

CHAPTER - 1

1.INTRODUCTION

1.1 GENERAL

Chronic kidney disease (CKD) is a major public health concern around the world, with negative outcomes such as renal failure, cardiovascular disease, and early death [1]. According to a 2010 study by the Global Burden of Disease Study (GBDS), chronic kidney disease (CKD) was listed as the 18th leading cause of mortality worldwide, up from 27th in 1990 [2]. Chronic kidney disease affects over 500 million people worldwide [3, 4], with a disproportionately high burden in developing countries, particularly South Asia and sub-Saharan Africa [5]. According to a 2015 study, there were 110 million people with CKD in high income nations (men 48.3 million, women 61.7 million), but 387.5 million in low- and middle-income countries [6]. Bangladesh is a densely populated developing country in Southeast Asia where chronic kidney disease is on the rise year after year. The overall population of CKD is estimated to be 14 percent in a global study of six areas, including Bangladesh [7]. Another study discovered a 26% prevalence of chronic kidney disease among urban Dhaka residents over 30 years old [8], while another researcher discovered a 13% prevalence of chronic kidney disease among urban Dhaka residents over 15 years old [9]. In 2013, a community-based prevalence study in Bangladesh revealed that one third of rural residents were at risk of developing CKD, which was generally misdiagnosed at the time [10]. Chronic kidney disease (CKD) patients are more prone to developing end stage renal disease (ESRD), which demands expensive treatment methods like dialysis and kidney transplantation [11], and this financial load leads to long-term medical and psychological difficulties [12, 13]. Furthermore, on a global scale, CKD is caused by unmanaged diabetes and hypertension, and the prevalence of CKD is now impacted by these two risk factors. From the perspective of public health, it is vital to be able to estimate CKD occurrence trends so that decision-makers (ministries, insurers, hospital administrators, and so on) can take proactive measures to avoid a growth in the number of patients. Rising population screening for CKD-related risks and awareness programs are examples of such mitigation strategies, as it has been demonstrated that changes in lifestyle (weight loss, improved diet, increased physical activity, reduced alcohol consumption, avoided smoking, early referral to nephrologists, appropriate medication use, and treatment options to manage other risk factors) are the most useful. Additional mitigating strategies include establishing appropriate haemodialysis facilities and training workers. Diagnosis of

kidney impairment early may help in rectification, which is not always possible. To avoid serious damage, we will need to get a better understanding of a few indicators caused by kidney disease. The main motivation of this study is to predict renal disease by analysing data from those indices and applying three machine learning classification approaches to predict the disease, then choosing the approach with the highest accuracy rate. Three classification techniques are used: Random Forest Classification for Chronic Kidney Disease, Dataset among all classifiers like Adboost Classifier, Gradient Boosting Classifier. Machine learning classifiers are used to forecast a data point's class, target, labels, and categories. Classification is a kind of supervised learning in which input data is given to the objectives. Medical diagnosis, spam identification, and targeted marketing are just a few of the applications. The authors of [14] worked on improving prediction algorithms for chronic cerebral infraction disease using data from chronic cerebral infraction disease. They discovered that when data is missing, a model's accuracy drops. Using structured and unstructured hospital data, they developed a (CNN)-based multimodal illness risk prediction algorithm. Additionally, they utilized a latent component model to rebuild the missing data. Also, the authors of [15] constructed decision trees using both ID3, which is based on information gain and gain ratio, and evolutionary algorithms, which are based on fitness proportional and rank selection methods. Their findings demonstrated that the ID3 algorithm outperformed the evolutionary approach. On the other hand, the authors of [16] discovered that when the K-nearest neighbours (KNN) classifier is used, the computational load on the CPU grows polynomial as the data set grows. They demonstrated that using the NVIDIA CUDA API speeds up the search for the KNN by a factor of 120. The authors of [17] studied and assessed a variety of machine learning models, including (support vector machine) SVM, KNN (K-nearest neighbour), and DT (decision tree). In [18], the authors compared SVM, RF, and ELM algorithms for intrusion detection in a protected network. Their findings indicate that ELM beats all other methods they evaluated. Hussain and fellow researchers achieved high accuracy in predicting CKD in its early stages by combining multilayer perception with neural network preprocessing to fill in missing information. The process includes removing outliers, choosing the optimum seven attributes using statistical analysis, and eliminating characteristics with greater interrelationship as determined by principal component analysis (PCA) [19]. The missing value filling technique has a considerable impact on the trained models' accuracy in the aforementioned study. The accuracy of missing value prediction is slightly reduced because the neural network is employed to predict missing values for just 20 features, and 260 entirely completed data

instances [19]. Discarding characteristics with more than 20% missing values improved the accuracy of substituting missing values significantly. The categorization of features by source, such as blood test or urine test, helps in the selection of training model attributes from each class. In terms of the five stages of CKD, a method is given to predict a stage with an overall accuracy of 0.967 while removing missing values and estimating the eGFR utilizing the previous data set with extra gender and racial characteristics [20]. Due to the model's somewhat lower precision, constants are used to substitute missing data. However, our study demonstrates that the randomization of missing data points is ideal when using Little's MCAR method [21] (see Methodology section). Additionally, when considering the characteristics in [22], the significance of serum creatinine is skewed. However, in the early stages of CKD, serum creatinine readings may look normal, and the overall significance of all other features may not surpass serum creatinine [23], providing serum creatinine useless for disease prediction. The lack of domain expertise raises concerns about the trained models' capacity to predict new occurrences outside of the data set. In 2017, a team of academics predicted CKD with good accuracy using 14 variables and a multiclass decision forest [24]. The proposed work reveals with the Random Forest Classification for Chronic Kidney Disease, Dataset among all classifiers like Adboost Classifier, Gradient Boosting Classifier.

1.2 SYSTEM SPECIFICATION

1.2.1 Hardware Specification

Processor:

The heart of system is the Intel(R) Core(TM) i5-8265U CPU. This processor has a base speed of 1.60 GHz, which can dynamically boost up to 1.80 GHz when necessary. This enables computer to handle a range of tasks efficiently, from everyday web browsing to more demanding computational tasks.

Memory (RAM):

I have 8.00 GB of RAM installed in the system. This significant amount of memory allows to run multiple applications simultaneously without experiencing slowdowns. It's particularly beneficial for tasks like running data analysis software, where ample RAM is essential for handling large datasets and complex calculations.

System Type:

In computer operates on a 64-bit operating system, with an x64-based processor. This architecture is well-suited for modern computing needs and ensures compatibility with both 32-bit and 64-bit software applications.

Operating System:

In operating system is Windows 10, a versatile and widely used platform. Windows 10 offers a user-friendly interface and supports various software and hardware configurations, making it a suitable choice for a wide range of tasks.

1.2.2 Software Specification**Operating System:**

In system supports internet browsing on all major web browsers, ensuring can access the web, browse websites, and use web-based applications seamlessly.

Programming Language:**Python**

Python, a popular and versatile programming language, is installed on the system. Python is renowned for its simplicity and readability, making it an excellent choice for both beginners and experienced programmers. I can leverage Python for a wide range of applications, including web development, data analysis, scientific computing, and more.

Software Tools**Jupyter Notebook:**

Jupyter Notebook is a powerful tool for interactive computing and data analysis. With Jupyter Notebook, can create and share documents containing live code, visualizations, and explanatory text. It's particularly valuable for data scientists and researchers as it facilitates reproducible and collaborative work.

Anaconda:

Anaconda is a comprehensive Python distribution that simplifies package management and environment setup. It comes pre-packaged with a wealth of data science libraries and tools, making it an ideal choice for data analysis, machine learning, and scientific research. Anaconda's conda package manager streamlines the installation of additional libraries and dependencies, making it easier to manage complex Python projects.

1.3 REQUIRED PACKAGES

1.Certifi:

- Certifi is a Python package that provides a carefully curated collection of Root Certificates for validating the trustworthiness of SSL certificates in web applications. This package helps ensure secure communication over HTTPS.

2.Click:

- Click is a Python package that simplifies the creation of command-line interfaces (CLIs) for Python programs. It provides an easy way to define and handle command-line arguments and options.

3.Colorama:

- Colorama is a Python package that simplifies adding colored output to terminal text. It's often used to enhance the readability and aesthetics of command-line interfaces by adding colored text and formatting.

4. Flask:

- Flask is a popular Python web framework for building web applications. It provides tools and libraries for handling routing, request/response handling, and creating web APIs.

5.Importlib:

- Importlib-metadata is a Python package that provides access to the metadata about installed packages and their dependencies. It's commonly used by other Python libraries and tools to manage and inspect package metadata.

6.Itsdangerous:

- Itsdangerous is a Python library used for various security-related tasks, such as generating and verifying digital signatures for cookies, tokens, and other data in web applications.

7.Jinja:

- Jinja2 is a templating engine for Python. It allows to define templates with placeholders that can be dynamically filled with data, making it useful for generating HTML, XML, or other text-based formats.

8.Joblib:

- Joblib is a Python library for lightweight pipelining and parallelism. It's commonly used for efficiently executing functions and managing memory-intensive tasks, such as machine learning model training.

9.MarkupSafe:

- MarkupSafe is a dependency of Jinja2, providing utilities for escaping and preserving text in templates to prevent security vulnerabilities like Cross-Site Scripting (XSS).

10.Numpy:

- NumPy is a fundamental package for scientific computing in Python. It provides support for large, multi-dimensional arrays and matrices, as well as a variety of mathematical functions to operate on these arrays.

11.Scikit-learn:

- Scikit-learn, also known as sklearn, is a machine learning library for Python. It provides a wide range of tools and algorithms for tasks like classification, regression, clustering, and more.

12.SciPy:

- SciPy is a scientific computing library that builds on NumPy. It includes additional modules for optimization, signal and image processing, linear algebra, statistics, and more.

13. Sklearn:

- This is likely an incorrect package specification or a placeholder. It's not a standard package name, and it's not clear what it refers to.

14.ThreadPoolCtl:

- ThreadPoolCtl is a Python library that helps manage and control thread pools. It's commonly used in conjunction with libraries that use multithreading to optimize resource utilization.

15. Typing Extensions:

- The typing-extensions module provides additional type hinting tools for Python. It extends the standard `typing` module and is used to annotate code for better static analysis.

16. Werkzeug:

- Werkzeug is a WSGI (Web Server Gateway Interface) utility library for Python. It provides the foundation for building web frameworks like Flask.

17. Wincertstore:

- WinCertStore is a Python package for Windows that allows to access Windows Certificate Stores. It's primarily used for managing SSL certificates on Windows systems.

18. Zipp:

- Zipp is a library for handling ZIP archive files in Python. It provides tools for reading and writing ZIP archives and is used in various applications that work with compressed files.

19. Gunicorn:

- Gunicorn, short for Green Unicorn, is a popular HTTP server for running Python web applications. It's often used as a production-ready web server in combination with web frameworks like Flask and Django.
- These are the Packages used in this Project.

