Approximately 2 billion people are overweight or obese, and over 400 million adults have diabetes. These conditions increase the probability of having hepatocellular carcinoma and NAFLD [25]. Acute hepatitis is primarily caused by drug-induced liver injury, however viral hepatitis is still very common worldwide. While liver replacement is the second most popular solid organ transplant technique, at present rates fewer than 10% of the world's transplant needs are satisfied. Although these numbers are concerning, they also indicate a great opportunity to improve public health, as most causes of LD are preventable [26].

1.5 CLASSIFIERS

In the latest modern technology and the most efficient method for data analyzing and pattern detection are ML and Deep Learning (DL). Doctors are using ML classifier methods can detect the illness more quickly than any other technique currently in use. These authors proposed various algorithms like,

- Decision tree classifier (DT)
- Random forest classifier (RF)
- Naïve Bayes classifier (NB)
- AdaBoost classifier (AB)
- Cat-Boost classifier (CB)
- Light-GBM classifier (LGBM)
- Support Vector Machine (SVM)
- K-nearest neighbor classifier (KNN)
- Logistic Regression (LR)
- Voting classifier (VC)

to obtain an optimum result of prediction. ML methods can analyze patient data such as blood pressure, age, sex, and test results to forecast the chance of CKD. This can assist in identifying those who might gain from early intervention. Early detection of CKD may allow for improved medication and preventative measures, which may help patients avoid unfavorable outcomes like dialysis or transplantation. These algorithms can assist healthcare providers in diagnosing CKD by examining the medical history of a patient, symptoms, and test results. This also predicts a patient's GFR over time, helping clinicians monitor disease progression.

1.6 CONTRIBUTIONS

In order to effectively identify CKD and non-CKD, this study applies ten different machine learning methods, which helps to produce a reproducible hybrid feature section technique. To identify any condition that is similar, the suggested technique is also applicable to a different comparable health dataset. Test outcomes and comparisons with pertinent research showed that the fused feature section technique used by the AdaBoost classifier performed better than the others in almost all dataset models.

In the context of this project for LD predictions, we shall be particularly concerned about the occurrence of LD. The applied methodology has made the following important contributions:

- The initial steps in data preprocessing are to remove duplicate instances, handle null values, and SMOTE. Because the instances in the dataset are distributed in this manner, we can develop effective classification models and predict the earliest signs of liver illness.

- The Pearson Correlation matrix feature selection approach has been used to determine the important features.
- Several ML model’s performances are compared and evaluated using standard metrics like Precision, Accuracy, F1 force, Recall, and AUC. Ensemble ML technique the LGBM classifier outperformed the other models that made up the study work's proposal, according to the experimental results.
- Our major idea, the LGBM classifier, performed better in terms of accuracy when compared to published works that used use of the same dataset.

1.7 SUMMARY OF THE WORK

In this research endeavor, our focus centers on the development of advanced ML models for the prediction of CKD and LD. Leveraging diverse clinical, demographic, and laboratory datasets, our primary objective is to enhance the accuracy and interpretability of predictive models through the integration of sophisticated feature selection techniques.

The project unfolds in several key phases, beginning with meticulous dataset selection and preprocessing to ensure the inclusion of comprehensive and relevant information. Notably, our approach incorporates advanced feature selection methodologies, such as Pearson correlation matrix, Mutual Information (MI), and Chi-squared test (Chi2) is applied, aimed at identifying and prioritizing the most influential features.

Motivated by the imperative for personalized healthcare and early disease detection, our models integrate these selected features to create a robust predictive framework. We delve into the complex multifactorial nature of CKD and LDs, recognizing the significance of identifying motifs or recurring patterns within the data.

Moreover, our research contributes to the interpretability of ML models, fostering trust among healthcare professionals. By providing insights into the key factors influencing predictions, we aim to bridge the gap between data-driven analytics and actionable clinical decisions.

The validation process involves rigorous assessment using diverse datasets, ensuring the generalizability and reliability of our models. Performance metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) serve as benchmarks for evaluating the effectiveness of our predictive models.

Ultimately, our project not only seeks to advance the field of predictive medicine but also emphasizes the integration of feature selection techniques and motif analysis to optimize model performance. This subsection sets the stage for a comprehensive exploration of our methodology, results, and implications, contributing valuable insights to the intersection of ML and healthcare.

Chapter 2 RELATED WORKS

2.1 INTRODUCTION

A comprehensive review of related work lays the foundation for understanding the current landscape of predictive modeling for CKD and LD. In the pursuit of refining our approach, we delve into existing literature and research initiatives that contribute valuable insights to the intersection of ML and healthcare.

Scholars and practitioners alike have endeavored to address the challenges posed by the multifaceted nature of CKD and LDs. Previous works have explored diverse ML algorithms, ranging from traditional methods like logistic regression and decision trees to more complex deep learning architectures. These studies showcase a spectrum of methodologies employed for disease prediction, each offering unique advantages and considerations.

Feature selection, a key component of our project, has garnered attention in related works. Researchers have explored various techniques to identify the most discriminative features, enhancing model efficiency and interpretability. Understanding the nuances of feature relevance has emerged as a crucial aspect of building robust predictive models.

Motif analysis, though more prevalent in genomics, has been sparingly applied in the context of disease prediction. We aim to build upon existing research in motif discovery, adapting and extending these approaches to uncover recurring patterns that may hold significance in understanding the progression and susceptibility of CKD and LDs.

Validation strategies and performance metrics employed in prior studies offer valuable benchmarks for evaluating model effectiveness. Notably, the incorporation of diverse datasets and external validations in related works guides our own validation process, ensuring the reliability and generalizability of our predictive models.

As we embark on this research journey, a synthesis of the insights gained from the related work informs our methodology, guiding us toward innovative approaches and potential avenues for improvement. This section serves as a bridge between existing knowledge and our unique contribution to the evolving landscape of predictive medicine.

2.2 CHRONIC KIDNEY DISEASE

Using different categorization methods, many studies have been conducted to predict the phases of CKD. Table 2.1 lists the most recent approaches to CKD diagnosis utilizing performance metrics and ML algorithms.

J. Qin et al. [27] developed data claim and a diagnostic sample. KNN is employed to verify data. Authors employed six ML algorithms to improve diagnosis accuracy. Such as Naive Bayes classifier, K-nearest neighbor, Logistic Regression, Feed forward neural network, Random Forest and Support Vector Machine to diagnose CKD. The authors proposed hybrid models that used perception to integrate the LOG and RF. Which, following several replications, could reach a typical precision of 99.83%.

Table 2.1 Existing recent work on CKD prediction and methods.

Existing work	Dataset	Modality	Method	Validation
J. Qin et al. [27]	CKD = 250 & not CKD = 150	CKD diagnosis	LOG + RF	Random Patient Level
Silveira et al. [28]	60 real-world recordings and 54 manually enhanced data records provide the dataset.	CKD diagnosis	MA, SMOTE	Random Patient Level
Ebiaredoh-Mienye et al. [29]	CKD = 250 & not CKD = 150	CKD diagnosis	LR, SAE + Softmax	Random Patient Level
Yashfi, S.Y. et al. [30]	Results of clinical tests on 455 patients	CKD diagnosis	RF, ANN	Random Patient Level
Almansour et al. [31]	CKD = 250 & not CKD = 150	CKD or not CKD Detection	ANN, SVM	Random Patient Level
Kim et al. [32]	741 ultrasound pictures in total.	Severe CKD, mild and moderate CKD and normal	ANN	Random Patient Level
Sara et al. [33]	CKD = 250 & not CKD = 150	CKD diagnosis	SVM+HWFFS	Random Patient Level
Rady et al. [34]	25 factors were present in the clinical test findings for 361 Indian patients with CKD.	CKD diagnosis	PNN, MLP, SVM, RBF	Random Patient Level
polat et al. [35]	CKD = 250 & not CKD = 150	CKD diagnosis	SVM	Random Patient Level
P. Ghosh et al. [36]	CKD = 250 & not CKD = 150	CKD diagnosis	GB	Random Patient Level

Silveira et al. [28] established a CKD prediction system with ML algorithms and many resampling techniques. Two resampling techniques are SMOTE and Borderline-SMOTE. Models that are implemented include DT algorithms, RF algorithms, and multi-class Ada-Boosted DTs. For dynamic classifier selection, they also used the k-nearest oracles-union, k-nearest oracles-eliminate, and META-DES techniques for dynamic ensemble selection in addition to the overall local accuracy and local class accuracy approaches. The DT model with Manual Augmentation (MA) and SMOTE achieved the highest accuracy of (98.99%), according to the experiment results.

Ebiaredoh-Mienye et al. [29] approached for better medical diagnosis through the use of a soft max layer in an upgraded sparse autoencoder (SAE) network. In contrast to conventional thin-film autoencoders, which impose a weight penalty on the hidden layer activations, the neural network attained sparsity. The proposed SAE demonstrated 98% accuracy in CKD prediction. Which shows better execution analyzed than other AI algorithms.

Yashfi, S.Y. et al. [30] developed a method for CKD risk prediction using data from the UCI Machine Learning repository, 455 patients with the Khulna City Medical College real-time dataset. A 10x cross-validation was used to train and test the ANN and RF on the data. The ANN achieved accuracy of 94.5% and RF achieved accuracy of 97.12%. The early detection of chronic renal disorders will be assisted by this technology.

Almansour et al. [31] were proposed to make an early CKD diagnosis. In order to develop their classification algorithms, researchers analyzed a dataset of 400 patients and 24 parameters associated with the diagnosis of chronic renal illness. The accuracy of ANN and SVM, which the researchers used to diagnose CKD, was found to be 99.75% and 97.75%, respectively.

Kim et al. [32] have proposed a genetic algorithm (GA) based on Neural Networks (NN) for the diagnosis of CKD. Their suggested GA enhanced the weight vectors for NN training. 741 ultrasound images were used in the study, comprising 251 images of kidneys in good health, 328 images of individuals with mild to moderate CKD, and 162 images of those with severe CKD. The authors used an Artificial Neural Network (ANN) to classify data with 95.4% accuracy.

Sara et al. [33] employed two techniques, such as Hybrid Wrapper and Filter-Based FS (HWFFS) and Feature Selection (FS), to significantly select the features associated with CKD and reduce the dimensionality of the dataset. After combining the information from the two methods, an SVM classifier was used to determine the hybrid traits.

Rady et al. [34] proposed a technique for diagnosis a dataset for CKD. The researchers used a dataset with 25 characteristics and 361 clinical test results from Indian patients with CKD. They applied Probabilistic Neural Networks (PNN), Multilayer Perceptions (MLP), Radial Basis Functions (RBF), and SVM algorithms. The authors discovered that SVM, MLP, and RBF algorithms outperformed the PNN approach with an accuracy of 96.7%.

Polat et al. [35] employed the SVM approach to predict CKD. They were working on a significant feature in order to get the right outcome. To find the right feature, they used an SVM method in conjunction with a two-approach wrapper and filter. In the Wrapper technique, there was a greedy stepwise search engine for the classifier subset evaluator and a best first search engine for the Wrapper subset evaluator. The filtered subset evaluator in the filter technique had the best first search engine, while the correlation feature section subset feature had a greedy stepwise search engine. The SVM Classifier, which combines the optimal filtered subset evaluator feature selection method for search engines, has a higher accuracy rate of 98.5% in the diagnosis of CKD, according to the comparison of the findings of all techniques.

P. Ghosh et al. [36] used a reliable dataset and the assessment process of a ML system could identify this illness considerably more quickly. Linear Discriminant Analysis (LDA), Gradient Boosting (GB), AB, and SVM methods have been used in the study to produce accurate results for CKD prediction. The UCI ML repository's data was used to evaluate the models' efficacy. With an accuracy rating of 99.80%, the gradient boosting classifier was the most accurate.

2.3 LIVER DISEASE

In Table 2.2 the models that have been suggested by previous research works on LD using the same dataset [37]. In comparison to the other works, our suggested model LGBM performs better, with an accuracy of 83.74%.

Research based on the Indian Liver Patients' Records dataset is presented in this section [37] that utilize several ML models to forecast the occurrence of LD. In particular, With 72% accuracy, the Gradient Tree Boosting classifier for a balanced dataset performed best in [38]. Further research was conducted by the authors at [39] to see whether KNN using feature selection techniques (KNNWFST) performs better than any other comparable. The authors of [40,41] offered the LR model, which had a 75% accuracy rate. While [42] shows that the SVM achieves 75.04% accuracy. The authors of [43] suggested a hybrid ML model named Mathematical Approach on Multilayer Feedforward Neural Network with Backpropagation (MAMFFN). The author of [44] proposed voting classifier and achieved accuracy of 80.10%

Table 2.2 Proposed models are illustrated using the same dataset as the source publications.

Research work	Modality	Proposed Model	Accuracy
A. Sokoliuk et al. [38]	LD diagnosis	Gradient Tree Boosting	72%
M. Azam et al. [39]	LD diagnosis	KNNWFST	74%
A. Srivastava et al. [40]	LD diagnosis	LR	75%
R. Choudhary et al. [41]	LD diagnosis	LR	75%
C. Geetha et al. [42]	LD or not LD Detection	SVM	75.04%
G. Gajendran et al. [43]	LD diagnosis	MAMFFN	75.30%
E. Dritsas et al. [44]	LD diagnosis	Voting	80.10%

2.4 SUMMARY

A survey of the existing body of work in the field of CKD and LD prediction using ML provides a rich tapestry of methodologies and findings. Numerous studies have employed various algorithms, ranging from classical models like logistic regression to more intricate DL architectures, showcasing the versatility in approach.

Chapter 3 PREPOSED METHODOLOGY

3.1 INTRODUCTION

Our proposed methodology represents a strategic synthesis of advanced ML techniques tailored to the unique challenges of predicting CKD and LD. Informed by insights from related work, our approach is designed to optimize model accuracy, interpretability, and clinical relevance.

Data Preprocessing: The initial phase involves meticulous dataset selection and preprocessing, ensuring the inclusion of comprehensive clinical, demographic, and laboratory parameters. Cleaning, normalization, and feature engineering are applied to prepare the data for subsequent analysis.

Feature Selection: Building on the importance highlighted in related work, our methodology incorporates advanced feature selection techniques. Univariate selection, recursive feature elimination, and feature importance analysis contribute to identifying and prioritizing the most influential features for disease prediction.

Model Development: Leveraging the insights from feature selection and motif analysis, our methodology integrates diverse ML models. This includes both DL architectures for discerning complex patterns and traditional algorithms for interpretability, striking a balance between model complexity and clinical utility.

Performance Metrics: The evaluation of our models is guided by comprehensive performance metrics, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). These metrics provide a holistic assessment of our models' effectiveness in predicting CKD and LDs.

Interpretability Measures: Recognizing the importance of model interpretability, our methodology includes measures to elucidate the significance of selected features and motifs. This transparency enhances the trustworthiness of our models for healthcare practitioners.

Here Figure 3.1 is the combaine proposed methodology's workflow for CKD and LD. First we collected a valid dataset from well know resource. Checked duplicate rows as multiple same row can wrongly predict the diagnosis. Some cell of data was NULL. For this reason calculation couldn’t done properly. Filled the NULL cells using fillna function. Some column was contained string value. As string can’t be train or test by machine so we did label encoding technique for converting string to numerical. After that most important has be done, that is feature selection. By feature selection we found most important features from the dataset that is most vital cause for CKD or LD. As output for both dataset wasn’t balanced that’s why SMOTE has been applied on both cases. Specially for LD we did a data scaling using min-Max scaling. After this we split the dataset as 75% for train and 25% for testing and trained with 10 different ML classifiers for better performance. After comparison performance and comparison finally we build a simple web app for both CKD and LD.

