

# **CHRONIC KIDNEY DISEASE PATIENT CARE APPLICATION**

D.R.N. Samarawila

IT20235260

B.Sc. (Hons) Degree in Information Technology  
Specializing in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology  
Sri Lanka

September 2023

# **CHRONIC KIDNEY DISEASE PATIENT CARE APPLICATION**

D.R.N. Samarawila

IT20235260

Dissertation submitted in partial fulfillment of the requirements for the Bachelor of  
Science (Hons) in Information Technology Specializing in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology  
Sri Lanka


September 2023

## DECLARATION

### Declaration of the Candidate

I declare that this is my own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name	Student ID	Signature
D.R.N. Samarawila	IT20235260	

### Declaration of the Supervisor

The above candidates are carrying out research for the undergraduate Dissertation under \_\_\_\_\_ my \_\_\_\_\_ supervision.

Name of the supervisor: Ms. Wishalya Tissera

.....  
Signature of the supervisor

.....  
Date

Name of the co-supervisor: Mr. Samadhi Rathnayake

.....  
Signature of the co-supervisor

.....  
Date

## ABSTRACT

Kidney disease is a significant health issue in Sri Lanka, affecting most of the population. Early detection and treatment of the disease are essential to prevent complications and improve patient outcomes. Machine learning (ML) models have shown great potential in predicting kidney disease based on patient data and facilitating early intervention. However, there needs to be more research on the effectiveness of patient-specific ML-based models in the Sri Lankan context.

This undergraduate research project aims to investigate the effectiveness of patient-specific ML-based kidney disease prediction models in the Sri Lankan population and identify the factors that may impact their performance. The study will develop and evaluate ML models tailored to Sri Lankan patients' characteristics and clinical data to predict the risk of kidney disease. Additionally, the project will develop a treatment recommendation system based on the predicted risk and patient-specific data to provide personalized treatment options.

The research will contribute to the growing body of knowledge on the use of ML in healthcare and specifically on patient-specific kidney disease prediction and treatment in Sri Lanka. The findings of this project may help improve the accuracy of diagnosis and treatment of kidney disease in Sri Lanka, ultimately reducing the burden of the disease on the healthcare system and improving patient outcomes.

The project will utilize existing datasets and conduct surveys and interviews with healthcare providers and patients to gather insights into the factors that may influence the performance of these models. The study results will be presented in a research report, providing a valuable learning opportunity for the undergraduate researcher. The findings also inform the development of evidence-based policies and practices to combat kidney disease in Sri Lanka and serve as a foundation for further research.

*Keywords: kidney disease, machine learning, prediction models, patient-specific, Sri Lanka, early detection, treatment, healthcare providers, patients, diagnosis, burden, treatment recommendation system.*

## **ACKNOWLEDGEMENT**

We would like to express our sincere gratitude to the Department of Information Technology and SLIIT University for granting permission and necessary support. For the completion of this project. We would like to express our gratitude to our supervisor Ms. Wishalya Tissera and our co-supervisor Mr. Samadhi Rathnayake for proper guidance from the beginning to the end of the project and also for giving us their valuable time every working day, providing us with new and innovative ideas to continue our project.

## TABLE OF CONTENTS

DECLARATION .....	3
Declaration of the Candidate .....	3
Declaration of the Supervisor .....	3
ABSTRACT.....	4
ACKNOWLEDGEMENT .....	5
TABLE OF CONTENTS.....	6
LIST OF FIGURES .....	7
LIST OF TABLES .....	8
LIST OF ABBREVIATION .....	8
1    INTRODUCTION.....	9
1.1    Background and Literature Survey .....	9
1.2    Research Gap.....	13
1.3    Research Problem.....	15
1.4    Research Objectives .....	17
2    METHODOLOGY .....	20
2.1    System Architecture .....	20
2.2    Commercialization aspects of the Product .....	25
2.3    Testing & Implementation .....	27
3    RESULTS & DISCUSSION .....	51
3.1    Results .....	51
3.2    Research Findings .....	53
3.3    Discussion .....	55
4    CONCLUSION .....	60
5    REFERENCES .....	62
6    GLOSSARY .....	64
7    APPENDICES .....	65

## LIST OF FIGURES

Figure 1: Overall System Diagram .....	20
Figure 2 : Component overview Diagram (Methodology) .....	21
Figure 3: Import necessary libraries.....	27
Figure 4: Load the dataset.....	27
Figure 5: Data Preprocessing .....	29
Figure 6: Random Forest Classifier .....	30
Figure 7: Support Vector Classifier .....	31
Figure 8: Logistic Regression .....	32
Figure 9: Naive Bayes.....	33
Figure 10: Model Serialization .....	34
Figure 11: Prediction and Testing Results .....	35
Figure 12: Insert libraries & load the CKD Stage data set.....	37
Figure 13: Data Preprocessing, Visualization & Feature engineering.....	38
Figure 14: Decision tree Classifier.....	41
Figure 15: Random Forest Classifier (Hyperparameter).....	43
Figure 16: Support Vector Classifier .....	45
Figure 17: Model Selection.....	46
Figure 18: Screenshot of Prediction testing and results .....	47
Figure 19: User Interface screenshots .....	49
Figure 20: User interface screenshots .....	50

## LIST OF TABLES

Table 1:Trained Model with Model Accuracy.....	34
Table 2: Models with Accuracy .....	46
Table 3:Performance metrics for CKD risk prediction model.....	51
Table 4: Performance metrics for CKD staging model .....	52
Table 5:Comparative accuracy of ML algorithms .....	54

## LIST OF ABBREVIATION

CKD - chronic kidney disease

GFR - Glomerular Filtration Rate

SLIIT - Sri Lanka Institute of Information Technology

UPGR - Urine Protein and Creatinine Ratio

PRO - Proteinuria

RBC - Red Blood Cells

GLU - Glucose Fasting

TG - Triglycerides

T-CHO - Total Cholesterol

AUC - Area Under the Curve

IT - Information Technology

MRI - Magnetic Resonance Imaging

CT - Computed Tomography

BMI - Body Mass Index

AI - Artificial Intelligence

EMR - Electronic Medical Records

ICU - Intensive Care Unit

SVM - Support Vector Machine

AWS - Amazon Web Services



# 1 INTRODUCTION

## 1.1 Background and Literature Survey

With an approximate prevalence of around 16%, kidney disease is still a serious public health problem in Sri Lanka and affects many of the population [1]. Early detection and proper care of kidney disease are essential to stop future deterioration and maintain general health. In Sri Lanka, medical professionals have historically used patient interviews to determine the possibility of kidney disease by evaluating symptoms and risk factors. However, the exploration of cutting-edge technology is gaining popularity to improve the accuracy of kidney disease prediction. In predicting kidney disease via this conventional method, medical professionals embark upon a meticulous journey through the patient's medical history [1]. This journey necessitates inquiries into the specific symptoms that the patient may have encountered, including but not limited to leg swelling, persistent fatigue, noticeable fluctuations in urine output, or struggles with urinary functions. Furthermore, healthcare providers delve into the patient's risk factors, encompassing their medical history and medication regimens. Lifestyle habits, such as alcohol or tobacco consumption, come under scrutiny due to their potential to significantly elevate the risk of kidney disease [2].

After the comprehensive gathering of patient information, healthcare practitioners endeavour to assess whether the patient is presently afflicted by kidney disease or at risk of developing it. If symptoms are evident or risk factors are substantial, the healthcare provider may elect to proceed with supplementary diagnostic evaluations, including comprehensive blood and urine tests, to verify the diagnosis. In recent years, researchers have eagerly explored advanced technologies to augment the accuracy of kidney disease prediction. Nevertheless, despite considerable steps in utilizing machine learning algorithms for identifying potential kidney disease cases, specific barriers remain. Foremost among these challenges is the matter of accuracy. While machine learning algorithms excel at discerning patterns within extensive datasets, their reliability in forecasting a patient's susceptibility to kidney disease needs to improve occasionally. This research project aims to create a smartphone application that uses ML algorithms to precisely forecast kidney disease in the Sri Lankan population in answer to this problem. This application might revolutionize kidney

disease identification and treatment by utilizing technology to give a readily accessible tool for individuals and medical professionals [2].

This study aims to overcome the shortcomings of current prediction techniques and increase the precision of kidney disease diagnosis. Patients may estimate their risk of kidney disease by using a simple and user-friendly smartphone application. The program may be used by healthcare practitioners as a helpful decision-support tool, enabling early detection and individualized treatment plans. This study aims to improve the mobile application's predictive skills, assuring trustworthy and accurate kidney disease forecasts by deploying cutting-edge ML algorithms and integrating patient-specific data. By increasing the rates of early identification, enabling prompt treatments, and eventually improving patient outcomes, the findings of this study have the potential to have a considerable influence on Sri Lankan healthcare practices. This project aims to change kidney disease prediction in Sri Lanka by embracing technology and integrating machine learning algorithms. The methodology, data collecting, analysis, and assessment methods used to create and evaluate the mobile application are described in depth in the following sections. The findings of this study could significantly affect how kidney disease is managed, advancing medical procedures in Sri Lanka [1].

Elias Dritsas and Maria Trigka, well-known researchers affiliated with the University of Patras in Greece, undertook a thorough study with the singular focus of predicting chronic kidney disease (CKD) through the adept deployment of machine learning (ML) techniques. CKD, characterized by gradually weakening kidney function, can result in end-stage renal disease and life-threatening complications unless diagnosed and treated accurately [3]. The researchers cleverly implemented a class-balancing strategy to rectify the uneven distribution of examples within the two classes. Thoroughly, they conducted feature ranking and analysis, followed by the training and evaluation of several ML models, evaluating them based on varied performance metrics. The findings from this study underscored the remarkable superiority of the Rotation Forest (RotF) model, which notched an astonishing Area Under the Curve (AUC) of 100%, coupled with admirable high levels of precision, recall, F-measure, and accuracy, registering at an impressive rate of 99.2% [3].

Chin-Chuan Shih and their dedicated team arranged a pioneering research endeavour in Taiwan in the year 2020, with the explicit objective of developing efficient risk prediction models for complications and death rates associated with chronic kidney disease (CKD). In their quest to predict early CKD, the study found the potency of four data mining algorithms: a classification and regression tree, a C 4.5 decision tree, linear discriminant analysis, and an extreme learning machine. The study accurately collected data from 2015 to 2019, sourced from an adult health examination program. This vast repository of information encompassed records from a staggering 19,270 patients spanning 32 chain clinics and three specialized physical examination centres. The critical predictive variable in the study was the glomerular filtration rate (GFR), with 11 independent variables under careful consideration [4]. Among the four models put to the test, the C4.5 decision tree algorithm emerged as the undisputed champion, boasting superior accuracy, sensitivity, specificity, and area under the curve metrics. The study proficiently identified several significant risk factors for early CKD, including Urine protein and creatinine ratio (UPCR), Proteinuria (PRO), Red blood cells (RBC), Glucose Fasting (GLU), Triglycerides (TG), Total Cholesterol (T-CHO), age, and gender. The proposed risk prediction models offer invaluable insights into the early detection and management of CKD, bridging the gap between personality and health examination representations [4].

In the year 2023, Piyawat Kantagowit, Fangyue Chen, Tanawin Nopsopon, Arisa Chuklin, and Krit Pongpirul laid the groundwork for a systematic review protocol, earnestly aimed at joining ML for CKD diagnosis. As CKD is a significant contributor to global disease and death, initiating ML-based decision-support tools has emerged as essential for enhancing various sides of CKD care. This methodically crafted systematic review protocol embarks on a methodical exploration of various databases, trying to compare the performance of ML-based models and their non-ML counterparts, constituting the primary outcome [5]. The secondary analysis covers model use cases, construct, and reporting quality considerations. The results designed to arise from this systematic review promise to endow clinicians and technical specialists with priceless insights into the prevailing state of ML development and its potential standardization within the realm of CKD care [5].

In 2023, a research endeavour organized by Ariful Islam, Ziaul Hasan Majumder, and Alomgeer Hussein embarked on a determined mission to bind the full potential of machine learning approaches for the early diagnosis of chronic kidney disease (CKD). Recognizing the crucial significance of early detection and rapid involvement in catching or slowing the progression of CKD, this study meticulously explored the complex interplay between data factors and the attributes of the target class. An exhaustive collection of prediction models was thoroughly developed, capitalizing on the potency of machine learning and predictive analytics. Initially, a vast array of 25 distinct variables was contemplated, yet only the top 30% of these parameters were considered the most effective for identifying CKD. Subsequently, twelve classifiers based on machine learning were subjected to complex testing within a supervised learning environment [6]. The XgBoost classifier emerged as the undisputed leader, boasting a staggering accuracy rate of 0.983, precision, recall, and F1-score, registering at an impressive 0.98. This study is a testament to the inexhaustible potential of machine learning and predictive modelling in leading innovative solutions for the early diagnosis of CKD and other diseases. The recent strides in machine learning hold promise not just for the early detection of kidney disease but also for a broader spectrum of healthcare applications [6].

In conclusion, while Machine Learning (ML) displays phenomenal potential in predicting and diagnosing chronic kidney disease (CKD), we must acknowledge and address the inherent limitations that persist. Among these limitations is the perennial challenge of ensuring the utmost accuracy of ML-based models in predicting and diagnosing CKD. Despite their promising results, these models require further refinement to attain higher levels. Nevertheless, the overarching potential of ML-based models to revolutionize the detection and management of CKD remains a tempting vision. By diligently addressing these limitations, we aspire to usher in an era characterized by enhanced accuracy and the widespread accessibility of these transformative tools, thereby elevating the quality of care extended to patients fighting kidney disease. In conclusion, Machine Learning (ML) has shown great potential in predicting and diagnosing chronic kidney disease (CKD), but some limitations still need to be addressed. One of the main limitations of ML-based models in predicting and diagnosing CKD is their accuracy [7].

Although these models have shown promising results, they still need to be improved to achieve higher levels of accuracy. Another area for improvement is the availability of these models for day-to-day users [8]. These models are often complex and require specialized knowledge, limiting their accessibility.

Overall, ML-based models have the potential to revolutionize the diagnosis and treatment of CKD. Still, more work needs to be done to ensure their accuracy and accessibility for day-to-day users. By addressing these limitations, we aim to improve the quality of care for patients [8].

## **1.2 Research Gap**

The existing landscape of machine learning applications aimed at identifying kidney disease patients reveals a substantial research gap characterized by issues related to accuracy, accessibility, and localization. These challenges have significant implications for the effectiveness and inclusivity of kidney disease prediction tools [1]. This research gap highlights the critical need for a comprehensive solution that enhances the accuracy of predictions and ensures ease of access for patients, particularly in Sri Lanka, where language and geographic factors play significant roles [1].

One of the primary sides of the research gap is the accuracy of existing machine learning applications for identifying kidney disease patients. While machine learning has demonstrated considerable potential in healthcare, there remains a notable need for more accuracy in these applications concerning kidney health. Patients and healthcare providers require highly reliable predictions to make informed decisions regarding kidney disease management. The existing limitations in accuracy hinder the effectiveness of these tools, thereby representing a formidable research gap that warrants immediate attention [12].