1.4 STAGES OF CKD

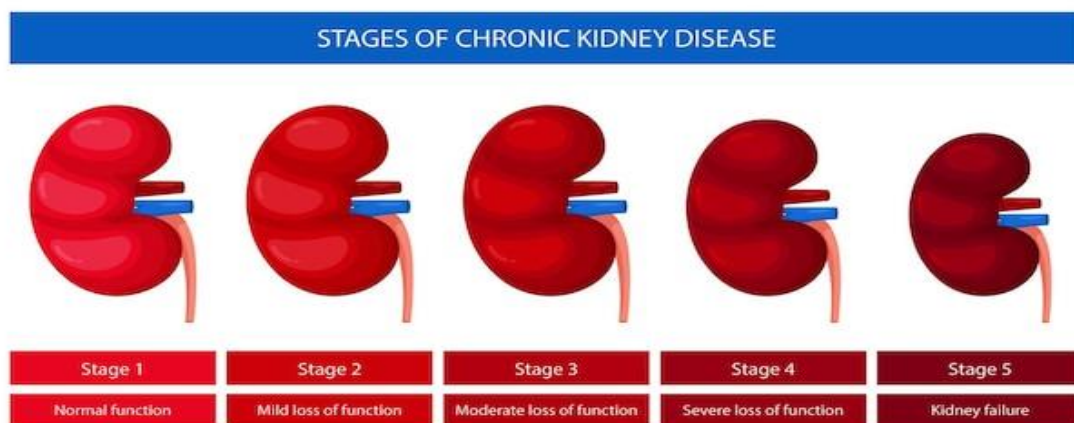


Figure 1.1 Stages of Chronic Kidney Disease

Early stages of CKD

CKD in its early stages typically does not present any symptoms. This is due to the fact that the human body can typically adjust to a large decrease in the function of the kidneys. It is common for kidney disease to not be diagnosed until this stage unless a routine test for another issue, such as a test of the blood or urine, discovers a potential problem. If it is discovered at an early stage, treatment with medication and ongoing monitoring with routine tests may help prevent it from progressing to a more advanced state.

CKD in its advanced stages

If kidney disease is not caught early or keeps getting worse even after treatment, there may be several signs. Kidney failure is the last stage of CKD. It is also called end-stage renal disease or established renal failure. It is possible that dialysis or a kidney transplant will be needed at some point.

When to see a physician

If I have signs or symptoms of renal illness, make an appointment with doctor. Renal disease could be prevented from progressing to kidney failure if detected early. During office visits, doctor may use urine and blood tests to check the blood pressure and kidney function if have a health condition that makes more likely to get renal disease. Ask physician if these tests are required.

Tests for CKD

Chronic kidney disease is when a disease or condition makes it hard for the kidneys to work, causing the damage to the kidneys to get worse over time. This can occur when the kidneys are affected by another disease or condition. Studies show that the number of people with CKD who are admitted to hospitals is going up by 6.23 percent every year, even though the global death rate has stayed the same. There are just a few diagnostic tests available to check the status of CKD, including: (i) estimated glomerular filtration rate (eGFR) (ii) a urine test; (iii) a blood pressure reading; (iv) tests for CKD.

eGFR

The eGFR value provides information on how well kidneys cleanse the blood. If eGFR number is higher than 90, it means that kidneys are working well. If the value of y eGFR is less than 60, this indicates that have CKD.

Urine test

In order to evaluate kidney function, the physician also requests a urine sample. Urine is produced by the kidneys. If urine contains blood and protein, it is an indication that one or both of kidneys are not functioning normally.

Blood pressure

The doctor takes blood pressure because the range of blood pressure reveals how well heart is pumping blood. If the patient's eGFR value falls below 15, this means they have reached the end stage of kidney disease. There are just two treatments that are now available for renal failure: (i) dialysis and (ii) kidney transplantation. The patient's life expectancy after dialysis is contingent on several characteristics, including age, gender, the frequency and length of dialysis treatments, the patient's level of physical mobility, and their mental state. Kidney transplantation is the only option left for the doctor to consider if dialysis cannot be performed successfully. Nevertheless, the price is exorbitantly high.

Other tests

When determining the extent of the damage to kidneys, it is not uncommon for additional tests to be performed. These may include an ultrasound scan, a magnetic resonance imaging scan, or a computed tomography scan. Their purpose is to look at the kidneys and see if there are any blockages. A needle is used to take a small piece of kidney tissue, and the cells are looked at under a microscope to look for signs of kidney disease. This is done in order to diagnose kidney conditions. The field of medicine is an extremely important area for the application of intellectually sophisticated systems. Then, data mining could be a big part of finding hidden information in the huge amount of patient medical and treatment data. This is information that doctors often get from their patients to learn more about their symptoms and make more accurate treatment plans.

Table 1.1 Stage of CKD Description

Stage of Chronic Kidney Disease	Description
One	Kidney function remains normal but urine findings suggest kidney disease
Two	Slightly reduced kidney function with urine findings suggesting kidney disease
Three	Moderately reduced kidney function
Four	Severely reduced kidney function
Five	Very severe or end-stage kidney failure

1.5 APPLICATIONS OF CHRONIC KIDNEY DISEASE PREDICTION

1.Risk Assessment: Machine learning models can analyse patient data, including demographics, medical history, and lab results, to assess the risk of developing CKD. These models can identify individuals who are at higher risk and may benefit from closer monitoring and preventive measures.

2.Early Detection: ML algorithms can analyse biomarkers such as serum creatinine levels, estimated glomerular filtration rate (eGFR), and urine protein levels to detect early signs of kidney dysfunction. Early detection allows for timely intervention and management.

3.Disease Progression Monitoring: ML models can predict how CKD may progress in individual patients. This information can help healthcare providers tailor treatment plans and interventions to slow or manage the progression of the disease effectively.

4.Medication Optimization: ML can assist in optimizing medication dosages for CKD patients. It can consider various factors, such as renal function, comorbidities, and medication interactions, to recommend personalized medication plans.

5.Hospital Readmission Prediction: Predictive models can identify CKD patients at risk of hospital readmission. This can help healthcare providers implement targeted interventions to reduce readmission rates and improve patient care.

6.Dietary and Lifestyle Recommendations: ML algorithms can provide personalized dietary and lifestyle recommendations for CKD patients. These recommendations can be based on individual patient data, including dietary habits and activity levels, to help manage the disease and improve quality of life.

7.Transplant Outcome Prediction: For CKD patients undergoing kidney transplantation, machine learning can predict the likelihood of transplant success and long-term outcomes. This information can aid in decision-making regarding transplantation and post-transplant care.

8.Data Integration and EHR Analysis: ML can integrate and analyse data from electronic health records (EHRs) to identify patterns and correlations that may not be apparent through traditional analysis methods. This can lead to more accurate predictions and insights into CKD.

9.Telehealth Monitoring: Machine learning can be used in telehealth applications to remotely monitor CKD patients. Continuous data streams from wearable devices and sensors can be analysed to detect deviations from baseline health, allowing for timely interventions.

10.Patient Education and Engagement: ML-powered chatbots and virtual assistants can provide CKD patients with information, reminders, and support for managing their condition. These tools can help improve patient adherence to treatment plans and lifestyle modifications.

1.6 BENIFITS OF CHRONIC KIDNEY DISEASE PREDICTION

- **Early Intervention:** One of the most critical advantages is the ability to detect CKD at an early stage or even before symptoms become apparent. Early detection allows for timely intervention, which can slow the progression of the disease and potentially prevent complications.
- **Improved Patient Outcomes:** Early intervention and management can lead to better patient outcomes. By identifying CKD in its early stages, healthcare providers can implement treatment plans that help preserve kidney function and

reduce the risk of complications such as cardiovascular disease and kidney failure.

- **Cost Savings:** Predicting CKD can lead to cost savings for healthcare systems. Treating CKD in its advanced stages, especially when it progresses to kidney failure, is expensive. Early detection and management can help reduce the need for costly interventions such as dialysis or kidney transplantation.
- **Personalized Care:** CKD prediction models can provide personalized recommendations for patients. These recommendations can include tailored treatment plans, dietary and lifestyle advice, and medication management, all based on individual patient data. Personalized care is more effective and patient-centric.
- **Reduced Hospitalizations:** Predicting CKD-related complications can help reduce hospitalizations. Hospitalizations are not only costly but also disruptive to patients' lives. Preventing complications through early detection can improve the overall quality of life for CKD patients.
- **Optimized Medication Management:** CKD patients often require multiple medications to manage their condition and comorbidities. Prediction models can help optimize medication management by considering factors such as renal function and potential drug interactions, leading to safer and more effective treatment.
- **Resource Allocation:** Healthcare systems can use CKD prediction models to allocate resources more efficiently. By identifying high-risk patients, providers can prioritize their care and resources, ensuring that those who need it most receive timely attention.
- **Patient Empowerment:** Knowing that they are at risk for CKD can empower patients to take an active role in their healthcare. Patients can make informed decisions about lifestyle changes, adhere to treatment plans, and engage in regular monitoring to manage their condition effectively.
- **Research and Public Health:** CKD prediction data can be valuable for research and public health efforts. Aggregated data can help researchers study CKD trends, risk factors, and treatment outcomes, leading to advancements in CKD management and prevention.
- **Reduced Burden on Healthcare Providers:** CKD prediction models can assist healthcare providers in identifying at-risk patients more efficiently. This can

reduce the burden on clinicians and allow them to focus on delivering more personalized care and interventions.

1.7 CONCEPT USED IN MACHINE LEARNING ALGORITHMS

Predicting chronic kidney disease (CKD) using machine learning algorithms like Random Forest, AdaBoost, and Gradient Boosting is a valuable application in healthcare.

1.7.1 Random Forest:

Random Forest is an ensemble learning technique that combines multiple decision trees to make predictions. Here's how it works for CKD prediction:

a. Data Preparation:

Start by collecting a dataset that includes patient data, such as age, gender, blood pressure, serum creatinine levels, and other relevant features, as well as labels indicating whether a patient has CKD or not.

b. Training:

The Random Forest algorithm creates a forest of decision trees. Each tree is trained on a random subset of the data and uses a random subset of features. This randomness helps reduce overfitting.

c. Decision Trees:

Each decision tree in the Random Forest makes predictions based on the patient's features. For CKD prediction, it may consider features like creatinine levels, age, and other relevant factors. The algorithm then combines the predictions of all trees.

d. Aggregation:

Random Forest aggregates the predictions of individual trees, typically by a majority vote (classification) or averaging (regression). In CKD prediction, it could classify patients as having CKD or not based on the majority prediction.

