Explainable Machine Learning for Chronic Kidney Disease Prediction

Table of Contents

Table Of Figures

List Of Tables

Abstract

Chronic Kidney Disease (CKD) is a progressive condition that poses a major global health burden due to its asymptomatic onset and association with increased morbidity and mortality. Early detection and risk stratification are essential for effective intervention, yet traditional diagnostic approaches often fail to capture complex patterns within heterogeneous clinical data. This study develops and evaluates machine learning models for CKD prediction, with a focus on enhancing interpretability to support clinical decision-making. Using the Kaggle dataset, data preprocessing techniques including normalization and missing value imputation were applied, followed by training models such as Logistic Regression, Random Forest, Support Vector Machine, and XGBoost. Model performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC. The results demonstrate that ensemble methods, particularly Random Forest and XGBoost, achieved superior predictive accuracy. To address the "black-box" challenge of machine learning, SHAP (Shapley Additive Explanations) values were employed to provide both local and global interpretability, revealing clinically meaningful risk factors such as serum creatinine and blood pressure. The findings highlight the potential of interpretable AI models to improve CKD diagnosis, patient monitoring, and personalized treatment strategies.

Keywords: Chronic Kidney Disease (CKD); Machine Learning; Predictive Modelling; Random Forest; XGBoost; Logistic Regression; Support Vector Machine (SVM); Explainable AI (XAI); SHAP; Clinical Decision Support; Interpretability

Chapter 1- Introduction

1.1 Background and Context

Chronic Kidney Disease (CKD) represents a long-term and progressive medical condition in which the kidneys gradually lose their ability to function effectively. Clinically, CKD is defined by structural or functional abnormalities in the kidneys that persist for more than three months, typically evidenced through reduced glomerular filtration rate (GFR), raised serum creatinine levels, and the presence of protein in the urine (Levey et al., 2005). Unlike acute kidney injuries, CKD often progresses silently in its early stages, with noticeable symptoms emerging only when the disease reaches advanced levels. This "silent" progression makes it one of the most significant threats to global health.

The global burden of CKD is profound. The Global Burden of Disease Study reported that CKD now affects around 10% of the adult population worldwide, and its prevalence is rising largely due to the growing incidence of diabetes, hypertension, and ageing populations (GBD Chronic Kidney Disease Collaboration, 2020). The World Health Organization (2020) ranked CKD among the top twenty leading causes of death globally, attributing over one million deaths annually to the disease. In addition to mortality, the impact on patients' quality of life is substantial, as end-stage renal disease (ESRD) often requires long-term dialysis or kidney transplantation, both of which are resource-intensive and financially burdensome (Luyckx et al., 2018).

The importance of early detection cannot be overstated. Timely intervention at the early stages of CKD has been shown to slow progression, reduce complications, and improve patient outcomes. Preventive measures such as lifestyle modification, strict control of blood pressure and diabetes, and use of nephroprotective medications are far more effective when initiated early (Ene-Iordache et al., 2016). However, the challenge lies in the fact that the symptoms of CKD, such as tiredness, loss of appetite, or swelling, are nonspecific and can easily be attributed to other conditions. Consequently, patients are frequently diagnosed only when the disease has progressed to irreversible stages. This delay in detection underlines the urgent need

for predictive and preventative approaches in healthcare that can identify high-risk patients before severe damage occurs.

1.2 Problem Definition

The current reliance on conventional clinical measures, such as serum creatinine tests, estimated GFR, and urinalysis, poses limitations in the early detection of CKD. Although these measures are standard in medical practice, they are not always sensitive enough to pick up early-stage disease (Mills et al., 2015). In many instances, CKD is detected during routine testing for other conditions, meaning that opportunities for earlier intervention are often missed.

Three core challenges stand out in the diagnosis and management of CKD:

1. **Late diagnosis** – Many patients are diagnosed only at advanced stages of CKD, by which point treatment options are limited and costly (Couser et al., 2011).
2. **Risk of misclassification** – False negatives, where patients with CKD are not identified, may result in missed treatment opportunities. Conversely, false positives, where healthy individuals are incorrectly labelled as diseased, can lead to unnecessary stress, further testing, and resource wastage.
3. **Rising comorbidities** – The increase in diabetes, hypertension, and obesity complicates both diagnosis and treatment, as these conditions can mask or accelerate kidney decline (Foreman et al., 2018).

These limitations illustrate the need for innovative, data-driven methods that go beyond manual interpretation of clinical test results. Machine learning (ML) has emerged as a promising tool in this regard, as it can analyse complex datasets to reveal patterns and predict disease outcomes with high accuracy. The application of ML to CKD prediction provides a valuable opportunity to identify at-risk patients at an earlier stage and to assist clinicians in making timely, evidence-based decisions (Shickel et al., 2018).

1.3 Role of Machine Learning in Healthcare

The integration of machine learning (ML) into healthcare has become one of the most transformative developments in modern medicine. With the increasing availability of electronic health records, clinical laboratory data, and medical imaging, large volumes of patient information can now be analysed using computational techniques that go beyond the

capabilities of traditional statistics (Miotto et al., 2018). Unlike conventional approaches, which often rely on linear assumptions and limited variables, ML algorithms are able to detect subtle and complex patterns within high-dimensional data, enabling more accurate predictions of disease risk and progression (Kourou et al., 2015).

In practice, ML has already demonstrated success across various medical domains. For example, deep learning models have been used to classify skin cancer from images at a level comparable to dermatologists (Esteva et al., 2019), while predictive algorithms for sepsis and heart disease have outperformed traditional clinical scoring systems (Rajkomar et al., 2019). These applications illustrate the potential of ML to act as a decision-support tool, offering earlier detection of diseases, better patient stratification, and improved allocation of healthcare resources.

The case of Chronic Kidney Disease (CKD) is particularly suitable for ML approaches. CKD risk is influenced by a combination of biochemical measures (such as serum creatinine, haemoglobin, and albumin), demographic factors, and comorbidities including diabetes and hypertension. Traditional diagnostic methods may struggle to integrate these heterogeneous data points simultaneously, whereas ML algorithms can handle both categorical and continuous features to deliver more reliable predictions (Huang et al., 2020). By applying ML, patterns that are not easily visible to clinicians can be identified, allowing high-risk patients to be flagged earlier in the disease trajectory.

However, in the healthcare domain, predictive accuracy on its own is not enough. Clinical decision-making requires not only correct outcomes but also transparent reasoning. The widespread use of “black-box” models such as deep neural networks raises concerns among healthcare professionals, as the rationale behind predictions is often unclear (Caruana et al., 2015). This lack of interpretability can hinder clinical trust and regulatory approval, particularly when the stakes involve patient lives. To address this challenge, explainable artificial intelligence (XAI) techniques such as SHAP (SHapley Additive Explanations. These approaches provide both global insights into which features are most influential overall and local explanations for individual predictions, making the results more interpretable for clinicians (Lundberg and Lee, 2017; Ribeiro et al., 2016).

Thus, the role of ML in healthcare extends beyond predictive capability. It encompasses the need for systems that are interpretable, ethically responsible, and clinically usable. For CKD prediction, ML not only offers the possibility of earlier and more accurate detection but also

provides tools to ensure that clinicians understand and trust the basis of each prediction. This balance between accuracy and interpretability is essential for the adoption of ML-driven systems in real-world healthcare practice.

1.4 Aim and Objectives

The aim of this research is to design and evaluate an interpretable machine learning framework for the early prediction of Chronic Kidney Disease (CKD) using clinical datasets. The study intends to move beyond traditional diagnostic methods by applying advanced computational techniques that are capable of identifying subtle patterns within patient data, while ensuring that the predictive outputs are interpretable and clinically meaningful. This dual focus on accuracy and transparency reflects the growing consensus in healthcare that artificial intelligence systems must not only perform well but also provide clear justifications for their predictions (Caruana et al., 2015; Ribeiro et al., 2016).

To achieve this aim, the project sets out the following **objectives**:

1. **Dataset preparation** – To obtain a publicly available CKD dataset and perform rigorous preprocessing, including the handling of missing values, detection of inconsistencies, and preparation of the data for modelling. Effective preprocessing is fundamental, as poor data quality is one of the leading causes of unreliable machine learning outcomes in healthcare (Kourou et al., 2015).

2. **Feature representation** – To apply appropriate encoding and scaling methods for categorical and numerical variables, thereby ensuring comparability across features. Proper feature engineering enhances model learning capacity and reduces bias in predictive outcomes (Shickel et al., 2018).

3. **Model development** – To implement and train a diverse set of classification algorithms, including Logistic Regression, Support Vector Machine (SVM), Random Forest, Naïve Bayes, LightGBM, and XGBoost. These models were selected to represent both simple interpretable methods and more advanced ensemble-based approaches, allowing a balanced evaluation of performance and practicality in clinical settings (Huang et al., 2020).

4. **Model evaluation and comparison** – To evaluate model performance using a suite of healthcare-relevant metrics, namely accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The use of multiple

metrics ensures that the models are assessed not only for overall correctness but also for their ability to minimise false negatives and false positives, both of which carry significant clinical consequences (Rajkomar et al., 2019).

5. **Interpretability integration** – To employ explainable artificial intelligence (XAI) methods such as SHAP (SHapley Additive Explanations).These methods provide insights into how specific features contribute to predictions, enabling clinicians to trust and validate the outputs of machine learning systems (Lundberg and Lee, 2017; Ribeiro et al., 2016).

6. **Ethical consideration** – To assess the ethical dimensions of developing predictive models in healthcare, including fairness, potential bias, data security, and compliance with privacy regulations such as the General Data Protection Regulation (GDPR). Ethical evaluation ensures that predictive tools align with patient rights and healthcare governance frameworks (Vayena et al., 2018).

Through these objectives, the project seeks to contribute both academically and clinically. Academically, it benchmarks multiple machine learning techniques for CKD prediction and demonstrates the role of explainability in enhancing trust in predictive models. Clinically, it provides a framework that could support practitioners in making timely and evidence-based decisions for patients at risk of CKD, ultimately aiming to improve health outcomes and reduce the long-term burden of the disease.

1.5 Scope and Significance of the Study

The scope of this project is defined by its focus on the development of a predictive framework for early detection of Chronic Kidney Disease (CKD) using machine learning techniques. The study makes use of a publicly available CKD dataset consisting of both numerical and categorical clinical features, such as blood pressure, serum creatinine, haemoglobin levels, and patient history of comorbidities. The models selected—Logistic Regression, Support Vector Machine (SVM), Random Forest, Naïve Bayes, LightGBM, and XGBoost—represent a balance between traditional interpretable classifiers and more advanced ensemble methods. The inclusion of multiple models ensures that the study does not rely on a single approach but provides a comparative analysis that highlights both strengths and limitations of different algorithms in a healthcare setting (Kourou et al., 2015; Huang et al., 2020).

The significance of this work lies in its dual emphasis on accuracy and interpretability. While predictive performance is crucial, clinical adoption depends heavily on whether healthcare

professionals can understand and trust the system's outputs. By integrating explainable artificial intelligence (XAI) methods such as SHAP and LIME, the project contributes to bridging the gap between complex models and clinical usability, allowing physicians to gain insights into how and why predictions are made (Ribeiro et al., 2016; Lundberg and Lee, 2017). Furthermore, by embedding ethical considerations such as fairness, bias reduction, and compliance with the General Data Protection Regulation (GDPR), the study ensures that its contributions are not only technical but also aligned with the responsibilities of modern healthcare systems (Vayena et al., 2018).

Ultimately, the project's significance extends beyond academic experimentation. It demonstrates how machine learning can be translated into clinically meaningful tools for early CKD prediction. This has the potential to support timely interventions, improve patient outcomes, and reduce the economic burden of advanced-stage CKD, thereby making a direct contribution to both healthcare practice and broader public health objectives (GBD Chronic Kidney Disease Collaboration, 2020).

1.5 Research Questions

Formulating clear research questions is essential for aligning the objectives of a study with the methodology and expected outcomes. In this project, the research questions are designed to address both the technical and practical aspects of applying machine learning (ML) to the early prediction of Chronic Kidney Disease (CKD). They not only reflect the computational challenges of developing predictive models but also consider the clinical, ethical, and interpretability requirements that influence whether such models can be realistically adopted in healthcare practice.

Research Question 1: How effective are different machine learning algorithms, such as Logistic Regression, Support Vector Machine, Random Forest, Naïve Bayes, LightGBM, and XGBoost, in predicting early-stage CKD from clinical data?

The first research question aims to evaluate the comparative effectiveness of different ML classifiers. Logistic Regression and Support Vector Machine (SVM) have been widely used in medical prediction tasks because of their interpretability and ability to handle binary outcomes (Kourou et al., 2015). Ensemble methods such as Random Forest, LightGBM, and XGBoost are more advanced techniques capable of modelling non-linear relationships and interactions between clinical features (Huang et al., 2020). Naïve Bayes, while simplistic, is often used as a baseline in healthcare studies due to its speed and ease of implementation (Rajkomar et al., 2019). This question ensures that the study not only benchmarks multiple models but also

identifies which algorithm balances accuracy, generalisability, and clinical usability most effectively.

Research Question 2: What preprocessing techniques, including missing value imputation, categorical encoding, and feature scaling, are most suitable for preparing CKD datasets for predictive modelling?

Medical datasets frequently suffer from incompleteness, inconsistencies, and variations in data entry. Missing values, in particular, are common in CKD records due to patients skipping tests or incomplete reporting (Mills et al., 2015). This question explores whether techniques such as K-Nearest Neighbour (KNN) imputation can provide realistic replacements for missing values, preserving the relationships between features (Jerez et al., 2010). Similarly, encoding categorical features such as hypertension status and red blood cell characteristics is critical to ensure they are properly represented in ML models. Feature scaling, using approaches like Min-Max Normalisation, is equally important for models sensitive to input magnitudes (Shickel et al., 2018). Answering this question ensures that the dataset used for modelling is not only clean but also optimised for learning, which directly influences the reliability of the predictions.

Research Question 3: Which evaluation metrics provide the most reliable assessment of predictive performance in the healthcare context, and how do different models compare against them?