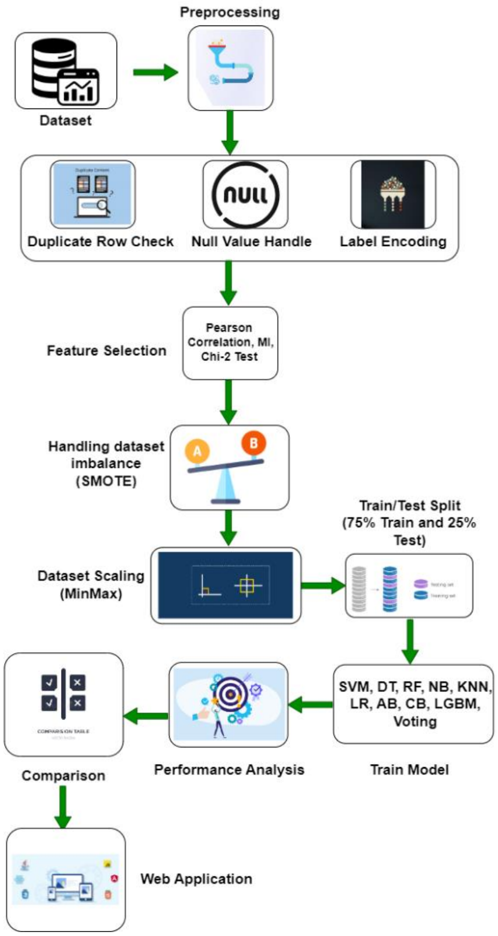


Figure. 3.1 Proposed methodology's workflow.

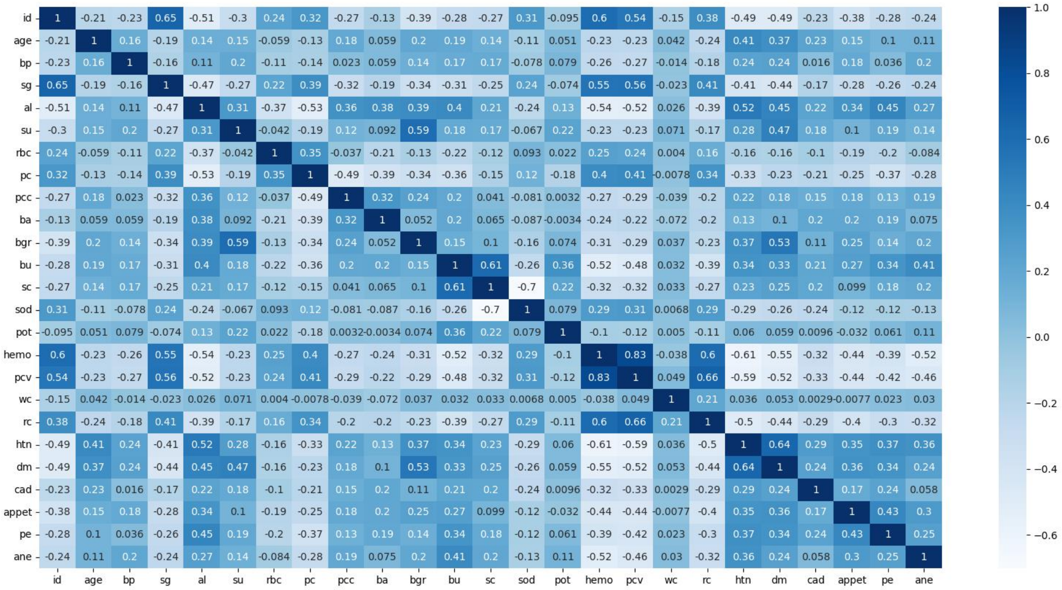


Figure. 3.2 Correlation matrix of all features for CKD.

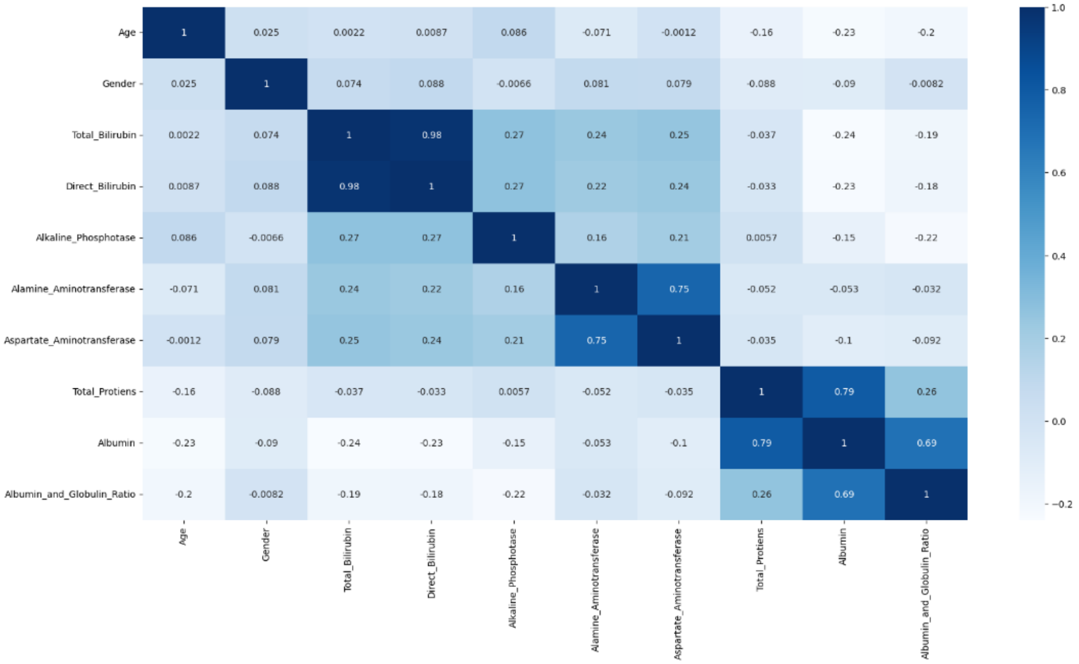


Figure. 3.3 Correlation matrix of all features for LD.

Figure 3.2 is showing the correlation of all features of CKD. Some features are very correlative with other features. The value of correlation is closer to 1 means highly correlative and the value of correlation is closer to -1 means uncorrelated. Figure 3.3 is correlation matrix of all features for LD.

3.2 MATERIALS AND METHODS

The different techniques and resources used in this study were covered in great detail in this section. Techniques for preparing datasets, several ML techniques for diagnosis, together with a feature selection method, are explained and looked at. The initial phase of this study was focused on the CKD dataset, so several experiments were carried out using 10 ML algorithms, such as K-Nearest Neighbors, Random Forest, Gaussian Naive Bayes, Decision Tree, Support Vector Machine, AdaBoost, Logistic Regression, Cat-Boost, Voting Classifier, and Light-GBM. The missing numerical values were calculated using the Median method. For preprocessing filling null values, label encoding, and the SMOTE is used. Pearson correlation matrix, Mutual Information (MI), and Chi-squared test (Chi2) for feature selection approach.

3.2.1 Dataset Collection

The ML Repository at the University of California, Irvine (Anon, 2022a) provided the CKD dataset used in this study, which contained information on 400 patients with CKD. The dataset includes class characteristics, like "CKD" and "not-CKD," for classification, as well as 24 features, split into 11 numerical traits and 13 categorical features. In the diagnostic class, the values "CKD" and "not-CKD" are present. A summary of the features of the dataset is provided in Table 3.1. To get the dataset ready for the model, several different data pretreatment techniques have been used.

The dataset of Indian Liver Patients' Records served as the foundation for our study [37]. There are 583 participants in this particular dataset, of whom 441 (75.64%) are men and 142 (24.36%) are women. The target class shows if the individual has received a diagnosis with whether or not they have LD. There are 416 participants (71.35%) who have been diagnosed with LD.

3.2.2 Dataset Description

The dataset's available features and their labels, which have an effect on how CKD is categorized, are listed in Table 3.2. Data indicates that there are currently 150 instances of "not-CKD" and 250 instances of the "CKD" class, making the dataset imbalanced. The features or characteristics of the dataset consist of one binary classification attribute and 24 unique features.

Table 3.1 A broad synopsis of the UCI ML repository's CKD dataset (Anon, 2022a).

Characteristics of Dataset	**Multivariate**
Characteristics of Attribute	Real
The total number of features(Col in general)	26
Number of cases	400
Number of characteristics have numerical values	14
Amount of categorical features	12
Amount of attributes we have regarded as independent variables	16
Name of target column	Classification
Instances are categorised as	CKD/Not CKD
The number of classes that make up CKD	250
How many classes are designated as NOT CKD	150
Source of data	UCI ML Repo

Table 3.2 Features and properties are described together with the various dataset datatypes (Anon, 2022a)

SL#	Features	Description	Data type
1	age	People's ages in the year	Numerical
2	bp	Blood Pressure (in mm/Hg)	Numerical
3	sg	Specific Gravity	Nominal
4	al	Value of Albumin in Urine	Nominal
5	su	Value of Sugar in Blood	Nominal
6	rbc	Red Blood Cells (Normal or Abnormal)	Nominal
7	pc	Pus Cell (Normal or Abnormal)	Nominal
8	pcc	Pus Cell Clumps (Present or Not Present)	Nominal
9	ba	Bacteria (Present or Not Present)	Nominal
10	bgr	Blood Glucose Random (in mgs/dl)	Numerical
11	bu	Blood Urea (mgs/dl)	Numerical
12	sc	Value of Serum Creatinine in Blood (mgs/dl)	Numerical
13	sod	Value of Sodium (mEq/L)	Numerical
14	pot	Value of Potassium (mEq/L)	Numerical
15	hemo	Value of Hemoglobin (gms)	Numerical
16	pcv	Value of Packed Cell Volume	Numerical
17	wbcc	Count of White Blood Cells (Cells/cumm)	Numerical
18	rbcc	Count of Red Blood Cells (millions/cumm)	Numerical
19	htn	Hypertension (Yes or No)	Nominal
20	dm	Diabetes Mellitus (Yes or No)	Nominal
21	cad	Coronary Artery Disease (Yes or No)	Nominal
22	appet	Appetite (Good or Poor)	Nominal
23	pe	Pedal Edema (Yes or No)	Nominal
24	ane	Anemia (Yes or No)	Nominal
25	Classification	Target Column (CKD or notCKD)	Binary

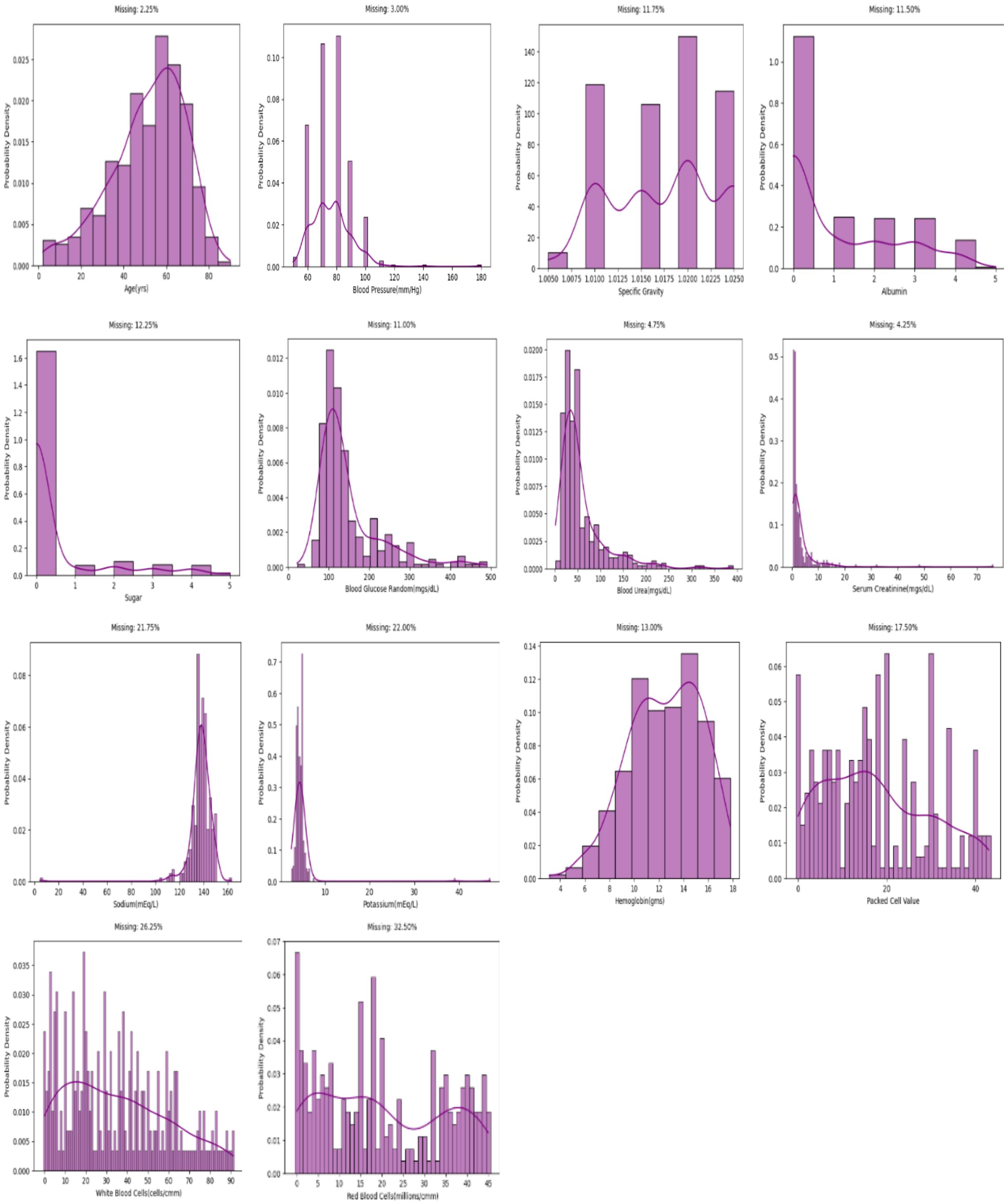


Figure 3.4. Displaying the dataset's missing percentage in the feature’s distribution of its numerical characteristics.

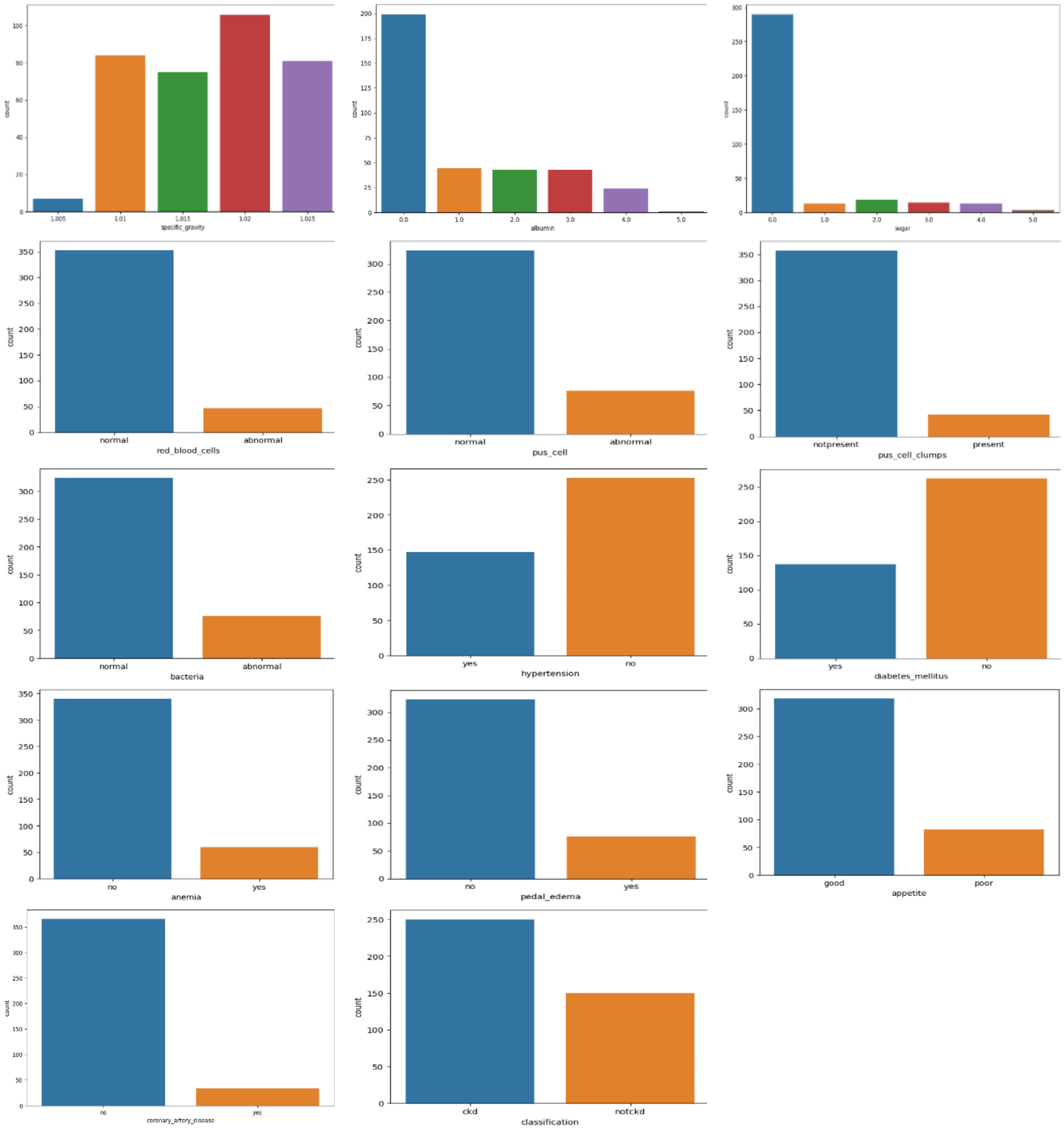


Figure 3.5. Displaying the dataset's nominal feature distribution visually.