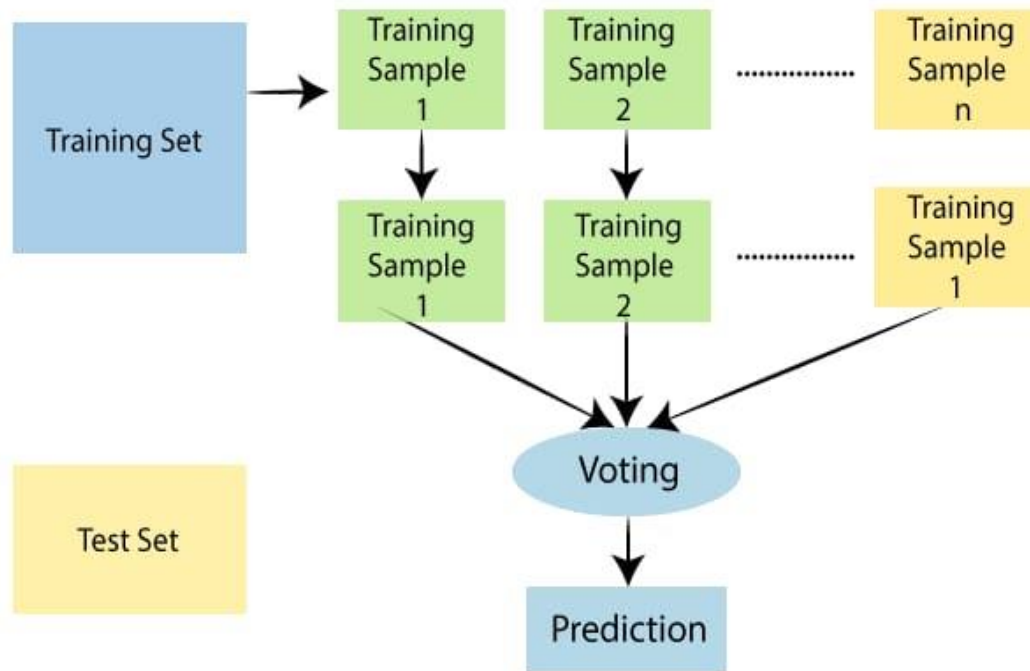


Figure 1.2 Flow Chart for simplified for Random Forest Algorithm

1.7.2 AdaBoost (Adaptive Boosting): AdaBoost is another ensemble learning algorithm that focuses on improving the accuracy of weak learners (often simple decision trees) by giving more weight to misclassified samples. Here's how AdaBoost works for CKD prediction.

a. Data Preparation: Similar to Random Forest, start with a dataset containing patient features and CKD labels.

b. Weighted Training: AdaBoost initially assigns equal weights to all samples. It trains a weak learner (e.g., a shallow decision tree) and calculates its error rate.

c. Weight Update: AdaBoost increases the weights of the misclassified samples, making them more important in subsequent iterations.

d. Iteration: AdaBoost repeats steps b and c multiple times, creating multiple weak learners. Each learner focuses on the mistakes made by the previous ones.

e. Combining Predictions: AdaBoost combines the predictions of all weak learners, giving more weight to those that performed better during training.

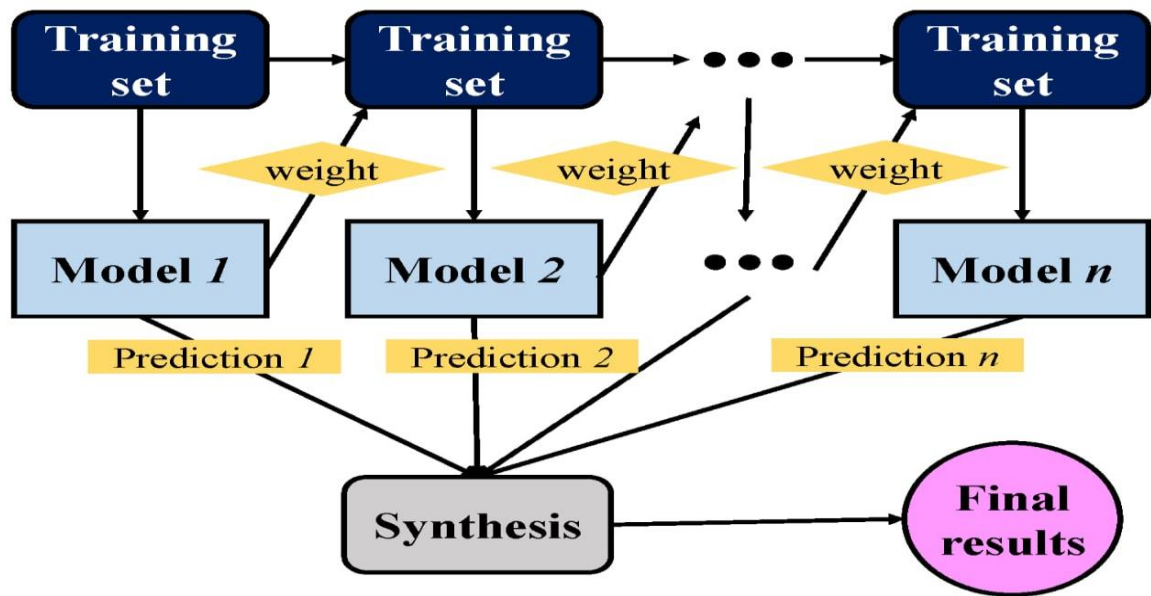


Figure 1.3 : Flow Chart for simplified AdaBoost Algorithm

1.7.3 Gradient Boosting

Gradient Boosting is another ensemble technique that builds an ensemble of decision trees sequentially. It aims to minimize a loss function iteratively. Here's how Gradient Boosting works for CKD prediction:

- a. Data Preparation:** Prepare the CKD dataset as before.
- b. Initial Model:** Start with an initial model, often a simple one like a single decision tree.
- c. Residuals:** Calculate the residuals (differences between actual and predicted values) from the initial model.
- d. Fit Next Model:** Train a new decision tree model to predict the residuals from the previous step.
- e. Update Predictions:** Update the predictions by adding the output of the new model to the previous predictions.
- f. Iteration:** Repeat steps c to e multiple times, with each new model trying to correct the errors of the previous ones.
- g. Final Prediction:** The final prediction is the cumulative sum of the predictions made by all the models.

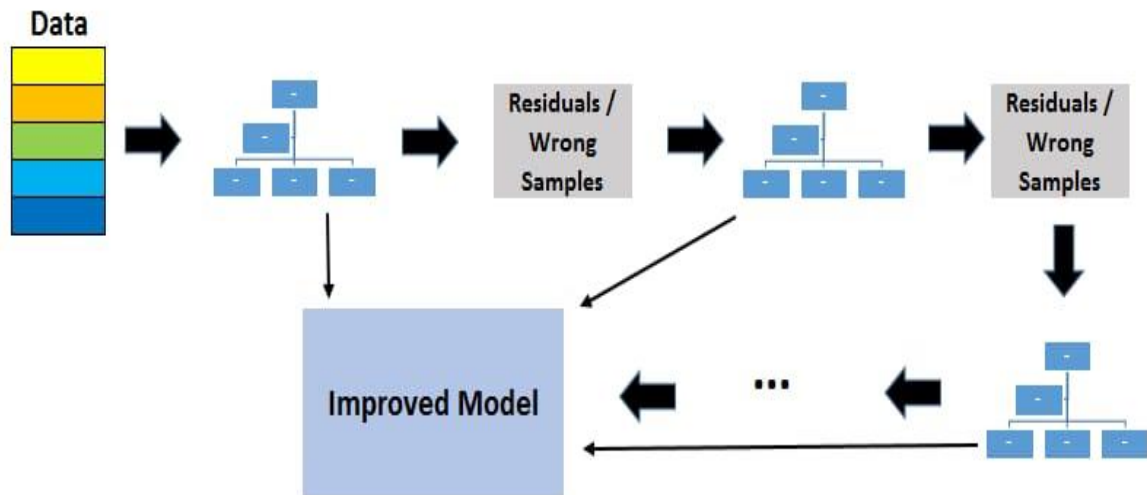


Figure 1.4 Flow Chart for simplified Gradient Boosting Algorithm

These algorithms are implemented using libraries like scikit-learn in Python. Detailed diagrams for real-world CKD prediction models using these algorithms would require data and software tools. These algorithms should be adapted and fine-tuned based on the specific CKD dataset and the desired performance metrics. Additionally, ethical considerations and data privacy regulations must be observed when working with patient data for healthcare applications.

1.8 OBJECTIVES OF THIS RESEARCH

1. To Develop screening and diagnostic methods to identify CKD at its earliest stages, when interventions are most effective.
2. To Implement Machine Learning Algorithms and data analytics to refine CKD detection, improving accuracy and speed.
3. To identify genetic factors that predispose individuals to CKD and inform personalized treatment.
4. To Improve the sensitivity and specificity of CKD diagnostic tests, reducing the risk of false positives and negatives.
5. To identify genetic factors that predispose individuals to CKD and inform personalized treatment.
6. To Detect CKD early to enable prompt medical interventions and lifestyle modifications, ultimately improving patient outcomes.
7. To create a user-friendly interface for healthcare providers to input patient data and receive risk assessments, making it accessible and practical for clinical use.

1.9 PROBLEM DEFINITION

Chronic diseases are becoming one of the leading causes of death around the world. A growing percentage of the world's population is suffering from the negative health repercussions of living. In general, doctors must thoroughly examine the patient's reports in order to make a disease diagnosis. It can be difficult for clinicians to treat patients efficiently when the diagnosis is manual. The number of persons affected by chronic diseases is steadily increasing. The traditional health-care system is a passive one. Patients may die as a result of a lack of effective treatment during crises such as cardiac arrest related to this type. The key to increasing health-care efficiency is to lower mortality rates due to a lack of effective treatment and to turn a passive health-care programme into a continuous, low-cost.