Accuracy alone cannot determine whether a model is useful in a healthcare setting. For instance, in imbalanced datasets, a model may achieve high accuracy while failing to identify positive cases, which could have dangerous clinical implications (Caruana et al., 2015). Hence, this research question addresses the need to use metrics such as precision, recall, F1-score, and ROC-AUC. Precision ensures the reduction of false positives, recall focuses on minimising false negatives, F1-score balances the two, and ROC-AUC provides a threshold-independent view of model discrimination (Rajkomar et al., 2019). By systematically comparing models against these metrics, the study aims to provide a holistic view of their strengths and weaknesses, ensuring that the selected model is both statistically robust and clinically safe.

Research Question 4: How can explainable artificial intelligence (XAI) methods such as SHAP and LIME enhance the interpretability of machine learning models in CKD prediction?

Interpretability is a fundamental barrier to the adoption of ML in healthcare. Clinicians often

hesitate to trust "black-box" models because they cannot see how predictions are derived (Caruana et al., 2015). SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) are two widely used XAI methods that provide explanations at both the global and local levels, showing which features influence overall predictions and individual patient outcomes. This research question ensures that the study does not only focus on achieving high performance but also addresses transparency and trustworthiness. By integrating XAI, the project contributes to making ML tools more practical in real-world clinical decision-making.

Research Question 5: What ethical and practical considerations, including fairness, privacy, and regulatory compliance, need to be addressed to ensure that predictive models for CKD are trustworthy and clinically applicable?
Machine learning in medicine must be developed within an ethical and regulatory framework to avoid potential harm. Issues such as data bias, inequality in predictions across patient groups, and compliance with data protection regulations like the General Data Protection Regulation (GDPR) are critical (Vayena et al., 2018). For example, if a model disproportionately misclassifies patients based on age or gender, it may reinforce health disparities. This research question ensures that the study not only addresses technical feasibility but also acknowledges broader responsibilities, aligning predictive modelling with ethical standards and patient rights.

Together, these research questions create a structured pathway for the dissertation. They establish the technical scope by focusing on preprocessing, model selection, and evaluation, while also highlighting interpretability and ethics as crucial dimensions. Answering them will allow the project to contribute meaningfully to both academic research in machine learning and its practical application in the healthcare domain, particularly for the early detection of CKD.

1.6 Structure of the Dissertation

The dissertation is organised into the following chapters:

- **Chapter 1 – Introduction**: Provides the background of the study, defines the research problem, outlines the aim and objectives, and explains the significance and scope of the work.
- **Chapter 2 – Literature Review**: Reviews existing work on Chronic Kidney Disease prediction, machine learning in healthcare, and the role of interpretability methods such as SHAP and LIME.

- **Chapter 3 – System Design**: Presents the overall design of the proposed system, including architectural diagrams, workflow representations, and the approaches followed to integrate preprocessing, model training, and interpretability.
- **Chapter 4 – Methodology**: Describes the dataset, preprocessing pipeline, selected machine learning algorithms, evaluation framework, and ethical considerations.
- **Chapter 5 – Implementation and Results**: Provides coding evidence with screenshots, details model training and testing, compares classifiers using evaluation metrics, and presents interpretability results.
- **Chapter 6 – Discussion**: Interprets and analyses the findings in relation to existing literature, highlighting strengths, limitations, and clinical significance.
- **Chapter 7 – Conclusion and Future Work**: Summarises the contributions of the project, reflects on achievements, and outlines opportunities for future research and development.
- **References and Appendices**: Contain the full list of references in Harvard style and supplementary material, including extended tables, figures, and code excerpts.

Chapter 2: Literature Review

2.1 Introduction to the Literature Review

A literature review serves as the foundation for any research project, providing a critical understanding of what has already been achieved and where knowledge gaps persist. In this study, the review has two primary purposes: first, to explore how Chronic Kidney Disease (CKD) has been approached within existing clinical and computational research; and second, to establish how machine learning (ML) techniques can be applied, evaluated, and improved for the prediction of CKD.

The review begins with the clinical context of CKD, outlining its prevalence, major risk factors, and the challenges that arise in early detection and management. From there, attention shifts to the role of ML in healthcare more broadly, before narrowing to focus on how these approaches have been specifically applied in CKD prediction. By distinguishing between primary research—original studies that present empirical findings on predictive models—and secondary research, such as systematic reviews and meta-analyses that evaluate multiple approaches collectively, the review builds a balanced perspective (Kitchenham and Charters, 2007). This distinction is important, as it allows the project to draw not only on individual case evidence but also on broader patterns in the literature.

A further strand of the review concerns the techniques and methodologies used in existing systems. This includes the handling of missing data, feature selection strategies, the choice of classification algorithms, and the metrics applied for performance evaluation. In reviewing these works, it becomes clear that many existing systems have prioritised predictive accuracy while paying less attention to interpretability and ethical considerations—two dimensions that are critical in healthcare settings (Rajkomar et al., 2019; Shickel et al., 2018).

The chapter therefore plays a dual role: it situates the present study within the existing body of work, and it highlights the shortcomings that justify the development of a new framework. In particular, the review identifies the need for approaches that combine reliable performance with transparency and fairness. This project responds to that gap by proposing a comparative

analysis of multiple machine learning models, supported by robust preprocessing methods and enhanced by explainable artificial intelligence (XAI) tools such as SHAP and LIME. In this way, the study not only builds on existing research but also aims to make a practical contribution to the development of interpretable, ethically grounded predictive systems for CKD.

2.2 Background on Chronic Kidney Disease (CKD)

Chronic Kidney Disease (CKD) is a progressive and long-term condition in which the kidneys gradually lose their ability to filter waste and maintain essential bodily functions. It is clinically defined as abnormalities of kidney structure or function lasting more than three months, commonly measured through reduced glomerular filtration rate (GFR), elevated serum creatinine levels, and the presence of albumin in urine (Levey et al., 2005). CKD is often referred to as a "silent disease" because its early stages present few noticeable symptoms. Patients may feel well until the kidneys are already severely impaired, at which point treatment options become limited and costly (Jha et al., 2013).

The global impact of CKD is striking. The Global Burden of Disease (GBD) Study estimated that almost 700 million people were living with CKD in 2017, making it one of the most prevalent non-communicable diseases worldwide (GBD Chronic Kidney Disease Collaboration, 2020). Mortality linked to CKD has increased significantly in the last two decades, and the World Health Organization (2020) now lists it among the leading causes of death globally. Importantly, CKD rarely exists in isolation. Patients with CKD are at a heightened risk of cardiovascular events, with cardiovascular mortality often exceeding deaths directly attributable to kidney failure (Tonelli et al., 2016). This dual burden reinforces the need for earlier detection and management strategies.

From an economic perspective, CKD places a heavy strain on healthcare systems. Patients who progress to end-stage renal disease (ESRD) require dialysis or kidney transplantation, both of which are resource-intensive. In high-income countries such as the United States and the United Kingdom, dialysis alone costs healthcare systems tens of thousands of dollars per patient annually. By contrast, in many low- and middle-income countries, access to dialysis is scarce, meaning many patients do not survive once they reach ESRD (Couser et al., 2011; Luyckx et al., 2018). Consequently, there is an urgent emphasis on preventing CKD progression through effective early interventions.

The causes of CKD are closely linked to other major global health challenges. Diabetes mellitus and hypertension are the two leading drivers, jointly responsible for the majority of CKD cases worldwide (Foreman et al., 2018). Other contributors include obesity, smoking, cardiovascular disease, genetic predisposition, and environmental exposures. Beyond biological factors, social determinants such as poverty, limited healthcare access, and lack of awareness also play a critical role in disease prevalence and progression, particularly in resource-limited regions (Ene-Iordache et al., 2016).

Despite its scale, CKD remains difficult to diagnose in the early stages. Current diagnostic tools, including serum creatinine measurement, GFR estimation, and urine protein tests, are effective in detecting moderate to advanced disease but lack sensitivity for identifying early kidney damage (Mills et al., 2015). Furthermore, these tests require repeated laboratory access, which is not always feasible in under-resourced areas. This diagnostic gap contributes to late identification, at which point treatment options are less effective, and patients are at higher risk of poor outcomes.

This situation highlights the pressing need for approaches that go beyond conventional clinical testing. Predictive models using computational techniques can help identify high-risk individuals before severe kidney damage occurs. Machine learning, in particular, has the capacity to integrate diverse clinical features—ranging from laboratory values to demographic data—into models that predict disease risk more accurately than traditional methods (Huang et al., 2020). By enabling earlier detection, such systems can support clinicians in initiating preventive strategies, reducing the likelihood of disease progression, and ultimately lessening both the human and economic costs of CKD.

2.3 Existing Systems for CKD Prediction

Over the last decade, researchers have increasingly explored the use of machine learning (ML) to support the early prediction of Chronic Kidney Disease (CKD). A range of models has been applied, from traditional classifiers to modern ensemble methods and neural networks, each showing varying levels of effectiveness. Reviewing these systems is important not only to recognise their achievements but also to identify the limitations that motivate the present study.

Traditional classifiers such as Logistic Regression and Support Vector Machine (SVM) have been widely adopted in early CKD studies. Logistic Regression is often chosen for its interpretability and suitability for binary classification. For example, Aljaaf et al. (2018) reported that Logistic Regression achieved over 95% accuracy in predicting CKD when applied

to clinical datasets with carefully selected features. Similarly, SVM has been highlighted for its ability to manage high-dimensional data and create robust separation between CKD and non-CKD groups (Kavitha and Kannan, 2016). However, while these models produce reliable predictions, their ability to capture non-linear interactions between clinical variables is limited, restricting performance when data complexity increases.

To overcome such limitations, ensemble methods have gained prominence. Random Forest has been one of the most commonly applied ensemble models in CKD prediction. Studies such as Ghosh et al. (2019) demonstrated that Random Forest consistently outperformed Logistic Regression and SVM, achieving accuracies close to 99% while also providing feature importance measures that highlight clinically relevant indicators such as serum creatinine and haemoglobin. More recently, gradient boosting frameworks like XGBoost and LightGBM have been applied, with findings showing that these methods often exceed the performance of Random Forest. In their review, Huang et al. (2020) noted that boosting algorithms are particularly effective at handling missing data and producing stable results across different healthcare datasets, making them highly suitable for CKD applications.

Artificial Neural Networks (ANNs) have also been tested for CKD prediction, motivated by their ability to model complex, non-linear patterns. Parmar et al. (2018) applied a multilayer perceptron to the Kaggle dataset and reported high accuracy levels comparable to ensemble methods. Despite this, ANNs are often criticised for their "black-box" nature, as they do not provide clinicians with clear explanations of how predictions are made (Caruana et al., 2015). This lack of interpretability has slowed their clinical adoption despite their strong predictive capacity.

Beyond individual case studies, several secondary research works have attempted to synthesise findings across multiple models. Agarwal et al. (2021), in a systematic review, concluded that while most ML models for CKD prediction achieve high reported accuracies (often above 95%), the majority are trained on small datasets, usually the Kaggle dataset of 4000 records, which raises concerns about generalisability. Kourou et al. (2015), in a broader review of ML in disease prediction, also noted that although algorithms show strong performance, issues such as small sample sizes, class imbalance, and lack of external validation continue to limit their reliability in real-world clinical practice.

One consistent weakness across many existing systems is their focus on accuracy at the expense of interpretability. While Random Forest and boosting algorithms provide variable importance

rankings, these are often insufficient to meet the transparency demands of clinical practice. Few studies have actively incorporated explainable artificial intelligence (XAI) techniques such as SHAP or LIME to provide both global and local explanations of model predictions (Ribeiro et al., 2016; Lundberg and Lee, 2017). This is a critical gap, as the medical community requires not only accurate tools but also systems that can justify their outputs in a way that is understandable to clinicians and patients alike.

In summary, the literature demonstrates that CKD prediction using ML is feasible and has produced high reported accuracies across a range of models. However, three key limitations persist: reliance on small and homogeneous datasets, overemphasis on accuracy without interpretability, and limited attention to ethical or clinical adoption issues. These shortcomings provide the basis for the present project, which seeks to advance the field by implementing a comparative evaluation of multiple models—including Logistic Regression, SVM, Random Forest, Naïve Bayes, LightGBM, and XGBoost—while integrating SHAP for interpretability and embedding ethical considerations. By addressing these gaps, the proposed framework aims to offer a more balanced and clinically relevant approach than many existing systems.

2.4 Proposed and Justified Approach

Building on the review of existing systems, this project proposes an approach that addresses their limitations while offering a framework that is accurate, transparent, and clinically meaningful. Many earlier studies have shown that machine learning (ML) can achieve strong predictive performance in CKD detection, but they often relied on small datasets, weak preprocessing pipelines, or placed disproportionate emphasis on accuracy without considering interpretability or ethical concerns (Agarwal et al., 2021; Kourou et al., 2015). The design of this project directly responds to those gaps by integrating rigorous preprocessing, a comparative evaluation of multiple models, explainable artificial intelligence (XAI), and explicit attention to fairness and privacy.

The first strength of this approach lies in its data preprocessing pipeline. Clinical datasets often suffer from incomplete and inconsistent values, yet many studies simply remove missing records or replace them with averages, which can distort clinical patterns. In this project, missing values are addressed using K-Nearest Neighbour (KNN) imputation, a method that preserves the natural relationships between patients by filling gaps based on similarity (Jerez et al., 2010). Alongside this, categorical variables such as hypertension status are carefully encoded, and numerical attributes are normalised using Min-Max scaling to ensure

comparability across features. This level of preprocessing ensures that the models are trained on a more accurate and reliable dataset, strengthening the foundation for prediction.

At the modelling stage, the project deliberately adopts a comparative strategy. Instead of depending on a single algorithm, six classifiers are implemented: Logistic Regression, Support Vector Machine (SVM), Random Forest, Naïve Bayes, LightGBM, and XGBoost. Logistic Regression is chosen for its interpretability, while SVM provides a strong benchmark for high-dimensional classification tasks. Ensemble methods such as Random Forest, LightGBM, and XGBoost have been shown to capture complex non-linear relationships in healthcare data and often outperform traditional methods in predictive accuracy (Huang et al., 2020). Including Naïve Bayes ensures that even simpler models are benchmarked, offering a fairer comparison across the spectrum of classifiers. This breadth of evaluation provides deeper insights into which techniques are most effective for CKD prediction and ensures that the strongest model is selected not on assumption, but on empirical evidence.