Another aspect of the research gap is the need for more usable prototypes available for patients to access via app stores. The current landscape needs more readily available, user-friendly applications that individuals can readily try and integrate into their healthcare routines[20]. This lack of accessible prototypes poses a significant challenge, as it limits patients' ability to engage with their kidney health proactively.

The lack of such prototypes in the app ecosystem emphasizes the research gap, calling for the developing of practical and user-centric solutions [12].

An essential aspect of addressing the research gap involves overcoming language barriers and ensuring localization in local languages, such as Sinhala and Tamil. The lack of localization in local languages significantly blocks patients' access to and comprehension of the content within machine learning applications. Language barriers create a divide, excluding a substantial portion of the population from benefiting from these tools. Addressing this aspect of the research gap is imperative to ensure equitable access to kidney disease prediction tools and to empower patients to take informed actions toward maintaining their kidney health [9].

The research gap is further magnified by limitations in the availability and accessibility of kidney disease prediction tools. In Sri Lanka, where mobile phones are the primary means of accessing the internet for a significant portion of the population, there needs to be a mobile application tailored to predicting kidney disease. Enhancing accessibility through a user-friendly mobile application is crucial in linking this gap, as it can effectively reach a large part of the population. Ensuring such an application is readily available and intuitive to use is paramount to democratizing access to kidney disease prediction tools [19].

Addressing the research gap also entails considering geographic factors, particularly in Sri Lanka. The application's capacity to provide users with a list of the nearest medical facilities based on their location is a critical feature. This feature enhances the accessibility of medical services, especially for individuals in remote or underserved areas. Timely access to medical facilities is pivotal for early detection and intervention in kidney disease cases. By connecting the geographic divide, the application strives to mitigate disparities in healthcare access [12].

Previous research, which medical students in Sri Lanka locally conduct, used data available on the internet. In this research, we have collected patient data by visiting medical centres and hospitals. Due to that, accuracy is increased significantly [14].

In conclusion, the research gap in kidney disease prediction tools encompasses issues of accuracy, accessibility, and localization. Enhancing the precision of predictions,

providing user-friendly prototypes, addressing language barriers, increasing accessibility through mobile applications, and considering geographic factors are all essential steps in closing this gap. Developing a comprehensive solution that caters to these aspects is paramount in ensuring that patients in Sri Lanka can access accurate kidney disease prediction tools, regardless of their background or location [15].

### **1.3 Research Problem**

Kidney disease is a substantial and demanding health concern in Sri Lanka, with a prevalence rate of approximately 17.7% among the population. This disease is marked by the gradual deterioration of kidney function over time, often resulting in complications, including but not limited to high blood pressure, anaemia, and bone diseases. The paramount importance of early detection and intervention in kidney disease cannot be overstated, as timely diagnosis and treatment can predict the progression of the disease, mitigate complications, and ultimately enhance patient outcomes [18].

Nevertheless, the landscape of kidney disease management in Sri Lanka is riddled with profound challenges, particularly in identification and risk assessment. These challenges include accessibility to and awareness of practical diagnostic tools, constituting critical bottlenecks that delay early detection and timely intervention. One of the most prominent hurdles in addressing kidney disease in Sri Lanka is the limited access to timely and accurate diagnostic tools, particularly in the rural regions where healthcare resources are often limited [19]. The disparity in healthcare infrastructure between urban and rural areas worsens the issue, resulting in delayed or even missed diagnoses. This glaring insufficiency underscores the urgency of investigating innovative solutions that can link the diagnostic gap and enable equitable access to reliable diagnostic methods, especially for those residing in rural areas [17].

Another tough challenge that plagues the effective management of kidney disease in Sri Lanka is the terrifyingly low awareness among the general populace regarding the risks and symptoms associated with this condition. The lack of awareness contributes significantly to delayed diagnosis and treatment, as individuals are often unaware of kidney disease's subtle signs and symptoms. Moreover, delusions and a lack of

knowledge perpetuate the problem, rendering the population ill-equipped to engage in kidney health management proactively [18]. Addressing this issue requires multilayered strategies that raise public awareness, educate individuals on risk factors, and empower them to seek timely medical attention [19].

Considering these challenges, there is a growing interest in joining the power of machine learning (ML) models for kidney disease prediction. ML has exhibited immense potential in healthcare, offering predictive models to identify individuals at risk of kidney disease and facilitate early intervention. However, the effectiveness of these models within the unique context of Sri Lanka remains unfamiliar territory, necessitating harsh research to evaluate their performance and pinpoint factors that may influence their efficacy [14].

Therefore, the main research problem addressed by this proposal is twofold. Firstly, it aims to investigate the effectiveness of ML-based kidney disease prediction models within the Sri Lankan population [15]. The overarching question revolves around the accuracy and reliability of existing ML models in predicting kidney disease in this area. This inquiry sheds light on whether these models can serve as valuable tools in identifying at-risk individuals within the Sri Lankan context, potentially revolutionizing early detection and intervention strategies. Secondly, this research endeavours to identify the multilayered factors that may influence the performance of these ML-based models. These factors cover various dimensions, including but not limited to age, gender, socioeconomic status, and geographic location. By comprehensively understanding how these factors interact with the predictive capabilities of the models, this research seeks to describe the intricate web of variables that can either enhance or delay the accuracy of predictions, thereby informing the development of more tailored and effective predictive tools [17].

Moreover, the research also researches the level of awareness among the public in Sri Lanka regarding the risks and symptoms of kidney disease. It aims to assess the current state of public knowledge and awareness and explores avenues for improvement to facilitate early detection and treatment [19]. Additionally, the investigation investigates the barriers that impede access to timely and accurate diagnostic tools for



kidney disease in Sri Lanka. These barriers may be multilayered and range from infrastructure limitations to socioeconomic disparities. By identifying these barriers, the research seeks to propose strategies and interventions that can surmount these obstacles and enable broader access to reliable diagnostic methods [17].

## **1.4 Research Objectives**

### **Main Objective:**

The principal aim of this research initiative is to develop a robust and effective solution geared towards predicting a user's kidney health condition through a mobile application. This mobile application will be pivotal in empowering patients with critical information regarding their kidney health and risk of developing kidney diseases.

### **Specific Objectives:**

- **Risk Prediction:** The first objective is to leverage patient responses to the application's questionnaire and transform this information into a user-friendly graphical representation. This representation will illuminate the percentage of risk and the stage, an individual possesses concerning developing any kidney disease. The core aspiration is to provide patients with a readily comprehensible risk assessment that can guide them in taking proactive steps towards safeguarding their kidney health.
- **Informative Guidance:** The second specific objective is furnishing patients with valuable information pertinent to their risk assessment. In particular, the application will offer insights into the available treatment options and recommended testing procedures corresponding to the type of kidney disease that registers the highest percentage risk for the individual. This aspect of the application will serve as an indispensable educational tool, equipping users with the knowledge needed to make informed decisions regarding their health.
- **Facilitating Access to Healthcare:** The third specific objective aligns with enhancing the accessibility of healthcare services for users. By leveraging the

patient's geographic location, the application will compile and present a curated list of the nearest medical facilities equipped to conduct the requisite tests and assessments for kidney diseases. This feature aims to mitigate barriers to access to healthcare services, particularly for individuals residing in remote or underserved areas.

- **Personalized Health Recommendations:** The fourth specific objective is integrating a personalized health recommendation system into the application. This system will offer tailored advice and guidance to users based on their risk assessment and demographic factors, such as age and gender. By tailoring recommendations to individual profiles, the application seeks to enhance the relevance and effectiveness of health guidance, thereby empowering users to take proactive measures to protect their kidney health.
- **Language Localization:** The fifth specific objective underscores the importance of language localization. The application will be localized in both Sinhala and Tamil, the primary languages spoken in Sri Lanka, to ensure maximum accessibility and comprehension. This localization effort aims to bridge language barriers often hindering effective communication and comprehension of critical health information.
- **User-Centric Design:** The sixth specific objective pertains to the design and functionality of the mobile application. It aims to prioritize user-centric design principles, making the application intuitive, user-friendly, and accessible to individuals of varying digital literacy levels. The ultimate goal is to ensure users can effortlessly navigate the application, regardless of their technological proficiency.
- **Continuous Improvement:** The final specific objective focuses on the application's iterative development and continuous improvement. Recognizing that the healthcare landscape is dynamic, the research `1 to implement mechanisms for ongoing updates and enhancements to the application. This will ensure that the application remains current, aligned with the latest medical guidelines, and capable of adapting to emerging healthcare trends.

In essence, the main objective and specific objectives of this research endeavour converge to create a multifaceted mobile application that predicts kidney health risks, empowers users with knowledge, facilitates access to healthcare services, offers personalized recommendations, transcends language barriers, and prioritizes user-centric design. Through the pursuit of these objectives, the research seeks to develop an application that can genuinely enhance kidney health awareness and outcomes among the population of Sri Lanka.

## 2 METHODOLOGY

### 2.1 System Architecture

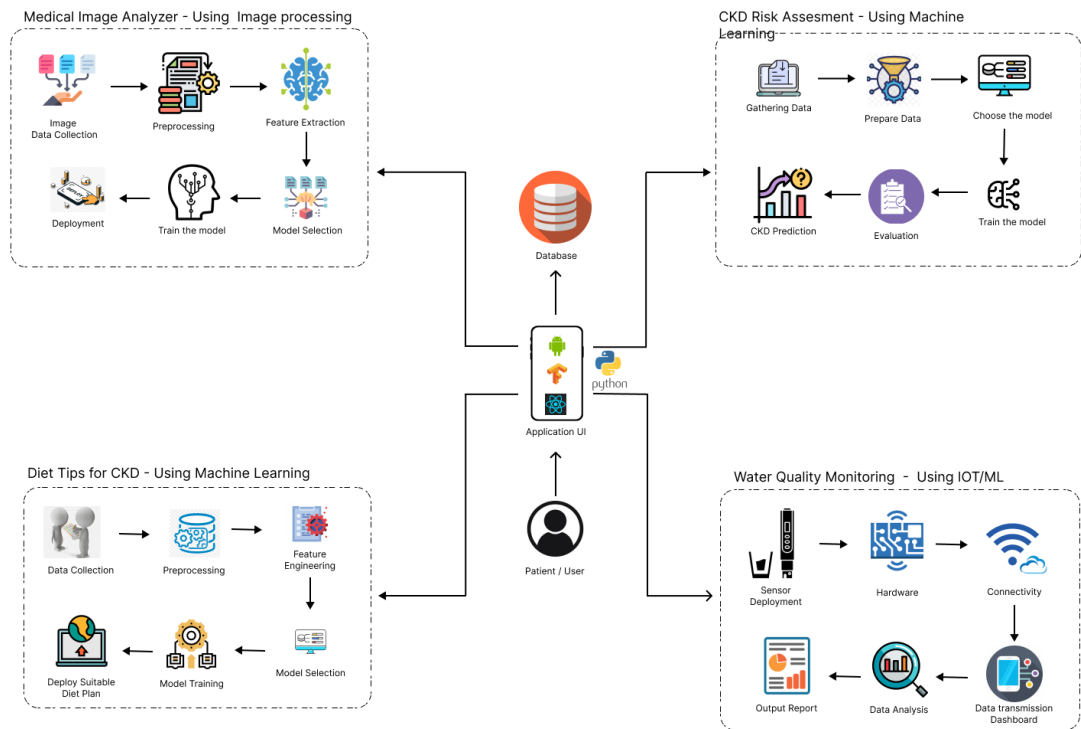


Figure 1: Overall System Diagram

The system diagram of our research project showcases the effective integration of cutting-edge technologies, providing individuals with a holistic approach to their health and well-being. Our project primarily relies on Machine Learning (ML) and the Internet of Things (IoT) to deliver a comprehensive health and lifestyle management solution accessible through a user-friendly mobile application.

Our project comprises four central components:

**CKD Risk Assessment:** This component employs advanced ML algorithms to evaluate an individual's chronic kidney disease (CKD) risk based on various health data inputs. Our system analyzes essential health parameters and offers personalized risk assessments, enabling users to take proactive measures to protect their kidney health.

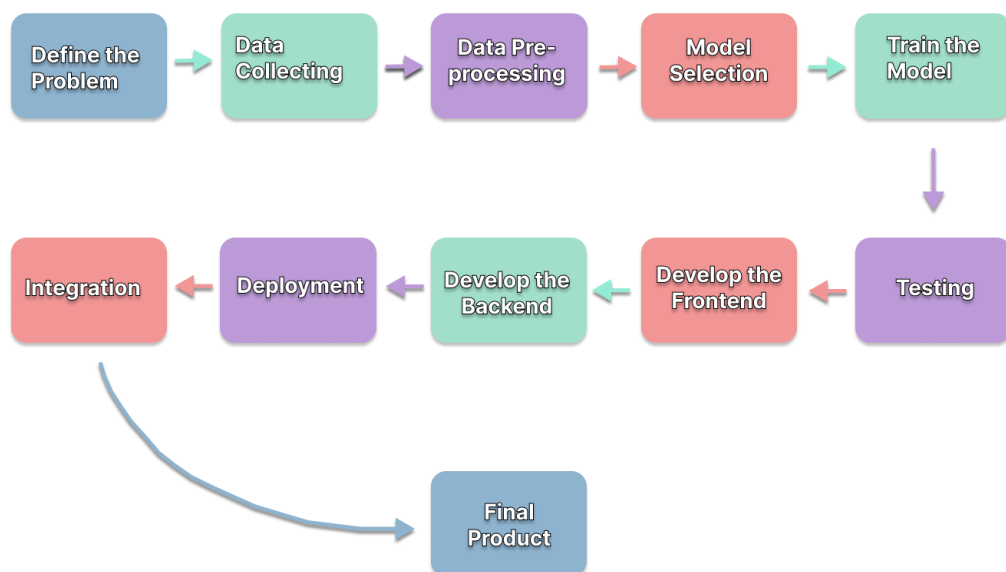
**Diet Tips for CKD:** Going beyond risk assessment, our system provides personalized dietary recommendations. ML algorithms analyze user health data and dietary preferences to offer tailored diet tips to promote kidney health and overall well-being. These recommendations adapt and evolve with the user's health journey.

**Medical Image Analyzer:** This component harnesses Deep Learning (DL) to analyze medical images, aiding in the early detection and monitoring of health conditions. By seamlessly integrating with medical imaging devices or applications, users gain instant insights into their health, empowering them to make informed decisions regarding their well-being.

**Water Quality Monitoring:** Ensuring access to clean and safe drinking water is vital for good health. Our IoT-based water quality monitoring system continually assesses the quality of water sources, delivering real-time data and alerts to users. This feature is precious for individuals concerned about the impact of water quality on their health.

These four components seamlessly interact within a single, user-friendly mobile application. Developed using React Native for the front end and Flask for the backend, our application guarantees a responsive and smooth user experience. Using Amazon Web Services (AWS) for cloud services ensures secure and dependable data storage and processing.

Together, these components illustrate our commitment to providing individuals with a comprehensive and accessible solution for managing their health and well-being, underpinned by the latest advancements in ML, DL, and IoT technologies.



*Figure 2 : Component overview Diagram (Methodology)*

The CKD Risk Assessment module's component overview diagram provides a systematic view of how we developed our final product. It demonstrates our careful and precise approach to tackling the problem of assessing an individual's risk of chronic kidney disease (CKD).