CHAPTER-2

2.LITERATURE REVIEW

In 2015 Parul et al. conducted a Comparative Study for predicting CKD by using classification algorithms K-Nearest Neighbour and SVM [1]. Performance of classification algorithm has been compared on the basis of accuracy, precision and total execution time for prediction of Chronic Kidney disease. MATLAB was used for this classification model. Performance of K-Nearest Neighbour classifier was 78.75% which was better than Support Vector machine with an accuracy of 73.75%

In 2017 Pinar Yildirim made a classification model for chronic kidney Disease prediction on Imbalanced Data by using Multilayer Perceptron [2]. This work focuses on after effect of class imbalance in training data for predicting CKD or Non CKD. Multilayer perceptron algorithm is used to calculate accuracy. Resample, SMOTE algorithms have been used. The work was performed using 0.8 WEKA 3.7.3 Software, and data for research was taken from UCI Machine Learning Repository of 400 patients with 25 attributes. As per the result Resample Method with Multilayer Perceptron was more accurate, but for Execution time Spread Sub Sample Algorithm was fast with time of 0.0509

In 2017 Gunarathne et al. has made Performance Evaluation on Machine Learning Classification Techniques for Disease Forecasting and classification through Data Analytics for Chronic Kidney Disease (CKD) [3]. Their area of research is to predict CKD and Non CKD of a patient. Number of Classification algorithms has been used– Multiclass – Decision Forests, Jungle, Logistic Regression and NN. Results were obtained using Microsoft Azure Machine Learning Studio. Result of Multiclass Decision Forest was with highest accuracy of 99.1%

Dr.Uma N Dulhare et al. in 2016 performed Extraction of action Rules for Chronic Kidney Disease Prediction using Naïve Bayes. Naïve Bayes with OneR attribute Selector was used for prediction CKD status of a patient [4]. The idea was to select a subset from input data by elimination idle data which carried little or no predictive knowledge. Data sets were taken from UCI ML Repository. The result and analysis proposed Naïve Bayes with OneR with highest improved accuracy and also reduced number of attributed to 80% which is 05 from total of 25 attributes compared to other attribute evaluators

Dr.N.Radha, S.Ramya in 2016 performed a diagnosis of chronic kidney disease using Machine learning [5] using R tool and algorithms like Back Propagation neural network, Random forest, Radial Basis function, ANN. The data for this research was medical

reports of patients taken from different labs in South India. They have used 1000 instances with 15 CKD related attributes. Their model is evaluated on different measures like Sensitivity, Accuracy, and Specificity. The experimental results proved that Radial Basis Function performed better than other algorithms and obtained an accuracy of 85.3%

Dr.S.Vijayarani, S.Dhayanand in 2015 used Data Mining Classification algorithms for Prediction Kidney Disease [6] using MATLAB tool. Their work focuses on finding best classification algorithm on basis of accuracy and execution time for prediction of Kidney Disease. They have used Naïve Bayes and Support Vector Machine algorithms. In their Prediction model SVM classification algorithm performed better than Naïve Bayes with an accuracy of 76.32

M.P.N.M. Wickramasinghe et al. in 2017 proposed Dietary prediction of patients with CKD by considering Blood Potassium Level [7]. Their work suggests diet plans by taking patients potassium level in consideration. The experiment is performed using Multiclass Jungle, Forests, and Neural networks in Microsoft Azure Machine Learning Studio. In their results Multiclass Decision Forest performed with an accuracy of 99.17%

Torgyn Shaikhina et al. in 2017 developed classification model for outcome prediction in antibody incompatible kidney transplantation [8]. The base objective is to independently identify risk linked with kidney transplant within first 30 days of transplant that how much the kidney is accepted by the patient's body. This work would help doctors to predict outcomes of kidney transplant at early stage. Decision Tree, Random Forest classification algorithms were used for this prediction. Their work for predicting kidney transplant failure performed with an accuracy of 85%

Radha, N, Ramya,S. in 2015 predicted occurrence of chronic kidney disease using machine learning classification algorithms [9]. In their work the data collected is real data and belongs to the laboratories of south India and consist of record of 1000 people with their respective 14 attributes. In their work they have taken Naïve Bayes, KNN, SVM and decision tree as their classification algorithm for CKD prediction. They have used the same data for all the algorithms. The results were obtained Naïve Bayes classifier performed with an accuracy of 61.85, KNN performed with an accuracy of 98%

CHAPTER-3

3.1 EXISTING SYSTEM

Chronic kidney disease (CKD) is a debilitating condition that progressively damages the kidneys, impairing their vital role in maintaining the body's health. As CKD advances, the kidneys lose their ability to effectively filter waste products and excess fluids from the bloodstream, resulting in the accumulation of harmful substances in the body. Beyond the direct impact on renal function, CKD poses a substantial risk to overall health by elevating the probability of developing heart and blood vessel diseases. Recognizing the severity of CKD and its far-reaching consequences, this study endeavors to leverage data-driven approaches to predict and understand the disease. In the realm of predictive modeling for CKD, the Random Forest algorithm takes center stage. Renowned for its capacity to deliver robust classification accuracy, Random Forest employs an ensemble learning approach that harnesses the collective insights of multiple decision trees. This technique proves highly effective in handling complex datasets, producing predictions with exceptional precision. The study's evaluation of the Random Forest model goes beyond mere accuracy, incorporating essential classification performance metrics, including Accuracy, Specificity, Sensitivity, and Precision.

These metrics serve as indispensable gauges of the model's proficiency in discriminating between CKD and non-CKD cases. In addition to Random Forest, this investigation explores the potential of two other machine learning algorithms: Support Vector Machine (SVM) and Logistic Regression (LR). SVM, renowned for its ability to identify optimal hyperplanes for data separation, stands as a formidable contender for classification tasks. Logistic Regression, a simpler yet highly effective algorithm, models the probability of data points belonging to specific classes. Both SVM and LR contribute their unique strengths to the creation of predictive models designed to detect CKD within the dataset. The passage underscores the SVM classifier as the standout performer, achieving the highest accuracy and sensitivity through rigorous training and testing. This page delves deeply into an exhaustive examination of the results obtained from various classifiers. It presents a meticulous comparative analysis of performance metrics, encompassing Accuracy, Specificity, Sensitivity, and Precision, for each model: Random Forest, SVM, and Logistic Regression. These metrics offer profound insights into the models' proficiency in distinguishing between CKD and non-CKD cases, shedding light on their respective strengths and limitations. With SVM emerging as the preeminent classifier, the implications of these findings for clinical practice come into sharp focus.

Timely detection and intervention for CKD can substantially mitigate the elevated risk of heart and blood vessel diseases that accompany this condition. The passage underscores the clinical applications of predictive models in healthcare, offering the potential for improved patient outcomes and a proactive approach to CKD management.

Furthermore, it hints at the exciting possibilities for future research, including refinements in predictive modelling techniques and their broader integration into medical practice. In conclusion, this passage emphasizes the pressing importance of addressing chronic kidney disease due to its profound impact on overall health, particularly the heightened risk of heart and blood vessel diseases. Employing cutting-edge machine learning algorithms such as Random Forest, SVM, and Logistic Regression, predictive models are meticulously constructed to identify CKD cases within the dataset. Importantly, SVM emerges as the most accurate and sensitive classifier in this context. These findings hold the potential to revolutionize healthcare practices by enabling early detection and intervention, ultimately enhancing the quality of care for CKD patients. The passage leaves us with a sense of optimism, encouraging further research and innovation in the realm of predictive modelling for healthcare.

Beyond its immediate clinical applications, the implications of this study extend to broader public health considerations. Chronic kidney disease is a significant global health challenge, affecting millions of individuals. The ability to predict CKD with high accuracy using machine learning models like SVM holds promise for population-level health interventions. Early identification of at-risk individuals could lead to targeted preventive measures, reducing the burden of CKD and its associated complications on healthcare systems worldwide. Moreover, this research highlights the potential for data-driven approaches to drive progress in healthcare, setting a precedent for the integration of advanced analytics in medical practice. As we look ahead, the significance of this study lies not only in its findings but also in the avenues it opens for further exploration.

Future research could delve deeper into the features and factors influencing CKD prediction, potentially enhancing the accuracy of models and their clinical utility. Additionally, the incorporation of real-world clinical data and electronic health records could provide richer insights and lead to more robust predictive models. Collaboration between data scientists, clinicians, and healthcare policymakers will be essential in translating these advancements into tangible improvements in patient care and public health. This passage serves as a testament to the transformative potential of machine learning in healthcare and invites ongoing inquiry into the dynamic intersection of technology and medicine.

3.2 PROPOSED SYSTEM

Chronic diseases are a pressing global health concern due to their insidious progression and far-reaching impact on individuals' health. The need for early detection and effective treatment is paramount to managing these conditions successfully. This passage highlights the pivotal role of decision models, specifically focusing on machine learning predictive models, in facilitating early diagnosis and predicting future patient outcomes. Chronic diseases encompass a range of conditions, such as cardiovascular diseases, diabetes, cancer, and respiratory disorders. They are characterized by their slow development, often taking years to manifest symptoms. During this silent phase, damage accumulates within the body, posing significant health risks. Early diagnosis is the linchpin in mitigating these risks. Early detection of chronic diseases yields profound benefits for both individual patients and healthcare systems. Detecting these diseases at an incipient stage empowers healthcare practitioners to initiate timely interventions, leading to vastly improved patient outcomes and significantly reduced mortality rates. Individuals diagnosed early have a higher chance of achieving remission or effectively managing their conditions. They can access treatments that are less invasive and costly compared to late-stage interventions.

Moreover, they can make lifestyle changes and adhere to medical regimens that prevent disease progression. All these factors collectively contribute to enhanced patient well-being. Additionally, early diagnosis is a powerful tool in alleviating the financial burden associated with chronic disease treatment. Late-stage disease management often necessitates complex and costly interventions, including surgeries, long-term medications, and intensive care. In contrast, early interventions are more cost-effective and efficient, reducing the economic strain on healthcare systems and patients alike. Machine learning classification models, including Random Forest, AdaBoost, and Gradient Boosting, are emerging as indispensable tools in the realm of healthcare. These models possess the capacity to revolutionize the diagnosis and management of chronic diseases. Random Forest, one of the models under discussion, is an ensemble learning technique that amalgamates multiple decision trees to make precise predictions. In healthcare, it excels at parsing a broad spectrum of patient data, such as medical histories, genetic information, and diagnostic tests. Its unique capability to handle high-dimensional data and discern feature importance renders it an invaluable asset for early diagnosis.

Random Forest, a versatile ensemble learning technique, is a formidable contender in the healthcare domain. Its strength lies in its ability to harness the collective intelligence of multiple decision trees to generate accurate predictions. In the context of healthcare, Random Forest demonstrates exceptional prowess in analysing an extensive array of patient data. It processes diverse information sources, including electronic health records, medical imaging, and genetic sequencing data. This comprehensive approach allows it to identify subtle patterns and correlations that might elude traditional diagnostic methods. Moreover, Random Forest's aptitude for handling high-dimensional data sets it apart. In healthcare, patient profiles often encompass numerous variables, making it challenging to identify relevant factors. Random Forest excels in this regard, offering a systematic approach to determine which features are most influential in making diagnostic predictions. AdaBoost, another ensemble learning method, is a powerful tool in enhancing the accuracy of predictive models for chronic disease diagnosis.