The distribution of 14 features and the percentages of missing data are also visualized in Figure 3.4 some of the characteristics in the distribution exhibit quite far-off outliers. Three qualities are regarded as continuous values in this context because they represent measurements of biological components that, in reality, behave as continuous variables. Some features, however, contain a significant percentage of missing values; as a result, measures of central tendency cannot be imputed to them because doing so would skew the distributions. The nominal feature distribution together with the percentages of characteristics contained in the dataset, however, is highlighted in Figure 3.5 instead. The dataset has issues with class imbalance, as seen by the visualization, and such issues have been fixed by using Synthetic Over-Sampling Techniques.

Table 3.3 shows the full description of dataset that we used for LD prediction.

Table 3.3 Description of the dataset.

Feature	Type	Description
Gender	Nominal	This characteristic displays the gender of the user.
Age (years)	Numeric	Participants range in age from 4 to 90 years old.
Total Protein—TP (g/L)	Numeric	The total protein of the individual is recorded in this feature.
Albumin—ALB (g/dL)	Numeric	The albumin of the participant is recorded by this function.
Albumin and Globulin Ratio—AGR	Numeric	The participant's albumin and globulin ratio are recorded in this feature.
Total Bilirubin—TB (mg/dL)	Numeric	The total bilirubin of the subject is recorded by this feature.
Direct Bilirubin—DB (mg/dL)	Numeric	This characteristic records the individual's direct bilirubin.
Aspartate Aminotransferase—SGOT (U/L)	Numeric	The participant's aspartate aminotransferase is captured by this characteristic.
Alkaline Phosphatase—ALP (IU/L)	Numeric	This characteristic records the alkaline phosphatase level of the individual.
Alanine Aminotransferase—SGPT (U/L)	Numeric	This characteristic records the alanine aminotransferase level of the individual.
Liver Disease	Nominal	This characteristic indicates if the individual has received a LD diagnosis or not.

3.3 DATASET PREPARATION

As we find the dataset directly use for ML that’s why we needed to prepare dataset for machine readable. For this reason various data preprocessing technique has been used. We will discuss about this in this section.

3.3.1 Data Preprocessing

To prevent data leaking, the duplicate sample cheeked was first used, however, no duplicate sample was found in the dataset for CKD. Next, using the testing approach, 75% of the data were allocated for training and 25% for testing. The following phase for both partitions involved developing and using data preparation algorithms that were exclusively dependent on the training set. After the dataset was cleaned, ten categorical (nominal) features were found. Additionally, these attributes were converted into numerical using label encoding technique.

In this work, the missing data are handled using the fillna approach, which comprised 400 records and 1008 missing values from 24 characteristics. The NULL values are changed to a specified value using the fillna technique. In the absence of the in-place option being True, a new Data Frame object is created by the fillna function. The original Data Frame's replacing is then handled by the fillna method. First, fillna methods are used to operate on the numerical missing data, which consist of pot, sc, al, bgr, su, bu, bp, sod, and age. The dataset's numerical features and statistical insights, which enabled this study to use a variety of available data sources to eradicate the missing values, are highlighted in Table 3.4.

Table 3.4 Statistical study of the dataset's eleven numerical features.

Characteristics	Mean	Std.	Min	Max	25%	50%	75%
age	51.4833	16.9749	2	90	42	54	64
bp	76.4691	13.7560	50	180	70	78	80
bgr	148.0370	76.5830	22	490	101	126	150
bu	57.4260	49.9870	1.5	391	27	44	61.750
hemo	12.5260	2.8150	3.1	17.8	10.87	12.53	14.625
pcv	38.8840	8.7620	9	54	34	38.77	44
pot	4.6270	2.9200	2.5	47	4	4.63	4.8
sc	3.0720	4.5120	0.4	76	0.9	1.4	3.070
sod	137.5260	9.9080	4.5	163	135	137.53	141
rc	4.7070	0.8900	8	2.1	4.69	5.1	8
wc	8406.1200	2823.3000	2200	26,400	6975	8377.63	9400

For Liver Disease, Firstly we checked for the presence of any duplicate rows. After inspection, we identified 13 duplicate rows and removed them. Next, we examined for any NULL cells and found 4. The missing data were handled using the fillna approach. Label Encoder technique was applied to convert the gender string type to numerical values. Their uneven distribution throughout the dataset may have an effect on the precise identification of LD and Non-LD cases. That's why an oversampling technique called SMOTE [45] is used to produce synthetic data on the minority class. The methodology of the proposed LD prediction technique is illustrated in Figure 3.1. The topics in the two classes are evenly distributed because the Non-LD class's occurrences are oversampled. There are 812 participants after SMOTE is implemented, with 554 (68.23%) of them being men and 258 (31.77%) of them being women. There are 406 LD and 406 Non-LD instances in the target class, and the dataset is now balanced.

3.3.2 Feature Selection

As stated by Qin et al. (2020) [46], figuring out which risk factors are most important for healthcare informatics can reduce the amount of time spent training ML algorithms, enhance data consistency, get rid of unnecessary features, and boost prediction accuracy. Because of this, using different methods to select qualities or remove less significant traits have become more popular over time. A range of approaches, such as the Chi-squared (Chi2) Test, Mutual Information (MI), Principal Component Analysis (PCA), and Recursive Feature Elimination (RFE) processes, have been employed by researchers in recent years to choose pertinent features. This study proposes a fused feature approach that is based on the Wrapper feature selection technique. The Chi^2 test and Mutual Information (MI) techniques' operating principles are demonstrated in Eqs (1) and (2). The degrees of freedom in the Chi-square test are denoted by d, the observed value by K, and the predicted value by F. Conversely, the joint probability density function of A and B is denoted by $p(x, y)$, while the marginal density functions are represented by $p(x)$ and $p(y)$. How comparable the joint distribution $p(x, y)$; is determined by the Mutual Information to the products of the factored marginal distributions.

$$\chi_d^2 = \sum \frac{(K_i - F_i)^2}{F_i} \qquad (1)$$

$$E(A; B) = \iint_{A,B} P(x, y) log \frac{P(x,y)}{P(x)P(y)} dxdy \qquad (2)$$

By removing unnecessary features, the method produces a subset λ_1 containing significant features, where λ_1 ε Features (F). Here is the mathematical justification:

$$\lambda_1 = \{\varepsilon_{\text{w}}\} - \{\varepsilon_{\text{correlation}} \qquad (3)$$

$$\varepsilon_{\text{w}} = \{ \text{z}|\text{z} \in \text{MI} \cap \text{Chi}^2 \} \qquad (4)$$

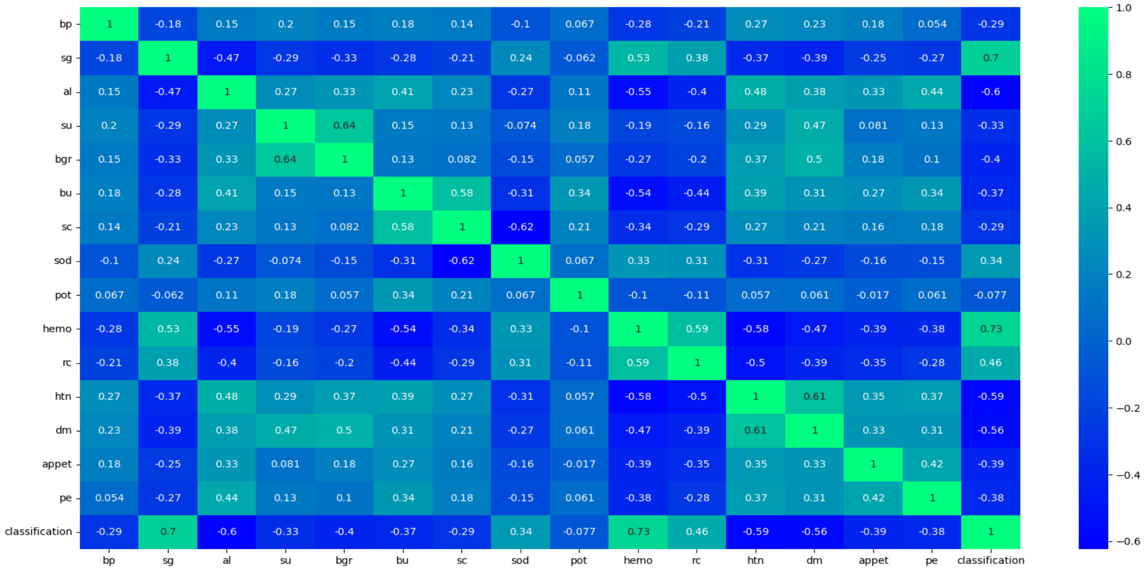
$$\varepsilon_{\text{correlation}} = \{\text{z}|\text{z } \varepsilon \text{ High correlation in features}\} \qquad (5)$$

$$\text{MI} = \{i|i \in \text{ significant elements discovered in the MI}\} \quad (6)$$

$$\text{Chi}^2 = \{i|i \in \text{ significant elements discovered in the Chi}^2 \text{ Test}\} \quad (7)$$

The traits were arranged according to priority in the two sets of Chi2 and MI. Reducing redundancy also requires identifying associated features. Additional Pearson correlation analysis is conducted, and characteristics that are 85% or higher connected with others are eliminated. Lastly, using Select Percentaile, the top 70% of important features were chosen and kept in $\varepsilon_{correlation}$. The characteristics for this investigation were $\varepsilon_{correlation}$ = {bp, sg, al, bgr, su, bu, sc, pot, sod, hemo, rc, htn, appet, dm, and pe}. The correlation matrix of selected features that was discovered through the Pearson correlation matrix check is shown in Figure 3.6. Eq. (8) illustrates the mathematical formula for the Pearson correlation matrix.

$$r = \frac{\sum(a_i - a)(b_i - \bar{b})}{\sqrt{\sum(a_i - \bar{a})^2 \sum(b_i - b)^2}} \quad (8)$$



	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo	rc	htn	dm	appet	pe	classification
bp	1	-0.18	0.15	0.2	0.15	0.18	0.14	-0.1	0.067	-0.28	-0.21	0.27	0.23	0.18	0.054	-0.29
sg	-0.18	1	-0.47	-0.29	-0.33	-0.28	-0.21	0.24	-0.062	0.53	0.38	-0.37	-0.39	-0.25	-0.27	0.7
al	0.15	-0.47	1	0.27	0.33	0.41	0.23	-0.27	0.11	-0.55	-0.4	0.48	0.38	0.33	0.44	-0.6
su	0.2	-0.29	0.27	1	0.64	0.15	0.13	-0.074	0.18	-0.19	-0.16	0.29	0.47	0.081	0.13	-0.33
bgr	0.15	-0.33	0.33	0.64	1	0.13	0.082	-0.15	0.057	-0.27	-0.2	0.37	0.5	0.18	0.1	-0.4
bu	0.18	-0.28	0.41	0.15	0.13	1	0.58	-0.31	0.34	-0.54	-0.44	0.39	0.31	0.27	0.34	-0.37
sc	0.14	-0.21	0.23	0.13	0.082	0.58	1	-0.62	0.21	-0.34	-0.29	0.27	0.21	0.16	0.18	-0.29
sod	-0.1	0.24	-0.27	-0.074	-0.15	-0.31	-0.62	1	0.067	0.33	0.31	-0.31	-0.27	-0.16	-0.15	0.34
pot	0.067	-0.062	0.11	0.18	0.057	0.34	0.21	0.067	1	-0.1	-0.11	0.057	0.061	-0.017	0.061	-0.077
hemo	-0.28	0.53	-0.55	-0.19	-0.27	-0.54	-0.34	0.33	-0.1	1	0.59	-0.58	-0.47	-0.39	-0.38	0.73
rc	-0.21	0.38	-0.4	-0.16	-0.2	-0.44	-0.29	0.31	-0.11	0.59	1	-0.5	-0.39	-0.35	-0.28	0.46
htn	0.27	-0.37	0.48	0.29	0.37	0.39	0.27	-0.31	0.057	-0.58	-0.5	1	0.61	0.35	0.37	-0.59
dm	0.23	-0.39	0.38	0.47	0.5	0.31	0.21	-0.27	0.061	-0.47	-0.39	0.61	1	0.33	0.31	-0.56
appet	0.18	-0.25	0.33	0.081	0.18	0.27	0.16	-0.16	-0.017	-0.39	-0.35	0.35	0.33	1	0.42	-0.39
pe	0.054	-0.27	0.44	0.13	0.1	0.34	0.18	-0.15	0.061	-0.38	-0.28	0.37	0.31	0.42	1	-0.38
classification	-0.29	0.7	-0.6	-0.33	-0.4	-0.37	-0.29	0.34	-0.077	0.73	0.46	-0.59	-0.56	-0.39	-0.38	1

Figure 3.6. Pearson correlation matrix of selected features after feature selection for CKD.

Here is the description of feature selection of LD.

The results of the Pearson correlation study are first shown in Figure 3.7 of selected features. This coefficient reflects the strength and direction of the association between two features, or a feature and the class of interest. Its values range from -1 to 1. When we concentrate on this coefficient, we find that TB and DB have a high correlation of 0.98. For the reason we can drop one features from this two. For this study we dropped DB. After that selected features are stored into df. The selected features are

$df = \{Age, Gender, TB, ALP, SGPT, SGOT, TP, ALB, AGR, Dataset\}$.