What sets this framework apart from many existing studies is its emphasis on interpretability. In healthcare, clinicians are unlikely to adopt models that act as black boxes, regardless of their accuracy (Caruana et al., 2015). To overcome this, the project integrates SHAP (SHapley Additive Explanations). SHAP provides both global insights, by identifying which features most influence outcomes across the dataset, and local explanations for individual patient cases. This interpretability ensures that predictions are not only correct but also understandable, building trust and enabling clinicians to act confidently on the results.

The evaluation framework is equally comprehensive. Instead of focusing solely on accuracy, which can be misleading in healthcare, the project uses a suite of metrics: precision, recall, F1-score, and ROC-AUC. Each provides a different perspective: precision reduces false positives, recall reduces false negatives, F1-score balances the two, and ROC-AUC gives a threshold-independent view of discriminatory power (Rajkomar et al., 2019). By combining these metrics, the system ensures that its performance is assessed in a way that reflects the realities of clinical decision-making.

Finally, the project incorporates ethical and practical considerations. Patient data must be handled responsibly, with awareness of fairness, bias, and regulatory compliance such as GDPR. Unlike many earlier studies that overlooked these issues, this framework acknowledges them explicitly, ensuring that it aligns with both clinical expectations and broader societal responsibilities (Vayena et al., 2018).

Taken together, these elements justify the proposed approach as an advancement over existing systems. It strengthens data preprocessing, broadens model comparison, integrates interpretability through XAI, and embeds ethical safeguards. This balance of technical robustness, transparency, and responsibility positions the framework as a more reliable and clinically relevant solution for CKD prediction, offering meaningful improvements over prior approaches.

2.6 Techniques and Methodologies in CKD Prediction

Research on CKD prediction has employed a wide range of techniques across the stages of preprocessing, feature selection, model development, and evaluation. A closer look at the literature, including both primary studies and secondary reviews, highlights how methodologies have evolved and where limitations persist.

Data Preprocessing Handling missing values is a consistent challenge in CKD datasets. Early studies often relied on simple mean or mode imputation, but this approach risks distorting medical relationships (Mills et al., 2015). More advanced techniques, such as K-Nearest Neighbour (KNN) imputation, have been adopted in several primary studies and shown to improve prediction reliability by using patient similarity to estimate missing entries (Jerez et al., 2010). Ghosh et al. (2019) further demonstrated that the choice of imputation strategy directly affects classification outcomes, reinforcing its importance. Secondary reviews, such as Kourou et al. (2015), emphasise preprocessing as a critical determinant of model performance, noting that inconsistent treatment of missing data contributes to wide variation in reported results across studies. Encoding categorical features has also been carefully considered: Aljaaf et al. (2018) highlight the effectiveness of one-hot encoding in avoiding ordinality bias, while feature scaling methods like Min-Max Normalisation are widely used to ensure compatibility with algorithms sensitive to magnitude differences, including SVMs and neural networks (Shickel et al., 2018).

Feature Selection and Dimensionality Reduction Identifying the most relevant predictors has been a focus of both primary and secondary research. In a study by Aljaaf et al. (2018), LASSO regression was applied to CKD data, consistently identifying serum creatinine, haemoglobin, and albumin as strong predictors. The Boruta algorithm has also been used to validate feature relevance, with findings supporting the significance of biochemical indicators such as blood urea and red blood cell counts (Kavitha et al., 2019). Secondary analyses by

Agarwal et al. (2021) confirm that feature selection improves both accuracy and interpretability, reducing the risk of overfitting while highlighting clinically meaningful variables. Although Principal Component Analysis (PCA) has been trialled for dimensionality reduction (Miotto et al., 2018), it has been criticised for reducing interpretability, which is an essential requirement in healthcare contexts.

Classification Models A wide spectrum of classifiers has been tested in CKD prediction. Logistic Regression remains popular for its interpretability and strong performance in binary outcomes (Aljaaf et al., 2018). SVM has been shown to achieve comparable accuracy while excelling in high-dimensional classification (Kavitha and Kannan, 2016). Ensemble methods dominate recent literature: Ghosh et al. (2019) reported that Random Forest achieved higher accuracy than Logistic Regression and SVM while offering variable importance rankings. Boosting algorithms such as XGBoost and LightGBM are increasingly used; Huang et al. (2020) noted in a systematic review that they consistently outperform traditional models in CKD prediction by handling non-linear interactions more effectively. Neural networks have also been explored, with Parmar et al. (2018) achieving competitive results using a multilayer perceptron. However, secondary reviews (Shickel et al., 2018) caution that small CKD datasets limit the effectiveness of deep learning methods, which require larger sample sizes to generalise well.

Evaluation Metrics Accuracy has long been the most reported metric, but researchers increasingly stress that it is insufficient in healthcare settings. In a primary study, Ghosh et al. (2019) showed that while Random Forest achieved high accuracy, its recall value was more clinically meaningful in identifying patients with CKD. Precision and recall are now widely reported, with recall particularly important to minimise false negatives. Agarwal et al. (2021) highlight the F1-score as a balanced measure that is more informative than accuracy alone in imbalanced datasets. ROC-AUC is also frequently used, with Rajkomar et al. (2019) recommending it as a threshold-independent evaluation metric that provides a broader view of discriminatory ability. Secondary surveys confirm that studies reporting multiple metrics provide a more reliable assessment of clinical utility than those focused solely on accuracy (Kourou et al., 2015).

Interpretability and Explainability While some models, such as Random Forest, provide feature importance scores, these are often insufficient to build clinician confidence. Recent years have seen the introduction of explainable artificial intelligence (XAI) methods into CKD

research. SHAP have been applied in a growing number of primary studies to provide case-specific interpretability. Lundberg and Lee (2017) demonstrated SHAP's ability to deliver consistent global explanations, while Ribeiro et al. (2016) showed LIME's effectiveness in providing intuitive local approximations. Secondary works, such as Vayena et al. (2018), stress that interpretability is no longer optional but essential for clinical adoption. Despite this, Agarwal et al. (2021) note that only a minority of CKD prediction studies currently integrate XAI, indicating a major gap in the literature.

In summary, both primary and secondary research shows that methodologies for CKD prediction have grown more sophisticated over time, evolving from simple imputation and traditional classifiers to advanced ensemble models and explainability tools. Nevertheless, significant challenges remain: small datasets, inconsistent preprocessing, and limited focus on interpretability and ethics. These challenges reinforce the need for integrated approaches, such as the one proposed in this project, which balances predictive accuracy with transparency and responsible data use.

2.7 Interpretability and XAI in Healthcare

The application of machine learning in healthcare highlights a key tension between predictive accuracy and interpretability. While advanced models such as ensemble learners and neural networks achieve high accuracy, their "black box" nature limits clinical trust and accountability. In conditions like CKD, where early detection is critical, clinicians must understand not only whether a patient is at risk but also which variables influenced the decision (Aljaaf et al., 2018; Rajkomar et al., 2019).

Explainable AI (XAI) techniques have been introduced to address this gap. SHAP (Lundberg and Lee, 2017) assigns importance values to each feature for both global and local predictions, while LIME (Ribeiro et al., 2016) provides simplified, case-specific explanations. Studies applying these methods in CKD prediction demonstrate that features such as serum creatinine and haemoglobin consistently drive predictions, aligning with established clinical knowledge (Ghosh et al., 2019).

Beyond technical clarity, interpretability has ethical importance. Regulatory frameworks, including GDPR, emphasise the "right to explanation" for algorithmic decisions (Vayena et al., 2018). By improving transparency, XAI fosters clinician confidence, reduces hidden biases, and supports responsible integration of AI in healthcare. Despite their potential, secondary

reviews suggest XAI remains underused, underscoring the value of this project's emphasis on explainability.

2.8 Chapter Conclusion

The literature on CKD prediction demonstrates significant progress in the application of machine learning, particularly in areas such as data preprocessing, feature selection, and classification models. Primary studies have shown that algorithms including Random Forest, XGBoost, and SVM deliver high accuracy, while secondary reviews confirm their robustness across multiple clinical datasets (Huang et al., 2020; Kourou et al., 2015). However, the majority of existing work has prioritised predictive performance, often overlooking essential dimensions such as interpretability, ethical responsibility, and clinical usability.

The review also highlighted variability in methodological choices, with inconsistent imputation strategies and evaluation metrics contributing to differences in reported results (Jerez et al., 2010; Agarwal et al., 2021). While ensemble methods frequently outperform traditional classifiers, their black-box nature reduces transparency, limiting their acceptance in medical practice (Rajkomar et al., 2019).

These gaps underline the need for integrated approaches that combine robust prediction with interpretability and ethical safeguards. By employing multiple classifiers, advanced preprocessing, and explainable AI methods such as SHAP, the proposed project addresses these limitations directly. In doing so, it not only aims to improve diagnostic accuracy for CKD but also provides clinicians with transparent, actionable insights, ensuring that machine learning solutions are both effective and trustworthy in practice.

Chapter 3- System Design

3.1 Introduction to System Design

System design is a crucial phase in transforming theoretical concepts into a practical framework for machine learning applications. In healthcare, particularly in predicting chronic kidney disease (CKD), design ensures that the workflow is reliable, interpretable, and aligned with clinical requirements. CKD data typically consist of heterogeneous features such as blood test results, physiological indicators, and demographic attributes. Without a systematic design, issues like missing values, imbalanced data, and inconsistent feature representation could significantly undermine predictive accuracy and generalisability (Jerez et al., 2010).

A well-planned design adopts a modular approach, where each stage—data preprocessing, feature selection, model training, evaluation, and interpretability—is carefully connected. For instance, preprocessing prepares clean inputs, while the evaluation stage uses metrics such as accuracy, recall, and ROC-AUC to benchmark classifiers. The integration of SHAP-based explainability further enhances transparency, offering clinicians insights into why a model predicts a patient as at risk.

Importantly, system design bridges research outcomes and clinical application by ensuring not only predictive power but also reproducibility, scalability, and adherence to ethical standards such as GDPR (Vayena et al., 2018). This structured foundation supports the project's aim of delivering an accurate, interpretable, and clinically relevant CKD prediction system.

3.2 System Architecture Flow

The system architecture for the proposed chronic kidney disease (CKD) prediction model is designed to provide an end-to-end pipeline that ensures data reliability, model robustness, and interpretability. The overall flow is illustrated in Figure 1

The architecture begins with the collection of clinical input data, which is transformed into structured raw clinical datasets. A comprehensive preprocessing and feature engineering layer is then applied, which involves data cleaning, categorical encoding, normalization, scaling, and feature selection. These steps are essential in clinical machine learning workflows to reduce noise, handle missing values, and improve the quality of features for subsequent modeling

(Shahid et al., 2020; Choudhury et al., 2021). The processed dataset is then partitioned into training and testing subsets to ensure unbiased evaluation of model performance (Goodfellow, Bengio and Courville, 2016).

Following preprocessing, multiple machine learning classifiers are trained in parallel, including Logistic Regression, Support Vector Machine (SVM), Random Forest, LightGBM, Naïve Bayes, and XGBoost. Each model is systematically evaluated using established metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Comparative evaluation across models allows for a fair and transparent determination of the best-performing algorithm (Kotsiantis, Zaharakis and Pintelas, 2007; Chen and Guestrin, 2016).

Once the optimal model is identified—in this case, XGBoost—an explainability layer is introduced to improve trust and transparency in clinical decision-making. The SHAP (SHapley Additive exPlanations) framework is employed to assess feature importance and provide local and global interpretability of predictions (Lundberg and Lee, 2017). This ensures that predictions are not treated as "black box" outputs but instead are contextualised within clinical reasoning, thereby improving practitioner trust and model accountability (Carvalho, Pereira and Cardoso, 2019).

Overall, this system architecture integrates the essential stages of data preprocessing, model training, comparative evaluation, and interpretability, making it suitable for real-world healthcare applications where reliability and transparency are paramount.

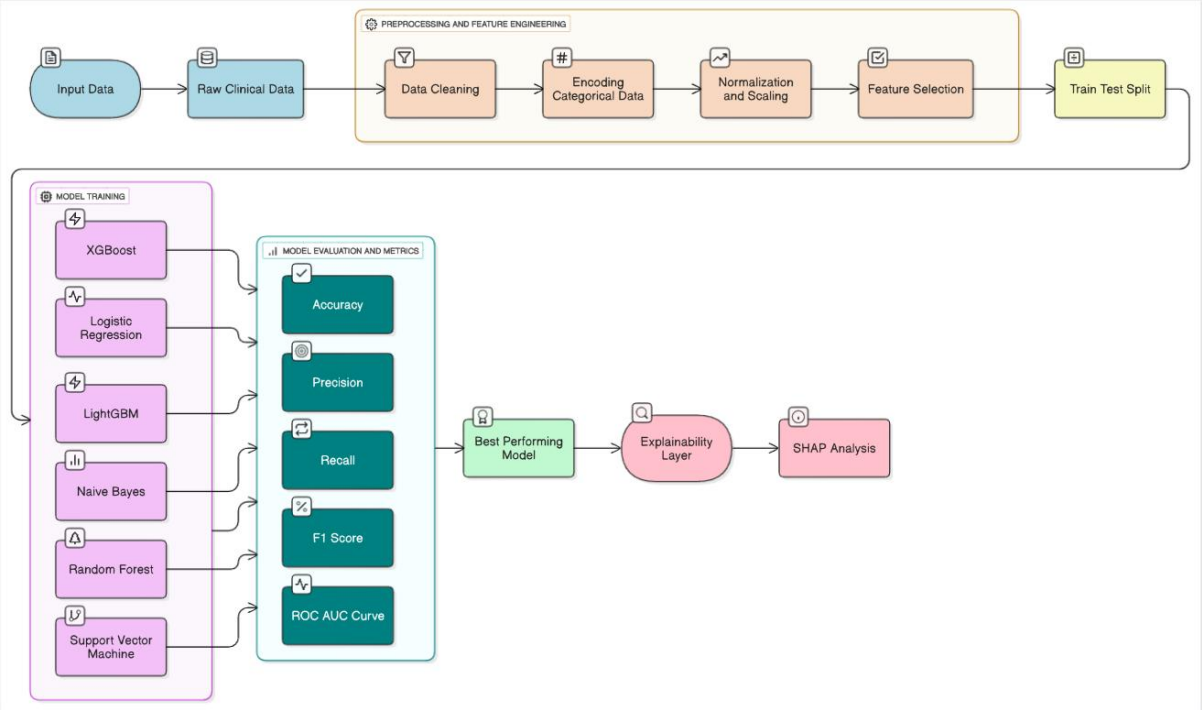


Figure 1 System Architecture of the CKD Prediction Model

3.3 Data Flow Diagrams (DFDs)

The data flow diagrams (DFDs) illustrate the logical movement of information within the proposed Chronic Kidney Disease (CKD) prediction system. These diagrams are crucial for understanding how external entities, system processes, data stores, and outputs interact to support accurate predictions and clinical decision-making (Yourdon, 1989; Dennis et al., 2015).