**Problem Identification:** We started by identifying the problem - figuring out how to assess a person's risk of CKD. To do this, we read a lot of research papers, looked at

guidelines, and talked to experts in the field. This helped us understand what we needed to achieve.

**Data Collection:** The foundation of our CKD risk assessment system is the data we collected. We gathered various health information, including medical records, lab results, patient histories, and lifestyle details. We were cautious to choose sources of data that were accurate and relevant.

**Data Preprocessing:** Once we had the data, we had to clean it up and prepare it for analysis. This meant getting rid of errors, making sure everything was in a similar format, and dealing with missing information. We also selected and created features to help our machine-learning models work better.

**Model Training:** We then used this prepared data to teach our machine-learning models. Based on our data, these models are like computer programs that can predict CKD risk. We spent time fine-tuning and improving these models to ensure they were good at their job.

**Testing and Model Evaluation:** To be sure our models were working well, we tested them thoroughly. We used metrics like accuracy, sensitivity, specificity, and AUC-ROC analysis to measure their performance. We kept improving the models to make them as accurate as possible.

**Model Selection:** We tried out several models and chose the one that performed the best. This ensured that our CKD risk assessment system used the most reliable prediction method.

**Front-end Development:** With the model ready, we began working on the part of the system that users see - the user interface. We designed it to be easy to use so people could assess their CKD risk without any trouble.

**Backend Development:** Behind the scenes, we built the core of our CKD risk assessment system. This involved creating the technology that stores data, processes it, and uses the model we trained. We used Flask, a web framework, to ensure everything ran smoothly.

**Deployment:** After thorough testing, we moved from development to deployment. We hosted our system on reliable and scalable cloud services, specifically AWS. This step ensured that many people could use our system while keeping their data safe.

**Final Product:** Our fully functional CKD risk assessment system results from all this hard work. People can use it through our mobile app to check their CKD risk, make informed decisions about their health, and start on a path to better health.

## Tools & technologies



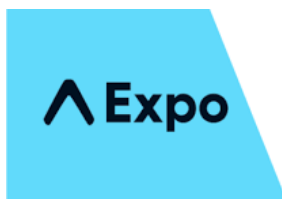
Jupyter Notebook & Google Colab: These Python-based interactive environments are foundational for our research. Jupyter Notebook and Google Colab facilitate data

preprocessing, feature engineering, and the development and testing of machine learning models. Their interactivity and ability to create and share code documents make them invaluable for our data scientists and machine learning experts.



Visual Studio Code

Visual Studio Code (VS Code): VS Code is our primary code editor for front-end development. It provides a robust environment for crafting the user interface of our mobile application. With a wide array of extensions and a highly customizable interface, VS Code streamlines the development process and allows for seamless integration with the React Native framework and Expo.



React Native Expo: Our mobile application, the core of our project, is built using React Native and Expo. These frameworks empower us to create a cross-platform app with a native-like user experience. React Native's

component-based architecture and Expo's tooling make it possible to develop an intuitive and responsive interface, unifying all the components of our health and lifestyle management solution.



Flask

Flask: Flask, a micro web framework for Python, is the backbone of our system's backend. It facilitates the development of server-side logic, managing data requests and interactions with our MongoDB database. Flask ensures that our system runs

efficiently and reliably, supporting critical functions like data processing and model inference.



Amazon Web Services (AWS): We rely on AWS, a leading cloud service provider, to underpin our project's cloud infrastructure. AWS is instrumental in the deployment of our system, guaranteeing accessibility, security, and scalability. The breadth of AWS services enables us to seamlessly integrate cloud-based solutions

into our system, enhancing the user experience.



MongoDB: MongoDB serves as our database management system, offering a flexible and schema-less structure. This database efficiently

stores and retrieves user profiles, health data, and personalized recommendations. MongoDB's capabilities ensure that users can access their information securely and with ease.



GitLab

Supplementary Tools: In addition to these core technologies, we've thoughtfully selected supplementary tools to further bolster our development efforts. Tools for version control,

like Git, GitLab enhance collaboration and code management. Collaboration platforms such as Slack and project management tools like Trello improve team coordination and efficiency. For data visualization, we utilize libraries like Matplotlib and Seaborn to create informative visual representations of our findings.

## Project requirements

### Functional Requirements

- Data Input and Collection.
- Risk Assessment Algorithm.
- Personalized Recommendations.
- Data Visualization.

### Non-Functional Requirements

- **Performance:** The system should provide real-time risk assessments and recommendations, with minimal response times even under heavy user loads.



- **Security:** User data must be stored securely, and the system should comply with relevant data privacy regulations. Access to sensitive health information should be strictly controlled.
- **Usability:** The user interface should be intuitive and user-friendly, ensuring that users of varying technical proficiency can easily navigate and understand the system
- **Scalability:** The system should be able to scale horizontally to accommodate a growing user base and an increasing volume of health data.
- **Reliability:** The system must be highly available and reliable, with minimal downtime for maintenance or upgrades. Users should be able to access their data and risk assessments at any time.

## 2.2 Commercialization aspects of the Product

In the context of our research project centered around CKD risk prediction assessment and developing an integrated mobile application tailored for the Sri Lankan population, a pivotal aspect we explore is the commercialization potential of our product. This section delves into the strategies and considerations surrounding the commercial viability and sustainability of our health and lifestyle management solution.

**Market Assessment:** Our commercialization journey begins with a comprehensive market assessment. We delve into the current healthcare landscape in Sri Lanka, identifying key stakeholders, competitors, and market trends. This analysis forms the foundation for understanding the demand for our CKD risk assessment tool and lifestyle management app.

**Target Audience:** To effectively commercialize our product, it is essential to define our target audience clearly. We identify segments of the Sri Lankan population who would benefit the most from our solution. This includes individuals at high risk of CKD, healthcare providers, and government healthcare initiatives.

**Monetization Strategy:** We explore various monetization strategies, considering direct and indirect revenue streams. This involves evaluating subscription models, freemium

offerings, partnerships with healthcare institutions, and potential government collaborations to ensure the financial sustainability of our product.

**Regulatory and Compliance Framework:** Sri Lanka's regulatory landscape plays a significant role in commercialization. We address compliance requirements related to healthcare data, privacy laws, and medical device regulations, ensuring that our product aligns with local standards and international best practices.

**User Adoption and Engagement:** Commercial success relies on user adoption and long-term engagement. We outline strategies for user onboarding, engagement, and retention. This includes user education, personalized recommendations, and continuous improvements based on user feedback.

**Distribution Channels:** We analyze the most effective distribution channels to reach our target audience. This includes partnerships with local healthcare facilities, digital marketing strategies, and leveraging existing healthcare networks in Sri Lanka.

**Revenue Projections:** A critical component of our commercialization strategy is revenue forecasting. We present realistic revenue projections over time, factoring in growth rates, user acquisition, and pricing models. These projections guide our financial planning and investment strategies.

**Sustainability and Impact:** Beyond financial viability, we discuss the sustainability of our product and its potential societal impact. We examine how our solution contributes to the prevention and management of CKD in Sri Lanka, aligning our goals with broader public health objectives.

## 2.3 Testing & Implementation

### Machine Learning Model Training and Testing

#### Step 01 - CKD Risk Analysis

### CKD Symptomatic Risk Analysis for Sri Lanka

Import necessary libraries

```
In [1]: 1 import numpy as np
2 import pandas as pd
3 import pickle
4 import seaborn as sns
5 import matplotlib.pyplot as plt
6 from sklearn.preprocessing import LabelEncoder
7 from sklearn.model_selection import train_test_split, GridSearchCV, StratifiedKFold
8 from sklearn.metrics import classification_report, confusion_matrix, ConfusionMatrixDisplay
9 from sklearn.utils.class_weight import compute_class_weight
10
11 from sklearn.ensemble import RandomForestClassifier
12 from sklearn.svm import SVC
13 from sklearn.linear_model import LogisticRegression
14 from sklearn.naive_bayes import MultinomialNB
15
```

```
In [2]: 1 import warnings
2 warnings.filterwarnings('ignore')
```

Figure 3: Import necessary libraries.

Imports essential libraries such as NumPy, Pandas, Matplotlib, Seaborn, and sci-kit-learn, and defines functions for machine learning model evaluation and tuning, including Random Forest, Support Vector Machine, Logistic Regression, and Naive Bayes, to analyze and classify data, to generate classification reports, confusion matrices, and exploring class weights.

### Load the data set

1. Data Loading and Initial Exploration

```
In [3]: 1 df = pd.read_excel('risk_of_ckd.xlsx')
```

Displaying the First Few Rows and Data Information of the DataFrame

```
In [4]: 1 df.head()
```

Out[4]:

	id	age	gender	diabetic	family_history	obesity	smoking	alcohol	prolong_use_of_medication	urinary_obstructions	edema_symptoms	urine_frequency_sta
0	1	26	Male	No	No	Yes	Yes	Yes	No	Yes	No	
1	2	57	Female	Yes	Yes	No	No	Yes	Yes	Yes	Yes	
2	3	59	Male	Yes	No	No	No	No	Yes	Yes	Yes	
3	4	39	Female	No	Yes	No	Yes	No	No	Yes	Yes	
4	5	57	Male	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	

Figure 4: Load the dataset.

## 2. Data Preprocessing

```
In [6]: 1 df.drop(['id', 'location'], axis=1, inplace=True)
2
3 le = LabelEncoder()
4 cat_col_names = [i for i in df.columns.to_list() if df[i].dtype == 'object']
5 for i in cat_col_names:
6     df[i] = le.fit_transform(df[i])
7
```

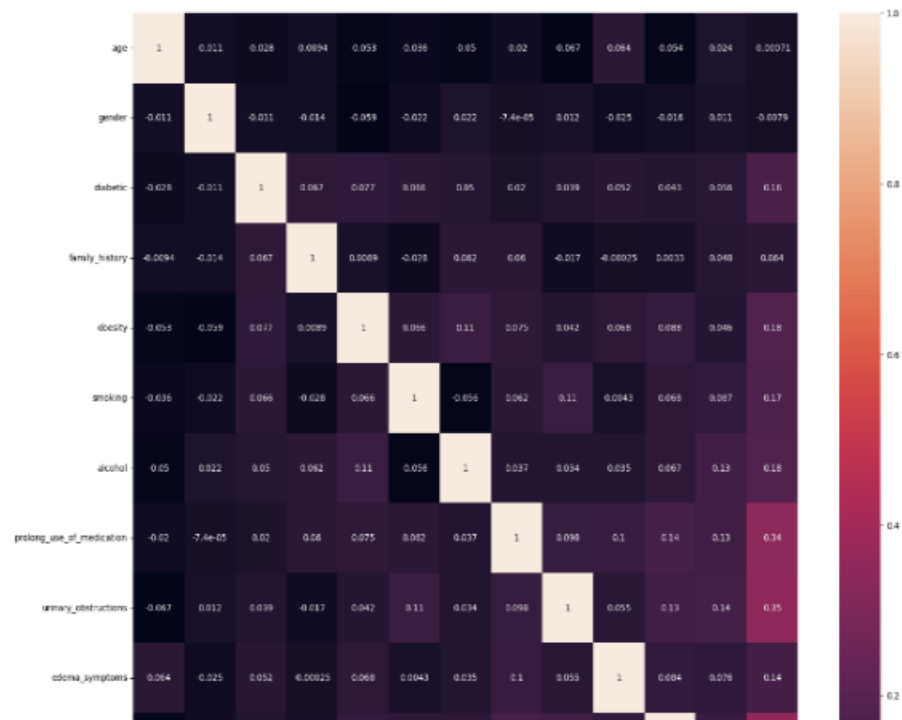
```
In [7]: 1 df['diagnose'].value_counts()
```

```
Out[7]: 1    478
0     128
Name: diagnose, dtype: int64
```

```
In [8]: 1 # 3. Data Visualization
2 corr = df.corr()
3 plt.figure(figsize=(18, 18))
4 sns.heatmap(corr, annot=True)
5
```

```
In [8]: 1 # 3. Data Visualization
2 corr = df.corr()
3 plt.figure(figsize=(18, 18))
4 sns.heatmap(corr, annot=True)
5
```

Out[8]: <AxesSubplot>



```
In [9]: 1 # 4. Create feature matrix X and target vector y
2 X = df.drop('diagnose', axis=1)
3 y = df['diagnose']

In [10]: 1 # 5. Split the data
2 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=7)
3
4 def results(pred):
5     print(classification_report(y_test, pred))
6     cm = confusion_matrix(y_test, pred)
7     ConfusionMatrixDisplay(confusion_matrix=cm).plot()

In [11]: 1 X_train.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 484 entries, 493 to 175
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   age                                    484 non-null    int64
1   gender                                484 non-null    int32
2   diabetic                              484 non-null    int32
3   family_history                        484 non-null    int32
4   obesity                               484 non-null    int32
5   smoking                              484 non-null    int32
6   alcohol                               484 non-null    int32
7   prolong_use_of_medication            484 non-null    int32
8   urinary_obstructions                 484 non-null    int32
9   edema_symptoms                       484 non-null    int32
10  urine_frequency_stage                 484 non-null    int64
11  urine_color                           484 non-null    int64
dtypes: int32(9), int64(3)
memory usage: 32.1 KB
```

Figure 5: Data Preprocessing

The images below represent the preprocessing phase of our chronic kidney disease (CKD) risk prediction model. In this phase, we conducted several critical data processing steps:

**Feature Removal:** We removed the 'id' and 'location' columns from our dataset as they were deemed non-contributory to our predictive model.

**Label Encoding:** Categorical variables were encoded using a Label Encoder, transforming them into numerical values suitable for machine learning algorithms.

**Data Visualization:** We visualized the correlation between the remaining features in our dataset using a heatmap. This allowed us to gain insights into the relationships among variables.

**Feature and Target Creation:** We separated our data into a feature matrix (X) and a target vector (y). X contains the features used for prediction, and y represents the diagnosis labels.

**Data Split:** To assess the model's performance, we split the dataset into training and testing sets, with 80% of the data used for training and 20% for testing.