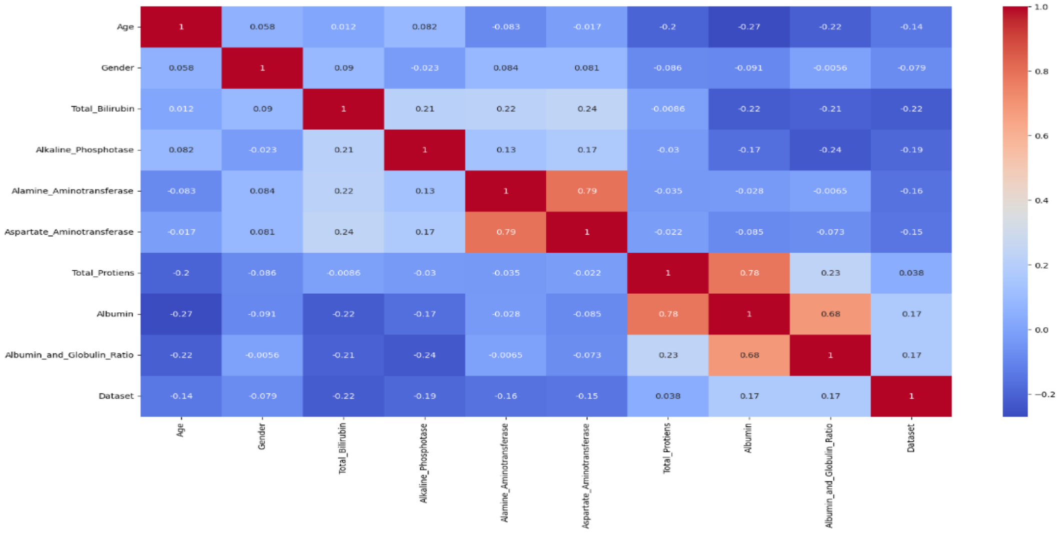


Figure 3.7. Pearson correlation matrix of selected features after feature selection for LD.

3.4 MACHINE LEARNING MODELS

The suggested model is specifically made to categorize CKD using ML techniques. To produce new and understandable patterns, classification templates were made using data mining techniques [47]. The creation of models based on historical analysis is required for both supervised and unsupervised learning approaches used in clinical and medical diagnostics for regression and classification. The classification methods discussed in this section are employed in this research.

3.4.1 Decision tree classifier

The discipline of ML is where the decision tree was invented. The nodes of a DT represent the input attributes. Based on the height data acquired for that specific feature, the output is forecasted. By combining the properties that weren't used in the previous improvements, a subtree is formed. In DT, leaf nodes indicate classes, branching represent outcomes, and internal nodes represent input factors [48]. Within its numerous layers of nodes, the root node is at the top and the leaves are at the bottom. Eq. (9) calculates each class's entropy. In contrast, Eq. (10) uses two features to estimate the entropy. Eq. (11) is utilized to compute the feature-level gain G. The total proportionate entropy is obtained by applying Eq. (12) to each branch's entropy following the serial separation of the data features. Here, G/I determine the height gain ratio for the feature level.

$$F(M) = \sum_{i=1}^{c} -T_i log_2 T_i \qquad (9)$$

$$F(M,H) = \sum_{c \in H} X(c).Y(c) \qquad (10)$$

$$G(M,H) = F(M) - F(M,H) \qquad (11)$$

$$I(M,H) = \sum_{j=1}^{v} \frac{H_i}{H} log_2 \frac{H_i}{H} \qquad (12)$$

3.4.2 Random forest classifier

This technique generates a huge number of interconnected decision trees. The cornerstones of this approach are the decision trees. The phrase "Random Forest" explains a collection of decision trees, the nodes of which are chosen during the preprocessing phase. Following the construction of numerous trees, the best feature is chosen at random from among the features. Making a decision tree is another result of the decision tree algorithm. These trees make up a random forest, which categories new objects based on a vector of inputs. Data is categorized using each decision tree that is produced. If votes are assigned to a class, the random forest will select the classification that has the most votes among all of the trees in the forest.

The RF algorithm can be mathematically represented as follows:

$$T_{ij} = m_j C_j - m_{left(j)} C_{left(j)} - m_{right(j)} C_{right(j)} \tag{13}$$

$Ti\ sub\ (j) = the\ significance\ of\ node\ j$

$m\ sub\ (j) = samples\ arriving\ at\ node\ j$

$C\ sub\ (j) = the\ node\ j's\ impurity\ value$

$left\ (j) = split\ on\ node\ j\ by\ the\ child\ node\ from\ the\ left$

$right\ (j) = split\ on\ node\ j\ by\ the\ child\ node\ from\ the\ right$

3.4.3 Naïve Bayes classifier

The Naive Bayes (NBs) classifier is based on conditional probability and is built using the Bayes theorem [49]. NBs surpasses the complex classification strategy that assumes the existence of a certain class without the presence of the same traits. The NBs classifier is that each characteristic contributes equally and independently to the result. This idea can be viewed as follows with regard to our dataset: We presume that no pair of features is dependent.

The NBs classifier uses the number of CKD and non-CKD samples in each distinct measurement interval to calculate the conditional probabilities of the sample falling into the interval. Due to its excellent performance on huge datasets and ability to produce insightful results, this classifier is used in medical statistical data analysis as a supervised ML technique. This classifier is mapped using Eq. (14) with $P(G|H)$ as the probability input hypothesis for a particular set of data based on $P(H|G), P(G),$ and $P(H)$ probabilities.

$$P(G|H) = \frac{P(H|G) \times P(G)}{P(H)} \tag{14}$$

$P(H|G)$ concentrates on the conditional probability distribution for the response variable L for each input instance $H = H1, H2, \ldots, H_n$. $P(G)$ stands for the response variable's marginal probability, while $P(H)$ stands for an input instance's marginal probability. The class label prediction made by the NBs classifier is given a function in Eq. (15).

$$G = argmax_L P(G) \prod_{k=1}^{n} P(H_k|G) \tag{15}$$

3.4.4 AdaBoost classifier

The AdaBoost classifier or algorithm is a ML method based on the idea of boosting. Turning weak learners into strong learners is the main goal of the boosting technique [50]. The objective of weak learning was to find a weak classifier capable of differentiating between positive and negative data. In order to ensure that the fewest examples are incorrectly identified, weak learning identified each feature's proper threshold value [51]. The AdaBoost algorithm uses two weights: The sample weight is one, and the weight of each learner who is weak is the other.

The Eq. (18) is used to calculate the weighting vector $w_{n.}$ Calculating the value of w_n requires the use of Eq. (16) and Eq. (17). $B_i(x_i)$ in Eq. (16) stands for the weak classifier, and the starting value of the observation weights is w_n, with $j = 1, 2, \ldots, N$. The weak classifier's value as well as Eq. (17) are used to determine $f_j(x)$ in Eq. (19). The final classifier output, a weighted linear combination of the classifiers generated at each step of the procedure serves as the representation of f, which is displayed in Eq. (20).

$$err_m = \frac{\sum_{j=1}^{N} w_j K(y_j \neq B_j(x_j))}{\sum_{j=1}^{N} w_j} \tag{16}$$

$$a_m = \log\left(\frac{1-err_t}{err_t}\right) \tag{17}$$

$$w_j \leftarrow w_j . \exp\left(a_t . K(y_j \neq B_j(x_j))\right) \tag{18}$$

$$f_j(x) = \sum_{m=j}^{M} a_m B_m(x) \tag{19}$$

$$f = sign(f_m(x)) \tag{20}$$

3.4.5 Cat-Boost classifier

Another ML method that forecasts category features is the Cat-Boost classifier. Binary decision trees are the primary predictors used by the gradient boosting method known as Cat-Boost [52]. Let's assume that the data consist of samples $D = (Xj, yj)\ j = 1, \ldots, m$, where $Xj = (x1\ j,\ x2\ j, \ldots xn\ j)$ is a vector with response feature and n attributes. $y_j \in \mathbb{R}$ is either a numerical property (0 or 1) or binary (yes or no). The samples (Xj, yj) are dispersed at random in accordance with an unknown distribution p (.,.). The goal of the learning job is to create a function $M: \mathbb{R}^n \to \mathbb{R}$ that lowers the estimated loss stated in Eq. (21).

$$\mathcal{L}(M) := \Psi L(y, M(X)) \tag{21}$$

The testing data selected from the training data D is represented by (X, y), and the smooth loss function is shown by $L(.,.)$.

3.4.6 Light-GBM classifier

For tasks involving classification, a ML method called LGBM Classifier is employed. It is a component of the open-source and free Light-GBM library, a distributed gradient-boosting ML system. An engine for scalable tree boosting called XGBoost was introduced by Chen et al. [53]. Despite XGBoost's superior accuracy, Daoud (2019) [54] found that the LightGBM ensemble exhibited more resilience, computational efficiency, and time efficiency. Consider a dataset having features x and a label y, denoted by $X = (xi, yi)$. Eq. (22), utilizing Γ as the loss function and K_0 as the goal for initial fit optimization.

$$\widehat{K} = arg_K \, minE_{x,y}[\Gamma(y, K(x)] \tag{22}$$

Eq. (23) provides the gradient or pseudo-relativistic a_m for the m^{th} iteration, this is fitted by the decision tree $hm(x)$.

$$b_m = -\frac{\partial \Gamma(\mathrm{y}_i,\ \widehat{K})}{\partial \mathrm{K}} \tag{23}$$

In order to minimize the loss function, GBDT uses an iterative criterion found in Eq. (24), where λ_m is a multiplier that is optimized using a linear search technique and serves as a step size. One can derive λ_m by using Eq. (25).

$$\mathrm{F}_m(\mathrm{x}) = \mathrm{F}_{m-1}(\mathrm{x}) + \lambda_m \mathrm{h}_m(\mathrm{x}) \tag{24}$$

$$\lambda_m = \mathrm{arg}_\lambda \mathrm{min} \sum_{i=1}^{N} \Gamma(\mathrm{y}_i, \mathrm{F}_{m-1}(\mathrm{x}_i) + \lambda \mathrm{h}_m(\mathrm{x}_i)) \tag{25}$$

3.4.7 Support Vector Machine

One of the main algorithms employed by data scientists is called SVM. It can be applied to regression and classification problems, but it is most commonly employed to classification problems. Its widespread use is due to the model's high accuracy and quick computation (depending on the volume of data). SVM are supervised learning models in ML that examine data and look for trends [55]. The basic SVM is a binary linear classifier that is non-probabilistic that predicts which of two possible classes each input will result in for the output.

The ideal boundary for dividing subjects into two classes is discovered by the SVM [56]. An instance with an unknown class can be best classified using any of the following kernel functions: linear, polynomial, radial basis, or quadratic. SVM used by the kernel technique, which converts inputs into high-dimensional feature spaces implicitly, allows for efficient non-linear classification. The chronic disease datasets are classified using all kinds of kernel functions, and excellent accuracy is obtained when radial basis function (RBF) and kernel function are combined. After using the kernel function, it was discovered that the RBF and kernel functions complement the RKM algorithm nicely to yield good accuracy. The mathematical model of the SVM classifier is as follows:

$$D = \{(x^1, y^1), \dots, (x^1, y^1)\}, x \in \mathbb{R}, y \in \{-1,1\} \tag{26}$$

The hyperplane was considered to have separated the vectors ideally if it had successfully and error-free separated the set of vectors and the distance between the nearest vectors to the hyperplane was maximal. Examining a canonical hyperplane, in which the parameters w and b are restricted by Eq. (27), because Eq. (26) contains some duplication.

$$(w, x) + b = 0 \tag{27}$$

$$min_i|\,(w, x^i) + b = 1 \tag{28}$$

A point x and the hyperplane (w, b) are separated by a distance $d\,(w, b;\ x)$ that is calculated using Eq. (29)

$$d(w, b;\ x) = \frac{|(w, x^i\,) + b|}{||w||} \tag{29}$$

The optimal hyperplane is obtained by maximizing the margin, in accordance with the constraints of Eq. (28). The margin is determined in this way:

$$\begin{aligned} p(s, b) &= min_{x^i:y^i=-1}\left(s, b; x^i\right) + min_{x^i:y^i=1}\left(s, b; x^i\right) \\ &= min_{x^i:y^i=-1} \frac{|\,(s,x^i)+b|}{||s||} + min_{x^i:y^i=1} \frac{|\,(s,x^i)+b|}{||s||} \\ &= \frac{1}{||s||}\left(min_{x^i:y^i=-1} \frac{|\,(s,x^i)+b|}{||s||} + min_{x^i:y^i=1} \frac{|\,(s,x^i)+b|}{||s||}\right) \\ &= \frac{|2|}{||s||} \end{aligned} \tag{30}$$

Therefore, the hyperplane that minimizes using Eq. (31) is the optimal one for separating the data.

$$\varphi(s) = \frac{1}{2}||s||^2 \tag{31}$$

Assume the following limit to demonstrate how minimizing Eq. (31) is comparable to putting the structural risk minimization (SRM) idea into practice.

$$||s|| < A \tag{32}$$

After that, use Eq. (28) and Eq. (29) to generate a new equation.

$$d(s, b;\ x) \geq \frac{1}{A} \tag{33}$$

3.4.8 K-nearest neighbor classifier

The goal of this technique is to maintain comparable objects close to one another. This approach makes use of a data set's feature vectors and class labels [57]. KNN maintains track of every case and classifies new ones using a similarity score. Text is represented as a spatial vector in KNN. $S = S(L_1, W_1; L_2, W_2; \ldots\ldots\ldots L_n, W_{n2};)$. The training text can be used to make any text similar, and the texts that are most similar to one another are chosen. Finally, K neighbors are used to determine the classes.

The feature vectors of each training text and the incoming text are compared using the formula below:

$$sim\ (Vi, Vj) = \frac{\sum_{k=1}^{N} LikLjk}{\sqrt{\sum_{k=1}^{N} Lik^2}\sqrt{\sum_{k=1}^{N} Ljk^2}} \qquad (34)$$

In this case, the training text's feature vector is Vj, while the incoming text's feature vector is Vi. N determines the feature vector's dimension. L_{ik} and L_{jk} are the k^{th} constituents of vectors Vi and Vj, respectively. The actual formulas for the CNN are as follows:

$$Q(Vi, Cm) = \sum_{j=1}^{k} sim(Vi, Vj)\ \delta(Vi, Cm) \qquad (35)$$

$$\delta(Vi, Cm) = \begin{cases} 1\ if\ Vi\ \in Cm \\ 0\ if\ Vi \in Cm \end{cases} \qquad (36)$$

3.4.9 Logistic Regression

Linear regression and logistic regression share a similar goal of estimating the values of the parameter's coefficients [58]. Instead of using linear regression to change the outcome prediction, the logistic function-a nonlinear function-is employed. Gradient descent is a useful technique for estimating logistic function coefficients. Eq. (37) displays the logical regression equation. In this case, the bias is $d0$, the predicted output is z, and the one input value's coefficient, y, is $d1$.

$$z = \frac{e^{(d0 + d1 \times y)}}{1 + e^{(d0 + d1 \times y)}} \qquad (37)$$

3.4.10 Voting classifier

A voting classifier uses an ensemble technique with many models to forecast the class with the highest probability inside the chosen category. There are primarily two types of classifiers. There's hard voting and soft voting. Hard and soft both voting [59,60] is used in this study. Hard voting is computed when the anticipated outputs class receives the majority of votes. Formula for voting classifier is shown in Eq. (38).

$$\sim\ a = mode(\,M_c\,(\,y_t^1)\,, M_c\,(\,y_t^2)\,, \ldots, M_c(\,y_t^n)\,) \qquad (38)$$

Where $M_c(y_t)$ is the most votes received.

3.5 SUMMARY

We used different ML algorithms to get most accurate result of our research study. Before that we needed to apply feature selection, data preprocessing, and some others data preparation techniques. That helps use to get our desire result.

Chapter 4

DESIGN AND IMPLEMENTATION

4.1 INTRODUCTION

This section provides a description of the CKD and LD prediction web application. This online application was developed using Flask, HTML, and CSS. After using PyCharm to train the models, a pklfile is generated. The ML model is stored in this pkl file on the cloud storage platform Heroku. The flask application is used to produce an API. Figure 4.1 depicts the web application's working flow.

4.2 IMPLEMENTATIONS TOOLS

Flask is developed using Python code, which causes a gap to appear between the webpage and the ML model. Anyone who has the necessary input parameters in the form can predict the ailment. The flask receives the inputs and requests that the ML models estimate the outcome using. Following the CDK and LD prediction, the result is delivered to the flask, which then communicates it again to the internet.

4.3 WORKING PROCEDURES

Here is the work flow of web application. When a user sends information or value of test to wen application then it will send a query to flask server. After that flask server send a request to ML model that perform better than other classifiers predict the result of the query and send it to the web app. And finally a dialogue box open that shows if the patient affected with the disease or not.