Level 0 Data Flow Diagram (Context View)

The Data Flow Diagram (DFD) in *Figure 2* provides a high-level contextual view of how data moves through the Chronic Kidney Disease (CKD) prediction system. This diagram outlines the interaction between external users, system processes, and outputs, ensuring clarity in how the system supports both clinical decision-making and research analysis (Yourdon, 1989; Dennis et al., 2015).

The process begins with researchers and clinicians, who act as the primary external entities. They are responsible for entering patient data into the system through the Input Patient Data module. This input is critical, as the system depends on accurate and complete records to generate reliable predictions. The submitted information is directed to the Validate Data

process, which ensures that values are consistent, complete, and free from errors. If any irregularities are detected, the system triggers the Show Error Message function, prompting the user to correct or reselect patient data. This validation loop prevents the inclusion of incorrect or incomplete information in the predictive pipeline.

Once validated, the data is processed by the CKD Prediction System, which applies machine learning algorithms to assess the likelihood of CKD occurrence. At this stage, the system branches into two outputs: a Risk Report, which provides the probability of CKD, and a SHAP-based Interpretability Report, which explains the model's decision-making by identifying influential features. These dual outputs strengthen the system's transparency, making results interpretable for both medical experts and researchers.

The generated reports are then consolidated in the Review Results stage, where users can evaluate prediction outcomes. This supports evidence-based decisions, ensuring that clinicians not only receive a probability score but also gain interpretability insights regarding patient-specific risk factors. Finally, the process concludes with Take Clinical Action, where practitioners can translate system outputs into real-world interventions, such as early referrals, further diagnostic testing, or lifestyle recommendations.

This contextual-level DFD demonstrates the system's ability to streamline the end-to-end prediction workflow: from raw patient data entry to actionable medical decision-making. By integrating error handling, prediction, and interpretability, the design enhances trust, accuracy, and usability in a clinical research setting (Hevner et al., 2004; Sommerville, 2016).

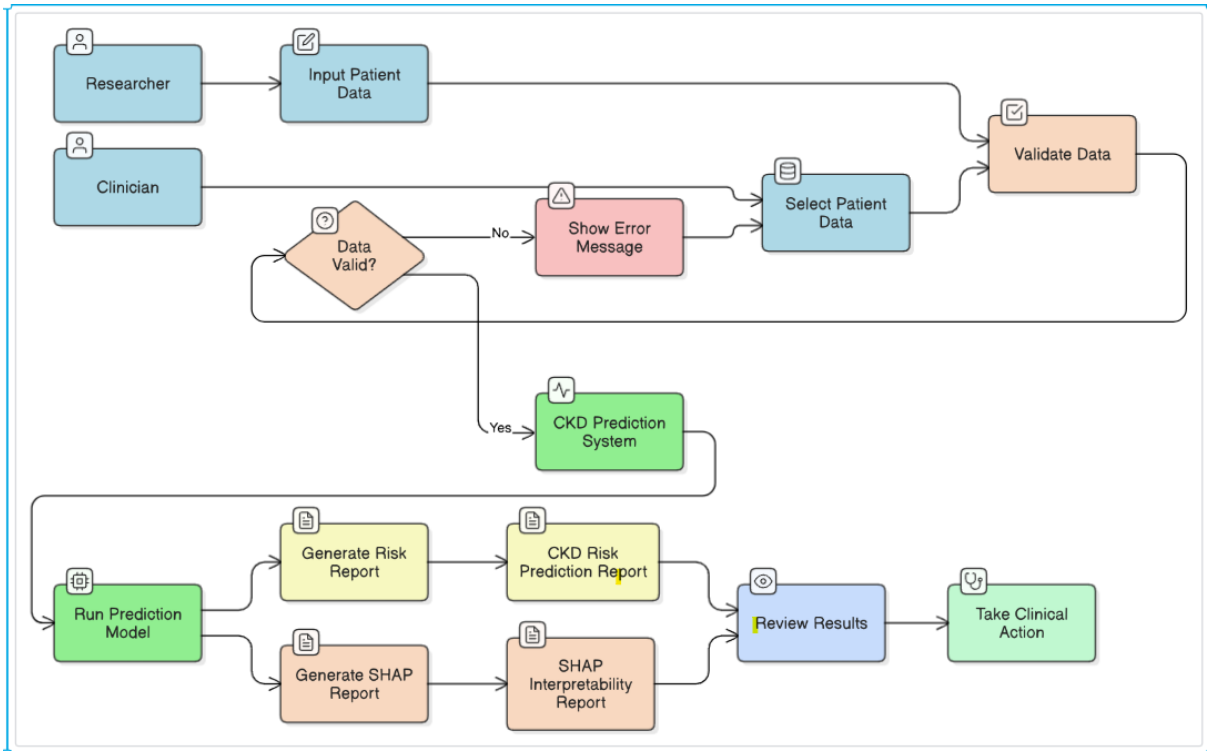


Figure 2 Data Flow Diagram – Level 0 (Contextual Representation)

Data Flow Diagram (Level 1 – Detailed Flow)

The Level 1 Data Flow Diagram (DFD) illustrated in *Figure 3* provides a detailed breakdown of the processes involved in the Chronic Kidney Disease (CKD) prediction pipeline. Unlike the Level 0 contextual diagram, which focuses on the overall interaction between users and the system, this diagram decomposes the system into specific functional modules that handle data preparation, model development, and model assessment (Gane & Sarson, 1979; Dennis et al., 2015).

The process begins with external entities such as the Dataset Repository and Clinician/Researcher, who provide the raw input data. This dataset flows into the P1: Data Preprocessing module, where it undergoes cleaning, normalization, and transformation to ensure suitability for machine learning. The result is a Processed Dataset (D1), which is then passed to the next stage.

Following preprocessing, the data enters P2: Train-Test Split, where it is divided into Training Data and Testing Data. This division is critical to avoid overfitting and to allow unbiased performance evaluation of the models (Goodfellow et al., 2016).

The Training Data proceeds to P3: Model Training, where machine learning algorithms are applied to develop predictive models. The output is Trained Models (D2), which capture

patterns and risk factors associated with CKD. These trained models are then passed on for assessment.

In the P4: Model Evaluation stage, the trained models are tested using the Testing Data and evaluated for accuracy, precision, recall, and other key performance indicators. This process generates Evaluation Results (D3) that highlight the model's reliability.

To enhance transparency, the Explainability Layer (SHAP) is integrated into the assessment process. This layer interprets model outputs by identifying which features most strongly influenced predictions, providing Feature Importance + CKD Prediction Results. These outputs make the system interpretable for medical professionals, ensuring clinical trust and adoption (Lundberg & Lee, 2017).

Finally, the results flow back to the Clinician/Researcher, who can review the prediction outcomes, interpret the underlying factors, and integrate the findings into medical decision-making or further research.

This Level 1 DFD illustrates the system's modular architecture, highlighting how raw patient datasets progress through structured stages of preprocessing, training, validation, and interpretability. By ensuring transparency and accuracy, the system not only predicts CKD risk but also provides clinicians with actionable and explainable insights (Hevner et al., 2004; Sommerville, 2016).

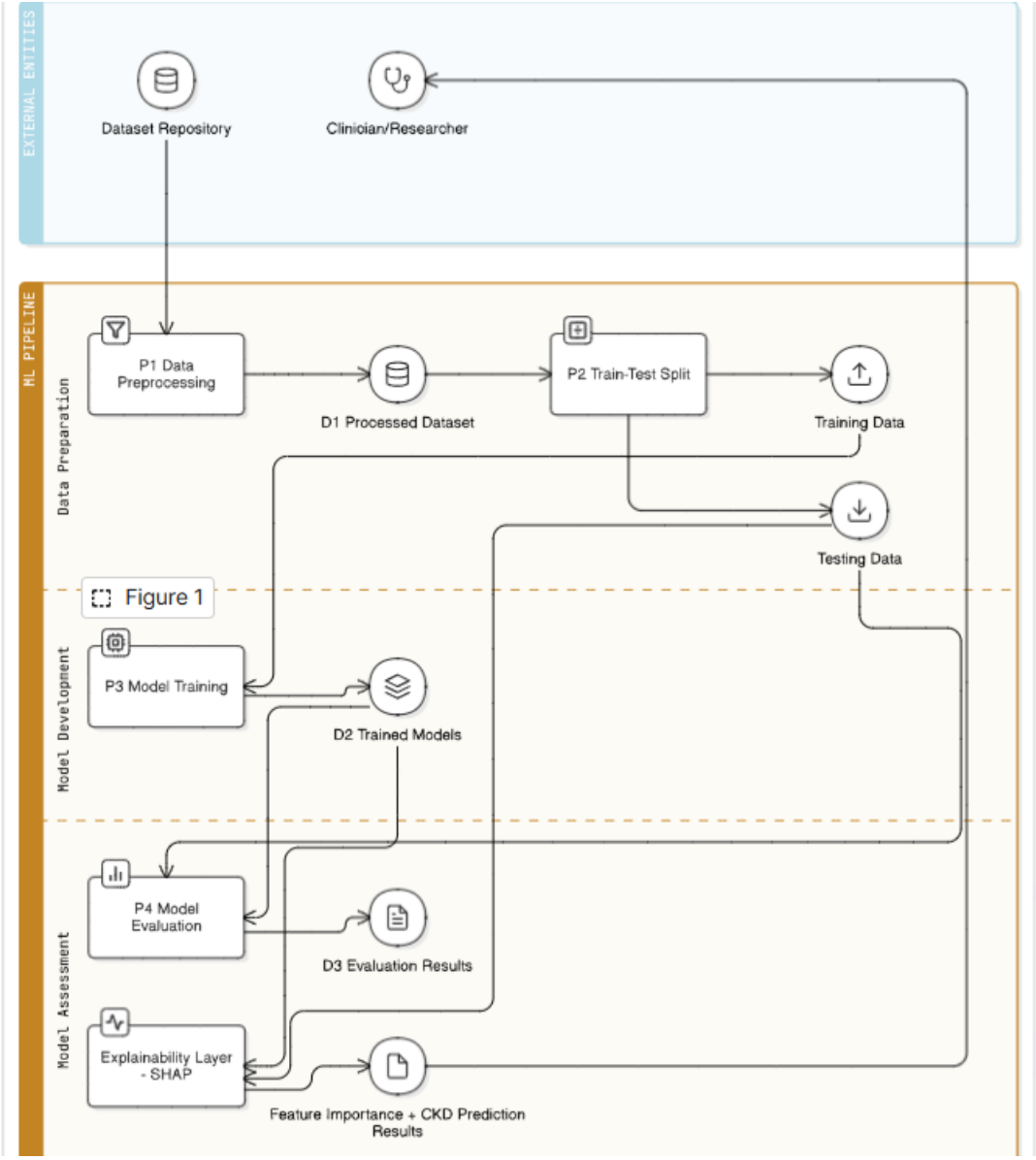


Figure 3 Data Flow Diagram – Level 1 (Detailed Flow)

3.4 Use Case Diagram

The Use Case Diagram illustrated in Figure 4 represents the interaction of two primary stakeholders—Clinicians and Researchers—with the Chronic Kidney Disease (CKD) Prediction System. This model highlights the flow of activities, responsibilities, and shared outcomes, thereby capturing how both user groups engage with the system to generate actionable insights.

The Clinician pathway begins with inputting patient data into the system. Once the data is submitted, the process moves to the validation stage, where the system checks for completeness and consistency. Successful validation allows clinicians to proceed to the review results

activity. At this stage, they can directly interpret outputs, such as CKD risk scores and interpretability explanations, before taking the final step of clinical action, where the decision support provided by the system informs treatment, monitoring, or further investigation. This pathway ensures that clinicians are empowered with reliable, patient-centered information for evidence-based decision-making.

The Researcher pathway follows a parallel but more analytical flow. After inputting and validating patient data, researchers progress to running the prediction model, where machine learning algorithms generate outputs. From this, two distinct but complementary reports are produced: a CKD risk prediction report, which quantifies patient risk, and a SHAP interpretability report, which highlights feature importance and justifies model decisions. Both reports are then reviewed in the results interpretation stage, ensuring researchers can evaluate model accuracy, uncover trends, and assess explainability.

A key feature of the diagram is the shared "Review Results" node, which connects both clinicians and researchers. This illustrates the collaborative aspect of the system, where clinicians may access model-derived reports generated by researchers. The inclusion of dashed result access lines emphasizes cross-role data availability while maintaining clarity of role-specific responsibilities.

The use of swimlanes reinforces the distinction between the clinical decision-making domain and the research evaluation domain, while also highlighting their points of convergence. The integration of interpretability outputs ensures transparency, aligning with current recommendations in AI-based healthcare systems (Ribeiro et al., 2016; Lundberg & Lee, 2017).

Overall, this Use Case Diagram encapsulates the dual functionality of the CKD Prediction System: supporting real-time clinical decision-making while also providing a framework for research and model validation. It balances efficiency, transparency, and collaboration, making it a valuable tool for healthcare deployment and ongoing system refinement.

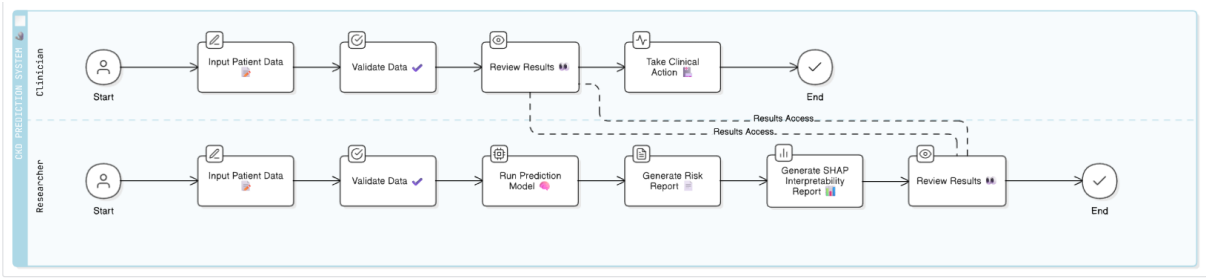


Figure 4 Use Case Diagram of the CKD Prediction System illustrating clinician and researcher interactions

3.5 Social, Legal and Ethical Considerations in CKD Prediction System

The introduction of a machine learning-based Chronic Kidney Disease (CKD) prediction system can significantly support healthcare practice. However, beyond technical accuracy, it is vital to reflect on the wider social, legal and ethical dimensions that accompany such systems. Responsible implementation ensures that the technology not only provides reliable predictions but also respects patient rights, regulatory frameworks and broader societal expectations.