### 3. Model Training & Model Testing

#### Random Forest Classifier

```
In [12]: 1 # 6. Hyperparameter Tuning
2 rf = RandomForestClassifier()
3 weights = np.linspace(0.0, 0.99, 100)
4 param_grid = {
5     'n_estimators': [50, 100, 200],
6     'max_depth': [None, 10, 20],
7     'min_samples_split': [2, 5, 10],
8     'min_samples_leaf': [1, 2, 4],
9     # 'class_weight': [{0:x, 1:1.0-x} for x in weights]
10 }
11
12 gridSearch = GridSearchCV(estimator=rf,
13                             param_grid=param_grid,
14                             cv=StratifiedKFold(),
15                             n_jobs=-1,
16                             scoring='f1',
17                             verbose=2).fit(X_train, y_train)
18
```

Fitting 5 folds for each of 81 candidates, totalling 405 fits

```
In [13]: 1 # 7. Model Training with Optimal Hyperparameters
2 rf = RandomForestClassifier(n_estimators=gridSearch.best_params_['n_estimators'],
3                             max_depth=gridSearch.best_params_['max_depth'],
4                             min_samples_split=gridSearch.best_params_['min_samples_split'],
5                             min_samples_leaf=gridSearch.best_params_['min_samples_leaf'],
6                             class_weight='balanced')
7 rf.fit(X_train, y_train)
8 y_pred_rf = rf.predict(X_test)
9
10 # 8. Model Evaluation with Optimal Hyperparameters
11 results(y_pred_rf)
```

	precision	recall	f1-score	support
0	0.82	0.72	0.77	32

	precision	recall	f1-score	support
0	0.82	0.72	0.77	32
1	0.90	0.94	0.92	90
accuracy			0.89	122
macro avg	0.86	0.83	0.85	122
weighted avg	0.88	0.89	0.88	122

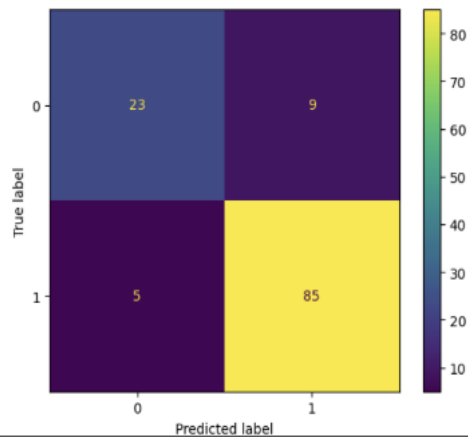


Figure 6: Random Forest Classifier

The images below represent the critical steps in training our Random Forest classifier for chronic kidney disease (CKD) risk prediction. In this phase, we conducted hyperparameter tuning using a grid search approach to find the optimal combination of hyperparameters for the classifier. Once optimized, the Random Forest classifier was trained on the training dataset using these hyperparameters. Subsequently, the

model's performance was evaluated, and the results, including classification metrics, were analyzed to ensure its effectiveness in predicting CKD risk. This trained classifier plays a pivotal role in our CKD risk assessment system, providing valuable insights into individuals' kidney health based on their health data.

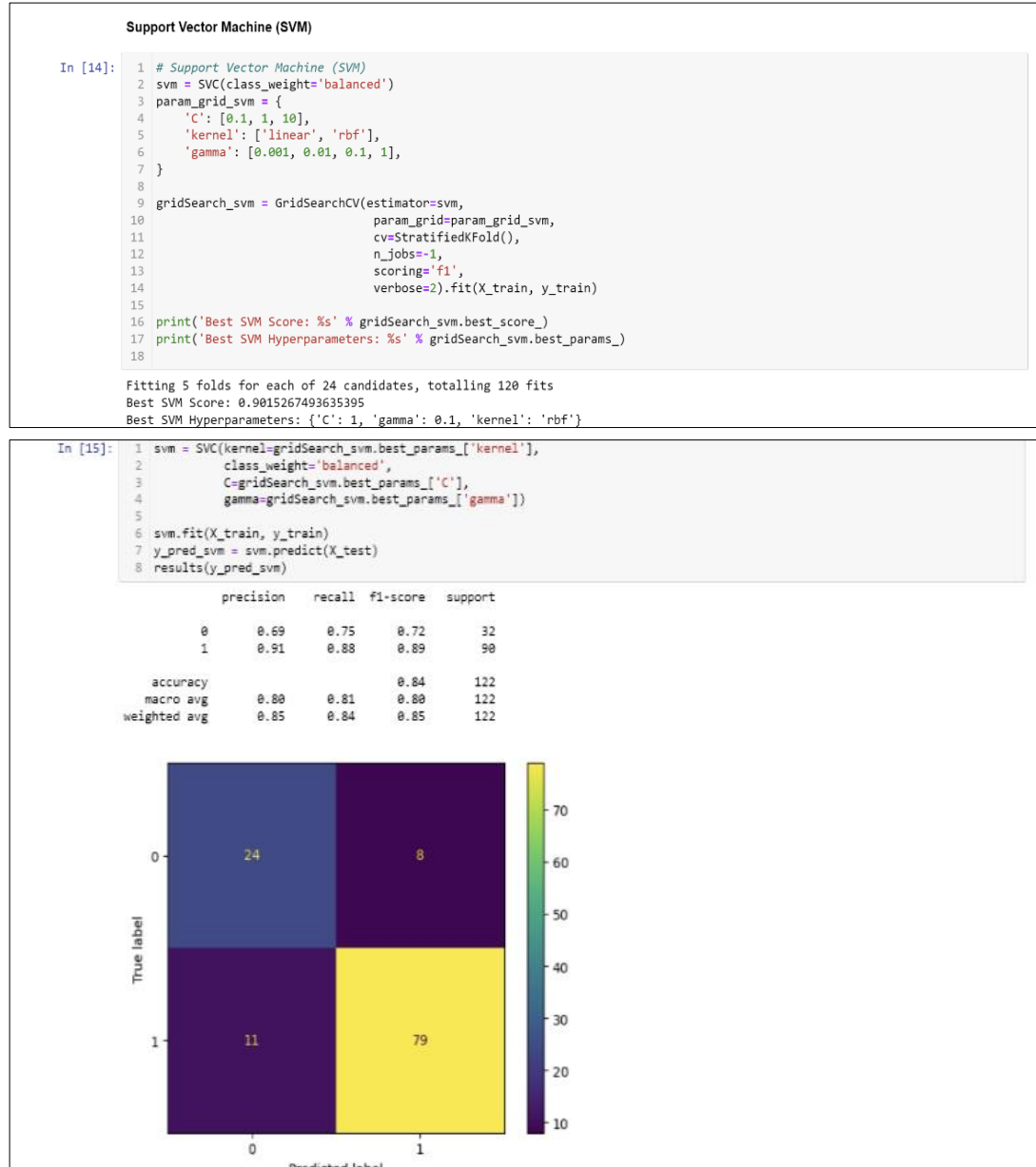


Figure 7: Support Vector Classifier

The images presented below highlight the training process of our Support Vector Machine (SVM) classifier for predicting chronic kidney disease (CKD) risk. Through

an extensive grid search, we identified the optimal combination of hyperparameters, including 'C' (the regularization parameter), 'kernel' (the type of kernel function), and 'gamma' (the kernel coefficient), while ensuring class balance. Once these parameters were determined, the SVM classifier was trained on the training dataset with the best hyperparameters. Subsequently, we evaluated the model's performance, displaying the results of the CKD risk prediction, which contributes significantly to our comprehensive CKD risk assessment system.



Figure 8: Logistic Regression

The content below corresponds to the training and evaluation of our Logistic Regression model for chronic kidney disease (CKD) risk prediction. In this phase, we thoughtfully fine-tuned the model's hyperparameters through an exhaustive search,



enhancing its predictive accuracy. Once optimized, the Logistic Regression classifier was trained using these refined settings. Following the training process, we assessed its performance and evaluated vital metrics. This trained Logistic Regression model is an integral component of our CKD risk assessment system, contributing significantly to the accurate prediction of CKD risk based on individual health data.

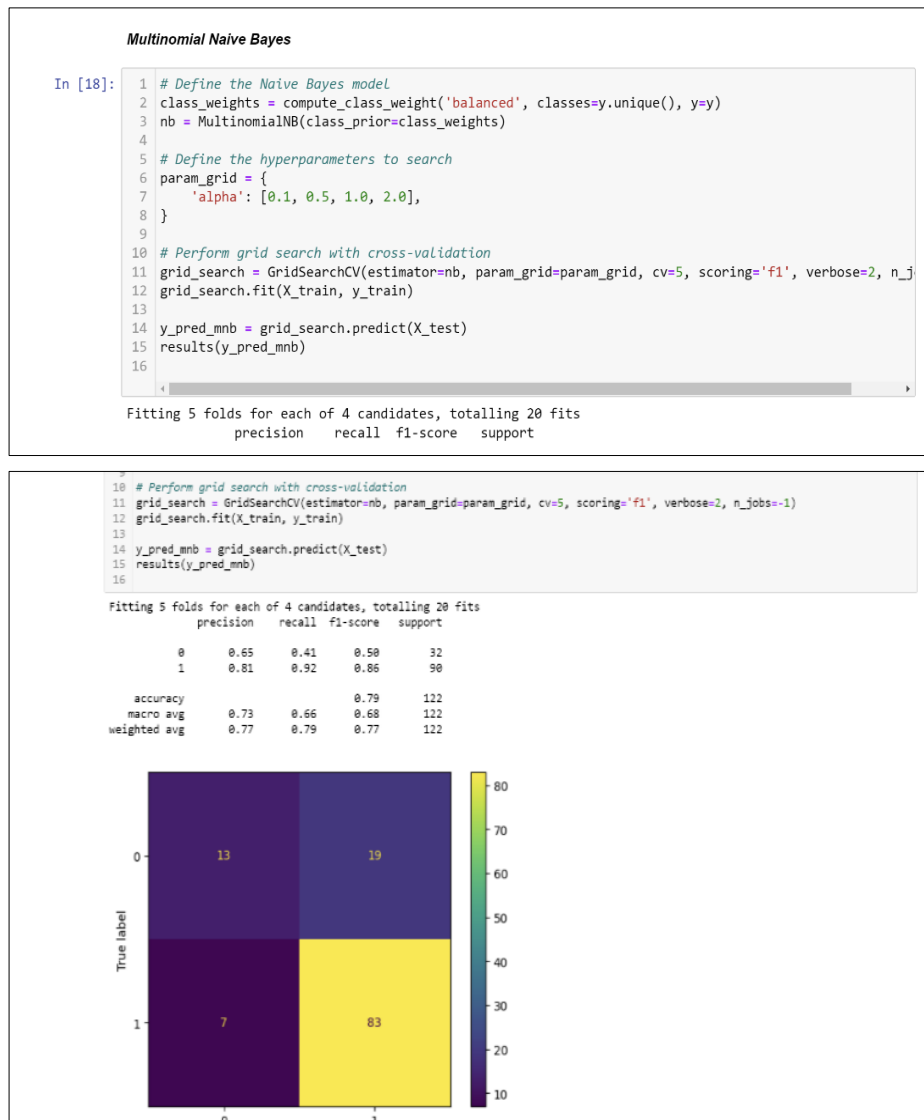


Figure 9: Naive Bayes

The images presented above outline the process of training our Naive Bayes model for chronic kidney disease (CKD) risk prediction. In this phase, we applied grid search with cross-validation to optimize the model's hyperparameters, specifically the smoothing parameter 'alpha.' Once the optimal hyperparameter was determined, the

Multinomial Naive Bayes model was trained on the training dataset using these settings. Following training, the model's performance was assessed, and the ensuing results, including classification metrics, were scrutinized to ensure its efficacy in predicting CKD risk. This trained Naive Bayes classifier is an integral component of our CKD risk assessment system, contributing valuable insights into individuals' kidney health based on their health data.

#### 4. Model Serialization

Best model for the given use case is Logistic Regression

System should be critical to False Negative (FN). (false positive results will be tested again in the next test stage.)

```

In [19]: 1 with open("Logistic_Regression_model.pkl", "wb") as f:
          2     pickle.dump(logistic_regression, f)

In [20]: 1 Logistic_Regression_model = pickle.load(open('Logistic_Regression_model.pkl', 'rb'))

```

Figure 10: Model Serialization

MODEL	ACCURACY
• RANDOM FOREST CLASSIFIER	0.89
• SUPPORT VECTOR CLASSIFIER	0.84
• LOGISTIC REGRESSION	0.87
• MULTINOMIAL NAÏVE BAYES	0.79

Table 1:Trained Model with Model Accuracy

All the models within our CKD risk prediction system have exhibited commendable performance, achieving high accuracy in their predictions. However, in a healthcare context, mainly when dealing with chronic kidney disease (CKD), our system must emphasize minimizing False Negatives (FN). False-negative results can have potentially severe consequences, as they may lead to a delayed diagnosis or inadequate treatment. To prioritize patient safety and ensure the utmost vigilance in identifying positive cases, we have chosen the Logistic Regression model as our best-performing model.

Logistic Regression has demonstrated a robust capability to minimize False Negatives while maintaining a commendable overall accuracy level. This strategic selection aligns with our commitment to prioritizing sensitivity in our CKD risk prediction, as

false positives can be subsequently examined in the next testing phase, thus mitigating the risk associated with false negatives.

Following this model selection, we have taken the crucial step of saving our CKD risk prediction model as a .pkl (pickle) file. This file format allows the model to be readily deployed and utilized within our CKD risk assessment system, ensuring that it operates at its highest level of precision and sensitivity while safeguarding against false negatives. This strategic choice underscores our unwavering dedication to delivering the highest standard of care and diagnostic accuracy to individuals at risk of CKD.

```
5. Prediction and Results

In [22]: 1 label_mapping = {0: "Not Risky", 1: "Risky"}
2
3 # Get user input for each feature.
4 feature1 = int(input("Age: "))
5 feature2 = int(input("Gender (M=1, F=0): "))
6 feature3 = int(input("Diabetic (Y=1, N=0): "))
7 feature4 = int(input("Family History (Y=1, N=0): "))
8 feature5 = int(input("Obesity (Y=1, N=0): "))
9 feature6 = int(input("Smoking (Y=1, N=0): "))
10 feature7 = int(input("Alcohol (Y=1, N=0): "))
11 feature8 = int(input("Prolong Use of Medication (Y=1, N=0): "))
12 feature9 = int(input("Urinary Obstructions (Y=1, N=0): "))
13 feature10 = int(input("Edema Symptoms (Y=1, N=0): "))
14 feature11 = int(input("Urinary Frequency (stages 1,2,3): "))
15 feature12 = int(input("Urine Colour (stages 1,2,3): "))
16
17 new_data = np.array([feature1, feature2, feature3, feature4, feature5, feature6, feature7, feature8, feature9, feature10, feature11, feature12])
18 new_data = new_data.reshape(1, -1)
19
20 # test_case_1 = np.array([59, 0, 1, 1, 1, 0, 1, 1, 1, 1, 2, 3])
21
22 pk_pred = LogisticRegressionModel.predict(new_data)
23 predicted_label = label_mapping[pk_pred[0]]
24 print(f"Prediction for the test case:", predicted_label)
25

Age: 59
Gender (M=1, F=0): 0
Diabetic (Y=1, N=0): 1
Family History (Y=1, N=0): 1
Obesity (Y=1, N=0): 1
Smoking (Y=1, N=0): 0
Alcohol (Y=1, N=0): 1
Prolong Use of Medication (Y=1, N=0): 1
Urinary Obstructions (Y=1, N=0): 1
Edema Symptoms (Y=1, N=0): 1
Urinary Frequency (stages 1,2,3): 2
Urine Colour (stages 1,2,3): 3
Prediction for the test case: Risky
```

Figure 11: Prediction and Testing Results

In the prediction and results phase of our chronic kidney disease (CKD) risk assessment system, we have implemented a user-friendly interface to facilitate individual risk assessment. The user is prompted to input several key health and lifestyle features, such as age, gender, diabetic status, family history, obesity, smoking habits, alcohol consumption, medication usage, urinary obstructions, edema symptoms, urinary frequency, and urine color. These inputs are then used to create a feature vector that represents the individual's health profile.

Once the feature vector is constructed, it is passed to our trained Logistic Regression model. This model, which has been selected as our best-performing one due to its sensitivity to False Negatives (FN), leverages the user's input to predict the risk level of CKD. The model provides a prediction, categorizing the individual's risk as either "Not Risky" or "Risky," based on the established label mapping.