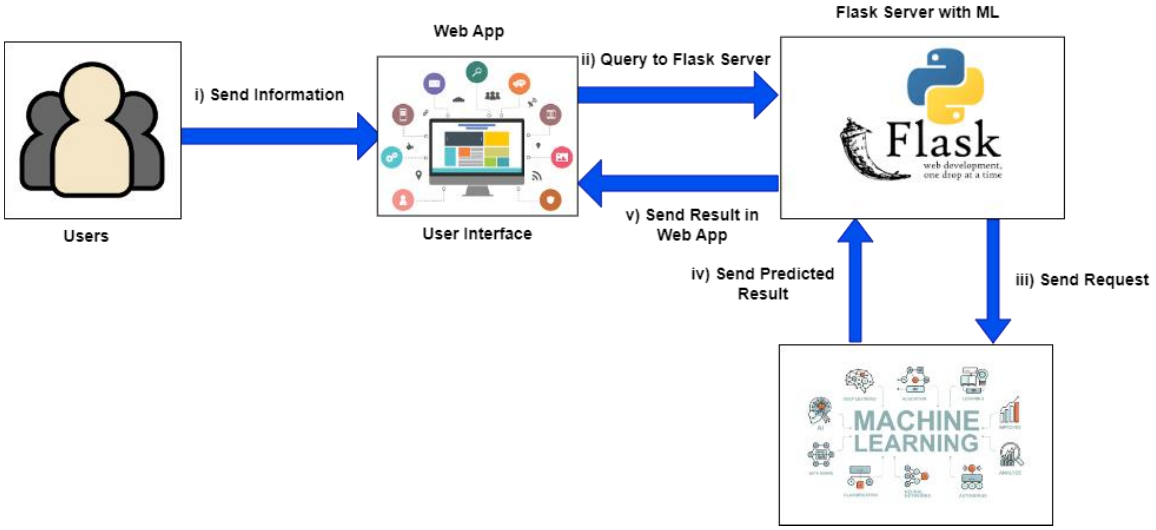
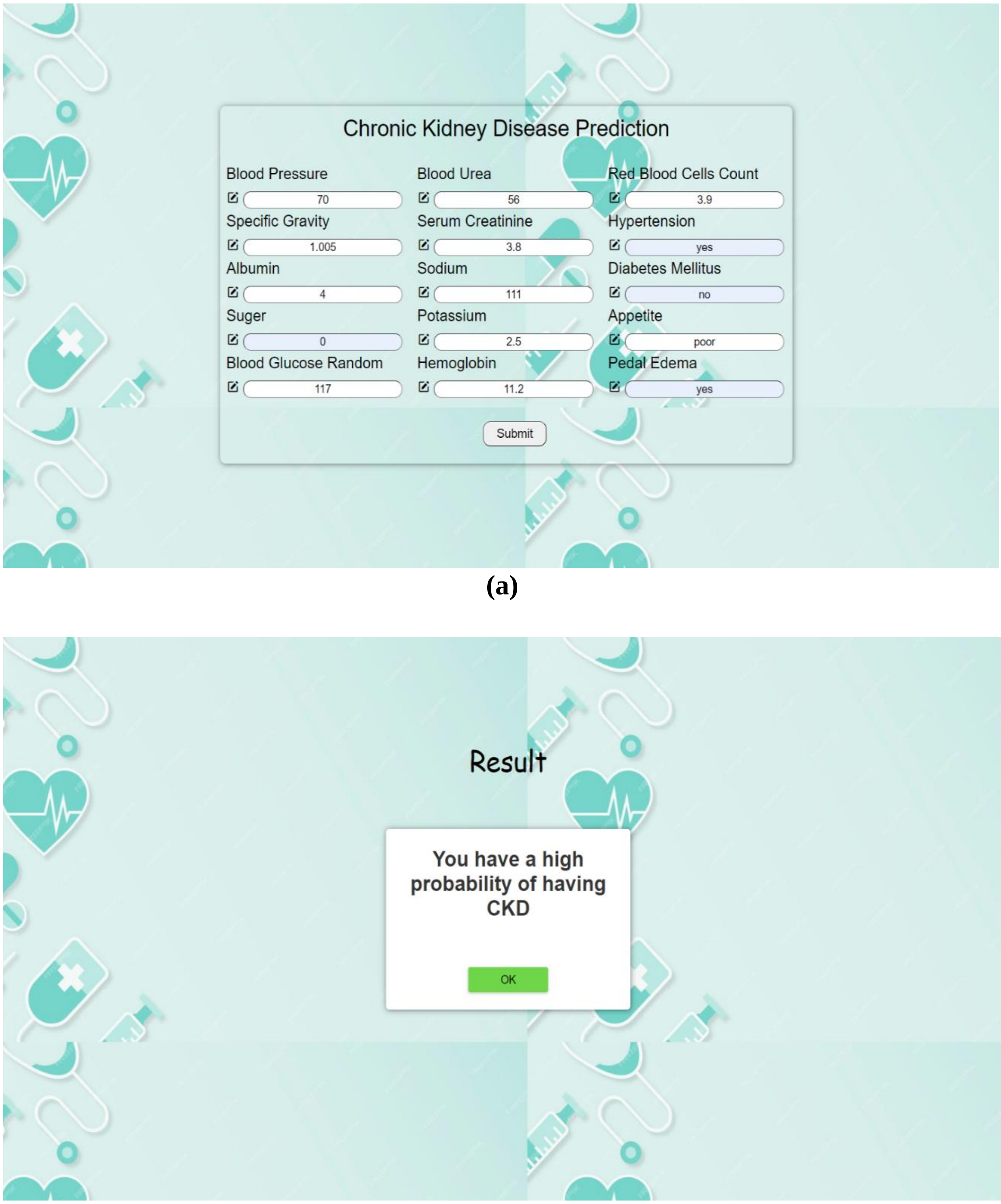

Users
i) Send Information
Web App
User Interface
ii) Query to Flask Server
Flask Server with ML
Flask
web development, one drop at a time
v) Send Result in Web App
iv) Send Predicted Result
iii) Send Request
MACHINE LEARNING

Figure 4.1. Working flow of the web application.

4.4 USER INTERFACE

Here is the sample input field and output filed of user interface for CKD. Taking the value of the test from a user. User must provide correct values along with required format.



Chronic Kidney Disease Prediction

Blood Pressure	Blood Urea	Red Blood Cells Count
70	56	3.9
Specific Gravity	Serum Creatinine	Hypertension
1.005	3.8	yes
Albumin	Sodium	Diabetes Mellitus
4	111	no
Suger	Potassium	Appetite
0	2.5	poor
Blood Glucose Random	Hemoglobin	Pedal Edema
117	11.2	yes

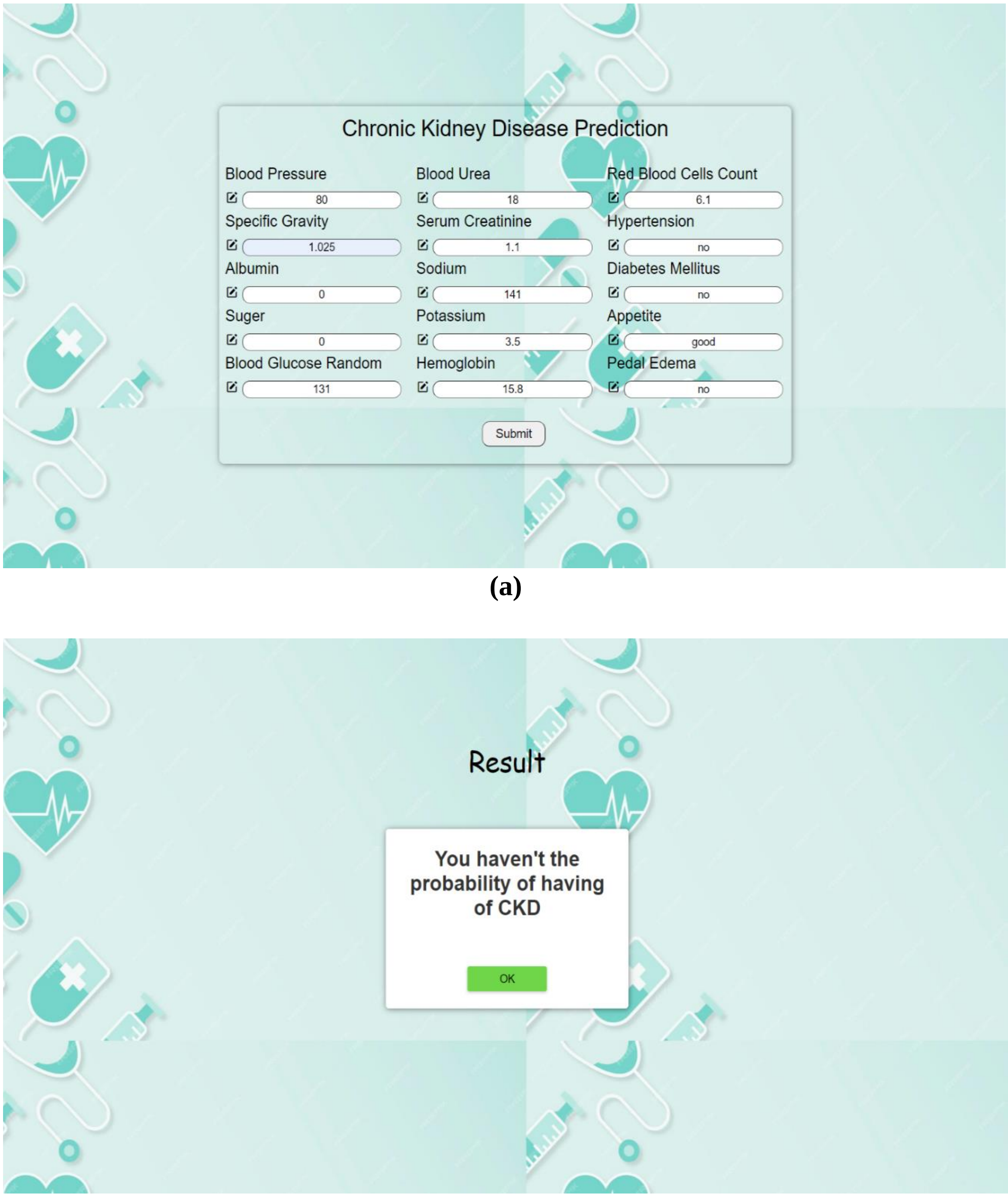
(a)

Result

You have a high probability of having CKD

(b)

Figure 4.2. Prediction outcome of CKD (a) input form and (b) The result (Have CKD)



Chronic Kidney Disease Prediction

Blood Pressure	Blood Urea	Red Blood Cells Count
80	18	6.1
Specific Gravity	Serum Creatinine	Hypertension
1.025	1.1	no
Albumin	Sodium	Diabetes Mellitus
0	141	no
Suger	Potassium	Appetite
0	3.5	good
Blood Glucose Random	Hemoglobin	Pedal Edema
131	15.8	no

(a)

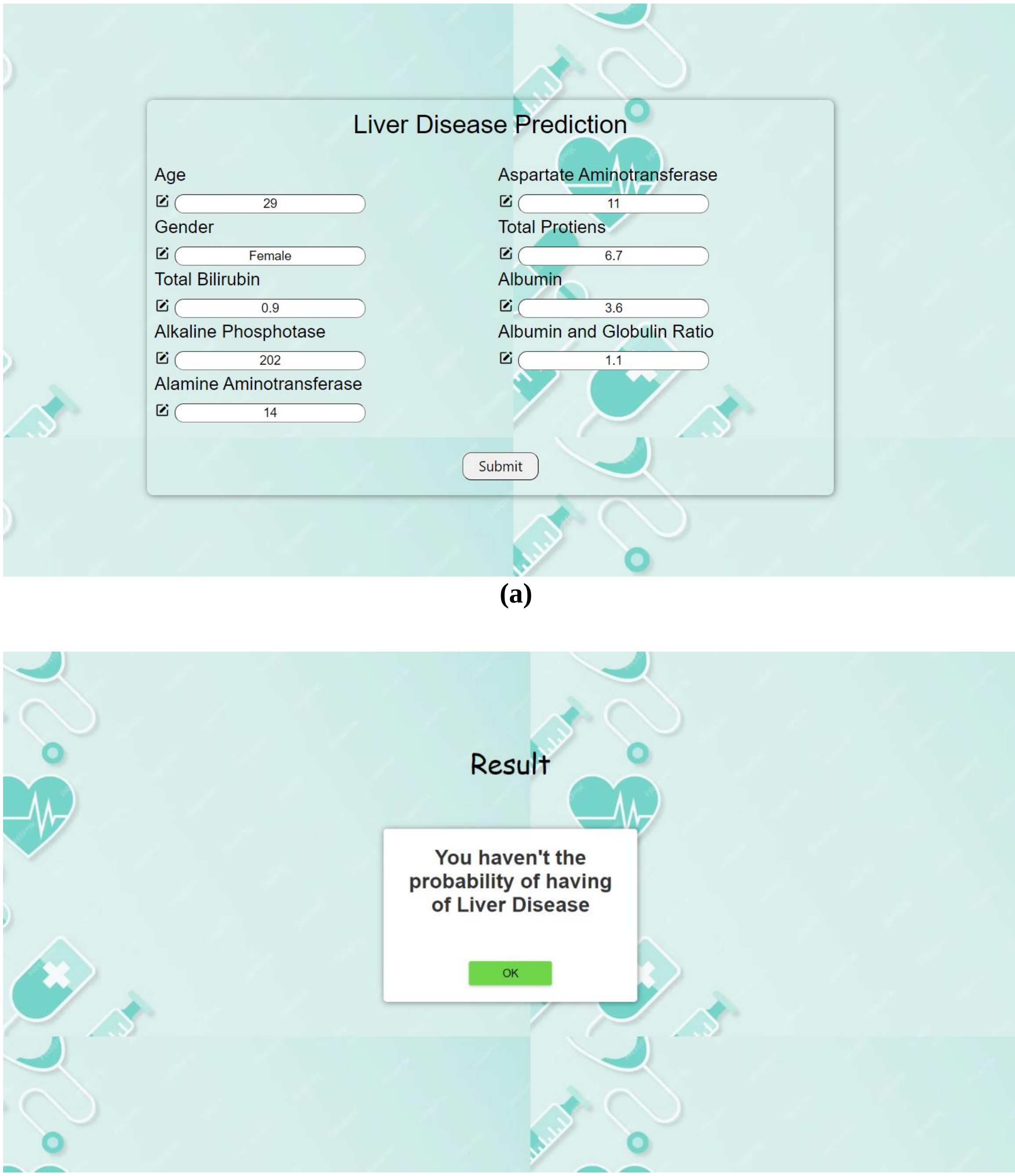
Result

You haven't the probability of having of CKD

(b)

Figure 4.3. Prediction outcome of CKD (a) input form and (b) The result (Haven't CKD)

As a result, a user can display the outcome depending on the inputs. Figure 4.2(a) displays the input form for CKD, and Figure 4.2(b) displays the anticipated outcome. Figure 4.3(a) displays the input form for not CKD, and Figure 4.3(b) displays the anticipated outcome.