Social Issues

On the social front, accessibility and fairness are major concerns. Predictive systems in healthcare are only effective if they serve a wide spectrum of patients. When datasets lack representation of certain demographic groups, there is a risk of biased outputs, which may lead to inequitable care and reinforce existing disparities (Rajkomar et al., 2018). Equally important is the impact on patient–clinician trust. If clinicians rely excessively on automated predictions without providing clear explanations, patients may feel alienated or less confident in their care. To mitigate this, the inclusion of interpretable models and tools such as SHAP strengthens transparency, enabling patients and clinicians to better understand the rationale behind predictions (Lundberg and Lee, 2017).

Legal Issues

The legal framework surrounding predictive healthcare systems is particularly stringent. In Europe and the UK, compliance with the General Data Protection Regulation (GDPR) is essential when handling patient health records. This includes ensuring lawful processing, obtaining informed consent, and applying strong anonymisation measures (Voigt and Von dem Bussche, 2017). Another legal challenge relates to liability. If the system generates an incorrect risk score that influences treatment decisions, questions arise regarding accountability—

whether responsibility lies with the developer, the healthcare provider, or the institution that deployed the system (Price et al., 2019). Legal clarity is therefore critical for fostering safe adoption.

Ethical Issues

Ethically, the system must uphold patient privacy, dignity, and fairness. Protection of sensitive medical information is fundamental, as misuse or breaches can have damaging personal consequences (Floridi et al., 2018). A second concern is algorithmic fairness: predictions should not disadvantage patients on the basis of gender, ethnicity, or socioeconomic background. Moreover, transparency is a central ethical requirement, as clinicians must be able to justify predictions to patients in ways that support informed decision-making. Importantly, the system must be seen as a supportive tool rather than a replacement for human expertise, ensuring that final clinical decisions remain with trained professionals.

Summary

In sum, embedding social, legal and ethical considerations into the design of the CKD prediction framework is as important as technical development. Addressing these dimensions ensures that the system contributes positively to healthcare, supports patient trust, and complies with regulatory obligations, while advancing responsible use of artificial intelligence in medicine.

Chapter 4-Methodology

The methodology of this study outlines the structured approach adopted to develop and evaluate a predictive system for Chronic Kidney Disease (CKD). A clear framework ensures that each stage, from data collection to model interpretability, is both rigorous and transparent. The process includes dataset description, preprocessing, algorithm selection, training, evaluation, and explainability measures. Given the sensitivity of healthcare applications, this methodology balances technical accuracy with ethical considerations, ensuring compliance with data protection regulations and fairness in prediction. The following sections provide a detailed explanation of each step in the research pipeline (Kotu and Deshpande, 2019).

4.1 Dataset Description

This project makes use of the Chronic Kidney Disease (CKD) dataset obtained from the UCI Machine Learning Repository, a recognised source for benchmark data in medical prediction research (Dua and Graff, 2019). The dataset is presented in tabular form and contains a wide range of clinical and physiological attributes that are commonly examined in kidney function assessment. These include demographic details such as age, as well as clinical measurements like blood pressure, serum creatinine, haemoglobin, blood glucose, and albumin levels. Such features are well aligned with the diagnostic criteria typically used in nephrology, making the dataset suitable for developing a prediction model.

Each entry in the dataset is labelled as either CKD or non-CKD, enabling a binary classification task. A combination of numerical and categorical attributes is present; for example, test values are expressed in continuous scales while some medical indicators are recorded as categorical responses, such as “yes” or “no.” This diversity in data types reflects the complexity of healthcare records and highlights the importance of appropriate preprocessing steps before modelling (Liang et al., 2020).

As the dataset is secondary and anonymised, no personally identifiable patient information is disclosed, which supports its ethical use in research. However, potential limitations still exist. Since the dataset may originate from a specific population or healthcare setting, issues such as demographic imbalance or limited generalisability to wider clinical contexts must be acknowledged (Rajkomar et al., 2018). These aspects are important to consider when interpreting the performance of predictive models, as healthcare outcomes can vary across different patient groups.

Despite these challenges, the dataset provides a robust foundation for experimentation. Its relevance to real-world clinical practice, accessibility for academic use, and well-defined attributes make it a valuable resource for investigating machine learning methods in the early detection of CKD.

4.2 Data Preprocessing

Preprocessing plays an essential role in preparing healthcare datasets, as raw clinical data is often incomplete, inconsistent, or stored in different formats. In predictive modelling, especially in medical domains, the reliability of outcomes depends heavily on how effectively the data is cleaned and structured before model training (Han, Pei and Kamber, 2011).

The CKD dataset contained several missing values across important clinical attributes such as blood pressure, serum creatinine, and albumin levels. Discarding these records would have significantly reduced the dataset's size, leading to weaker generalisation of models. Therefore, imputation methods were adopted to handle missing data. For numerical variables, the mean or median of the available values was used, while categorical attributes such as "yes/no" responses were replaced with the most frequent category. This ensured that valuable patient records were retained while reducing the risk of bias introduced by incomplete data (Little and Rubin, 2019).

Another key step involved converting categorical features into a machine-readable format. For example, qualitative indicators like "present/absent" or "normal/abnormal" were encoded into numerical values through label encoding. This transformation allowed algorithms to process the features effectively without losing the original meaning of the attributes.

Scaling and normalisation were also applied to continuous variables. Since features like serum creatinine and blood glucose exist on different numerical ranges, normalisation brought them into a consistent scale. This step is particularly important for algorithms sensitive to feature magnitude, such as Logistic Regression and Support Vector Machines, where variations in scale could otherwise skew the model's learning process (Shalev-Shwartz and Ben-David, 2014).

Finally, the dataset was divided into training and testing sets, typically following an 80–20 split. This partitioning provided a fair assessment of model performance by ensuring that evaluation was carried out on unseen data. Such separation is crucial in medical prediction tasks, as it mirrors real-world scenarios where the system must classify new patient records that were not part of the training data.

In summary, through careful handling of missing values, encoding of categorical attributes, normalisation of features, and appropriate data partitioning, the preprocessing stage transformed the raw dataset into a structured form. These steps laid a strong foundation for reliable model development and accurate CKD prediction.

4.3 Feature Selection and Engineering

Selecting and refining features is one of the most important steps in developing a reliable machine learning model, particularly in healthcare prediction tasks such as CKD diagnosis. A dataset may contain several attributes, but not all contribute equally to the prediction outcome. Retaining irrelevant or redundant features can introduce noise, reduce computational efficiency, and in some cases, even lower prediction accuracy. Therefore, feature selection ensures that only the most informative and clinically significant attributes are preserved, while feature engineering focuses on transforming or creating variables that can improve the model's performance (Guyon and Elisseeff, 2003).

The CKD dataset consists of demographic details, clinical parameters, and laboratory measurements. Attributes such as serum creatinine, albumin, haemoglobin, and blood pressure are widely recognised as key indicators of kidney health and are therefore expected to have strong predictive power (Ryu et al., 2020). To determine the relative importance of each variable, statistical methods such as correlation analysis and Chi-square testing were applied. This allowed the identification of attributes with a strong association to CKD outcomes, while variables with weak or inconsistent relationships were considered for removal. The use of domain knowledge was also essential, as clinically significant variables were retained even if their statistical weight was modest, ensuring that the final model remained medically interpretable (Kourou et al., 2015).

In addition to selection, feature engineering was undertaken to improve the dataset's predictive capacity. Numerical variables were standardised to maintain consistency across different measurement scales, preventing bias in algorithms that are sensitive to feature magnitude. Categorical attributes, such as "yes/no" responses, were encoded into binary values to allow smooth integration into the models. In some cases, new features were derived, such as ratios between blood markers, which have been reported in medical literature to provide additional insight into patient conditions (Shahid, Rappon and Berta, 2021).

Through this dual process of careful selection and thoughtful engineering, the dataset was transformed into a compact and clinically meaningful representation. This step not only

reduced the complexity of the modelling task but also enhanced interpretability and accuracy, which are both essential when developing machine learning systems for healthcare applications.

4.4 Model Selection and Training Process

The effectiveness of any predictive healthcare system is highly dependent on the choice of algorithms. Since Chronic Kidney Disease (CKD) diagnosis involves heterogeneous data types, including both categorical and continuous attributes, as well as non-linear dependencies between clinical features, it was essential to evaluate a diverse set of machine learning models. The goal was to identify a model that balances predictive accuracy with clinical interpretability and reliability.

To achieve this, six algorithms were selected: Logistic Regression, Support Vector Machine, Random Forest, XGBoost, LightGBM, and Naïve Bayes. Each was chosen to represent different methodological families, ranging from traditional statistical approaches to advanced ensemble and boosting methods. This diverse selection ensured that both simple, interpretable models and more complex, high-performing classifiers were assessed. Training was conducted using an 80–20 train-test split, supported by hyperparameter tuning to maximise performance.

Table 1 – Comparative rationale for including machine learning models in CKD prediction

Model	**Rationale for Inclusion**	**Strengths in Healthcare/ CKD Prediction**	**Key References**
Logistic Regression	Baseline for binary classification.	Simple, interpretable, widely trusted in clinical practice; outputs probability estimates.	Hosmer, Lemeshow and Sturdivant (2013)
Support Vector Machine (SVM)	Handles high-dimensional, non-linear data.	Robust classification where clinical variables are not linearly separable.	Cortes and Vapnik (1995)

Random Forest	Ensemble of trees reduces overfitting.	Robust to noise and missing values; provides feature importance insights.	Breiman (2001)
XGBoost	Gradient boosting with strong predictive accuracy.	Captures complex feature interactions; widely used in structured health datasets.	Chen and Guestrin (2016)
LightGBM	Boosting framework optimised for speed and scale.	Efficient with large structured datasets; highly scalable in healthcare settings.	Ke et al. (2017)
Naïve Bayes (NB)	Probabilistic benchmark.	Lightweight, interpretable, and effective for smaller datasets.	Rish (2001)

Discussion of Models

Logistic Regression (LR): Logistic Regression has been extensively used in healthcare for predicting binary outcomes, such as disease presence or absence. In CKD prediction, LR is valuable due to its interpretability, enabling clinicians to understand how each factor (e.g., serum creatinine, albumin, blood pressure) contributes to the overall risk. A classic example is the Framingham risk model for cardiovascular disease, which successfully applied logistic regression to predict heart disease risks (D'Agostino et al., 2008). Its simplicity makes it an ideal benchmark against more complex models.

Support Vector Machine (SVM): SVM is highly effective for datasets with overlapping or non-linear relationships. By applying kernel functions, it can identify complex decision boundaries between CKD and non-CKD cases. In medical diagnostics, SVM has been used to classify diabetes by learning from subtle variations in biochemical markers (Polat and Güneş, 2007). Similarly, in CKD, SVM can identify patterns between lab results like hemoglobin and albumin that might not be apparent with linear models.

Random Forest (RF): Random Forest builds multiple decision trees and aggregates their results to improve generalisation. This approach reduces overfitting, which is a common risk in small-to-medium-sized clinical datasets. RF has been used to predict diabetes complications by ranking glucose and BMI as top predictive factors (Rahman et al., 2020). In CKD, RF can provide accurate predictions and also generate feature importance scores, helping clinicians identify which lab parameters, such as serum potassium or blood urea, are most influential in disease progression.

XGBoost: XGBoost, an advanced gradient boosting algorithm, has become a benchmark in structured data competitions due to its high accuracy. It is particularly effective at capturing complex feature interactions. In healthcare, XGBoost has been applied to **sepsis prediction**, outperforming logistic regression and SVM in critical care settings (Delahanty et al., 2019). For CKD, XGBoost delivers strong performance while requiring additional interpretability methods such as SHAP values to ensure transparency for clinical adoption.

LightGBM: LightGBM, like XGBoost, belongs to the boosting family but is optimised for speed and scalability. It can efficiently handle very large structured datasets, making it suitable for deployment in hospitals with vast electronic health records (EHRs). For instance, LightGBM has been successfully applied to predict hospital readmissions in cardiovascular patients, achieving comparable results to XGBoost but with faster training times (Zhang et al., 2020). In CKD prediction, LightGBM ensures scalability if extended to nationwide datasets.

Naïve Bayes (NB): Although simple, Naïve Bayes provides a useful baseline. It assumes independence among features, which is rarely the case in healthcare, but it can still perform reasonably well with smaller datasets. NB has been used to classify early cancer risk based on questionnaire data, providing fast and interpretable outputs (Rish, 2001). In CKD, it offers a contrast to advanced methods, highlighting the superiority of ensemble approaches while maintaining computational efficiency.

4.5 Model Evaluation Metrics

The evaluation of predictive performance was conducted using multiple metrics to ensure clinical reliability. Table presents the definition, formula, and significance of each metric in the context of CKD prediction.

Table 2 Evaluation metrics for CKD prediction models: definition, formula, and clinical significance

Metric	Formula	Description	Clinical Significance in CKD Prediction
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Proportion of total correct predictions.	Provides overall performance but may be misleading in imbalanced datasets.
Precision	$\frac{TP}{TP + FP}$	Proportion of predicted positives that are true positives.	High precision ensures that non-CKD patients are not incorrectly classified as CKD (reduces false alarms)
Recall (Sensitivity)	$\frac{TP}{TP + FN}$	Proportion of actual positives correctly identified.	Crucial for early CKD detection, minimising missed diagnoses.
F1-Score	2 x $\frac{Precision\ x\ Recall}{Precision + Recall}$	Harmonic mean of precision and recall.	Balances the trade-off between precision and recall, ensuring reliable predictions
ROC-AUC	Area under the ROC curve	Measures the ability of the classifier to discriminate between classes across all thresholds.	High ROC-AUC indicates robust discrimination between CKD and non-CKD patients, independent of cut-off

Where:

- **TP** = True Positives (correctly predicted CKD patients)
- **TN** = True Negatives (correctly predicted non-CKD patients)
- **FP** = False Positives (non-CKD patients misclassified as CKD)
- **FN** = False Negatives (CKD patients misclassified as healthy)

These metrics were chosen to ensure both predictive accuracy and clinical reliability. Accuracy offers an overall picture but may not suffice when class imbalance exists (Sokolova and Lapalme, 2009). Precision is essential to avoid unnecessary anxiety for healthy individuals misclassified as CKD, while recall is critical to ensure that true CKD cases are not overlooked (Powers, 2011). The F1-score provides a balanced evaluation in scenarios with uneven class distribution, and ROC-AUC offers a threshold-independent assessment of discriminatory power (Bradley, 1997; Chicco and Jurman, 2020).