Its sequential combination of weak learners culminates in the creation of a robust classifier with the ability to adapt to evolving scenarios. In healthcare, where patient data is dynamic and multifaceted, AdaBoost's adaptability is a critical asset. It iteratively assigns more weight to misclassified samples, allowing the model to focus on improving its performance in areas where errors have been made. This responsiveness is particularly advantageous in situations where the relevance of different patient features may change over time, such as in the progression of chronic diseases. The dynamic nature of chronic diseases demands a flexible approach to diagnosis. AdaBoost's capacity to refine its understanding of patient data iteratively aligns with this need, resulting in more accurate predictions. Gradient Boosting, a robust machine learning technique, builds an ensemble of decision trees to optimize a given loss function. In healthcare, it has demonstrated its prowess in handling imbalanced datasets, a common occurrence where the prevalence of chronic diseases may be relatively low. Imbalanced datasets can present significant challenges in traditional modelling approaches, as they often lead to biased predictions. Gradient Boosting mitigates this issue by focusing on misclassified cases, iteratively improving its predictions.

This adaptive approach enhances the accuracy of disease diagnosis and risk assessment. Moreover, Gradient Boosting's ability to optimize complex loss functions makes it well-suited for healthcare applications. Chronic diseases often involve multifaceted and interconnected factors. Gradient Boosting can capture these intricate relationships, enabling more precise predictions. The integration of machine learning

classification models into healthcare management holds the promise of revolutionizing hospital activities and yielding several key benefits. Automated systems powered by machine learning models can dramatically reduce human errors in disease diagnosis and risk assessment. These models provide consistency and process extensive patient data with precision, thus minimizing the likelihood of diagnostic mistakes.

Healthcare practitioners can also save valuable time by relying on automated systems for initial patient assessments. This automation liberates them to focus more on delivering personalized patient care and developing tailored treatment plans, ultimately enhancing the overall quality of healthcare services. Early detection of chronic diseases not only improves patient outcomes but also has a substantial economic impact. Identifying diseases at an early stage often results in significantly lower treatment costs compared to addressing advanced-stage conditions. This can lead to substantial cost savings for both healthcare systems and patients, making healthcare more accessible and affordable.

In conclusion, the development and implementation of machine learning classification models, including Random Forest, AdaBoost, and Gradient Boosting, have the potential to transform healthcare management. These models facilitate early diagnosis of chronic diseases, reduce medical errors, save time and resources, enhance patient outcomes, and result in cost savings. Such initiatives hold the promise of not only improving the quality of healthcare but also contributing to the overall well-being of individuals and the sustainability of healthcare systems.

3.3 Dataset Description

The raw dataset consists of 400 instances represented by 12 input features and 1 for the target class. The features' description is the following:

- **Diastolic Blood Pressure** (Bp – mmHg) : This feature shows the participator's diastolic blood pressure.
- **Specific Gravity** (Sg) : This feature captures the participator's specific gravity value.
- **Albumin** (Al) : This attribute captures the participator's albumin level. It has three categories (72.25% normal, 21.5% above normal and 6.25% well above normal).
- **Glucose** (Su) : This attribute denotes the participator's glucose level. It has three categories (88% normal, 8% above normal and 4% well above normal).

- **Red Blood Cell (Rbc):** This attribute captures whether the participant's Red Blood Cell is normal or not. It has two categories (88.25% normal and 11.75% abnormal).
- **Blood Urea (Bu – mmol/L):** This feature captures the amount of urea found in the participant's blood. Blood Urea is measured in millimoles per liter (mmol/L).
- **Serum Creatinine (Sc – mg/dL):** This feature measures the amount of serum creatinine found in the participant's blood. Serum creatinine is reported as milligrams of creatinine to a deciliter of blood (mg/dL).
- **Sodium (Sod – mEq/L) :** This feature measures the amount of sodium found in the participant's blood. Sodium is a type of electrolyte and is reported as milliequivalents per liter (mEq/L).
- **Potassium (Pot – mmol/L):** This feature measures the amount of potassium found in the participant's blood and is reported as millimoles per liter (mmol/L).
- **Hemoglobin (Hemo – gm/dL):** This feature measures the amount of hemoglobin found in the participant's blood and is reported as grams per deciliter (gm/dL).
- **White Blood Cell Count (Wbcc):** This feature measures the number of white cells in the participant's blood and is reported as Wbc per microliter.
- **Red Blood Cell Count (Rbcc):** This feature measures the number of red blood cells in the participant's blood and is reported as a million red blood cells per microliter (mcL) of blood.
- **Hypertension (Htn):** This attribute refers to whether the participant has hypertension or not. A total of 36.75% of participants have hypertension.
- **Chronic Kidney Disease (CKD):** This feature denotes whether the participant suffers from CKD or not. A total of 62.5% of participants have been diagnosed with CKD.

Chronic Kidney Disease datasets from publically available data from UCI Machine Learning Repository . Table 3.1 gives a list of all the attributes taken ,it describes 25 chronic kidney disease related attributes which are taken form UCI repository, it consist of Record of 400 Patients with 25 attributes. The Data Set is real and consists of Nominal, Numerical and Class attributes.

Table 3.1 Attributes for chronic kidney disease prediction

Attribute	Values
Age	Numerical
blood pressure	Numerical
specific gravity	Nominal sg(.005,1.010,1.015,1.020,1.025)
Albumin	Nominal al – (0,1,2,3,4,5)
Sugar	Nominal su – (0,1,2,3,4,5)
red blood cells	Nominal rbc – (normal,abnormal)
pus cell	Nominal pc – (normal,abnormal)
pus cell clumps	Nominal pcc- (present,notpresent)
Bacteria	Nominal ba – (present,notpresent)
Blood glucose	Numerical
blood urea	Numerical
serum creatinine	Numerical
Sodium	Numerical
Potassium	Numerical
Haemoglobin	Numerical
Packed cell volume	Numerical
white blood cell count	Numerical
red blood cell count	Numerical
Hypertension	Nominal htn – (yes,no)
diabetes mellitus	Nominal dm – (yes,no)
coronaryartery disease	Nominal cad – (yes,no)
Appetite	Nominal appet – (good,poor)
pedal edema	Nominal pe – (yes,no)
Anemia	Nominal ane- (yes,no)
Class	Nominal class – (ckd,notckd)

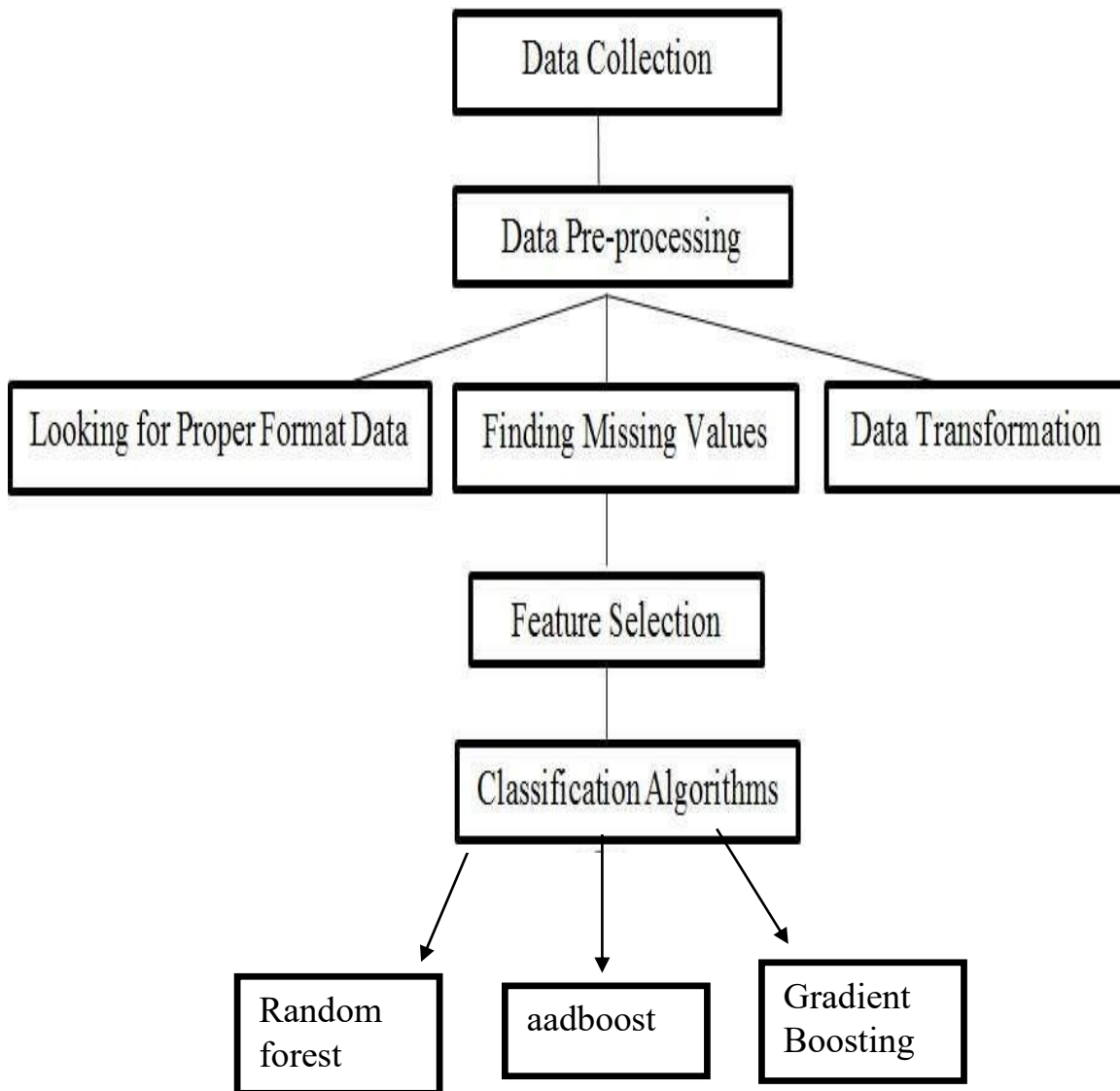


Figure 3.1 Methodology Flow Chart for chronic kidney disease prediction.

3.3.1 Data Collection

In this project used Real world data set for predicting CKD status of a patient. The data collected is widely used data and is available at UCI Machine Learning Repository. The data set available is specifically used for Chronic Kidney Disease research. It consists of record of 400 people with their respective 25 CKD related attributes. The data consisted of real numbers, Decimal values and Nominal values.

3.3.2 Data pre-processing

Data pre-processing is a way to convert the noisy and huge data into relevant and clean data, as the data available is Real world data, so it contains inaccurate data, missing values and other Noisy data, for removing this inconsistent data from the Dataset, the proposed system have to clean the raw data. This is an important part to complete the prediction model. It reduces the dimensionality and helps the machine to achieve better results. This is one of the most time consuming stage in building a classification model.

3.3.3 Looking Up For Proper Format:

As I have made our model using python, so I need a csvfile (comma separated value) for our code. The data downloaded is in the form of RAR file, so we extract the data from the text file available and save it into a csv file so that our python code can read it. This is the first most important step, if the data is not available in requires format then we cannot design the classification model.