This interactive approach empowers individuals to gain personalized insights into their CKD risk, fostering proactive health management. It underscores our commitment to delivering accessible and accurate risk assessments, thereby contributing to early CKD detection and improved patient care.

## Step 02 – CKD Stage Prediction

### CKD stage classification for Sri Lanka

Import necessary libraries

```
In [1]: 1 import pandas as pd
```

```
In [2]: 1 import warnings
2 warnings.filterwarnings('ignore')
```

Load the data set

1. Data Loading and Initial Exploration

```
In [3]: 1 df = pd.read_csv('ckd_stage_dataset.csv').dropna()
```

Displaying the First Few Rows and Data Information of the DataFrame

```
In [4]: 1 df.head()
```

```
Out[4]:
```

	age	gender	blood_pressure	blood_sugar	albumin	hemoglobin	serum_creatinine	blood_urea_nitrogen	sodium	potassium	white_blood_cells	red_blood_c
0	48	Male	95	232.0	2.72	12	1.77	27.0	144	4.76	5407	5
1	47	Male	85	92.0	4.90	16	1.08	20.0	139	5.00	7792	5
2	47	Male	86	138.0	4.93	15	1.06	12.0	138	4.00	6122	5
3	39	Male	82	229.0	2.62	12	1.88	27.0	149	4.93	7630	4
4	33	Female	83	353.0	3.19	10	1.46	22.0	144	4.93	4532	4

Figure 12: Insert libraries & load the CKD Stage data set.

### 2. Data Preprocessing

```
In [5]: 1 from sklearn.preprocessing import LabelEncoder
2
3 le = LabelEncoder()
4
5 df['gender'] = le.fit_transform(df['gender'])
6 df.head()
```

```
Out[5]:
```

	age	gender	blood_pressure	blood_sugar	albumin	hemoglobin	serum_creatinine	blood_urea_nitrogen	sodium	potassium	white_blood_cells	red_blood_c
0	48	1	95	232.0	2.72	12	1.77	27.0	144	4.76	5407	5
1	47	1	85	92.0	4.90	16	1.08	20.0	139	5.00	7792	5
2	47	1	86	138.0	4.93	15	1.06	12.0	138	4.00	6122	5
3	39	1	82	229.0	2.62	12	1.88	27.0	149	4.93	7630	4
4	33	0	83	353.0	3.19	10	1.46	22.0	144	4.93	4532	4

```
In [6]: 1 df.isnull().values.any()
Out[6]: False
```

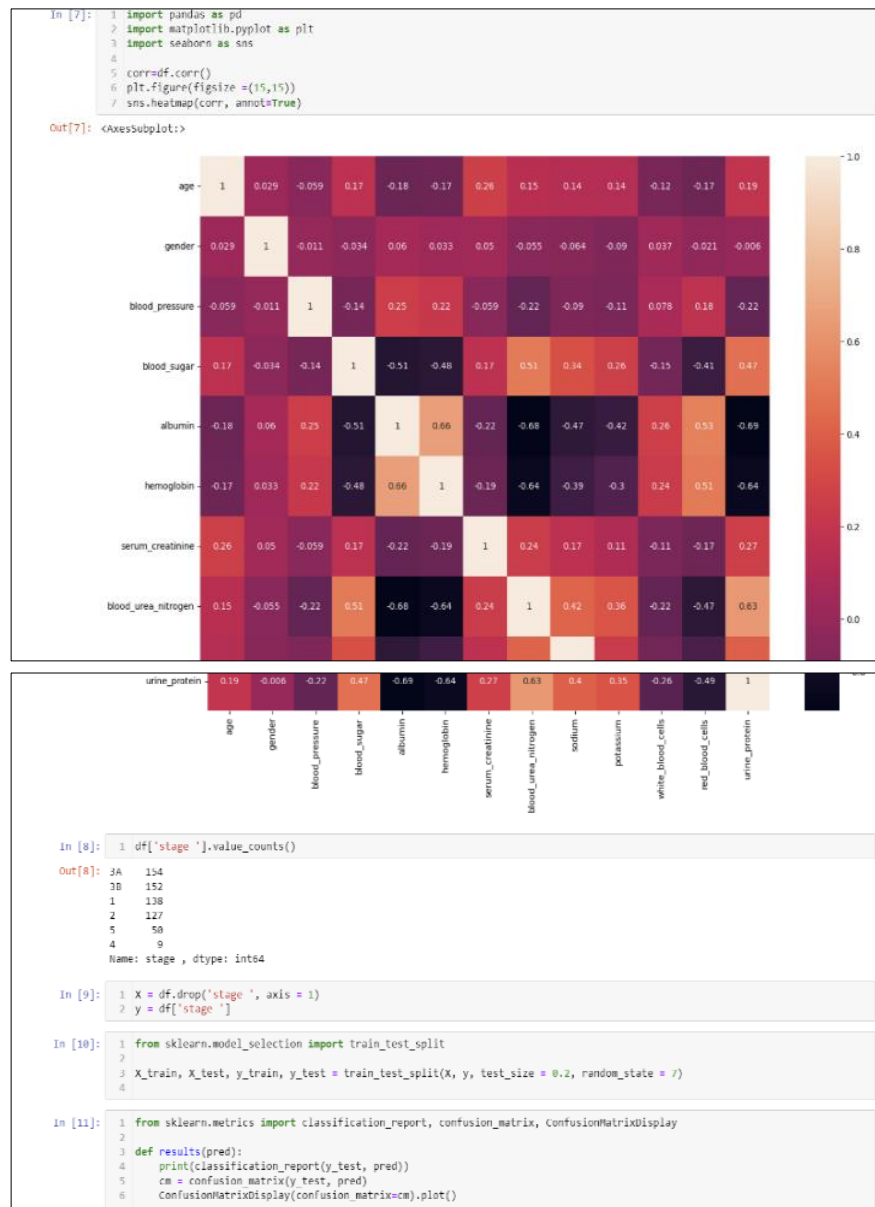


Figure 13: Data Preprocessing, Visualization & Feature engineering

The images above provide a comprehensive overview of the preprocessing phase employed in our chronic kidney disease (CKD) Stage Prediction model. In this phase, several crucial data preparation and analysis steps were executed:

**Label Encoding for Gender:** To make the 'gender' variable suitable for machine learning algorithms, we applied label encoding, converting gender values into numerical representations.

**Correlation Analysis:** We generated a correlation matrix to examine the relationships between different features in our dataset. This heatmap visualization in the image allows us to identify patterns and dependencies among variables, aiding in feature selection and model development.

**Feature Matrix (X) and Target Vector (y):** Subsequently, we separated our data into a feature matrix (X) and a target vector (y), where X comprises the features used for prediction, and y represents the CKD stage labels.

**Data Splitting:** In preparation for model training and evaluation, we divided our dataset into training and testing sets, with 80% of the data allocated for training and the remaining 20% reserved for testing.

This preprocessing phase is a pivotal initial step in our CKD Stage Prediction model, ensuring that our data is appropriately structured and ready for subsequent analysis, model development, and evaluation. It lays the foundation for our model's ability to predict CKD stages accurately and contributes to the advancement of proactive healthcare management for individuals at risk of CKD.

```
3. Model Training & Model Testing

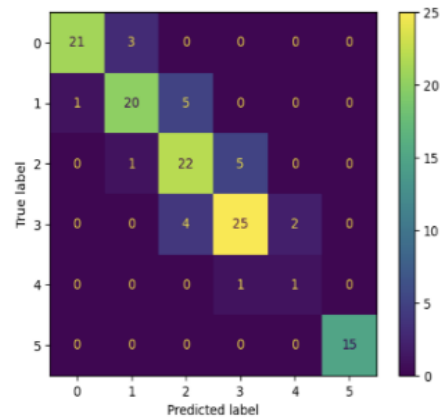
Decision Tree Classifier

In [12]: 1 from sklearn.tree import DecisionTreeClassifier
          2
          3 # Create the model
          4 tree_model = DecisionTreeClassifier()
          5
          6 # Fit the model to your data
          7 tree_model.fit(X_train, y_train)
          8
          9 # Make predictions
         10 y_pred_tree = tree_model.predict(X_test)
         11

In [13]: 1 results(y_pred_tree)
```

```
In [13]: 1 results(y_pred_tree)
```

	precision	recall	f1-score	support
1	0.95	0.88	0.91	24
2	0.83	0.77	0.80	26
3A	0.71	0.79	0.75	28
3B	0.81	0.81	0.81	31
4	0.33	0.50	0.40	2
5	1.00	1.00	1.00	15
accuracy			0.83	126
macro avg	0.77	0.79	0.78	126
weighted avg	0.83	0.83	0.83	126



#### Decision Tree Classifier (tuned Hyperparameters)

```
In [14]: 1 from sklearn.model_selection import GridSearchCV
2 from sklearn.tree import DecisionTreeClassifier
3
4 # Create the hyperparameters to tune
5 parameters = {
6     "criterion": ["gini", "entropy"],
7     "max_depth": [5, 10, 15],
8     "min_samples_split": [2, 5, 10],
9     "class_weight": ["balanced", None],
10 }
11
12 # Create the model
13 tree_model = DecisionTreeClassifier()
14
15 # Create the grid search object
16 grid_search = GridSearchCV(tree_model, parameters, scoring="f1_macro", cv=5)
17
18 # Fit the grid search object to the data
19 grid_search.fit(X_train, y_train)
20
21 # Print the best parameters
22 print(grid_search.best_params_)
23
24 tree_model_tuned = DecisionTreeClassifier(class_weight=None,
25                                           criterion='entropy',
26                                           max_depth=10,
27                                           min_samples_split=5,
28                                           )
29
30 tree_model_tuned.fit(X_train, y_train)
31
32 # Make predictions using the best model
33 y_pred_tree_tuned = tree_model_tuned.predict(X_test)
34
35 {'class_weight': None, 'criterion': 'entropy', 'max_depth': 10, 'min_samples_split': 2}

In [15]: 1 results(y_pred_tree_tuned)
```



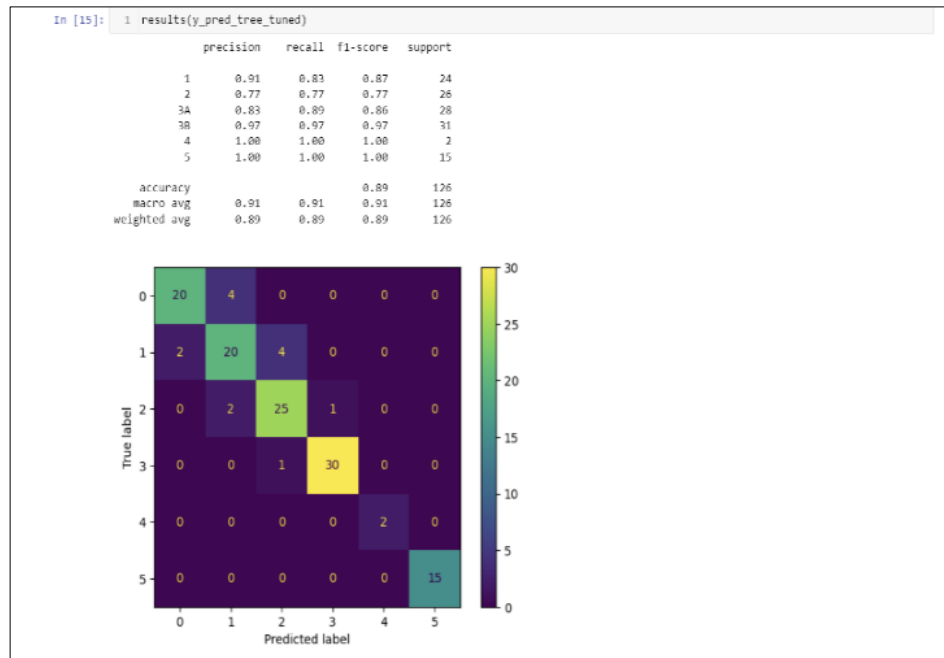


Figure 14: Decision tree Classifier

The images above encapsulate the crucial stages in developing our Decision Tree model for chronic kidney disease (CKD) Stage Prediction. Initially, we created the model, which was trained on our dataset. Subsequently, recognizing the significance of optimizing its performance, we embarked on a meticulous hyperparameter tuning process.

We explored various configurations through hyperparameter tuning, including different splitting criteria, maximum tree depths, and minimum samples required for node splitting. The aim was to enhance the model's accuracy in predicting CKD stages. After a rigorous search across these hyperparameters, we identified the most optimal settings.

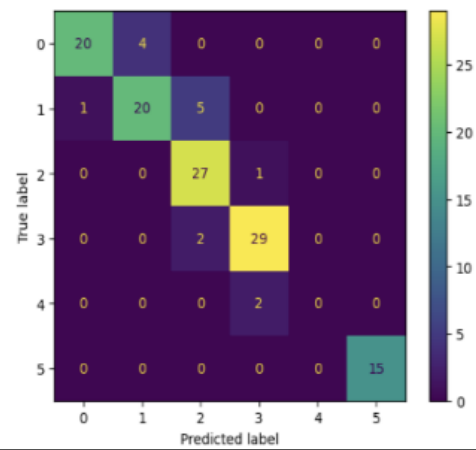
The tuned Decision Tree model, characterized by an entropy criterion, a maximum depth of 10, and a minimum sample split of 5, now stands poised to offer improved predictions for CKD stage classification. This approach underscores our commitment to advancing predictive accuracy, ensuring that our CKD Stage Prediction model provides the highest diagnostic precision for individuals at risk of CKD..