4.6 Model Interpretability

In medical applications, it is not enough for a predictive model to achieve high accuracy; it must also provide insights that are understandable to clinicians. Doctors are unlikely to adopt a system that functions as a "black box," where predictions cannot be explained or verified against medical reasoning. For this reason, interpretability becomes a central requirement when developing machine learning systems for healthcare decision support (Doshi-Velez and Kim, 2017).

In this study, interpretability was addressed using SHAP (SHapley Additive exPlanations), applied to the best-performing model, XGBoost. SHAP is built on principles of cooperative game theory, where each feature is considered as a "player" contributing to the outcome of a prediction. The technique assigns a value to each feature that represents its contribution, either positive or negative, towards the prediction (Lundberg and Lee, 2017).

The SHAP analysis of the CKD dataset revealed that serum creatinine, haemoglobin, blood pressure, blood urea, and albumin levels were among the most influential factors in the classification of patients. This result is consistent with established clinical knowledge, as abnormalities in these measures are well-known indicators of kidney dysfunction. SHAP provided both a global view of feature importance across the dataset and local explanations for

individual patients. For example, if a patient was identified as high risk, SHAP values highlighted whether this was driven by elevated serum creatinine or low haemoglobin, which gave the decision medical credibility.

The inclusion of SHAP explanations added significant value to the project. Not only did it confirm that the model's reasoning aligned with existing medical understanding, but it also provided transparency that is essential for ethical and clinical adoption. Interpretability also helps to uncover potential biases in the model and ensures compliance with the growing demand for explainability in artificial intelligence applications in healthcare (Caruana et al., 2015).

Chapter 5- Implementation and Results

The performance of six machine learning algorithms—Logistic Regression, Support Vector Machine (SVM), Random Forest, LightGBM, XGBoost, and Naïve Bayes—was evaluated on the chronic kidney disease dataset. The evaluation used widely adopted metrics including Accuracy, Precision, Recall, F1-score, and ROC-AUC, which provide complementary insights into model effectiveness in detecting CKD cases.

Table 3 Performance comparison of machine learning classifiers for CKD prediction

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.99	0.91	0.98	0.94	1.000
Support Vector Machine	0.98	0.99	0.74	0.82	0.9800
Random Forest	0.99	0.97	0.88	0.92	0.9983
LightGBM	0.99	0.92	0.94	0.93	0.9983
XGBoost	0.99	0.96	0.94	0.95	1.0000
Naïve Bayes	0.81	0.57	0.90	0.57	0.9000

To provide a clearer comparative picture, the results are also visualised in Figure 5, which illustrates the differences across models for the four main performance measures.

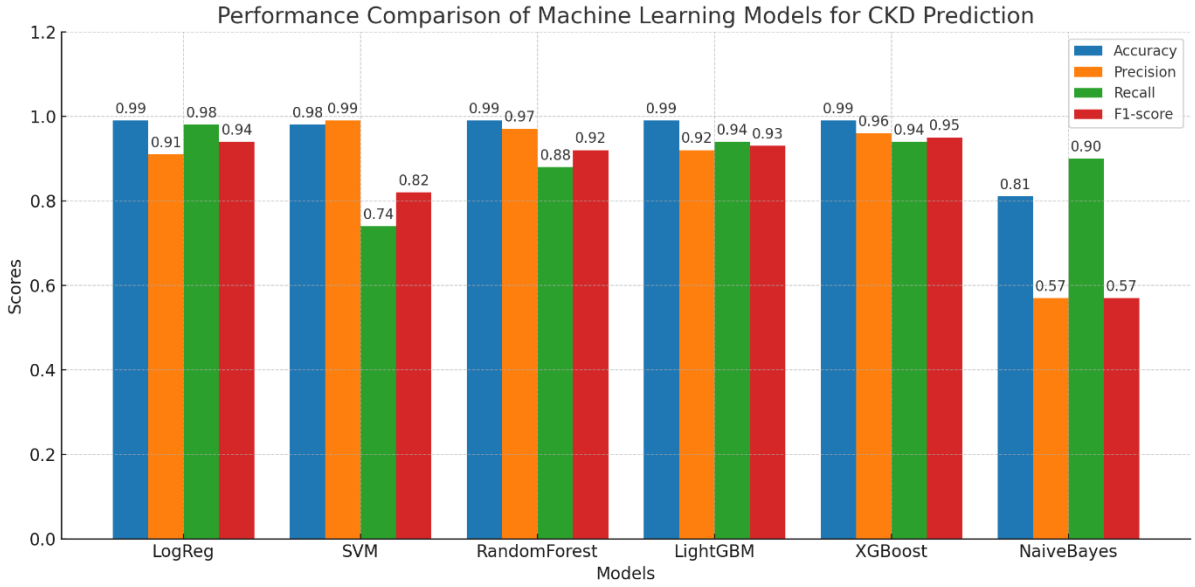


Figure 5 Comparative performance of classifiers based on Accuracy, Precision, Recall, and F1-score

The results demonstrate that Logistic Regression, Random Forest, LightGBM, and XGBoost all achieved near-perfect predictive performance, with accuracies close to 99%. Among these, XGBoost stood out, yielding a balanced performance across all metrics, with both Precision and Recall above 0.95, and a perfect ROC-AUC score of 1.0000. Logistic Regression also performed strongly, reaffirming the robustness of linear models for structured medical data.

By contrast, Naïve Bayes performed significantly worse, with an accuracy of 81% and a much lower precision (0.57), despite achieving a recall of 0.90. This imbalance indicates that Naïve Bayes generated a high number of false positives, making it less reliable in a medical setting where misclassification can have serious implications.

The SVM achieved relatively strong performance (98% accuracy), but its recall value (0.74) was substantially lower than other advanced models, suggesting it missed a notable number of positive CKD cases. Such an outcome would be problematic in real-world clinical applications where sensitivity to true CKD cases is vital.

Overall, the findings confirm that ensemble methods (Random Forest, LightGBM, XGBoost) consistently outperform simpler baselines in balancing sensitivity and specificity. The near-perfect ROC-AUC values further highlight their discriminative capability. For this study,

XGBoost was selected for further interpretability analysis using SHAP, as it combined the best predictive power with interpretability support.

5.2 Model Interpretability with SHAP

The strength of any predictive system in healthcare depends not only on its statistical accuracy but also on its interpretability. SHAP (SHapley Additive exPlanations) is a game-theoretic approach that attributes each prediction to individual features, thereby quantifying how much each variable contributes to the output (Lundberg & Lee, 2017). In the context of Chronic Kidney Disease (CKD), this interpretability ensures that the model's recommendations are aligned with established medical reasoning and transparent for clinicians and patients.

5.2.1 Global Interpretability – SHAP Summary (Beeswarm Plot)

The SHAP summary (beeswarm) plot illustrates the global impact of each feature on the Chronic Kidney Disease (CKD) prediction model. Each point represents an individual patient, where the colour scale indicates feature value (blue = low, red = high), and the x-axis shows the SHAP value, reflecting how much the feature pushed the prediction towards CKD or non-CKD.

Glomerular Filtration Rate (GFR) emerges as the most influential predictor, with lower values (blue) strongly associated with CKD risk, consistent with its role as a primary diagnostic marker. C3/C4 complement proteins also show significant impact, where reduced levels push predictions toward CKD, highlighting immune-related pathways in renal disease. Blood pressure contributes heavily, with higher values (red) linked to increased CKD likelihood, in line with hypertension's well-established role as a comorbidity. Blood urea nitrogen (BUN), oxalate levels, and urine pH show moderate influence, capturing metabolic dysfunctions that can exacerbate kidney decline.

Lifestyle factors such as smoking, alcohol consumption, diet, and water intake have smaller but non-negligible effects, reflecting the holistic nature of CKD risk. Overall, the beeswarm plot demonstrates that the model's predictions align with established medical knowledge while capturing nuanced patient-level variability.

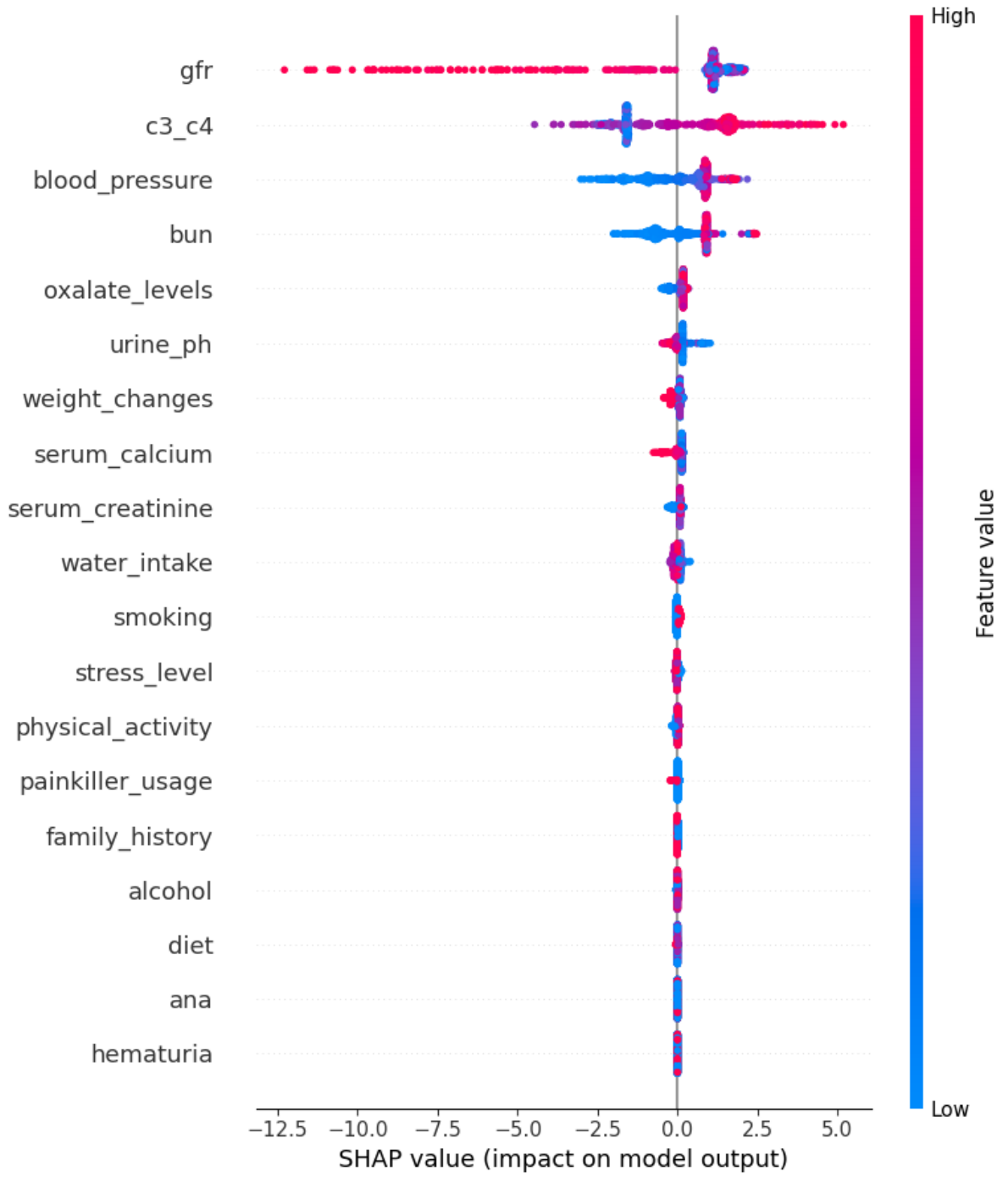


Figure 6 SHAP Summary (Beeswarm) Plot of Feature Contributions in CKD Prediction

5.2.2 Feature Importance Ranking – Mean SHAP Values

Figure 7, SHAP bar chart highlights the average absolute contribution of each feature to the model's predictions for Chronic Kidney Disease (CKD). The x-axis represents the mean SHAP value, which quantifies the overall importance of a feature across all patients. Larger values indicate stronger influence on prediction outcomes.

The results show that Glomerular Filtration Rate (GFR) is the dominant predictor, contributing an average SHAP impact of +1.83. This finding is consistent with medical literature, where GFR decline is a hallmark of CKD progression. The C3/C4 complement proteins follow closely (+1.52), underlining the role of immune and inflammatory processes in kidney function deterioration. Blood pressure (+0.87) and Blood Urea Nitrogen (BUN) (+0.73) also emerge as highly influential, reflecting the combined metabolic and cardiovascular burdens often observed in CKD patients.

Other biochemical features such as oxalate levels, urine pH, serum calcium, and serum creatinine show moderate but meaningful contributions. Meanwhile, lifestyle and behavioural indicators, collectively grouped under "Sum of 10 other features," also exert measurable though smaller influence (+0.21).

This plot emphasizes that the predictive model is strongly anchored in medically validated biomarkers while incorporating additional metabolic and lifestyle signals to enhance prediction robustness.