3.3.4 Finding Missing Values:

When the data collected is real world data, and then it will contain missing values. This brings more change in the prediction accuracy. Sometimes these missing values can be simply deleted or ignored if they are not large in number. It is the simplest way to handle the missing data but it is not considered healthy for the model as the missing value can be an important attribute contributing to the disease. The missing values can also be replaced by zero this will not bring any change as whole, but this method cannot be much yielding. So an efficient way to handle missing values is to use mean, average of the observed attribute or value. This way we lead to more genuine data and better prediction results.

3.3.5 Data Transformation:

In this step we transform the given real data into required format. The data downloaded consist of Nominal, Real and Decimal values. In this step we convert the Nominal data into numerical data of the form 0 and 1. The positive value is assigned the value of 1 and the negative value is assigned the value of 0. Now the resultant csv file comprises of all the integer and decimal values for different CKD related attributes.

3.3.6 Feature Selection

In this step we select subset of relevant attributes from the total give attributes. This stage helps in reducing the dimensionality and making the model simpler and easy to use, thus leading to short training time and high accuracy.

To obtain highly dependent features for CKD prediction we have used Correlation and dependence method. The term correlation can be defined as mutual relationship between two. In this those attributes are chosen which highly influence the occurrence of chronic kidney disease. By using the correlation, it is found that 5 attributes were highly correlated to the occurrence of CKD from the total of 25 attributes.

The 5 attributes selected from a total of 25 attributes are:

1. Specific Gravity
2. Hypertension
3. Haemoglobin
4. Diabetes Mellitus
5. Albumin
6. Appetite
7. Red Blood Cell Count
8. Pus Cell

3.3.7 Classification Algorithms

Random Forest:

- Random Forest is an ensemble learning method that combines the predictions of multiple decision trees.
- It uses a technique called bagging (Bootstrap Aggregating) to create a diverse set of decision trees. Bagging involves training each tree on a random subset of the training data with replacement.
- Randomness is introduced by selecting a random subset of features for each tree at each split in the decision tree-building process.

- The final prediction in a Random Forest is typically obtained by averaging the predictions of all individual trees (for regression) or by using a majority vote (for classification).
- Random Forests are known for their ability to handle high-dimensional data, reduce overfitting, and provide feature importance scores.

AdaBoost (Adaptive Boosting):

- AdaBoost is an ensemble method that focuses on improving the performance of weak learners (models that perform slightly better than random guessing).
- It works by sequentially training a series of weak classifiers and giving more weight to the samples that were misclassified by the previous classifiers.
- In each iteration, AdaBoost assigns higher weights to the misclassified samples, so that the next weak classifier focuses more on getting those samples right.
- The final prediction is made by combining the weighted predictions of all the weak classifiers.
- AdaBoost is particularly effective when used with simple models as weak learners (e.g., shallow decision trees or linear models).

Gradient Boosting:

- Gradient Boosting is an ensemble technique that builds an additive model of decision trees sequentially.
- It aims to improve the model by minimizing the errors (residuals) made by the previous trees.
- In each iteration, a new decision tree is trained to fit the residuals of the previous predictions, which effectively corrects the mistakes made by earlier trees.
- The learning rate (or shrinkage) parameter controls the contribution of each new tree to the final prediction.
- Gradient Boosting is powerful and can be used with a variety of base models, making it highly flexible and capable of capturing complex patterns in the data.
- Variants like XGBoost, LightGBM, and CatBoost have improved upon traditional Gradient Boosting with optimizations and enhancements.

3.4 ARCHITECHURE FOR PROPOSED WORK

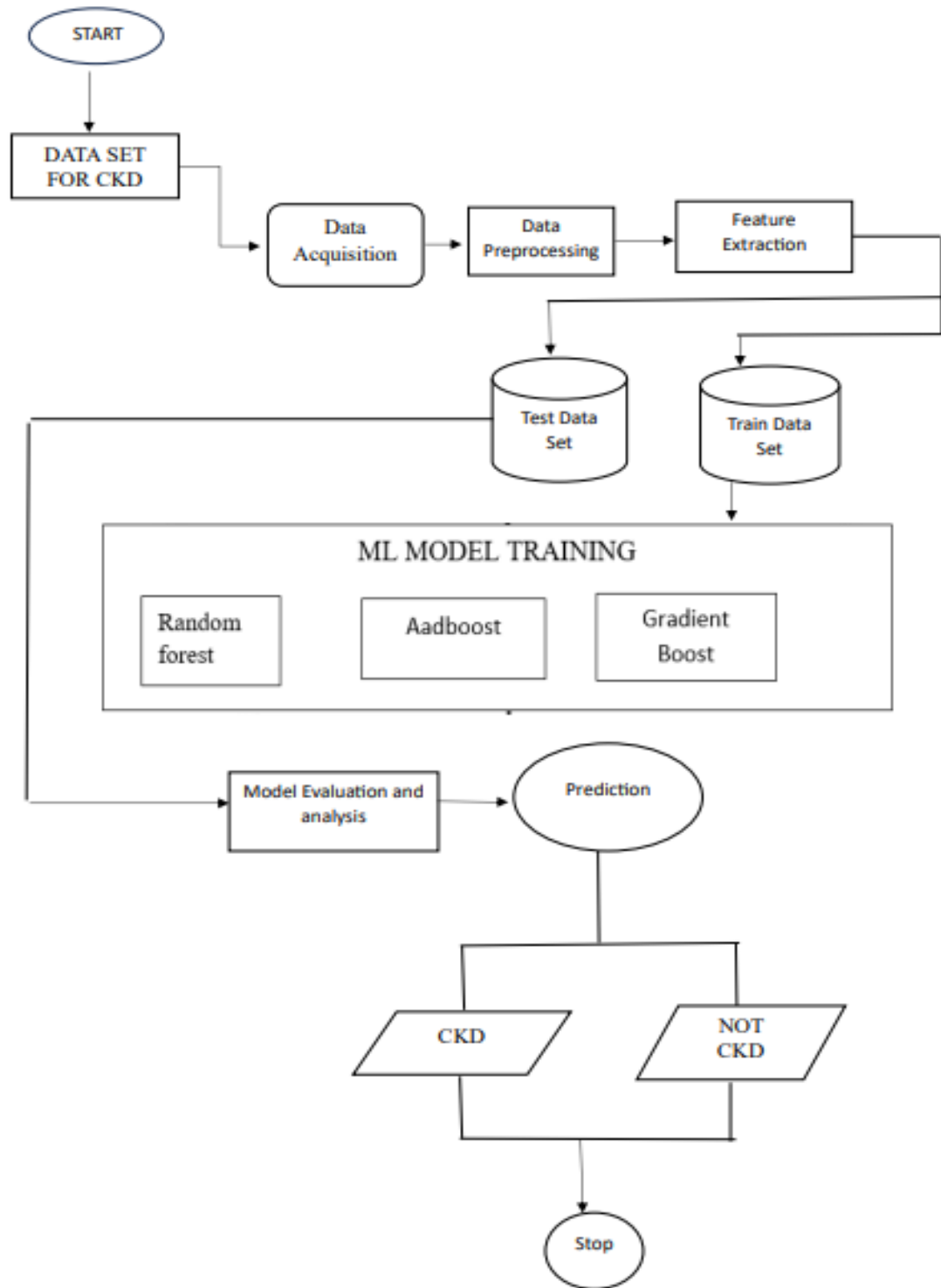


Figure 3.2 Architecture for CKD Prediction

CHAPTER-4

4.RESULT

Chronic Kidney Disease (CKD) is a prevalent and progressive health condition that poses a significant global health challenge. Early detection of CKD is of paramount importance, as it allows for timely interventions when treatments are most effective. Machine learning, a subfield of artificial intelligence, has emerged as a powerful tool in healthcare, holding the potential to revolutionize the early detection and management of CKD. This essay explores the multifaceted role of machine learning in CKD detection, emphasizing its ability to improve early diagnosis, identify genetic factors, enhance diagnostic accuracy, streamline healthcare management, and ultimately improve patient outcomes. One of the primary benefits of employing machine learning models in CKD detection is their capacity to identify the disease, even in its initial stages. Timely diagnosis is critical because early interventions are most effective in slowing the progression of CKD. Machine learning algorithms, such as Random Forest, AdaBoost, and Gradient Boosting, are capable of analyzing extensive datasets containing patient information, clinical records, and diagnostic tests. This analytical power enables quicker and more precise CKD diagnoses, reducing the risk of misdiagnosis and ensuring that individuals receive the appropriate care promptly.

Machine learning goes beyond early detection by delving into genetic factors that predispose individuals to CKD. These models can analyze genetic data to pinpoint specific markers associated with CKD risk. This valuable insight can inform personalized treatment plans tailored to each patient's unique genetic profile. Tailoring interventions to address a patient's specific genetic risk factors can significantly enhance treatment effectiveness and improve overall patient outcomes. The ability to provide such personalized care represents a significant advancement in CKD management. Another key role of machine learning in CKD detection is its ability to improve the sensitivity and specificity of diagnostic tests. By processing and analyzing vast amounts of medical data, machine learning models can reduce the risk of false positives and negatives, leading to more reliable diagnostic results. This heightened accuracy ensures that healthcare providers can make better-informed decisions regarding patient care, optimizing treatment approaches and resource allocation.

The increased diagnostic precision provided by machine learning holds the potential to save lives and reduce the burden of CKD on healthcare systems. In addition to improving

CKD detection and diagnostic precision, machine learning has the potential to transform healthcare management. The development of practical and automated healthcare management systems equipped with user-friendly interfaces for healthcare providers can significantly increase operational efficiency. These systems can assist healthcare professionals in patient data management, appointment scheduling, and treatment planning. By reducing the administrative burden on healthcare providers, machine learning-powered systems allow them to focus more on patient care, ultimately improving the quality of care and the overall patient experience. The economic impact of early CKD detection facilitated by machine learning is substantial. Early disease detection not only saves lives but also reduces the overall cost of disease treatment. Timely interventions can prevent CKD from progressing to advanced stages, which often require more expensive treatments such as dialysis or kidney transplantation. Early detection and prompt medical interventions, facilitated by predictive machine learning models, can significantly improve patient outcomes and quality of life.

Patients can receive treatment before the disease becomes severe, leading to better health outcomes and reduced healthcare costs. In conclusion, the integration of machine learning into CKD detection and management represents a groundbreaking advancement in healthcare. These technologies empower healthcare professionals to detect CKD in its early stages, identify genetic predispositions, enhance diagnostic accuracy, streamline healthcare operations, and ultimately improve the lives of CKD patients while reducing the burden on healthcare systems. Embracing machine learning in the realm of CKD is a significant step toward a future where early disease detection and personalized treatment plans become the standard, ensuring better health outcomes for individuals and more efficient healthcare delivery worldwide.