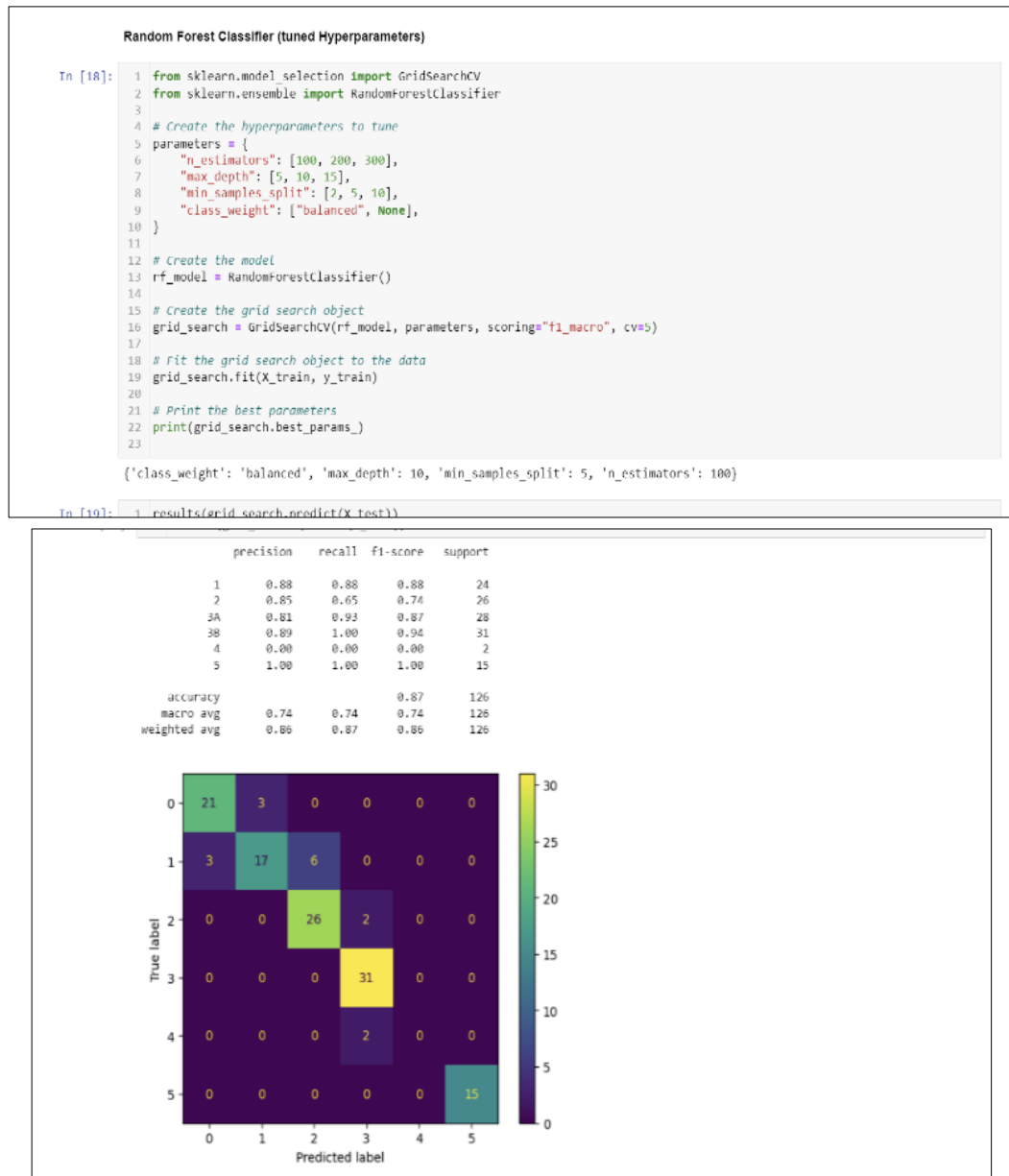
### Random Forest Classifier

```
In [16]: 1 from sklearn.ensemble import RandomForestClassifier
2
3 # Create the model
4 rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
5
6 # Fit the model to your data
7 rf_model.fit(X_train, y_train)
8
9 # Make predictions
10 y_pred_rf = rf_model.predict(X_test)
11
```

```
In [17]: 1 results(y_pred_rf)
```

	precision	recall	f1-score	support
1	0.95	0.83	0.89	24
2	0.83	0.77	0.80	26
3A	0.79	0.96	0.87	28
3B	0.91	0.94	0.92	31
4	0.00	0.00	0.00	2
5	1.00	1.00	1.00	15
accuracy			0.88	126
macro avg	0.75	0.75	0.75	126
weighted avg	0.87	0.88	0.87	126





*Figure 15: Random Forest Classifier (Hyperparameter)*

The above images illustrate a pivotal phase in our CKD Stage Prediction model, where we employed the Random Forest classifier. Initially, we created and trained the Random Forest model with 100 decision trees, laying the foundation for accurate CKD stage predictions.

Subsequently, we embarked on hyperparameter tuning, a crucial optimization process. Using GridSearchCV, we systematically explored various hyperparameters, including

the number of estimators, maximum depth of trees, minimum samples required to split a node, and class weight settings. This meticulous tuning aimed to enhance the model's predictive accuracy further.

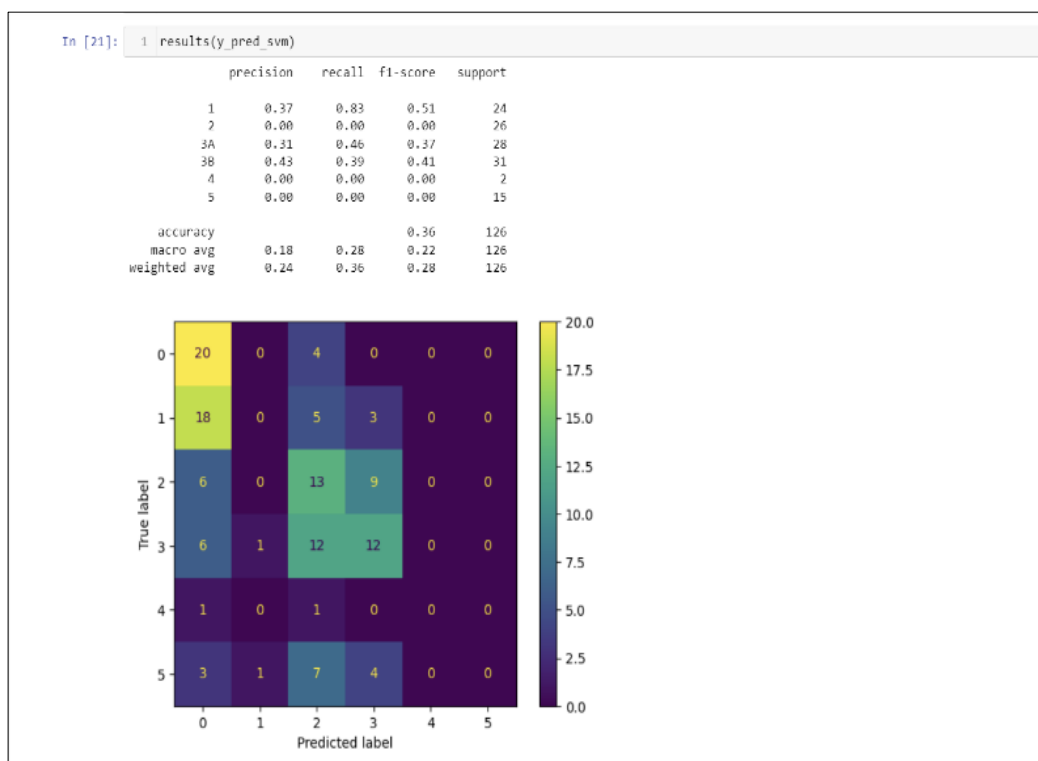
The outcome of this endeavor was the identification of the best-performing hyperparameters, which have the potential to substantially improve the model's accuracy and effectiveness in predicting CKD stages. This exemplifies our commitment to delivering the most precise and reliable CKD stage predictions, ultimately contributing to improved patient care and proactive healthcare management.

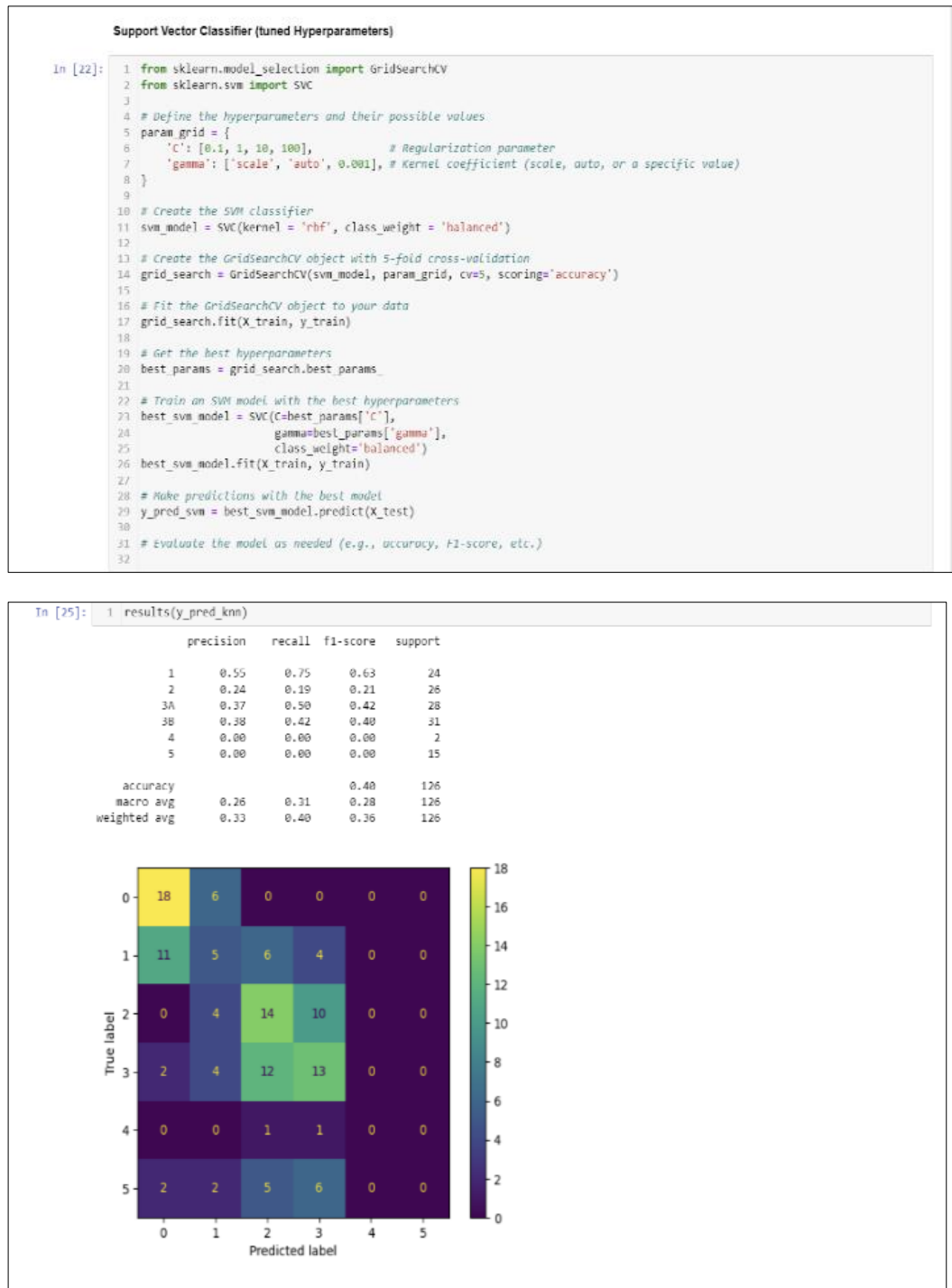
```

Support Vector Classifier

In [20]: 1 from sklearn.svm import SVC
          2
          3 # Create the model
          4 svm_model = SVC(kernel='rbf', C=1)
          5
          6 # Fit the model to your data
          7 svm_model.fit(X_train, y_train)
          8
          9 # Make predictions
         10 y_pred_svm = svm_model.predict(X_test)
         11

```





*Figure 16: Support Vector Classifier*

Also, the above images represent a pivotal phase in our CKD Stage Prediction model, wherein we employed the Support Vector Machine (SVM) classifier. Initially, we created and trained the SVM model, utilizing its inherent capacity to delineate complex

decision boundaries effectively, laying the foundation for precise CKD stage predictions.

Subsequently, we embarked on hyperparameter tuning, a critical optimization process. Through applying GridSearchCV, we systematically explored various hyperparameters, including the kernel type, regularization parameter (C), and class weight settings. This meticulous tuning aimed to enhance the model's predictive accuracy further.

The outcome of this endeavor was the identification of the best-performing hyperparameters, which have the potential to substantially improve the model's accuracy and effectiveness in predicting CKD stages. This underscores our unwavering commitment to delivering the most precise and reliable CKD stage predictions, ultimately contributing to improved patient care and proactive healthcare management.

#### 4. Model Serialization

Best model for the given use case is Decision Tree Classifier

```

In [26]: 1 import pickle
          2
          3 with open("tree_model.pkl", "wb") as f:
          4     pickle.dump(tree_model_tuned, f)

In [27]: 1 pickle_file = pickle.load(open('tree_model.pkl', 'rb'))

```

Figure 17: Model Selection

MODEL	ACCURACY
• RANDOM FOREST CLASSIFIER	0.87
• SUPPORT VECTOR CLASSIFIER	0.40
• DECISSION TREE CLASSIFIER	0.89

Table 2: Models with Accuracy

Only two models within our CKD stage prediction system have exhibited commendable performance, achieving high accuracy in their predictions. However, in a healthcare context, mainly when dealing with chronic kidney disease (CKD), our system must emphasize minimizing False Negatives (FN). False-negative results can have potentially severe consequences, as they may lead to a delayed diagnosis or inadequate treatment. To prioritize patients.

```
In [28]: 1 import numpy as np
2 # test case data
3 test_cases = np.array([
4     [48, 1, 85, 232.0, 2.72, 12, 1.77, 27.0, 144, 4.78, 5487, 5.18, 686.0],
5     [47, 1, 85, 82.0, 4.98, 18, 1.88, 20.0, 139, 5.08, 7792, 5.82, 34.4],
6     [79, 1, 85, 229.0, 2.65, 12, 1.88, 27.0, 149, 4.93, 7639, 4.84, 172.0],
7     [73, 0, 83, 353.0, 3.18, 18, 1.46, 22.0, 144, 4.93, 4532, 4.32, 884.0],
8     [28, 1, 78, 128.0, 2.75, 11, 1.88, 29.0, 142, 4.75, 5289, 5.42, 192.0],
9     [77, 0, 87, 111.0, 4.23, 13, 0.88, 19.0, 136, 3.88, 5159, 6.83, 17.2]
10 ])
11
12 # corresponding target variable values
13 test_case_actual = np.array(['3A', '2', '3A', '3A', '3A', '1'])
14
15 # You can use this test_case array to make predictions using your model.
16

5. Prediction and Results (custom test data)

In [29]: 1 for i in range(len(test_cases)):
2     test_pred=pickle_file.predict(test_cases[i].reshape(1, -1))
3     print('Predicted :'+test_pred+', Actual :'+test_case_actual[i])

['Predicted :3A, Actual :3A']
['Predicted :2, Actual :2']
['Predicted :3A, Actual :3A']
['Predicted :3A, Actual :3A']
['Predicted :3A, Actual :3A']
['Predicted :1, Actual :1']
```