Liver Disease Prediction

Age: 29

Gender: Female

Total Bilirubin: 0.9

Alkaline Phosphotase: 202

Alamine Aminotransferase: 14

Aspartate Aminotransferase: 11

Total Protiens: 6.7

Albumin: 3.6

Albumin and Globulin Ratio: 1.1

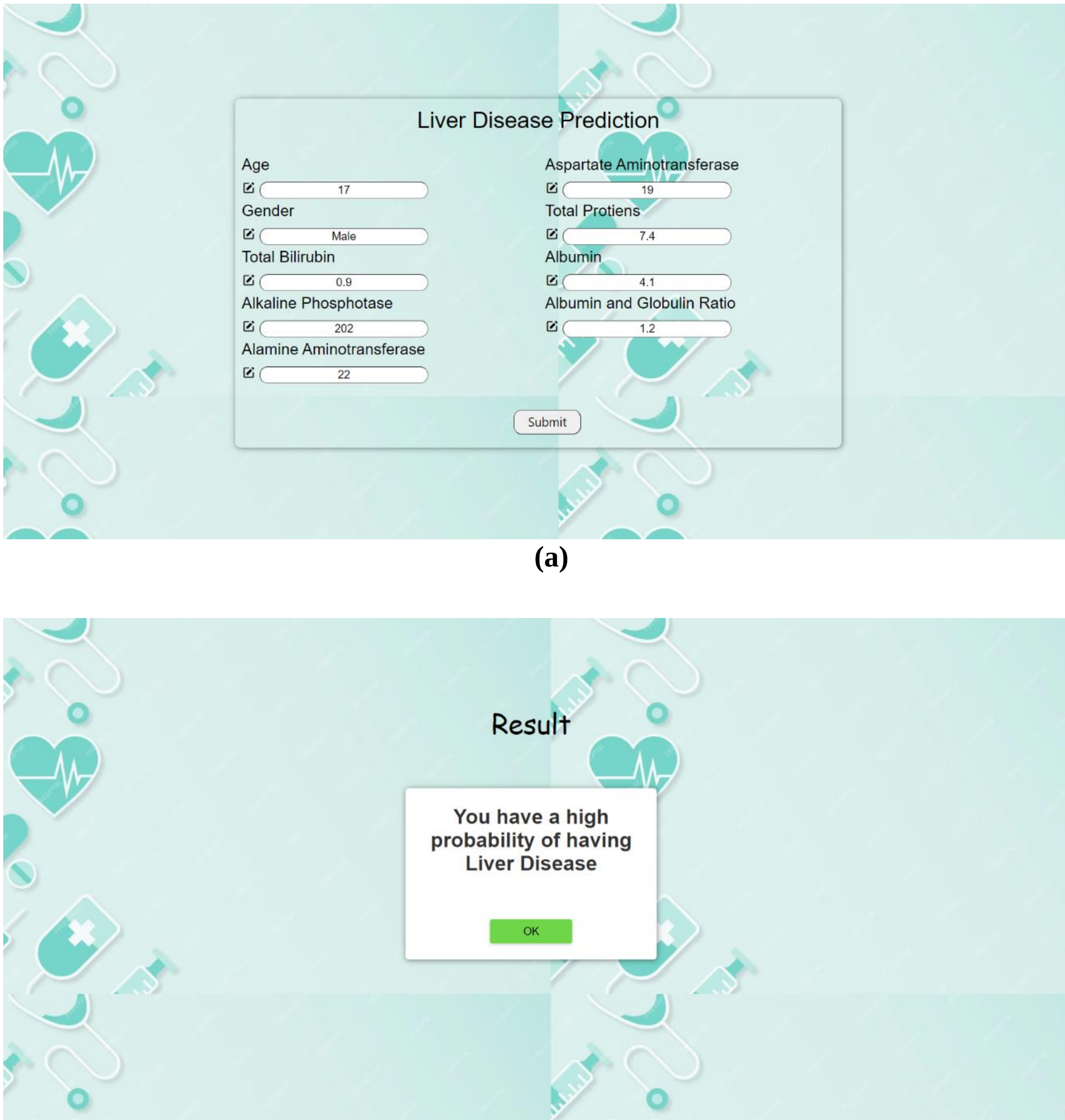
(a)

Result

You haven't the probability of having of Liver Disease

(b)

Figure 4.4. Prediction outcome of LD (a) input form and (b) The result (Haven’t LD)



Liver Disease Prediction

Age	Aspartate Aminotransferase
17	19
Gender	Total Protiens
Male	7.4
Total Bilirubin	Albumin
0.9	4.1
Alkaline Phosphotase	Albumin and Globulin Ratio
202	1.2
Alamine Aminotransferase	
22	

(a)

Result

You have a high probability of having Liver Disease

(b)

Figure 4.5. Prediction outcome of LD (a) input form and (b) The result (Have LD)

Figure 4.4(a) displays the input form for LD, and Figure 4.4(b) displays the anticipated outcome. Figure 4.5(a) displays the input form for having LD, and Figure 4.5(b) displays the anticipated outcome.

4.5 SUMMARY

The main idea for making this web application is to interact with users. That will help for patient to access the services from anywhere through internet.

Chapter 5 RESULT ANALYSIS

5.1 INTRODUCTION

This section gives a thorough summary of the relevant experiment results. There are training and testing sets inside the dataset after it has been randomized. Before being applied to the testing and training sets, a number of data preparation techniques were first fitted exclusively to the training set to avoid data leaking and over-fitting. This research has used a variety of situations to obtain the results. The environment setting needed for the experiment is highlighted in Table 5.1. All of the methods, which included the use of Python notebooks and libraries like Matplotlib, NumPy, Pandas, Seaborn, and Scikit-Learn, were carried out locally utilizing Python. For code we used google colab.

5.2 ENVIRONMENT SET UP

For getting the best result for ML environmental setup is important. The main concern of the setup is CPU. Cause as per concern of performance we must need a fast updated CPU that can calculate fast and train fast enough. For this reason we used updated 11th Gen and core i5 processor. Side factors are also important such a good amount of RAM. Here are the full details of environment setup we used for our project.

Table 5.1 The Specifications of System.

CPU	1 × 11TH Gen Intel®Core™i5-1135G7 @ 2.40GHz
RAM	8GB
GPU	Intel iRIS$_{X}^{e}$
Cache	46MB
GPU Memory	8GB
Disk Space	256GB
Session Limit	10h

5.3 PERFORMANCES MATRIX

The suggested system is tested and assessed using the performance metrics. We applied the most commonly utilized metrics found in the relevant literature, such as F-Measure, AUC, accuracy, precision, and recall [61,62], to evaluate the ML models' performance. Additionally, various assessment criteria were employed to evaluate the performance of the existing ML classifier. The method for determining additional Confusion Matrix evaluation metrics is shown in the following equations, Eq. (39) – (42). The metrics Accuracy (AC), Precision (P), Recall (R), and F1-score are produced by computing the samples that were correctly classified (True positive (TP) and True Negative (TN)) and the samples that were incorrectly classified (False positive (FP) and False Negative (FN)).

Here is a definition of the previously stated measures:

- *Accuracy:* The accuracy of the classification indicates the rate of right predictions. The confusion matrix is the source of computation.

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \times 100\% \qquad (39)$$

- *Precision:* An essential matrix for evaluating the performance of a model is precision. In relation to all the examples that were obtained, it is the percentage of linked instances. Its expected value is positive.

$$Precision = \frac{TP}{TP+FP} \times 100\% \qquad (40)$$

- *Recall:* The model performance evaluation matrix also includes recall as a crucial component. Out of all the instances that were obtained, this is the percentage of connected instances.

$$Recall = \frac{TP}{TN+TP+FN+FP} \times 100\% \qquad (41)$$

- *F-Measure:* F Score or F1-force is another name for it. The F-measure is computed to assess the test's accuracy.

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} * 100\% \qquad (42)$$

The dataset is divided into two parts: 25% is utilized for testing and the remaining 75% is used for training. The AdaBoost classifier outperforms the other 9 ML classifiers, produces the maximum accuracy of 99.19% with 100% precision, 99% recall and 99% F1-score. On other hand the LGBM Ensemble technique performs better than the other models, with accuracy of 83.74%, precision of 78%, recall of 85%, F1-score of 81% and AUC of 90.3%.

Ultimately, in CKD the AdaBoost classifier achieved the maximum accuracy (99.19%), while the SVM classifier obtained the lowest accuracy (88.62%), according to the data. In LD LGBM classifier achieved the maximum accuracy (83.74%), while the SVM classifier obtained the lowest accuracy (67.98%).

5.4 ACCURACY TABLE

Table 5.2 is the accuracy table for CKD. From the table we can see AdaBoost classifier got the accuracy of 99.19%. Which is the most among all other classifers. For precision AdaBoost also got the highest precision of 100% alone with Random Forest. But as Random forest's accuracy didn't surpass the accuracy of AdaBoost so we can ignore it. For recall, although Naïve Bayes got the 100% recall and AdaBoost got 99%. As I mention before AdaBoost got the highest accuracy for that reason we can simply ignore this factor also. As comparing with other classifier we can come to a conclusion that AdaBoost performs better that other classifiers.

Table 5.2 The ML model's performance measures are based on the entire dataset.

ML algorithms	Accuracy	Precision	Recall	F1-score
Support Vector Machines (SVM)	88.62%	84%	97%	90%
Decision Tree Classifier (DT)	95.93%	97%	96%	96%
Random Forest Classifier (RF)	96.75%	100%	94%	97%
Naïve Bayes Classifier (NB)	97.56%	96%	100%	98%
KNN Classifier (KNN)	93.50%	93%	96%	94%
Logistic Regression (LR)	97.56%	97%	99%	98%
Ada Boost Classifier (AB)	99.19%	100%	99%	99%
Cat Boost Classifier (CB)	96.75%	98%	96%	97%
LGBM Classifier (LGBM)	98.37%	98%	96%	97%
Voting Classifier	98.37%	99%	99%	99%

Figure. 5.1. is the bar chat of accuracy classifers for CKD that has beed used for project. It's giving us a visual represenation that how different classifiers perform interm of accuracy. SVM got the lowest accuracy and AdaBoost got the highest. The second highest accuracy comes from LGBM and Voting classifiers which is 98.37%.

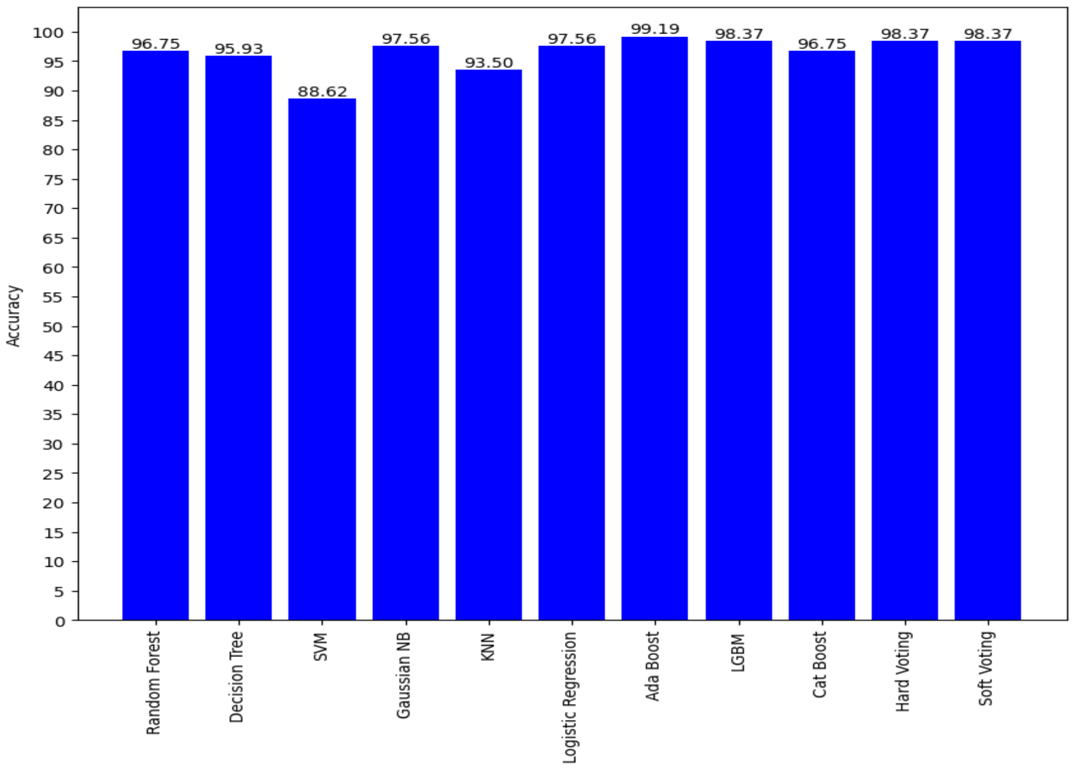


Figure 5.1. Overall percentage of classification accuracy across all methods applied.

CKD is essential for enabling decision-makers (such as ministries, insurers, and hospital managers) to take the necessary precautions against an impending patient overpopulation. ML approaches are being used more and more in the healthcare sector to develop disease prediction models because biomedical laboratory data is now readily available. Out of the 10 distinct ML models used in this work, the AdaBoost classifiers fared better than other models.

Figure. 5.2 is the visual representation of ROC curve of different ML classifers that has been used for the project. That value of ROC is in between 0 to 1. The value of ROC is more closer to 1 means the better one. And the value of ROC is closed to 0 means not better one. From the figure we can clearly see that AdaBoost got the value of ROC is 1. Which means it does better perform. As AdaBoost perform interm of accuracy, precision, recall and now for ROC so we can say that it is more better for precition of CKD.