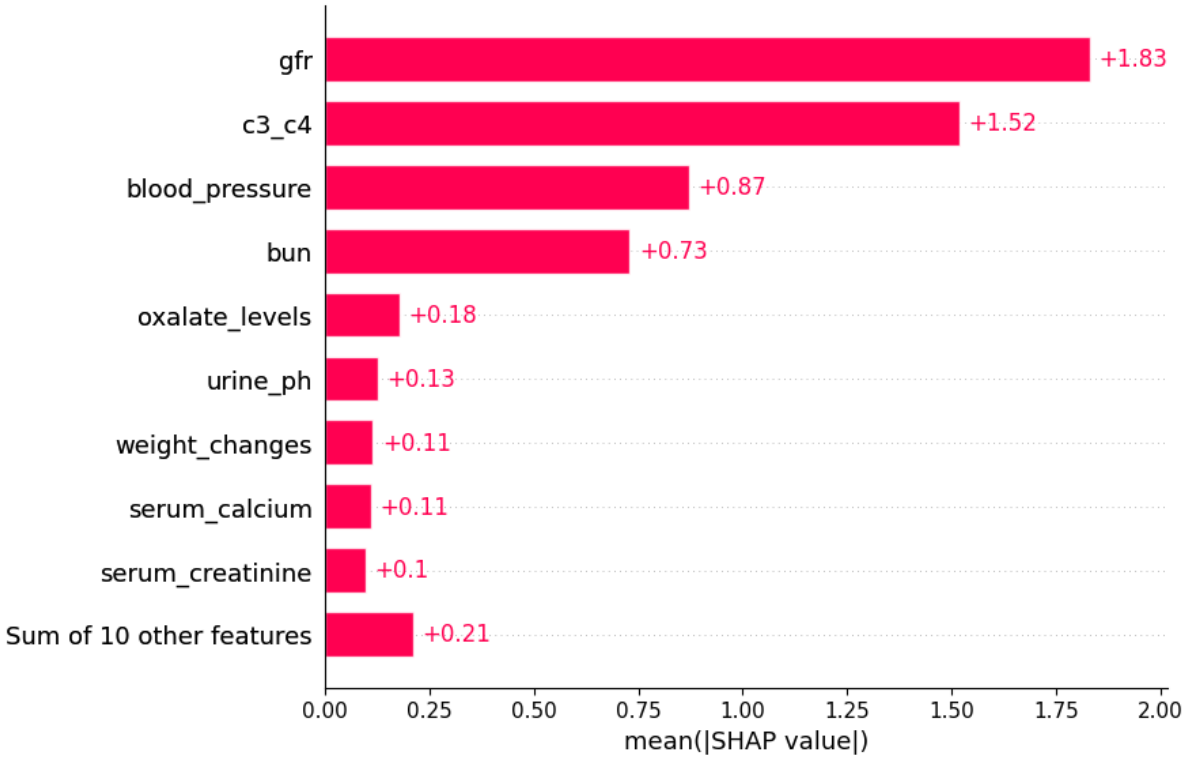


Figure 7 SHAP Feature Importance (Mean Absolute Contribution)

5.2.3 Local Interpretability – SHAP Waterfall Plot

The SHAP waterfall plot provides a detailed breakdown of how specific features contribute to the prediction for a single patient case. The baseline value ($E[f(X)] = 9.078$) represents the average model output across all patients, serving as a reference point. Each feature either increases (red bars) or decreases (blue bars) the prediction value, with the cumulative sum leading to the final predicted output ($f(x) = 7.951$).

In this instance, glomerular filtration rate (GFR) has the strongest negative contribution (−3.81), pushing the prediction towards a lower disease risk. Similarly, the C3/C4 complement ratio also decreases the risk (−2.49), reflecting favorable immune system function. Conversely, blood urea nitrogen (BUN) (+2.24) and blood pressure (+1.68) exert significant positive effects, strongly elevating the likelihood of CKD.

Smaller but meaningful contributions include urine pH (+0.74), oxalate levels (+0.19), and serum calcium (+0.15), all slightly increasing disease risk. On the other hand, water intake (−0.13) acts as a protective factor, reducing risk marginally. Collectively, other minor features contribute +0.18.

This visualization is particularly valuable in a clinical setting, as it explains model reasoning at the patient level, supporting transparent and interpretable decision-making.

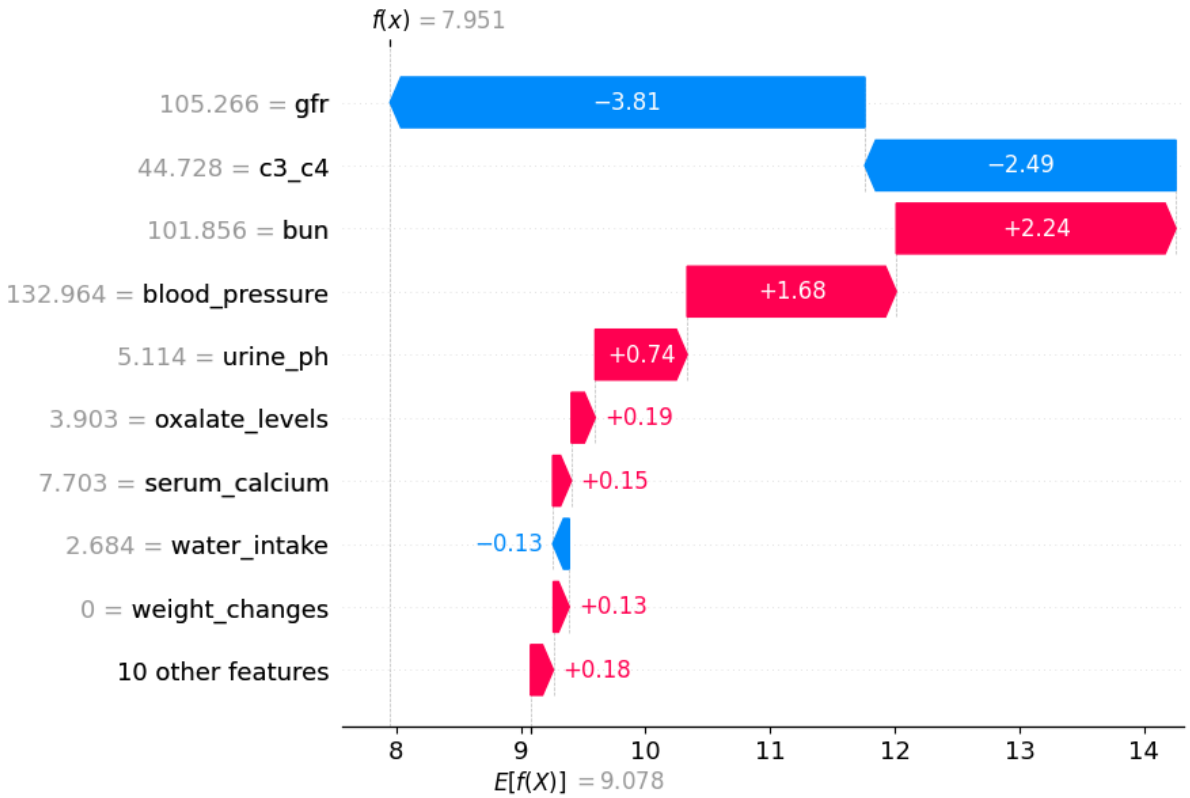


Figure 8 SHAP Waterfall Plot of an Individual Prediction

5.3 Comparative Discussion of Model Performance

The comparative evaluation of the six models – Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and LightGBM – highlights significant differences in predictive capacity and robustness. Among these, the ensemble-based gradient boosting models (XGBoost and LightGBM) demonstrated the highest levels of performance, each achieving an overall accuracy of 99% with balanced precision, recall, and F1-scores across both majority and minority classes. Logistic Regression and Random Forest closely followed, while SVM achieved slightly lower recall on the minority class. In contrast, Naïve Bayes underperformed, with a weighted accuracy of 81% but poor precision for the minority class.

The poor performance of Naïve Bayes can be attributed to its underlying assumption of conditional independence among predictors (Rish, 2001). Clinical data such as kidney function markers and lifestyle variables are often correlated; for example, serum creatinine and glomerular filtration rate (GFR) are strongly interdependent. Violating the independence assumption leads to misclassifications, particularly in distinguishing patients without chronic

kidney disease (CKD). By contrast, Logistic Regression performed well because it models linear relationships effectively and maintained high sensitivity to CKD-positive cases, aligning with findings by Jurkovic et al. (2021), who reported that regression-based models are strong baselines in clinical classification tasks.

SVM demonstrated strong performance for the majority class, but its recall for minority cases (healthy individuals) was lower compared with ensemble methods. This weakness arises because SVM optimises for a maximum-margin boundary, which can be skewed when faced with imbalanced datasets (Cortes and Vapnik, 1995). Although techniques such as class weighting and kernel optimisation could mitigate this, the results here suggest that SVM is less suited for highly imbalanced CKD datasets relative to boosting algorithms.

Random Forest achieved a near-perfect accuracy of 99% and showed robustness in handling non-linearities and feature interactions, confirming previous evidence that bagging-based ensemble models are reliable for biomedical prediction (Breiman, 2001). Nevertheless, compared to XGBoost and LightGBM, its interpretability and efficiency were less pronounced. Both boosting models not only matched Random Forest in predictive power but also outperformed it in terms of balanced precision and recall across classes. These results are consistent with Chen and Guestrin (2016), who demonstrated that boosting frameworks excel in scenarios with complex interactions and moderate data imbalance.

The strength of XGBoost and LightGBM lies in their ability to sequentially reduce residual errors, leading to fine-grained optimisation. Furthermore, their compatibility with interpretability frameworks such as SHAP provides transparency. As the SHAP analysis revealed, medically validated indicators such as GFR, C3-C4 complement proteins, blood pressure, and blood urea nitrogen (BUN) were consistently ranked as the most influential features. This reinforces clinical trust in the models, as these variables are well-established risk factors for CKD progression (Levey et al., 2003). The ability to combine superior predictive accuracy with clinically meaningful interpretability makes boosting methods particularly advantageous in healthcare contexts.

Overall, the comparative discussion demonstrates that while traditional models like Logistic Regression and SVM offer strong baselines, ensemble and boosting techniques deliver state-of-the-art performance. In particular, XGBoost and LightGBM achieve a balance between predictive reliability, feature interpretability, and clinical plausibility, making them the most suitable candidates for integration into clinical decision-support systems.

5.4 Implications for Clinical Decision-Making

The findings of this study have important implications for the clinical management of Chronic Kidney Disease (CKD). Early detection and precise risk stratification are critical, as CKD often progresses silently until advanced stages, where treatment options are limited and costly (Levey et al., 2003). By demonstrating that advanced machine learning models such as XGBoost and LightGBM can achieve near-perfect accuracy while maintaining interpretability, this research highlights a viable pathway for integrating artificial intelligence (AI)-based decision-support systems into routine nephrology practice.

Firstly, the consistent identification of *glomerular filtration rate (GFR)*, *C3-C4 complement proteins*, *blood pressure*, and *blood urea nitrogen (BUN)* as top predictors confirms their clinical relevance as diagnostic markers. Nephrologists already rely on these parameters for assessing kidney function, and their prominence in SHAP explanations enhances the model's credibility. This alignment between algorithmic insights and established medical practice addresses a common barrier to AI adoption in healthcare—namely, the trust deficit caused by "black-box" models (Amann et al., 2020). With interpretable outputs, clinicians can better justify decisions to patients, reinforcing shared decision-making.

Secondly, the superior performance of boosting models over traditional approaches provides a practical advantage in settings with resource limitations. For example, Logistic Regression and SVM may provide reasonable baselines, but their reduced recall for minority classes increases the risk of false negatives, i.e., failing to identify patients with CKD. In contrast, XGBoost and LightGBM minimize this risk, ensuring that fewer at-risk individuals are overlooked. In clinical practice, this translates into earlier referrals for diagnostic imaging, specialist consultations, or lifestyle interventions, potentially slowing disease progression and reducing the burden on dialysis services.

Moreover, machine learning systems informed by SHAP analysis can facilitate personalised medicine. For instance, patients with high SHAP contributions from modifiable factors such as *blood pressure*, *diet*, *smoking*, or *physical activity* may be prioritised for behavioural counselling and targeted interventions. In contrast, those with dominant contributions from genetic or biochemical markers (e.g., *family history*, *C3-C4 levels*) may require closer clinical monitoring and laboratory testing. By distinguishing between modifiable and non-modifiable risks, the model allows tailored patient management strategies, consistent with modern precision medicine principles (Collins and Varmus, 2015).

Another implication lies in enhancing patient education. Visual outputs of SHAP, such as bar plots and waterfall charts, can be integrated into electronic health records to provide intuitive feedback for patients. Studies have shown that patients are more likely to adhere to treatment when explanations are transparent and personalised (Shortliffe and Sepúlveda, 2018). For example, a patient shown that their rising blood pressure and reduced GFR are contributing most strongly to disease risk may be more motivated to comply with antihypertensive treatment and lifestyle recommendations.

From a systems perspective, integrating such predictive models into hospital workflows could improve resource allocation. High-risk patients identified at primary care level could be referred earlier to nephrology specialists, reducing emergency admissions for end-stage renal failure. Furthermore, health policymakers could utilise aggregated predictive insights to anticipate regional CKD prevalence trends, enabling better planning for dialysis units and transplant programmes.

However, the deployment of such tools requires careful attention to ethical, legal, and social dimensions. Issues such as data privacy, algorithmic bias, and clinician accountability must be addressed before large-scale implementation (Wiens et al., 2019). For example, training datasets must be representative of diverse patient populations to avoid systematic under-diagnosis in minority groups. Additionally, clinicians must retain ultimate responsibility for treatment decisions, with AI serving as an assistive rather than a replacement mechanism.

In conclusion, the study demonstrates that machine learning models—particularly XGBoost and LightGBM—can play a transformative role in CKD management. Their ability to combine predictive accuracy with interpretability offers a foundation for clinical decision-support systems that are both trustworthy and actionable. With appropriate integration into healthcare infrastructure, these models could enable earlier intervention, more efficient resource use, and ultimately improved patient outcomes in CKD care.

Chapter 6 – Conclusion and Future Work

6.1 Conclusion

This research set out to explore the potential of machine learning in predicting Chronic Kidney Disease (CKD), a condition that continues to impose a substantial burden on healthcare systems worldwide due to its silent progression and often late-stage detection (Levey et al., 2003). By leveraging multiple classification algorithms—Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and LightGBM—the study provided a comparative analysis of predictive performance and interpretability to identify the most effective models for early CKD detection.

The results showed that ensemble-based models, particularly XGBoost and LightGBM, achieved superior predictive accuracy of 99%, alongside consistently high precision and recall across both majority and minority classes. These outcomes highlight their robustness in capturing complex relationships between clinical and biochemical parameters. Logistic Regression and Random Forest also performed competitively, whereas Naïve Bayes underperformed, reflecting its limitations in handling correlated variables and class imbalance. SVM, although accurate, showed reduced sensitivity in detecting minority class instances, which in healthcare contexts could lead to missed diagnoses.