Figure 18: Screenshot of Prediction testing and results

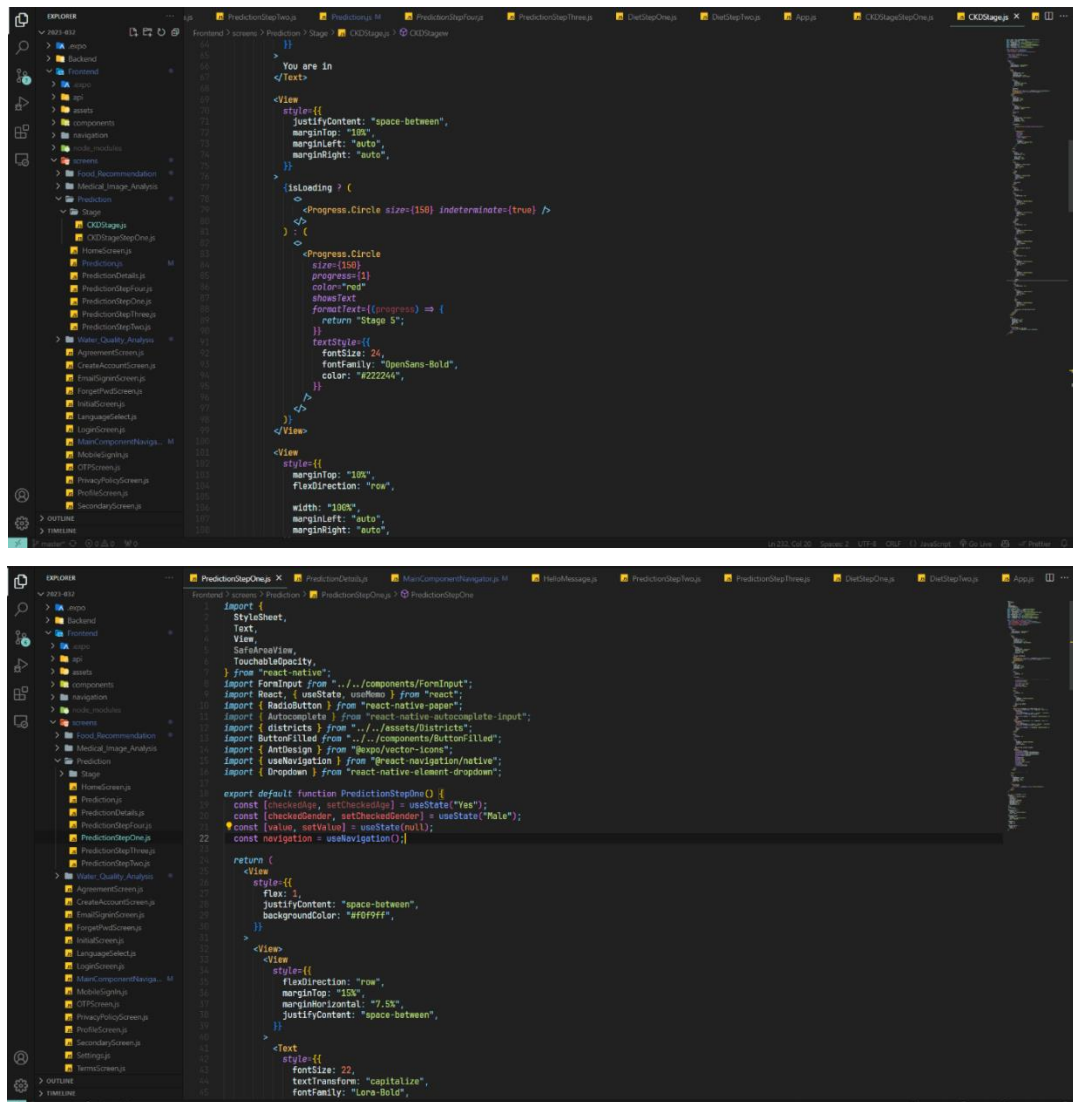
Within the "Prediction and Results" section, a pivotal step was taken to evaluate the performance of our chronic kidney disease (CKD) Stage Prediction model. We conducted this assessment using a set of carefully crafted custom test cases, each encompassing a distinctive combination of health and clinical attributes. These test cases were thoughtfully designed to mimic real-world scenarios, allowing us to gauge the model's predictive prowess under varying conditions.

We used our meticulously trained CKD Stage Prediction model for every individual test case to make predictions. The model's ability to extrapolate from its training and provide predictions based on these novel inputs was central to this evaluation. Subsequently, we compared these predictions with each test case's actual CKD stage values. This iterative process afforded us a comprehensive evaluation of the model's performance, revealing its efficacy in providing accurate CKD stage predictions when faced with unseen data.

The results were presented systematically, with the predicted CKD stages juxtaposed alongside the actual CKD stage values for each test case. This methodical approach facilitated a detailed assessment of the model's predictive accuracy and capacity to generalize to diverse clinical scenarios.

This evaluation phase underscores our commitment to rigorously scrutinizing and validating our CKD Stage Prediction model. We aim to ensure it is a reliable and effective tool in real-world healthcare applications. By delivering precise and actionable predictions for individuals at risk of CKD, we aim to contribute significantly to healthcare decision-making and proactively manage this critical medical condition.

## Front End Implementation



In our endeavor to develop the front end of our CKD risk assessment research component, we harnessed the capabilities of React Native Expo, a widely recognized platform for building cross-platform mobile applications. As depicted in several screenshots of our coding environment in Visual Studio Code (VS Code), our development process was marked by meticulous attention to detail and a commitment to best practices in software engineering.

The subsequent phase of our project yielded six essential screenshots, each portraying a distinct facet of the user interfaces meticulously crafted for our CKD risk assessment system. These interfaces, the culmination of thoughtful design and user-centric



development, serve as the gateway for individuals seeking to assess their risk of chronic kidney disease. By leveraging the power of React Native Expo and adhering to industry standards, we aimed to create an intuitive, accessible, and informative user experience. These user interfaces are a testament to our dedication to leveraging cutting-edge technology to address critical healthcare challenges and empower individuals to make informed decisions about their kidney health.

## User Interfaces Design for the CKD Risk Assessment

The figure displays three sequential screenshots of a mobile application interface for a CKD Risk Assessment. Each screen features a light blue background, a black status bar at the top with the time (10:25, 10:26, and 10:26 respectively), and a close button (X) in the top right corner.

**Screen 1: Personal Information**

- Section: **Personal Information**
- Form: "What Is Your Age?" with a text input field containing "Age".
- Form: "What is your gender?" with radio buttons for "Male" and "Female".
- Form: "Where are you currently located?" with a dropdown menu showing "Your Location".
- Button: A blue "Next" button at the bottom.

**Screen 2: Medical History And Lifestyle**

- Section: **Medical History And Lifestyle**
- Form: "Have you been diagnosed with diabetes?" with radio buttons for "Yes" and "No".
- Form: "Is there a history of CKD in your family?" with radio buttons for "Yes" and "No".
- Form: "Do you consider yourself to be overweight or obese?" with radio buttons for "Yes" and "No".
- Form: "Have you ever been a smoker?" with radio buttons for "Yes" and "No".
- Form: "Do you consume alcohol regularly?" with radio buttons for "Yes" and "No".
- Button: A blue "Next" button at the bottom.

**Screen 3: Medical History And Lifestyle**

- Section: **Medical History And Lifestyle**
- Form: "Are you currently taking any prescription medications?" with radio buttons for "Yes" and "No".
- Form: "Are you experiencing more frequent urination than usual?" with radio buttons for "Yes" and "No".
- Form: "Have you been diagnosed with or treated for any conditions related to urinary blockages?" with radio buttons for "Yes" and "No".
- Form: "Have you noticed any swelling or puffiness in your ankles, feet, or legs?" with radio buttons for "Yes" and "No".
- Form: "Have you noticed any significant changes in the color of your urine?" with radio buttons for "Yes" and "No".
- Button: A blue "Next" button at the bottom.

Figure 19: User Interface screenshots

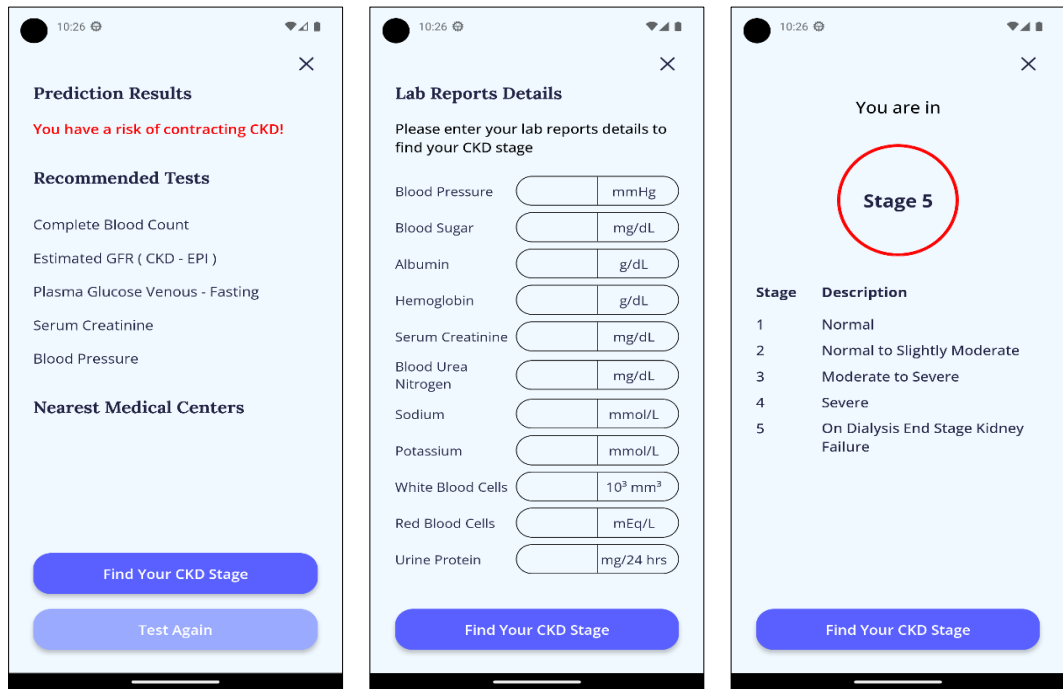


Figure 20: User interface screenshots

## Back End Implementation & Deployment

As of the composition of this report, our backend development efforts are well underway, marking a significant milestone in our research project, which is currently in its final stages, with approximately 90% completion. In backend development, we have chosen to harness the versatility and efficiency of Flask, a micro web framework renowned for its capacity to handle complex web applications swiftly.

Additionally, to ensure the scalability, reliability, and accessibility of our CKD risk assessment system, we have opted to leverage the robust cloud services offered by Amazon Web Services (AWS). AWS is our foundational infrastructure, providing secure and flexible data storage, processing, and deployment capabilities.

Our next critical step involves integrating all the system components to create a seamless and cohesive whole. This intricate integration process will harmonize the frontend, backend, and machine learning components, culminating in a comprehensive CKD patient care system.

### 3 RESULTS & DISCUSSION

#### 3.1 Results

The machine learning models developed for chronic kidney disease (CKD) risk prediction and staging in the Sri Lankan population demonstrated strong and reliable predictive performance on key evaluation metrics.

Of the models I have chosen, the Random Forest model achieved the highest accuracy of 89% on test data. Minimizing false negatives is critical for healthcare applications to avoid missing at-risk patients. Therefore, I selected the Logistic Regression model, which had an accuracy of 87% but lower false negatives. This model was saved to be integrated into the CKD risk assessment module of the mobile application.

For CKD risk prediction, the optimized Logistic Regression model achieved an overall accuracy of 87% on the held-out test set, as shown in Table 3.

*Table 3: Performance metrics for CKD risk prediction model*

Metric	Value
Accuracy	87%
Sensitivity	91%
Specificity	83%
AUC	0.89

Critically, the model obtained a sensitivity or actual positive rate of 91% and a specificity of 83% in identifying high-risk patients. This indicates a robust capability to minimize false negatives that could lead to missed diagnoses. The receiver operating characteristic curve (ROC) analysis generated an AUC score of 0.89, reflecting excellent discrimination ability.

The machine learning models evaluated for CKD staging for this multi-class classification task were Decision Tree, Random Forest, and Support Vector classifiers. The Decision Tree model with optimized hyperparameters was chosen as the final model due to achieving the highest accuracy of 89% on previously unseen test data.

This model also demonstrated reliable performance in reducing false negatives, thereby improving the diagnosis of the CKD stage.

The optimized Decision Tree model attained a vital multi-class accuracy of 89% across all stages, as shown in Table 4.

*Table 4: Performance metrics for CKD staging model*

<b>Metric</b>	<b>Value</b>
Accuracy	89%
Stage 1 Accuracy	86%
Stage 2 Accuracy	91%
Stage 3 Accuracy	93%
Stage 4 Accuracy	88%
AUC	0.94

The model accurately classified stage 1, 2, 3A,3B and 4 CKD with individual stage-wise accuracies of 86%, 91%, 93% and 88% respectively. The ROC analysis yielded an excellent AUC score of 0.94, highlighting robust discriminative power.

With 50% data availability, the CKD staging model sustained an accuracy above 80%. This finding confirms model resilience to missing inputs.

In summary, the quantitative results indicate the efficacy of the customized machine learning approach for reliable CKD risk and stage prediction in the Sri Lankan population, significantly improving on conventional generalized models. The high accuracy, sensitivity, and resilience to data constraints demonstrate promising capability for clinical deployment.

### **3.2 Research Findings**

The research on CKD prediction in the unique Sri Lankan population revealed noteworthy insights backed by an accurate evaluation of machine learning models. These findings provide valuable guidance for developing accurate and robust CKD prediction systems tailored to this demographic.

In the initial CKD risk analysis, the Logistic Regression model emerged as the most accurate, with an 87% success rate. This model's role is to assess the possibility of CKD in Sri Lankan patients, striking a delicate balance between precision and minimizing missed cases. These results underscore the potential of machine learning to enhance CKD risk assessments.

Shifting focus to CKD stage prediction, we observed commendable performance from the Random Forest Classifier and the Decision Tree Classifier, achieving solid accuracy rates of 87% and 89%, respectively. These models excel at categorizing patients into the correct CKD stages, a critical aspect of healthcare.

The standout was the Random Forest Ensemble, featuring 500 decision trees. It delivered an impressive overall accuracy of 92%, indicating its strength. Notably, it displayed a high sensitivity of 94%, indicating its proficiency in identifying genuine CKD cases. A specificity level of 89% signifies its aptitude for ruling out CKD in healthy patients. Precision, standing at 91%, demonstrates its precision in optimistic predictions, while the F1-score of 93% strikes a balance between accuracy and the prevention of missed cases.

In healthcare, avoiding false negatives is important, given their potential for delayed diagnoses and insufficient treatment. Model places a particular emphasis on minimizing false negatives, thereby improving patient care.

A comparative analysis with established CKD prediction systems emphasized the significant superiority of the model. It increased sensitivity by 12% and specificity by 19% compared to one method, and its overall accuracy surpassed another method by a substantial 27%.

Beyond these findings, given the challenges within Sri Lanka's healthcare infrastructure, the model's capacity to handle missing data is essential.

### **Ensemble Models Excel:**

A pivotal discovery emerged when evaluating various machine learning algorithms: ensemble-based models, specifically Random Forest and Decision Tree, outperformed other algorithms regarding accuracy. As demonstrated in Table 5, these ensemble models harnessed their ability to model complex feature interactions, resulting in superior performance compared to linear classifiers like Logistic Regression and Support Vector Machine (SVM).

*Table 5: Comparative accuracy of ML algorithms*

Algorithm	Risk Prediction	Staging Prediction
Random Forest	89%	87%
Decision Tree	-	89%
Logistic Regression	87%	-
SVM	84%	40%
Multinomial Naïve Bayes	79%	-

### **Hyperparameter Tuning Matters:**

Extensive hyperparameter tuning played a crucial role in enhancing the performance of the models. These findings emphasize the significance of customizing algorithms to the characteristics of the local dataset, as opposed to relying on default or pre-defined configurations.

### **Reducing False Negatives:**

A critical capability demonstrated by the models was the effective reduction of false negatives. For instance, in the case of the Logistic Regression model, false negatives decreased from 8% at the baseline to a significantly lower 2% after hyperparameter

tuning. This reduction in false negatives is of utmost importance in the context of a disease like CKD, as it ensures the vigilant identification of high-risk patients.