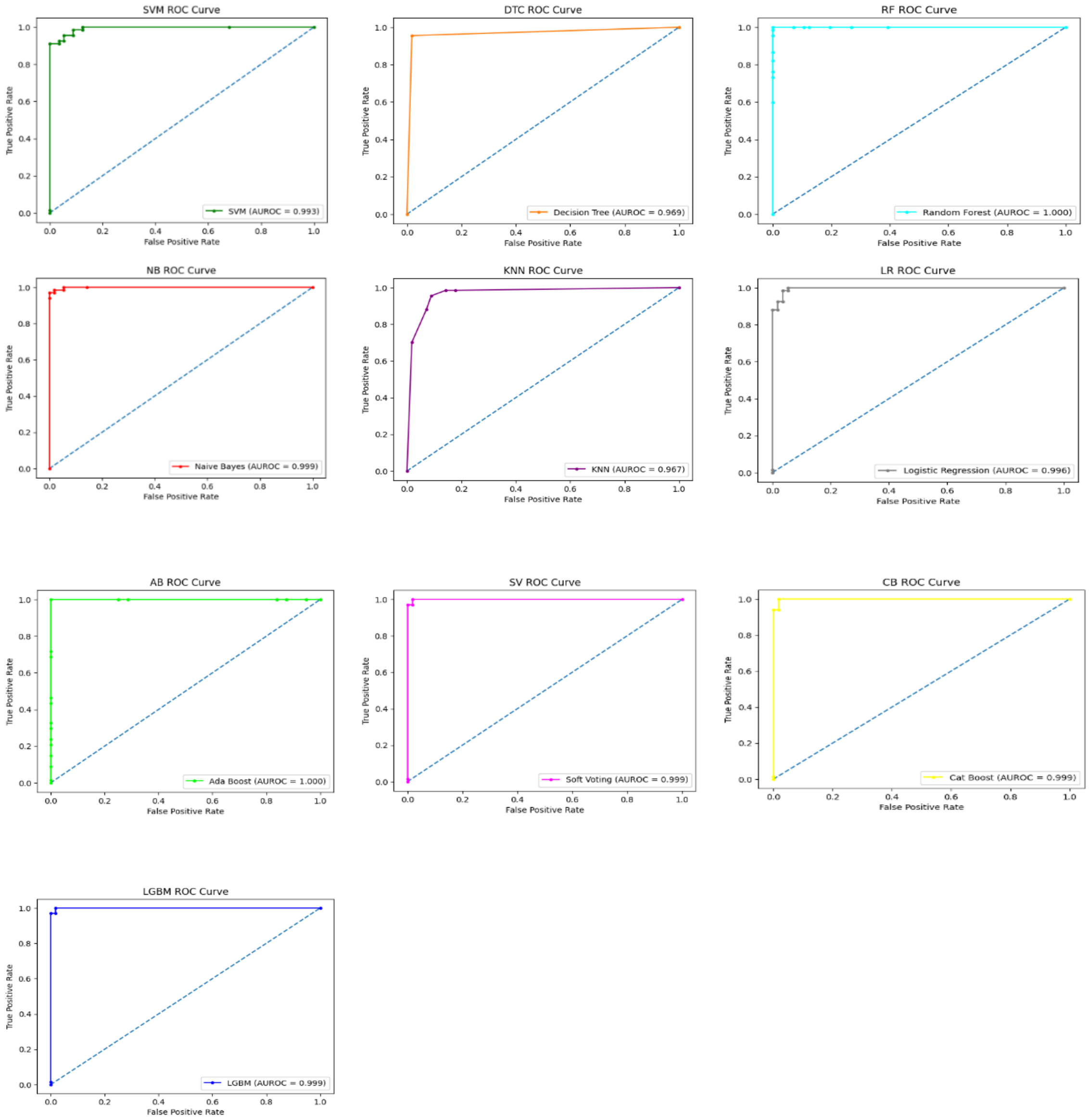
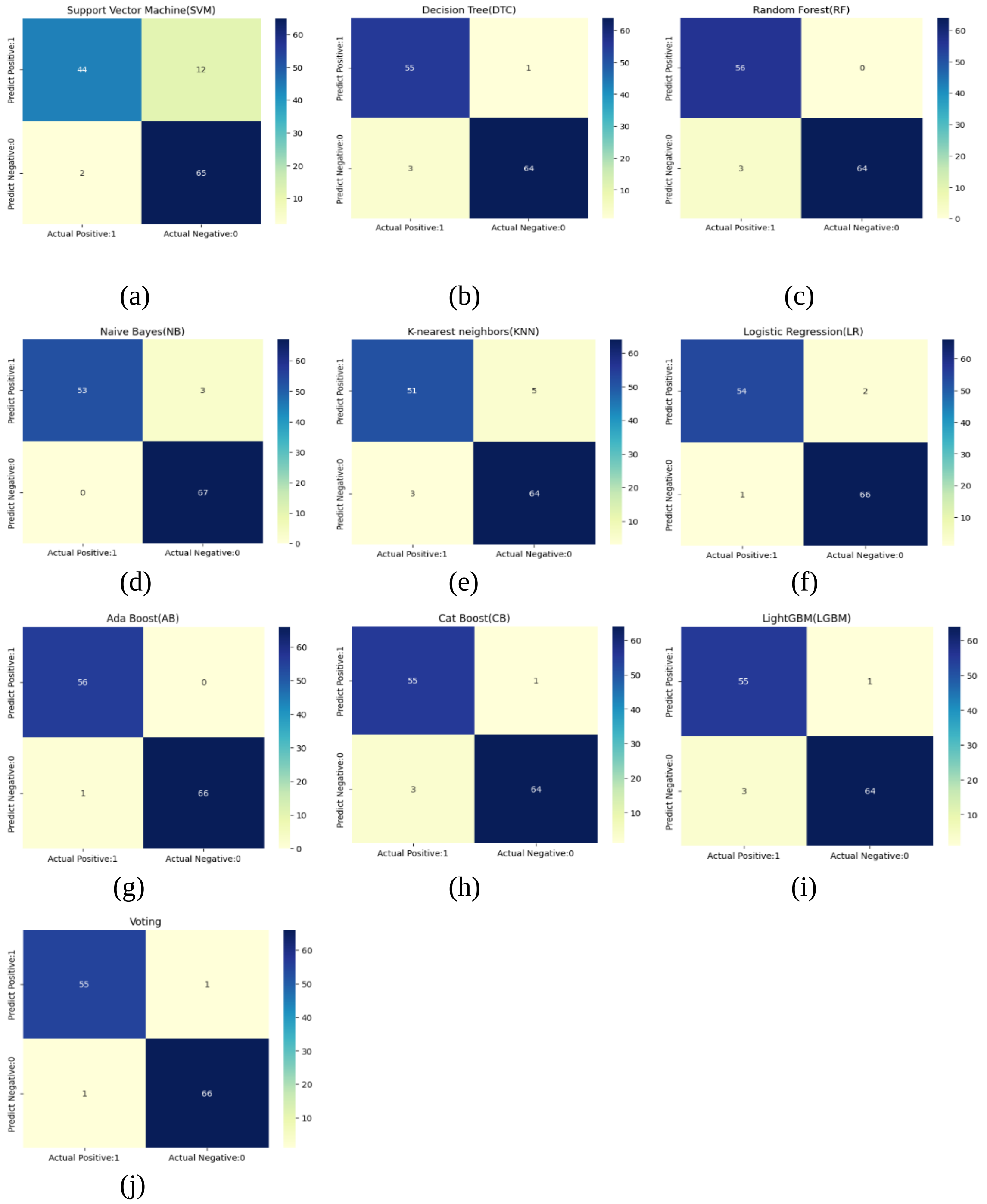


Figure 5.2. ROC Curve of different classifiers for CKD.

Figure. 5.3 is visual representation of confusion matrix. The value of true positive and true negative is closer to each other’s means the better one. Same goes for the false positive and false negative. From the confusion matrix we can clearly see that the differences between the value of the factor is closer for AdaBoost which is our most considerable classifier.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Figure 5.3. Confusion Matrix of (a) SVM (b) DT (c) RF (d) GNB (e) KNN (f) LR (g) AB (h) CB (i) LGBM (j) Voting for CKD.

Table 5.3 Evaluation of ML models' performance following SMOTE.

Classifiers	Accuracy	Precision	Recall	F-Measure	AUC
SVM	67.98%	67%	79%	73%	75.3%
KNN	68.97%	69%	79%	73%	71.6%
LR	68.97%	71%	72%	71%	75.3%
GNB	70.94%	66%	96%	78%	76.8%
DT	72.91%	75%	75%	75%	72.7%
AdaBoost	73.89%	74%	79%	77%	81.6%
RF	77.83%	79%	83%	81%	87.7%
CatBoost	78.82%	78%	85%	81%	88.0%
Voting	83.25%	83%	86%	85%	90.2%
LGBM	83.74%	78%	85%	81%	90.3%

Table 5.3 illustrates the models' average performance for LD. It is worth mentioning that the consistent class distribution contributed to enhancing the accuracy of identifying healthy individuals while maintaining higher average performance levels. This indicates that the models under investigation retained their effectiveness for the precise and accurate identification of the patients as well. The LGBM Ensemble technique performs better than the other models, with accuracy of 83.74%, precision of 78%, recall of 85%, F1-force of 81% and AUC of 90.3%. The voting classifier model also has very strong performance with accuracy of 83.25%, precision of 83%, recall of 86%, F1-force of 85% and AUC of 90.2%.

Figure 5.4 shows the overall accuracy of all classifiers for LD. From the bar chat we can clearly see that the tallest one is LGBM. And the lowest one is 67.98% which is SVM. The second highest bar is for soft voting, which is 83.25%. The evaluation of the ML model using AUC ROC curves after SMOTE is shown in Figure 5.5 in addition to the metrics mentioned previously. As mentioned before the value of ROC curve closer to 1 means more reliable for prediction. And we can see that the value for LGBM is 0.903 which is outperform the classifiers. Lowest ROC value for LD is 0.706, which is got from Decision Tree classifier.

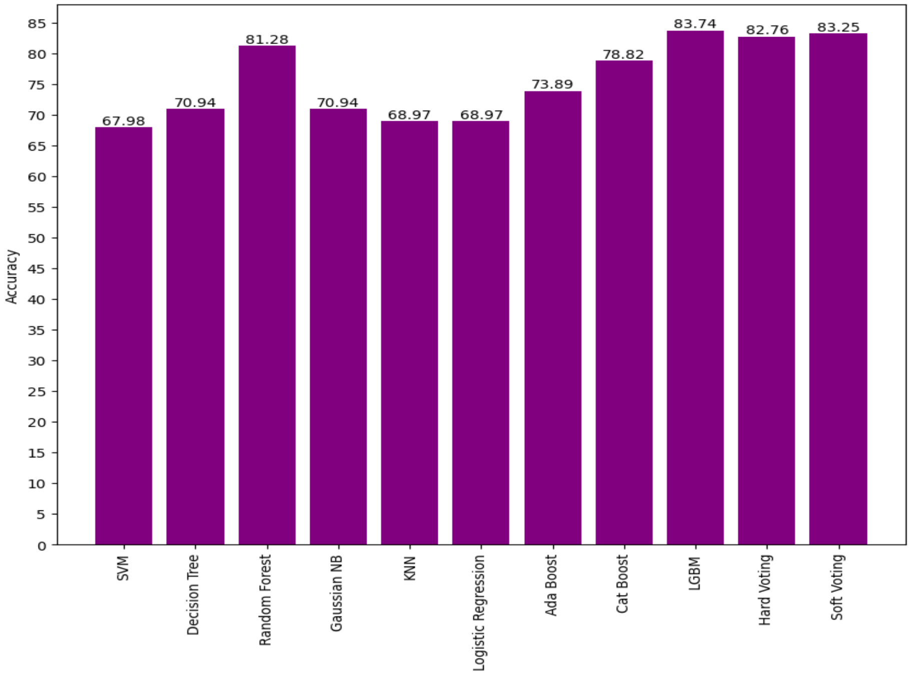


Figure 5.4. Overall classification accuracy percentage for all used algorithms.

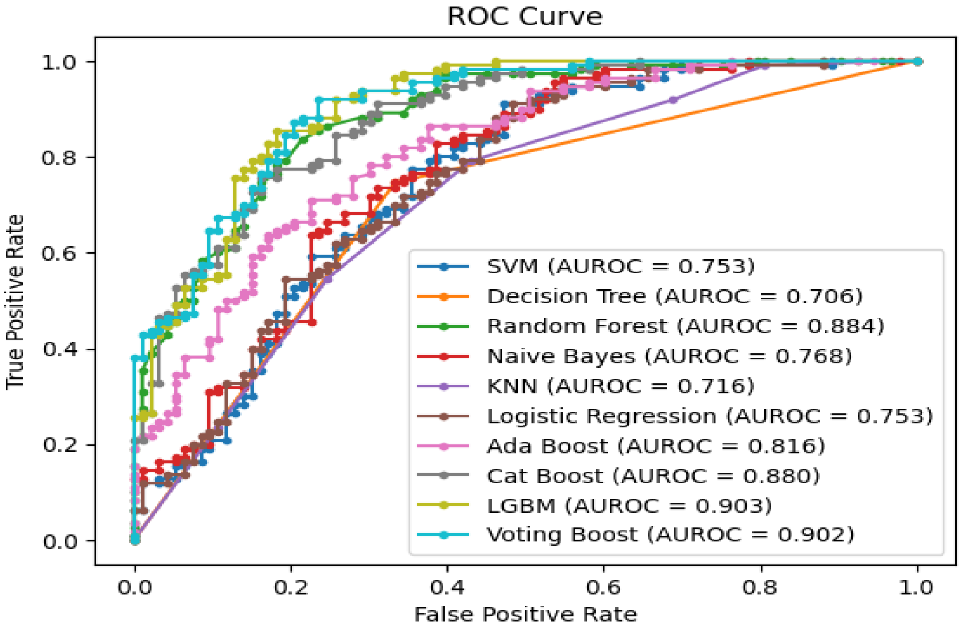


Figure 5.5. ML Model’s evaluation based on AUC ROC curves.

5.5 COMPARISON TABLE

Table 5.4 shows the comparison with recent previous related work. J. Qin et al. applied Random Forest Classifier and got the accuracy of 99.83% but didn't apply any feature selection approach. With the accuracy of 98.99% Silveira et al. using feature selection. By using Logistic Regression Ebiaredoh-Mienye et al. got the accuracy of 98%. A study by Yashfi, S.Y. et al. using feature selection approach and Random Forest classifier got the accuracy of 97.12%. Almansour et al. proposed a model with ANN and SVM and got the accuracy of 99.75% but didn't use any feature selection approach. Kim et al. also didn't use any feature selection approach but got the accuracy of 95.40% using ANN. Sara et al. got the accuracy of 90% and used a feature selection approach with SVM classifier. Rady et al. didn't use any feature selection approach and got the accuracy of 96.70% with SVM classifier. Polat et al. used feature selection approach with SVM and got the accuracy of 98.50%. With an outstanding accuracy of 99.80% P. Ghosh et al. proposed model with a feature selection. As compare to other our proposed model got the highest accuracy than most of the previous work. Those who got more accuracy than use their model either didn't use feature selection approach or got less precision or recall or F1-force. Here is the full comparison table to more clarify.

Table 5.4 Assessment of the proposed model's effectiveness in relation to previous studies.

Reference	Accuracy	Precision	Recall	F1-score	Applied feature selection
J. Qin et al.	99.83%	100%	100%	100%	No
Silveira et al.	98.99%	99%	99%	98%	Yes
Ebiaredoh-Mienye et al.	98%	97%	97%	97%	Yes
Yashfi, S.Y. et al.	97.12%	97%	97%	97%	Yes
Almansour et al.	99.75%	100%	99.60%	99.70%	No
Kim et al.	95.40%	-	-	-	No
Sara et al.	90%	95.24%	90.91%	93.02%	Yes
Rady et al.	96.70%	98.75%	100%	99.37%	No
Polat et al.	98.50%	97.90%	97.80%	-	Yes
P. Ghosh et al.	99.80%	98%	-	99%	Yes
Our proposed model	99.19%	100%	99%	99%	Yes

We present in Table 5.5 the models that have been suggested by previous research works using the same dataset [37]. A study by A. Sokoliuk et al. [38] identified the Gradient Tree Boosting classifier as the most effective, achieving a 72% accuracy on a balanced dataset. Subsequent research conducted by M. Azam et al. [39] explored the performance of KNN with feature selection techniques (KNNWFST) and achieved an accuracy of 74%, surpassing other comparable methods. A. Srivastava et al. [40] and R. Choudhary et al. [41] both advocated for the Logistic Regression (LR) model, which demonstrated a 75% accuracy rate. Additionally, C. Geetha et al. [42] highlighted that the Support Vector Machine (SVM) yielded an accuracy of 75.04%. G. Gajendran et al. [43] proposed a hybrid ML model named Mathematical Approach on Multilayer Feedforward Neural Network with Backpropagation (MAMFFN), achieving an accuracy of 75.30%. Lastly, E. Dritsas et al. [44] introduced a voting classifier, attaining an accuracy of 80.10%. In comparison to the other works, our suggested model LGBM performs better, with an accuracy of 83.74%.

Table 5.5 Proposed models are illustrated using the same dataset as the source publications.

Research work	**Proposed Model**	**Accuracy**
A. Sokoliuk et al. [38]	Gradient Tree Boosting	72%
M. Azam et al. [39]	KNNWFST	74%
A. Srivastava et al. [40]	LR	75%
R. Choudhary et al. [41]	LR	75%
C. Geetha et al. [42]	SVM	75.04%
G. Gajendran et al. [43]	MAMFFN	75.30%
E. Dritsas et al. [44]	Voting	80.10%
Our Proposed Method	LGBM	83.74%

5.6 SUMMARY

By comparing with others work we can see that for CKD it is very much well ahead our work int term of accuracy, precision, recall and F-1 force. For liver disease we can see that the main concern which is the accuracy that is well ahead from previous works. So we can rely on this work more that compared one.

Chapter 6 CONCLUSION AND FUTURE SCOPE

6.1 DISCUSSION

CKD is characterized by a progressive loss of kidney function over time. The quality of life is reduced for CKD patients and diminished kidney function for the duration of their illness. A large financial burden is borne by patients, healthcare providers, and the government as a result of CKD. Since most patients don't exhibit any symptoms, it is an inaudible disease. Reinforcement therapy (RRT) may be expensive or difficult for individuals with end-stage renal disease (ESRD). Patients with such conditions often need hemodialysis, peritoneal dialysis, and occasionally transplantation. The medical community takes the important task of early diagnosis and treatment of CKD very seriously, and ML theory is being utilized to design an effective solution.