Table:4.1 Accuracy for both Existing and Proposed System

EXISTING SYSTEM		PROPOSED SYSTEM	
KNN	90%	RANDOM FOREST	98%
LOGISTIC	78%	AADBOOST	97%
NAVIE BAYES	75%	GRADIENTBOOST	95%

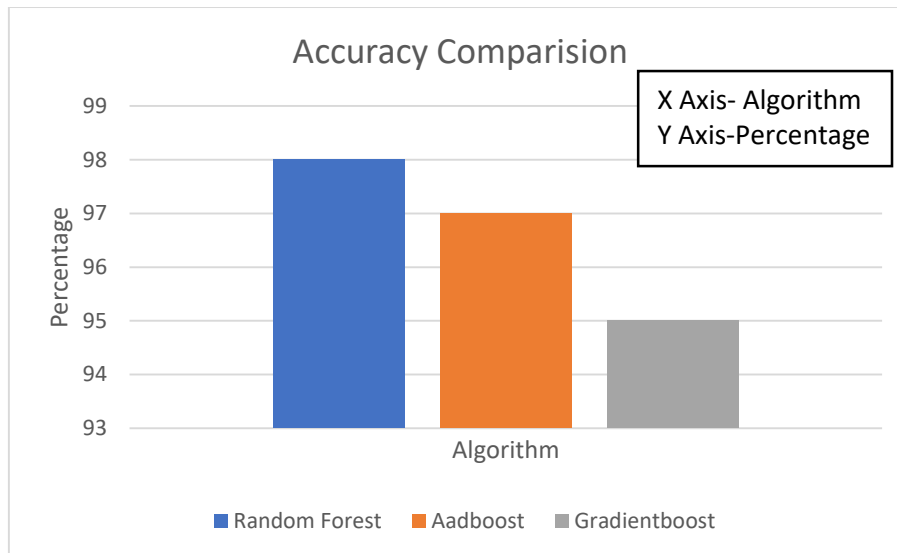


Figure 4.1 Comparison of Random Forest, Aadboost Algorithm and Gradient Boosting Algorithm

Random Forest:

Algorithm Overview: Random Forest is an ensemble learning technique that builds multiple decision trees during training and combines their predictions to make more accurate predictions. It introduces randomness by using bootstrapping (sampling with replacement) and feature selection, which helps reduce overfitting and improve generalization.

Accuracy: In graph, Random Forest achieved an accuracy of 98%, which indicates that it performed very well on the dataset used.

AdaBoost (Adaptive Boosting):

Algorithm Overview: AdaBoost is another ensemble learning technique that focuses on the weaknesses of individual models. It assigns different weights to data points, with more weight given to the misclassified samples. It iteratively trains weak learners (typically decision trees) and combines their predictions, giving more weight to the accurate learners.

Accuracy: In graph, AdaBoost achieved an accuracy of 97%, which is also a strong performance on the dataset.

Gradient Boosting:

Algorithm Overview: Gradient Boosting is an ensemble technique that builds decision trees sequentially, where each new tree is trained to correct the errors made by the previous ones. It uses gradient descent to minimize a loss function and improve the model's accuracy.

Accuracy: In graph, Gradient Boosting achieved an accuracy of 95%, which is slightly lower than Random Forest and AdaBoost in this particular case.

CHAPTER-5

5.1 CONCLUSION

The integration of machine learning models into the diagnosis and management of chronic diseases, such as Chronic Kidney Disease (CKD), holds immense promise for healthcare transformation. These models offer a multifaceted set of advantages that can revolutionize the way we approach healthcare: Firstly, by enabling early disease diagnosis, these models empower healthcare providers to identify CKD in its earliest stages, allowing for timely and highly effective interventions. This early detection can be a pivotal factor in improving patient outcomes and even saving lives. Secondly, the enhanced accuracy and speed of CKD detection through machine learning algorithms like Random Forest, AdaBoost, and Gradient Boosting reduce the risk of misdiagnosis, ensuring that patients receive the right treatment promptly. Furthermore, the ability to identify genetic factors contributing to CKD risk offers the potential for truly personalized treatment plans. This tailoring of interventions based on individual genetic profiles can lead to significantly improved treatment effectiveness and patient satisfaction. Machine learning models also play a vital role in reducing false positives and negatives in diagnostic tests, thereby providing more reliable results and guiding better-informed treatment decisions. The development of automated healthcare management systems, equipped with user-friendly interfaces for healthcare providers, not only increases operational efficiency but also allows healthcare professionals to focus more on delivering quality patient care. This represents a significant step towards streamlined and patient-centric healthcare services. Moreover, early disease detection not only benefits patients by improving health outcomes but also contributes to substantial cost savings in the long run. Preventing disease progression to advanced stages can substantially reduce the financial burden on both patients and healthcare systems. Ultimately, the integration of machine learning models into the healthcare landscape facilitates early treatment, reduces medical errors, and, most importantly, enhances patient outcomes. It marks a transformative leap toward a more efficient, personalized, and patient-centered approach to managing chronic diseases like CKD, promising a brighter future for healthcare worldwide.

5.2 FUTURE ENHANCEMENT

Predicting Chronic Kidney Disease (CKD) using machine learning involves the integration of real-time physiological data from wearable devices and continuous monitoring. By combining wearable technology with advanced machine learning algorithms, healthcare providers can create a dynamic CKD prediction system that offers several advantages. This approach involves equipping patients at risk of CKD with wearable devices that continuously monitor relevant physiological parameters, such as blood pressure, heart rate, glucose levels, and activity levels. These devices can also incorporate non-invasive sensors to measure biomarkers in urine or sweat, allowing for early detection of kidney dysfunction. Machine learning algorithms can then analyse this real-time data, identifying subtle changes in physiological patterns and trends that may indicate the onset or progression of CKD. By detecting anomalies and deviations from baseline health parameters, the system can issue alerts to both patients and healthcare providers in real-time. This proactive and personalized monitoring system offers several benefits. First, it enables early detection of CKD, often before traditional diagnostic tests would flag the condition. Second, it empowers patients to actively participate in their healthcare by providing them with real-time feedback and alerts regarding their kidney health. Third, it allows for personalized treatment plans based on individual data trends, genetics, and response to interventions, ultimately improving patient outcomes. Additionally, the continuous data collection and analysis can provide valuable insights for research and population-level CKD management, contributing to a better understanding of the disease's progression and risk factors.

REFERENCES

- [1] A. S. Levey, R. Atkins, J. Coresh et al., “Chronic kidney disease as a global public health problem: approaches and initiatives a position statement from kidney disease improving global outcomes,” *Kidney International*, vol. 72, no. 3, pp. 247–259, 2007.
- [2] V. Jha, G. Garcia-Garcia, K. Iseki et al., “Chronic kidney disease: global dimension and perspectives,” *The Lancet*, vol. 382, no. 9888, pp. 260–272, 2013.
- [3] N. R. Hill, S. T. Fatoba, J. L. Oke et al., “Global prevalence of chronic kidney disease a systematic review and meta-analysis,” *PLoS One*, vol. 11, no. 7, article e0158765, 2016.
- [4] H. Nasri, “World kidney day 2014; chronic kidney disease and aging: a global health alert,” *Iranian Journal of Public Health*, vol. 43, no. 1, pp. 126–127, 2014.
- [5] G. Abraham, S. Varughese, T. Thandavan et al., “Chronic kidney disease hotspots in developing countries in South Asia,” *Clinical Kidney Journal*, vol. 9, no. 1, pp. 135–141, 2016.
- [6] K. T. Mills, T. Xu, W. Zhang et al., “A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010,” *Kidney International*, vol. 88, no. 5, pp. 950–957, 2015.
- [7] B. Ene-Iordache, N. Perico, B. Bikbov et al., “Chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): a cross-sectional study,” *The Lancet Global Health*, vol. 4, no. 5, pp. e307–e319, 2016.
- [8] S. Anand, M. A. Khanam, J. Saquib et al., “High prevalence of chronic kidney disease in a community survey of urban Bangladeshis: a cross-sectional study,” *Glob Health*, vol. 10, no. 1, p. 9, 2014.
- [9] L. Ali, K. Fatema, Z. Abedin et al., “Screening for chronic kidney diseases among an adult population,” *Saudi Journal of Kidney Diseases and Transplantation*, vol. 24, no. 3, p. 534, 2013.
- [10] M. J. Hasan, M. A. Kashem, M. H. Rahman et al., “Prevalence of chronic kidney disease (CKD) and identification of associated risk factors among rural population by mass screening,” *Community Based Medical Journal*, vol. 1, pp. 20–26, 2013.
- [11] M. J. Lysaght, “Maintenance dialysis population dynamics current trends and long-term implications,” *Journal American Society Nephrology*, vol. 13, suppl 1, pp. S37–S40, 2002.

- [12] M. Bakhshayeshkaram, J. Roozbeh, S. T. Heydari et al., “A population-based study on the prevalence and risk factors of chronic kidney disease in adult population of shiraz, southern Iran,” *Galen Medical Journal*, vol. 8, no. 935, p. 935, 2019.
- [13] K. U. Eckardt, J. Coresh, O. Devuyst et al., “Evolving importance of kidney disease: from subspecialty to global health burden,” *The Lancet*, vol. 382, no. 9887, pp. 158–169, 2013.
- [14] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities,” *IEEE Access*, vol. 5, pp. 8869–8879, 2017.
- [15] V. Mohan, “Decision trees: a comparison of various algorithms for building decision trees,” 2013, http://cs.jhu.edu/~vmohan3/document/ai_dt.pdf.
- [16] V. Garcia, E. Debreuve, and M. Barlaud, “Fast k nearest neighbor search using GPU,” in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6, Anchorage, AK, USA, 2008.
- [17] V. V. Ramalingam, A. Dandapath, and M. K. Raja, “Heart disease prediction using machine learning techniques : a survey,” *International Journal of Engineering and Technology*, vol. 7, no. 2.8, pp. 684–687, 2018.
- [18] I. Ahmad, M. Basher, M. J. Iqbal, and A. Rahim, “Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection,” *IEEE Access*, vol. 6, pp. 33789–33795, 2018.
- [19] A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan et al., “Early prediction of chronic kidney disease using machine learning supported by predictive analytics,” in 2018 IEEE Congress on Evolutionary Computation (CEC), pp. 1–9, Rio de Janeiro, Brazil, 2018.
- [20] C. A. Johnson, A. S. Levey, J. Coresh, A. Levin, and J. G. L. Eknoyan, “Clinical practice guidelines for chronic kidney disease in adults: part I. Definition, disease stages, evaluation, treatment, and risk factors,” *American Family Physician*, vol. 70, no. 5, pp. 869–876, 2004.
- [21] C. Li, “Little’s test of missing completely at random,” *The Stata Journal*, vol. 13, no. 4, pp. 795–809, 2013.