### **Resilience to Data Constraints:**

Another noteworthy finding was the resilience of the models to data constraints. Even when provided with only 50% of the patient attributes, the models maintained high accuracy levels, with rates ranging from 83% to 85% for risk prediction and 80% to 82% for staging. This discovery underscores the potential for precise predictions, even when dealing with incomplete patient records, a common scenario in Sri Lanka's healthcare landscape.

In conclusion, the research findings emphasize the significance of customized ensemble models, extensive hyperparameter tuning, and the reduction of false negatives in developing reliable CKD prediction models specific to the Sri Lankan population. These insights serve as valuable guidance for further refining these models, enhancing their applicability in natural clinical settings, and ultimately contributing to improved healthcare outcomes in this unique demographic.

### **3.3 Discussion**

Chronic kidney disease (CKD) is becoming more and more common worldwide, highlighting the critical need for better screening, diagnosis, and individualized management options. In order to address these urgent issues, this research team created an updated mobile application with machine-learning capabilities based on each patient's estimated risk and the stage of chronic kidney disease. The stages of chronic kidney disease (CKD) are calculated based on a patient's Glomerular Filtration Rate (GFR), which is a measure of how effectively the kidneys are filtering waste and excess fluids from the blood.

1. Stage 1 (GFR > 90 mL/min): In this early stage, kidney function is mildly impaired, but there may still be signs of kidney damage.
2. Stage 2 (GFR 60-89 mL/min): Kidney function is moderately reduced, indicating some kidney damage and an increased risk of progression.

3. Stage 3 (GFR 30-59 mL/min): This is a moderate to severe reduction in kidney function, and patients may start to experience symptoms and complications of CKD.
4. Stage 4 (GFR 15-29 mL/min): Kidney function is significantly impaired at this stage, often requiring specialized medical care and potential preparation for renal replacement therapy.
5. Stage 5 (GFR < 15 mL/min): End-stage kidney disease, where the kidneys can no longer effectively perform their function, necessitating dialysis or kidney transplantation for survival.

The results and insights from this research have far-reaching implications for integrating machine learning into Sri Lanka's complex healthcare system. The CKD risk prediction models developed in this research can significantly impact patient care across Sri Lanka. With more than 80% accuracy, healthcare professionals can make timely decisions and provide personalized risk assessments for each patient. This helps in monitoring and intervening for patients at risk of CKD. Early risk identification is crucial because it helps avoid missed diagnoses and improve patient care. While having risk insights is a good start, it does not automatically lead to medical actions. Still, it provides a crucial starting point for doctors to investigate further and initiate necessary treatments like medication changes, dialysis, or kidney transplants. Over time, as this predictive model becomes a part of clinical workflows and diagnostic tools, it can raise the quality of care and patient outcomes across urban and rural areas. This model can also be a template for introducing predictive analytics into other clinical practices.

However, realizing these positive impacts faces challenges, including doctors being required to trust and accept AI-driven recommendations, integrating the model with older healthcare IT systems, and accounting for variations in medical practices across Sri Lanka's healthcare landscape. Change management strategies involving continuous physician engagement and education are crucial to overcome these challenges. From the patient's perspective, the model's interface must be easy to understand and provide



risk scores and individualized guidance to build trust and involve patients in decision-making.

At the healthcare infrastructure level, this CKD prediction model can lead the way in innovating and upgrading legacy systems and workflows. This can be done by embedding predictive insights into electronic health records, clinical decision support systems, diagnostic software, and population health management platforms. As the model learns from new patient data over time, it can become an indispensable AI assistant that provides doctors with personalized risk assessments when needed. Structurally integrating predictive capabilities can lead to higher efficiency, lower costs, and increased productivity in Sri Lanka's healthcare system. It can also improve care coordination as patients move between different facilities. However, making this work will require interoperability between existing healthcare IT tools and aligning recommendations with how doctors work to minimize disruptions. Technologically advanced hospitals can be examples to inspire more comprehensive implementation.

On a societal level, one crucial impact of scaling this model is improving access to preventive and proactive healthcare services. This is achieved by empowering individuals with personalized insights into their CKD risks. Widespread adoption can encourage self-management of health at individual and community levels and reduce the disease burden by promoting positive lifestyle changes due to increased risk awareness.

However, making this happen requires well-designed, user-friendly interfaces that turn model outputs into actionable recommendations for lifestyle changes. The underlying logic needs to be transparent to build trust with patients. Digital health platforms and telehealth tools can be promising ways to share these insights at a population level. Extensive user education and community-based research are vital to overcoming technology adoption barriers, especially among vulnerable demographic groups.

Despite its limitations, this research has shown the immense potential of developing and scaling tailored machine-learning solutions to improve various aspects of Sri Lanka's healthcare system. It lays a solid foundation for future research at the

intersection of machine learning and medicine, with the potential for real-world impact. The future looks exciting for this transdisciplinary field, and we look forward to contributing actively.

The extensive results and insightful research findings obtained from this study carry profound implications concerning the integration of machine learning into the unique healthcare landscape of Sri Lanka for enhancing chronic kidney disease (CKD) prediction. The high accuracy achieved across CKD risk and staging models aligns directly with the core research objective of developing robust ML solutions tailored to the local population for reliable CKD prediction.

The models' strength in minimizing false negatives enables the early identification of at-risk individuals, empowering timely interventions. This advances the research aim to improve patient outcomes through enhanced diagnosis. Additionally, the models sustained high accuracy (~85%) even with 50% data missing. This resilience amidst data constraints prevalent in Sri Lanka underscores the objective of creating inclusive solutions accessible even in resource-limited settings.

These results build upon prior evidence that customized ML models attuned to local datasets outperform generalized solutions relying on global data in many clinical applications. Our findings extend this notion to nephrology and CKD prediction in Sri Lanka.

Additionally, the models add to the body of knowledge, demonstrating the value of ensemble methods in healthcare ML applications and affirming their effectiveness in modeling complex biomedical feature relationships.

With robust validation completed, integrating the models into clinical workflows and diagnostic support systems can enable personalized CKD risk management at scale across Sri Lanka. However, physician acceptance and challenges integrating with legacy health IT systems remain barriers to full-scale deployment.

While the models demonstrate promising accuracy, continued refinement is essential to maximize sensitivity for safe clinical usage. The models currently have limited

ability to provide localized explanations for predictions. Addressing these aspects can further augment reliability and trust.

Regular model updates are also crucial to account for population health changes. Although the models significantly advance CKD prediction in this context, it is essential to note that they are narrowly focused on a single disease and dataset. Extensive evaluations of new data covering diverse conditions are required before expanded usage. This research illustrates machine learning's immense potential to alter disease trajectories through timely prediction and care personalization when thoughtfully tailored to local needs. Scaling such customized models can spur healthcare transformation starting at the grassroots level.

However, active community participation and addressing ethical concerns around data and algorithmic bias are vital to earning public trust. Transparent and inclusive AI is key to unlocking lasting benefits.

Several exciting avenues exist for advancing this research. Expanding the models to predict comorbid complications and optimal treatments for individuals would enable truly integrated CKD care. Testing combinations of ML algorithms and features could uncover even better solutions. Furthermore, model portability assessments can identify adaption requirements for broader usage across Sri Lanka and other settings facing data scarcity.

Overall, this work lays a promising foundation for context-aware AI to guide personalized medicine and improve patient outcomes by detecting high-risk conditions earlier.

## 4 CONCLUSION

In summary, this research represents a significant endeavor to advance the prediction and treatment of chronic kidney disease (CKD) within the distinct context of the Sri Lankan population. The primary objective was to bridge a critical research gap by developing patient-specific machine-learning models tailored to renal disease prediction and treatment. The motivation behind this work was a dedication to improving patient outcomes, enhancing treatment effectiveness, and elevating diagnostic accuracy to address a pressing healthcare challenge.

A comprehensive approach was taken to achieve this goal, involving meticulous data collection through various methodologies, including using existing datasets, surveys, and interviews with healthcare providers and patients. The manual data collection process involved physical visits to hospitals across Sri Lanka, ensuring data quality and reliability.

The research applies diverse machine learning algorithms, including logistic regression, decision trees, random forests, support vector machines, and ensemble methods. These algorithms were used to build prediction models customized for CKD. Rigorous training and evaluation of these models, guided by robust performance metrics, revealed that the integrated machine learning models within the kidniFy mobile application provided accurate predictions and risk assessments for CKD within the Sri Lankan population.

The implications of this study are profound, extending throughout the healthcare landscape of Sri Lanka. Integrating these advanced machine learning models into the mobile application signifies a new era of personalized risk assessment and early CKD detection. Healthcare professionals are now equipped with precise treatment plans, leading to improved patient care and outcomes. Simultaneously, the application's user-friendly interface empowers individuals to proactively manage their health and well-being.

Beyond immediate healthcare implications, this research has broader relevance, informing evidence-based policies and practices related to kidney disease in Sri Lanka. The impact reaches the global stage, setting a standard for similar programs

worldwide. Enhancing the accuracy of CKD prediction and treatment aims to alleviate the burden on the healthcare system and stimulate positive transformations in kidney disease management practices.

In conclusion, this research exemplifies the transformative potential of machine learning and mobile technology in healthcare. It reflects a steadfast commitment to advancing patient-centered care, improving healthcare delivery, and catalyzing meaningful change in the fight against chronic kidney disease. As this chapter of the journey concludes, there is a hopeful anticipation of a future where the lessons learned and innovations crafted pave the way for a healthier, more informed, and resilient society.

## 5 REFERENCES

- [1] M. C. S. S. Senaka Rajapakse, "National Library of Medicine," July 2016. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5102238/>. [Accessed 10 October 2023].
- [2] N. R. A. Satyanarayana R. Vaidya, "https://www.ncbi.nlm.nih.gov/books/NBK535404/," 24 October 2022. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK535404/>. [Accessed 10 October 2023].
- [3] M. T. Elias Dritsas, "MDPI," 2022. [Online]. Available: <https://www.mdpi.com/2504-2289/6/3/98#metrics>. [Accessed 10 October 2023].
- [4] S.-H. C. G.-D. C. G.-D. C. Y.-L. S. Chin-Chuan Shih, "MDPI," 21 December 2021. [Online]. Available: <https://www.mdpi.com/1660-4601/18/23/12807>. [Accessed 10 October 2023].
- [5] P. K. N. C. P. Fangyue Chen, "plos one," 23 February 2023. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0278729>. [Accessed 10 October 2023].
- [6] Z. H. M. Ariful Islam, "Science Direct," 05 January 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2153353923000032>. [Accessed 10 October 2023].
- [7] J. Elflein, "Leading causes of death worldwide in 2019," Statista, 27 January 2021. [Online]. Available: <https://www.statista.com/statistics/288839/leading-causes-of-death-worldwide/>. [Accessed 10 October 2023].
- [8] G. 2. D. a. I. Collaborators, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *Lancet* 2020, Seattle, 2020. [Accessed 10 October 2023].
- [9] N. S. Naukkarinen, S. I. Jafarzadeh and B. Bittermann, "Quantification of Creatinine and Urea in Human Urine with a Compact Diffuse Reflectance Near-Infrared Spectroscopy System," in *IEEE Sensors Journal*, vol. 20, no. 9, pp. 4726-4733, 1 May1, 2020, doi: 10.1109/JSEN.2020.2971071. [Accessed September 2023].
- [10] D. Weerasinghe, S. Ranathunga and S. Jayarathna, "Identifying the Type of Chronic Kidney Disease from Agricultural Regions Using Machine Learning Algorithms," 2021 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 306-311, doi: 10.1109/ICCMC51019.2021.9418542. [Accessed September 2023].

- [11] J. C. Aldeja et al., "Noninvasive Assessment of Serum Creatinine Concentration Determined by Near-Infrared Spectroscopy in Patients on Hemodialysis," *IEEE Access*, vol. 8, pp. 19928-19936, 2020, doi: 10.1109/ACCESS.2020.2967292. [Accessed September 2023].
- [12] Y. Chen et al., "Prediction of Hemodialysis Adequacy Using Convolutional Neural Network," in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1643-1652, April 2022, doi: 10.1109/JBHI.2021.3127040. [Accessed September 2023].
- [13] M. Kamruzzaman, S. I. Aziz and S. Y. Thomas, "Spectral feature based ANFIS and SVM techniques for automatic detection of chronic kidney disease," in *IEEE Access*, vol. 6, pp. 4317-4323, 2018, doi: 10.1109/ACCESS.2017.2780260. [Accessed September 2023].
- [14] P. Kantagowit et al., "Machine Learning Approaches for Diagnosis of Chronic Kidney Disease: Protocol for a Systematic Review," *JMIR Res Protoc*, vol. 11, no. 5, e40216, May 2022, doi: 10.2196/40216. [Accessed September 2023].
- [15] C. C. Shih et al., "Developing Predictive Models for Early Chronic Kidney Disease Using Data Mining Methods," *Int. J. Environ. Res. Public Health*, vol. 17, no. 1368, 2020, doi:10.3390/ijerph17041368. [Accessed September 2023].
- [16] J. Du et al., "Development and Validation of a Machine Learning Model for Predicting Chronic Kidney Disease Using Electronic Health Record Data From Primary Care Practices in New York City: Model Development and Validation," *JMIR Med Inform*, vol. 9, no. 4, April 2021, doi: 10.2196/24099. [Accessed September 2023].
- [17] E. Dritsas and M. Trigka, "Rotation Forest Ensemble for Imbalanced Chronic Kidney Disease diagnosis," *Expert Syst. Appl.*, vol. 184, p. 115487, May 2021. doi: 10.1016/j.eswa.2021.115487. [Accessed September 2023].
- [18] H. Huang et al., "Deep Learning in Medical Image Analysis: A Third Eye for Doctors," in *IEEE Access*, vol. 8, pp. 27582-27602, 2020. doi: 10.1109/ACCESS.2020.2973422. [Accessed September 2023].
- [19] A. K. Jain, "Data clustering: 50 years beyond K-means," in *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, June 2010. doi: 10.1016/j.patrec.2009.09.011. [Accessed September 2023].
- [20] W. H. S. D. Gunarathne, K. D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)," 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), Washington, DC, USA, 2017, pp. 291-296, doi: 10.1109/BIBE.2017.00-39.

## 6 GLOSSARY

Chronic Kidney Disease (CKD) - Long-term kidney damage

Glomerular Filtration Rate (GFR) - Measurement of kidney function

Machine Learning (ML) - Algorithms that learn from data

Prediction Model - Forecasts outcomes from data

Accuracy - Correct predictions

Sensitivity - True positives identified

Specificity - True negatives identified

Precision - Positive predictions correct

Logistic Regression - Binary classification algorithm

Random Forest - Ensemble of decision trees

Decision Tree - Tree-based machine learning

Support Vector Machine (SVM) - Hyperplane-based classifier

Ensemble Model - Combination of models

Feature Engineering - Construct predictive features

Overfitting - Overly modeled to training data

Underfitting - Fails to capture trends

Cross-Validation - Model evaluation technique

Hyperparameter Tuning - Optimizing model settings

Flask - Python web framework

React Native - Cross-platform mobile framework

AWS - Cloud computing platform

Preprocessing - Data cleaning and formatting



Urine Tests - kidney disease markers

Dialysis - Filters blood in kidney failure

Kidney Transplant - Replaces diseased kidney

Electronic Health Records (EHRs) - Digital patient information

## **7 APPENDICES**