6.2 OUTCOME

This research highlights how ML algorithms may identify CKD with the least amount of tests or characteristics. Several ML techniques were employed in this work to detect CKD early on. For feature selection Mutual Information (MI), Pearson correlation matrix, and Chi-squared test (Chi2) was applied. Out of all other ML techniques AdaBoost classifier came out as the best performer. With the exception of SVM, using this model on the provided dataset, 99.19% accuracy, 100% precision, 99% recall, and 99% F1-score were achieved. SVM provided a comparatively low accuracy of 88.62% when compared to the other models. According to the findings of this study, ML-based models have the potential to be a useful tool concerning public health and resource development initiatives including early detection of CKD and close patient monitoring. The future focus of this research project is Development of a web application to help clinicians diagnose patients with renal failure in real time. It also aims to create a dataset with ultrasound imaging data from CKD patients and using DL methods to determine the illness.

LD is a dangerous illness that needs immediate medical attention since it poses a threat to human life. Medical practitioners use pathological procedures to document a patient's status and provide a medical report. This study aimed to predict LD early using ML approaches.

To be more precise, a variety of ML models, including GNB, KNN, DT, RF, SVM, AB, CB, LR, LGBM, and voting, were assessed based on their precision, recall, F-Measure, accuracy, and AUC for predicting the probability of LD. Based on the experimental data, the LGBM classification method is the major proposition of this study since it performs better than the other methods with an accuracy of 83.74%, precision of 78%, recall of 85%, F1-force of 81% and AUC of 90.3% after SMOTE. Finally, our proposed model LGBM demonstrates greater accuracy when compared to similar published research efforts based on the dataset [37].

6.3 LIMITATIONS

As per concern limitation that we haven’t use 10-fold cross validation for our work. Although we did manage to get valid dataset but as those are not from a hospital or clinic that’s why we can’t say those are 100% reliable. But as previously many author work on those dataset and get better result for that reason we also work on those dataset.

6.4 FUTURE WORK

As mentioned in the limited section we didn’t do a cross validation and didn’t work with real life hospital data so we will work with real life hospital data more precisely image data to get more accurate result. And we want to build a professional web app with most advances features and options that will help the user for interact with more easily.

6.5 CONCLUSION

In concluding this research endeavor, we have undertaken a comprehensive exploration into the predictive modeling of CKD and LD using advanced ML techniques. Our methodology, forged through meticulous data preprocessing, innovative feature selection, and the unique incorporation of motif analysis, represents a robust framework aimed at optimizing model accuracy, interpretability, and clinical relevance.

Motif analysis, though less conventional in disease prediction, provides an additional layer of understanding, uncovering recurrent patterns that may hold key insights into the progression and susceptibility of CKD and LDs.

Validation, a cornerstone of our methodology, ensures the reliability and generalizability of our predictive models. Rigorous assessment using diverse datasets and comprehensive performance metrics serves as a robust benchmark, affirming the effectiveness of our approach.

As we look back on the journey from conceptualization to execution, this project stands as a testament to the intersection of innovation and practicality in healthcare. The insights gained from our research contribute not only to the refinement of predictive models for CKD and LD but also underscore the importance of interpretability and feature relevance in the realm of healthcare analytics.

Moving forward, the implications of our work extend beyond the confines of this project. The integration of motif analysis, the optimization of feature selection, and the strategic fusion of diverse ML models pave the way for future advancements in predictive medicine. By fostering a deeper understanding of disease complexities, our research holds promise in shaping the landscape of personalized and effective healthcare interventions for individuals at risk of CKD and liver diseases.

REFERENCES

[1] Kidney Location References: Kidneys: Location, function, anatomy, pictures, and related diseases (medicalnewstoday.com)

[2] Levey, A.S.; Coresh, J. Chronic kidney disease. Lancet 2012, 379, 165–180. [CrossRef]

[3] Koye, D.N.; Magliano, D.J.; Nelson, R.G.; Pavkov, M.E. The global epidemiology of diabetes and kidney disease. Adv. Chronic Kidney Dis. 2018, 25, 121–132. [CrossRef]

[4] Anon, Estimated Glomerular Filtration Rate (eGFR), National Kidney Foundation. (2015). (accessed February 4, 2022). [CrossRef]

[5] Razavi, H. Global epidemiology of viral hepatitis. Gastroenterol. Clin. 2020, 49, 179–189. [CrossRef]

[6] Ginès, P.; Krag, A.; Abraldes, J.G.; Solà, E.; Fabrellas, N.; Kamath, P.S. Liver cirrhosis. Lancet 2021, 398, 1359–1376. [CrossRef]

[7] Ringehan, M.; McKeating, J.A.; Protzer, U. Viral hepatitis and liver cancer. Philos. Trans. R. Soc. B Biol. Sci. 2017, 372, 20160274. [CrossRef]

[8] Powell, E.E.; Wong, V.W.S.; Rinella, M. Non-alcoholic fatty liver disease. Lancet 2021, 397, 2212–2224. [CrossRef]

[9] Smith, A.; Baumgartner, K.; Bositis, C. Cirrhosis: Diagnosis and management. Am. Fam. Physician 2019, 100, 759–770. [CrossRef]

[10] Marchesini, G.; Moscatiello, S.; Di Domizio, S.; Forlani, G. Obesity-associated liver disease. J. Clin. Endocrinol. Metab. 2008, 93, s74–s80. [CrossRef]

[11] Seitz, H.K.; Bataller, R.; Cortez-Pinto, H.; Gao, B.; Gual, A.; Lackner, C.; Mathurin, P.; Mueller, S.; Szabo, G.; Tsukamoto, H. Alcoholic liver disease. Nat. Rev. Dis. Prim. 2018, 4, 1–22. [CrossRef]

[12] Åberg, F.; Färkkilä, M. Drinking and obesity: Alcoholic liver disease/nonalcoholic fatty liver disease interactions. In Seminars in Liver Disease; Thieme Medical Publishers: New York, NY, USA, 2020; Volume 40, pp. 154–162. [CrossRef]

[13] Bae, M.; Park, Y.K.; Lee, J.Y. Food components with antifibrotic activity and implications in prevention of liver disease. J. Nutr. Biochem. 2018, 55, 1–11. [CrossRef]

[14] Cai, J.; Zhang, X.J.; Li, H. Progress and challenges in the prevention and control of nonalcoholic fatty liver disease. Med. Res. Rev. 2019, 39, 328–348. [CrossRef]

[15] Bhaskar, N.; Suchetha, M.; Philip, N.Y. Time Series Classification-Based Correlational Neural Network With Bidirectional LSTM for Automated Detection of Kidney Disease. IEEE Sens. J. 2021, 21, 4811–4818. [CrossRef]

[16] Bikbov, B., Perico, N., Remuzzi, G., & O. behalf of the GBD Genitourinary Diseases Expert Group. (2018). Disparities in chronic kidney disease prevalence among males and females in 195 countries: Analysis of the Global Burden of Disease 2016 Study. NEF, 139, 313–318. [CrossRef]

[17] Lv, J.-C.; Zhang, L.-X. Prevalence and Disease Burden of Chronic Kidney Disease. In Renal Fibrosis: Mechanisms and Therapies; Liu, B.-C., Lan, H.-Y., Lv, L.-L., Eds.; Advances in Experimental Medicine and Biology; Springer: Singapore, 2019; pp. 3–15. ISBN 9789811388712. [CrossRef]

[18] Chothia, M.Y.; Davids, M.R. Chronic kidney disease for the primary care clinician. South Afr. Fam. Pract. 2019, 61, 19–23. [CrossRef]

[19] Stanifer, J.W.; Jing, B.; Tolan, S.; Helmke, N.; Mukerjee, R.; Naicker, S.; Patel, U. The epidemiology of chronic kidney disease in sub-Saharan Africa: A systematic review and meta-analysis. Lancet Glob. Health 2014, 2, e174–e181. [CrossRef]

[20] Olanrewaju, T.O.; Aderibigbe, A.; Popoola, A.A.; Braimoh, K.T.; Buhari, M.O.; Adedoyin, O.T.; Kuranga, S.A.; Biliaminu, S.A.; Chijioke, A.; Ajape, A.A.; et al. Prevalence of chronic kidney disease and risk factors in North-Central Nigeria: A population-based survey. BMC Nephrol. 2020, 21, 467. [CrossRef]

[21] Varughese, S.; Abraham, G. Chronic Kidney Disease in India: A Clarion Call for Change. Clin. J. Am. Soc. Nephrol. 2018, 13, 802–804. [CrossRef]

[22] Ali, S.I.; Bilal, H.S.M.; Hussain, M.; Hussain, J.; Satti, F.A.; Hussain, M.; Park, G.H.; Chung, T.; Lee, S. Ensemble Feature Ranking for Cost-Based Non-Overlapping Groups: A Case Study of Chronic Kidney Disease Diagnosis in Developing Countries. IEEE Access 2020, 8, 215623–215648. [CrossRef]

[23] Prof. Harun-Ur-Rashid Over 35,000 develop kidney failure in Bangladesh every year | The Daily Star

[24] National Institute of Health Chronic Kidney Disease (CKD) - NIDDK (nih.gov)

[25] Summet K. Asrani, Harshad Devarbhavi, John Eaton, Patrick S Kamath [CrossRef]

[26] Jurgen Rehm, Andriy V. Samokhvalov, Kevin D. Shield [CrossRef]

[27] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, ‘‘A machine learning methodology for diagnosing chronic kidney disease,’’ IEEE Access, vol. 8, pp. 20991–21002, 2020. [CrossRef]

[28] Silveira, A.C.M.D.; Sobrinho, Á.; Silva, L.D.D.; Costa, E.D.B.; Pinheiro, M.E.; Perkusich, A. Exploring Early Prediction of Chronic Kidney Disease Using Machine Learning Algorithms for Small and Imbalanced Datasets. Appl. Sci. 2022, 12, 3673. [CrossRef]

[29] Ebiaredoh-Mienye, S.A.; Esenogho, E.; Swart, T.G. Integrating Enhanced Sparse Autoencoder-Based Artificial Neural Network Technique and SoftMax Regression for Medical Diagnosis. Electronics 2020, 9, 1963. [CrossRef]

[30] S. Y. Yashfi et al., "Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-5. [CrossRef]

[31] Almansour, N. A.., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., … Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in Biology and Medicine, 109, 101–111. [CrossRef]

[32] Kim, D.-H., & Ye, S.-Y. (2021). Classification of chronic kidney disease in sonography using the GLCM and artificial neural network. Diagnostics, 11, 864. [CrossRef]

[33] Sara, S. A. B. V. J., & Kalaiselvi, K. (2018). Ensemble swarm behaviour based feature selection and support vector machine classifier for chronic kidney disease prediction. International Journal of Engineering & Technology, 7, 190–195. [CrossRef]

[34] Rady, E.-H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked, 15, Article 100178. [CrossRef]

[35] Polat H, Danaei Mehr H, Cetin A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. J Med Syst. 2017 Apr;41(4):55. [CrossRef]

[36] Ghosh, P.; Shamrat, F.J.M.; Shultana, S.; Afrin, S.; Anjum, A.A.; Khan, A.A. Optimization of prediction method of chronic kidney disease using machine learning algorithm. In Proceedings of the 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Bangkok, Thailand, 18–20 November 2020; pp. 1–6. [CrossRef]

[37] Indian Liver Patient Records. Available online: (accessed on 14 November 2022). [CrossRef]

[38] Sokoliuk, A.; Kondratenko, G.; Sidenko, I.; Kondratenko, Y.; Khomchenko, A.; Atamanyuk, I. Machine learning algorithms for binary classification of liver disease. In Proceedings of the 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, Ukraine, 6–9 October 2020; pp. 417–421. [CrossRef]

[39] Azam, M.S.; Rahman, A.; Iqbal, S.H.S.; Ahmed, M.T. Prediction of liver diseases by using few machine learning based approaches. Aust. J. Eng. Innov. Technol. 2020, 2, 85–90. [CrossRef]

[40] Srivastava, A.; Kumar, V.V.; Mahesh, T.; Vivek, V. Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms. In Proceedings of the 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 21–22 April 2022; pp. 1–4. [CrossRef]

[41] Choudhary, R.; Gopalakrishnan, T.; Ruby, D.; Gayathri, A.; Murthy, V.S.; Shekhar, R. An Efficient Model for Predicting Liver Disease Using Machine Learning. In Data Analytics in Bioinformatics: A Machine Learning Perspective; Wiley Online Library: Hoboken, NJ, USA, 2021; pp. 443–457. [CrossRef]

[42] Geetha, C.; Arunachalam, A. Evaluation based Approaches for Liver Disease Prediction using Machine Learning Algorithms. In Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 27–29 January 2021; pp. 1–4. [CrossRef]

[43] Gajendran, G.; Varadharajan, R. Classification of Indian liver patient's data set using MAMFFN. In Proceedings of the AIP Conference Proceedings, Coimbatore, India, 17–18 July 2020; Volume 2277, p. 120001. [CrossRef]

[44] Elias Dritsas, Supervised Machine Learning Models for Liver Disease Risk Prediction, 13 January 2023 [CrossRef]

[45] Maldonado, S.; López, J.; Vairetti, C. An alternative SMOTE oversampling strategy for high-dimensional datasets. Appl. Soft Comput. 2019, 76, 380–389. [CrossRef]

[46] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2020). A machine learning methodology for diagnosing chronic kidney disease. IEEE access: practical innovations, open solutions, 8, 20991–21002. [CrossRef]

[47] Aldhyani, T. H. H., Alshebami, A. S., & Alzahrani, M. Y. (2020). Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms. Journal of Healthcare Engineering, 2020, Article e4984967. [CrossRef]

[48] Zhao, L., Lee, S., & Jeong, S.-P. (2021). Decision tree application to classification problems with boosting algorithm. Electronics, 10, 1903. [CrossRef]

[49] Puga, J. L., Krzywinski, M., & Altman, N. (2015). Bayes' theorem. Nature Methods, 12, 277–278. [CrossRef]

[50] Shahraki, A.; Abbasi, M.; Haugen, Ø. Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. Eng. Appl. Artif. Intell. 2020, 94, 103770. [CrossRef]

[51] Wyner, A. J., Olson, M., Bleich, J., & Mease, D. (2017). Explaining the success of adaboost and random forests as interpolating classifiers. The Journal of Machine Learning Research, 18, 1558–1590. [CrossRef]

[52] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 6639–6649). Curran Associates Inc. [CrossRef]

[53] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery. [CrossRef]

[54] Daoud, E. A. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. International Journal of Computer and Information Engineering, 13, 6–10. [CrossRef]

[55] Zhang, Y. (2012). Support vector machine classification algorithm and its application. In C. Liu, L. Wang, & A. Yang (Eds.), Information computing and applications (pp. 179–186). Berlin, Heidelberg: Springer. [CrossRef]

[56] Pisner, D.A.; Schnyer, D.M. Support vector machine. In Machine Learning; Elsevier: Amsterdam, The Netherlands, 2020; pp. 101–121. [CrossRef]

[57] Chatzigeorgakidis, G., Karagiorgou, S., Athanasiou, S., & Skiadopoulos, S. (2018). FMLkNN: Scalable machine learning on Big Data using k-nearest neighbor joins. Journal of Big Data, 5, 4. [CrossRef]

[58] (S. Dre) S. Dreiseitl, & L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, J. Biomed. Inform. 35 (5-6) (2002) 352–359. [CrossRef]

[59] A Mahabub, MI Mahmud, MF Hossain, A robust system for message filtering using an ensemble machine learning supervised approach, ICIC Expr. Lett. Part B Appl. 10 (2019) 805–811. [CrossRef]

[60] Mahabub A Mahabub, AZSB Habib, A voting approach of modulation classification for wireless network, in: Proceedings of the 6th international

conference on networking, systems and security, ACM, 2019, pp. 133–138. [CrossRef]

[61] Handelman, G.S.; Kok, H.K.; Chandra, R.V.; Razavi, A.H.; Huang, S.; Brooks, M.; Lee, M.J.; Asadi, H. Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. Am. J. Roentgenol. 2019, 212, 38–43. [CrossRef]

[62] Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics 2021, 10, 593. [CrossRef]