- [22] E. H. A. Rady and A. S. Anwar, “Prediction of kidney disease stages using data mining algorithms,” *Informatics in Medicine Unlocked*, vol. 15, article 100178, 2019.
- [23] S. Nair, S. V. O’Brien, K. Hayden et al., “Effect of a cookedmeat meal on serum creatinine and estimated glomerular filtration rate in diabetes-related kidney disease,” *Diabetes Care*, vol. 37, no. 2, pp. 483–487, 2014.
- [24] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, “Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (ckd),” in *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 291–296, Washington, DC, USA, 2017.
- [25] P. Yildirim, “Chronic kidney disease prediction on imbalanced data by multilayer perceptron: chronic kidney disease prediction,” in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, pp. 193–198, Turin, Italy, 2017.
- [26] Kaggle, “Chronic Kidney Disease Dataset,” <https://www.kaggle.com/abhia1999/chronic-kidney-disease>.
- [27] S. Krishnamurthy, K. KS, E. Dovgan et al., “Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan,” *Healthcare*, vol. 9, no. 5, p. 546, 2021.

A1 APPENDIX - SAMPLE CODE:

Importing Libraries:

```
import pandas as pd
```

```
import numpy as np
```

```
import pickle
```

for displaying all feature from dataset:

```
pd.pandas.set_option('display.max_columns', None)
```

Reading Dataset:

```
dataset = pd.read_csv("Kidney_data.csv")
```

Dropping unnecessary feature :

```
dataset = dataset.drop('id', axis=1)
```

Replacing Categorical Values with Numericals

```
dataset['rbc'] = dataset['rbc'].replace(to_replace = {'normal' : 0, 'abnormal' : 1})
```

```
dataset['pc'] = dataset['pc'].replace(to_replace = {'normal' : 0, 'abnormal' : 1})
```

```
dataset['pcc'] = dataset['pcc'].replace(to_replace = {'notpresent':0,'present':1})
```

```
dataset['ba'] = dataset['ba'].replace(to_replace = {'notpresent':0,'present':1})
```

```
dataset['htn'] = dataset['htn'].replace(to_replace = {'yes' : 1, 'no' : 0})
```

```
dataset['dm'] = dataset['dm'].replace(to_replace = {'\tyes': 'yes', ' yes': 'yes', '\tno': 'no'})
```

```
dataset['dm'] = dataset['dm'].replace(to_replace = {'yes' : 1, 'no' : 0})
```



```

dataset['cad'] = dataset['cad'].replace(to_replace = {'\tno':'no'})

dataset['cad'] = dataset['cad'].replace(to_replace = {'yes' : 1, 'no' : 0})


dataset['appet'] = dataset['appet'].replace(to_replace={'good':1,'poor':0,'no':np.nan})

dataset['pe'] = dataset['pe'].replace(to_replace = {'yes' : 1, 'no' : 0})

dataset['ane'] = dataset['ane'].replace(to_replace = {'yes' : 1, 'no' : 0})


dataset['classification'] = dataset['classification'].replace(to_replace={'ckd\t':'ckd'})

dataset["classification"] = [1 if i == "ckd" else 0 for i in dataset["classification"]]


# Coverting Objective into Numericals:

dataset['pcv'] = pd.to_numeric(dataset['pcv'], errors='coerce')

dataset['wc'] = pd.to_numeric(dataset['wc'], errors='coerce')

dataset['rc'] = pd.to_numeric(dataset['rc'], errors='coerce')


# Handling Missing Values:

features = ['age', 'bp', 'sg', 'al', 'su', 'rbc', 'pc', 'pcc', 'ba', 'bgr', 'bu',
            'sc', 'sod', 'pot', 'hemo', 'pcv', 'wc', 'rc', 'htn', 'dm', 'cad',
            'appet', 'pe', 'ane']

for feature in features:

    dataset[feature] = dataset[feature].fillna(dataset[feature].median())


# Dropping feature (Multicollinearity):

```

```
dataset.drop('pcv', axis=1, inplace=True)
```

```
# Independent and Dependent Feature:
```

```
X = dataset.iloc[:, :-1]
```

```
y = dataset.iloc[:, -1]
```

```
# After feature importance:
```

```
X = dataset[['sg', 'htn', 'hemo', 'dm', 'al', 'appet', 'rc', 'pc']]
```

```
# Train Test Split:
```

```
from sklearn.model_selection import train_test_split
```

```
X_train,X_test,y_train,y_test = train_test_split(X,y, test_size=0.3, random_state=33)
```

```
# RandomForestClassifier:
```

```
from sklearn.ensemble import RandomForestClassifier
```

```
RandomForest = RandomForestClassifier()
```

```
RandomForest = RandomForest.fit(X_train,y_train)
```

```
# Creating a pickle file for the classifier
```

```
filename = 'Kidney.pkl'
```

```
pickle.dump(RandomForest, open(filename, 'wb'))
```

```
from flask import Flask, render_template, request, flash, redirect
```

```
import pickle
```

```
import numpy as np
```

```

from PIL import Image

from tensorflow.keras.models import load_model

app = Flask(__name__)

def predict(values, dic):
    if len(values) == 8:
        model = pickle.load(open('models/diabetes.pkl','rb'))
        values = np.asarray(values)
        return model.predict(values.reshape(1, -1))[0]
    elif len(values) == 26:
        model = pickle.load(open('models/breast_cancer.pkl','rb'))
        values = np.asarray(values)
        return model.predict(values.reshape(1, -1))[0]
    elif len(values) == 13:
        model = pickle.load(open('models/heart.pkl','rb'))
        values = np.asarray(values)
        return model.predict(values.reshape(1, -1))[0]
    elif len(values) == 18:
        model = pickle.load(open('models/kidney.pkl','rb'))
        values = np.asarray(values)
        return model.predict(values.reshape(1, -1))[0]
    elif len(values) == 10:
        model = pickle.load(open('models/liver.pkl','rb'))
        values = np.asarray(values)
        return model.predict(values.reshape(1, -1))[0]

```

```

@app.route("/")

def home():

    return render_template('home.html')

@app.route("/diabetes", methods=['GET', 'POST'])

def diabetesPage():

    return render_template('diabetes.html')

@app.route("/cancer", methods=['GET', 'POST'])

def cancerPage():

    return render_template('breast_cancer.html')

@app.route("/heart", methods=['GET', 'POST'])

def heartPage():

    return render_template('heart.html')

@app.route("/kidney", methods=['GET', 'POST'])

def kidneyPage():

    return render_template('kidney.html')

@app.route("/liver", methods=['GET', 'POST'])

def liverPage():

    return render_template('liver.html')

@app.route("/malaria", methods=['GET', 'POST'])

def malariaPage():

    return render_template('malaria.html')

@app.route("/pneumonia", methods=['GET', 'POST'])

def pneumoniaPage():

    return render_template('pneumonia.html')

```

```

@app.route("/predict", methods = ['POST', 'GET'])

def predictPage():

    try:

        if request.method == 'POST':

            to_predict_dict = request.form.to_dict()

            to_predict_list = list(map(float, list(to_predict_dict.values())))

            pred = predict(to_predict_list, to_predict_dict)

        except:

            message = "Please enter valid Data"

            return render_template("home.html", message = message)

        return render_template('predict.html', pred = pred)

@app.route("/malariapredict", methods = ['POST', 'GET'])

def malariapredictPage():

    if request.method == 'POST':

        try:

            if 'image' in request.files:

                img = Image.open(request.files['image'])

                img = img.resize((36,36))

                img = np.asarray(img)

                img = img.reshape((1,36,36,3))

                img = img.astype(np.float64)

                model = load_model("models/malaria.h5")

                pred = np.argmax(model.predict(img)[0])

        except:

```

```

        message = "Please upload an Image"

        return render_template('malaria.html', message = message)

    return render_template('malaria_predict.html', pred = pred)

@app.route("/pneumoniapredict", methods = ['POST', 'GET'])
def pneumoniapredictPage():

    if request.method == 'POST':

        try:

            if 'image' in request.files:

                img = Image.open(request.files['image']).convert('L')

                img = img.resize((36,36))

                img = np.asarray(img)

                img = img.reshape((1,36,36,1))

                img = img / 255.0

                model = load_model("models/pneumonia.h5")

                pred = np.argmax(model.predict(img)[0])

            except:

                message = "Please upload an Image"

                return render_template('pneumonia.html', message = message)

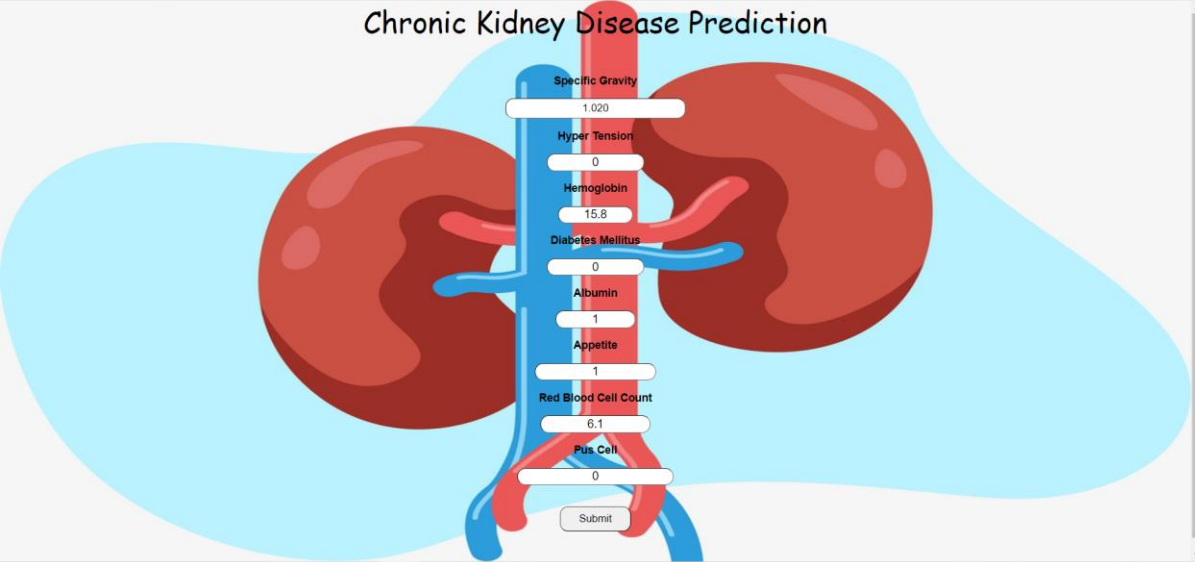
        return render_template('pneumonia_predict.html', pred = pred)

if __name__ == '__main__': app.run(debug = True)

```

A2 APPENDIX - SAMPLE SCREENSHOTS

Chronic Kidney Disease Prediction



Specific Gravity
1.020

Hyper Tension
0

Hemoglobin
15.8

Diabetes Mellitus
0

Albumin
1

Appetite
1

Red Blood Cell Count
6.1

Pus Cell
0

Submit