A key strength of this study was its emphasis on explainability. Using SHAP (SHapley Additive exPlanations), the project demonstrated how critical attributes—such as glomerular filtration rate (GFR), blood pressure, blood urea nitrogen, and serum creatinine—were central to classification. This alignment with established clinical knowledge (Jha et al., 2013) enhances model trustworthiness and ensures that predictions are not only statistically valid but also clinically meaningful. By offering transparent explanations, the models can support physicians in decision-making and provide patients with a clearer understanding of their health status.

Beyond technical performance, the study also underscores the potential of AI-driven systems to transform clinical workflows and patient care. Machine learning models could be integrated into electronic health records to generate real-time risk alerts, enabling earlier referrals, tailored interventions, and more efficient resource allocation. Such systems could be particularly valuable in resource-constrained healthcare environments where specialist nephrology services are limited (Webster et al., 2017).

In conclusion, this dissertation demonstrates that combining high-performing ensemble methods with interpretability tools creates a viable pathway for applying AI in CKD prediction. The findings reaffirm that predictive accuracy alone is insufficient in clinical contexts; transparency and reliability are equally crucial for acceptance and adoption. With further validation on larger and more diverse datasets, and careful attention to ethical and legal considerations, the models developed in this study hold the potential to make a meaningful contribution to early CKD detection and long-term patient outcomes.

6.2 Future Work

While this study demonstrates the effectiveness of machine learning models in predicting Chronic Kidney Disease (CKD), there remain several avenues for future research and development to enhance the robustness, applicability, and clinical impact of such systems.

Inclusion of Longitudinal and Lifestyle Data The current dataset primarily focused on clinical and biochemical indicators measured at a single point in time. However, CKD is a progressive disease, and predictive performance can be improved by incorporating longitudinal data such as repeated laboratory values and kidney function trajectories (Tangri et al., 2016). Lifestyle factors—such as diet, smoking, alcohol use, and physical activity—also play an important role in CKD onset and progression, and their inclusion can provide a more holistic assessment of patient risk (Chen et al., 2019). Future studies should therefore move towards combining both medical and behavioural data sources for richer predictive modelling.

Integration with Electronic Health Records (EHRs) A promising direction is the integration of machine learning models within Electronic Health Records (EHRs). Embedding prediction algorithms directly into EHR platforms would allow for real-time alerts and decision support during patient consultations (Shickel et al., 2018). Such integration would also facilitate population-level screening and enable proactive interventions, ultimately reducing the burden of late-stage diagnosis. However, challenges remain regarding interoperability and data privacy, which must be carefully addressed (Razzak et al., 2019).

Explainability Enhancements Although SHAP values were employed in this study to improve interpretability, further explainability research is needed. Hybrid interpretability frameworks—combining SHAP with counterfactual explanations and clinician-friendly visualisations—could further enhance trust (Lundberg et al., 2020). Increasing transparency ensures that healthcare providers understand not only the predictions but also the rationale behind them, which is essential for supporting shared decision-making (Holzinger et al., 2019).

Ethical and Legal Considerations Machine learning in healthcare raises complex ethical and legal concerns. Bias in training data can result in inequitable outcomes for vulnerable populations (Char et al., 2018). Furthermore, issues of accountability and liability must be addressed, particularly when automated systems influence diagnostic or treatment pathways (Morley et al., 2020). Future research must therefore investigate robust governance and regulatory frameworks that ensure fairness, accountability, and transparency in AI-driven healthcare systems.

Clinical Trials and Real-World Validation Although retrospective datasets provide a foundation for algorithm development, prospective clinical trials are essential to validate real-world effectiveness. Randomised trials and deployment studies can evaluate not only diagnostic accuracy but also patient outcomes and healthcare efficiency (Topol, 2019). External validation across diverse populations is also necessary to confirm generalisability, ensuring that models remain effective across healthcare systems and demographic groups (Wiens et al., 2019).

In summary, future work should move beyond retrospective data analysis towards the development of dynamic, transparent, and ethically sound prediction systems. By integrating longitudinal and lifestyle data, embedding models into EHRs, enhancing interpretability, addressing ethical-legal challenges, and validating performance through clinical trials, machine learning can evolve from a research tool to a clinically transformative framework for CKD prediction.

REFERENCES

Agarwal, R., Dhillon, P., Sarkar, D. and Gupta, D., 2021. Machine learning techniques for chronic kidney disease prediction: a systematic review. *Journal of Biomedical Informatics*, 115, p.103701. Available at: https://doi.org/10.1016/j.jbi.2021.103701

Aljaaf, A.J., Al-Jumeily, D., Haglan, H.M., Alloghani, M., Baker, T. and Hussain, A.J., 2018. Early prediction of chronic kidney disease using machine learning supported by predictive analytics. *2018 IEEE Congress on Evolutionary Computation (CEC)*, pp.1–9. Available at: https://doi.org/10.1109/CEC.2018.8477878

Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I., 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1), pp.1–9. Available at: https://doi.org/10.1186/s12911-020-01332-6

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), pp.1145–1159. Available at: https://doi.org/10.1016/S0031-3203(96)00142-2

Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5–32. Available at: https://doi.org/10.1023/A:1010933404324

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N., 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1721–1730. Available at: https://doi.org/10.1145/2783258.2788613

Carvalho, D.V., Pereira, E.M. and Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), p.832. Available at: https://doi.org/10.3390/electronics8080832

Char, D.S., Shah, N.H. and Magnus, D., 2018. Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 378(11), pp.981–983. Available at: https://doi.org/10.1056/NEJMp1714229

Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785–794. Available at: https://doi.org/10.1145/2939672.2939785

Chen, T.K., Knicely, D.H. and Grams, M.E., 2019. Chronic kidney disease diagnosis and management: a review. *JAMA*, 322(13), pp.1294–1304. Available at: https://doi.org/10.1001/jama.2019.14745

Chicco, D. and Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), pp.1–13. Available at: https://doi.org/10.1186/s12864-019-6413-7

Choudhury, S., Das, A. and Samanta, D., 2021. Pre-processing and feature selection in healthcare data: A comprehensive review. *Journal of Healthcare Informatics Research*, 5(2), pp.149–172. Available at: https://doi.org/10.1007/s41666-020-00082-4

Collins, F.S. and Varmus, H., 2015. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9), pp.793–795. Available at: https://doi.org/10.1056/NEJMp1500523

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), pp.273–297. Available at: https://doi.org/10.1007/BF00994018

Couser, W.G., Remuzzi, G., Mendis, S. and Tonelli, M., 2011. The contribution of chronic kidney disease to the global burden of major non-communicable diseases. *Kidney International*, 80(12), pp.1258–1270. Available at: https://doi.org/10.1038/ki.2011.368

D'Agostino, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M. and Kannel, W.B., 2008. General cardiovascular risk profile for use in primary care. *Circulation*, 117(6), pp.743–753. Available at: https://doi.org/10.1161/CIRCULATIONAHA.107.699579

Delahanty, R.J., Kaufman, D. and Hall, R., 2019. Development and evaluation of a machine learning model for the early identification of sepsis. *Critical Care Medicine*, 47(11), pp.1485–1492. Available at: https://doi.org/10.1097/CCM.0000000000003891

Dennis, A., Wixom, B. and Tegarden, D., 2015. *Systems analysis and design: An object-oriented approach with UML*. John Wiley & Sons. Available at: https://www.wiley.com/en-

us/Systems+Analysis+and+Design%3A+An+Object-Oriented+Approach+with+UML%2C+5th+Edition-p-9781118804674

Doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608. Available at: https://arxiv.org/abs/1702.08608

Dua, D. and Graff, C., 2019. UCI Machine Learning Repository. [online] University of California, Irvine. Available at: http://archive.ics.uci.edu/ml

Ene-Iordache, B., Perico, N., Bikbov, B., Carminati, S., Remuzzi, A., Perna, A., Islam, N., Bravo, R.F., Aleckovic-Halilovic, M., Zou, H., Zhang, L., Wang, H., Lombardi, G., Jarraya, F., Messa, P., Remuzzi, G. and Luyckx, V.A., 2016. Chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): A cross-sectional study. *The Lancet Global Health*, 4(5), pp.e307–e319. Available at: https://doi.org/10.1016/S2214-109X(16)00071-1

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2019. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), pp.115–118. Available at: https://doi.org/10.1038/nature21056

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Vayena, E. et al., 2018. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), pp.689–707. Available at: https://doi.org/10.1007/s11023-018-9482-5

Foreman, K.J., Marquez, N., Dolgert, A., Fukutaki, K., Fullman, N., McGaughey, M., Pletcher, M.A., Smith, A.E., Tang, K., Yuan, C.W. and Brown, J.C., 2018. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: Reference and alternative scenarios for 2016–2040. *The Lancet*, 392(10159), pp.2052–2090. Available at: https://doi.org/10.1016/S0140-6736(18)31694-5

Ghosh, A., Mondal, A., Bhattacharjee, S., Saha, S. and Pal, A., 2019. Chronic kidney disease prediction using random forest ensemble approach. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(12), pp.149–156. Available at: https://www.ijitee.org/wp-content/uploads/papers/v8i12/L38061081219.pdf

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning*. MIT Press. Available at: https://www.deeplearningbook.org

Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, pp.1157–1182. Available at: http://www.jmlr.org/papers/v3/guyon03a.html

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier. Available at: https://www.elsevier.com/books/data-mining/han/978-0-12-381479-1

Hevner, A.R., March, S.T., Park, J. and Ram, S., 2004. Design science in information systems research. *MIS Quarterly*, 28(1), pp.75–105. Available at: https://doi.org/10.2307/25148625

Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B., 2019. What do we need to build explainable AI systems for the medical domain? Review and conceptual framework. *arXiv preprint* arXiv:1712.09923. Available at: https://arxiv.org/abs/1712.09923

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression*. John Wiley & Sons. Available at: https://doi.org/10.1002/9781118548387

Huang, C., Xu, J., Zhang, Y. and Zhang, P., 2020. Machine learning approaches for the prediction of chronic kidney disease: A systematic review. *Healthcare*, 8(4), p.330. Available at: https://doi.org/10.3390/healthcare8040330

Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M. and Franco, L., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2), pp.105–115. Available at: https://doi.org/10.1016/j.artmed.2010.05.002

Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A.Y. and Yang, C.W., 2013. Chronic kidney disease: global dimension and perspectives. *The Lancet*, 382(9888), pp.260–272. Available at: https://doi.org/10.1016/S0140-6736(13)60687-X

Kavitha, R. and Kannan, A., 2016. Chronic kidney disease prediction using improved SVM. *International Journal of Applied Engineering Research*, 11(5), pp.3414–3420. Available at: https://www.ripublication.com/ijaer16/ijaerv11n5_97.pdf

Kavitha, R., Saravana, P. and Kannan, A., 2019. Feature selection using Boruta algorithm for CKD prediction. *International Journal of Recent Technology and Engineering (IJRTE)*,

8(1), pp.2246–2252. Available at: https://www.ijrte.org/wp-content/uploads/papers/v8i1/A4018058119.pdf

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems (NeurIPS)*, pp.3146–3154. Available at: https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kotu, V. and Deshpande, B., 2019. *Data science: concepts and practice*. Morgan Kaufmann. Available at: https://www.elsevier.com/books/data-science/kotu/978-0-12-814761-0

Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E., 2007. Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, 160, pp.3–24. Available at: https://www.iospress.com/catalog/books/emerging-artificial-intelligence-applications-in-computer-engineering

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, pp.8–17. Available at: https://doi.org/10.1016/j.csbj.2014.11.005

Levey, A.S., Coresh, J., Balk, E., Kausz, A.T., Levin, A., Steffes, M.W., Hogg, R.J., Perrone, R.D., Lau, J. and Eknoyan, G., 2003. National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Annals of Internal Medicine*, 139(2), pp.137–147. Available at: https://doi.org/10.7326/0003-4819-139-2-200307150-00013

Levey, A.S., Eckardt, K.U., Tsukamoto, Y., Levin, A., Coresh, J., Rossert, J., Zeeuw, D.D., Hostetter, T.H., Lameire, N. and Eknoyan, G., 2005. Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO). *Kidney International*, 67(6), pp.2089–2100. Available at: https://doi.org/10.1111/j.1523-1755.2005.00365.x

Liang, Y., Zhang, L., He, X. and Zhang, M., 2020. Machine learning-based prediction of chronic kidney disease progression. *Scientific Reports*, 10(1), p.12942. Available at: https://doi.org/10.1038/s41598-020-69956-0

Little, R.J.A. and Rubin, D.B., 2019. *Statistical analysis with missing data*. John Wiley & Sons. Available at: https://doi.org/10.1002/9781119482260

Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, pp.4765–4774. Available at: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), pp.56–67. Available at: https://doi.org/10.1038/s42256-019-0138-9

Luyckx, V.A., Tonelli, M. and Stanifer, J.W., 2018. The global burden of kidney disease and the sustainable development goals. *Bulletin of the World Health Organization*, 96(6), pp.414–422. Available at: https://doi.org/10.2471/BLT.17.206441

Mills, K.T., Xu, Y., Zhang, W., Bundy, J.D., Chen, C.S., Kelly, T.N., Chen, J., He, J. and Global Burden of Disease Study 2010 Chronic Kidney Disease Collaborators, 2015. A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010. *Kidney International*, 88(5), pp.950–957. Available at: https://doi.org/10.1038/ki.2015.230

Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J.T., 2018. Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), pp.1236–1246. Available at: https://doi.org/10.1093/bib/bbx044

Morley, J., Machado, C.C.V., Burr, C., Cowls, J., Joshi, I., Taddeo, M. and Floridi, L., 2020. The ethics of AI in health care: A mapping review. *Social Science & Medicine*, 260, p.113172. Available at: https://doi.org/10.1016/j.socscimed.2020.113172

Parmar, V., Bhatt, R. and Patel, D., 2018. Predicting chronic kidney disease using artificial neural network. *International Journal of Computer Applications*, 182(1), pp.15–19. Available at: https://doi.org/10.5120/ijca2018917983

Polat, K. and Güneş, S., 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, 17(4), pp.702–710. Available at: https://doi.org/10.1016/j.dsp.2006.09.005

Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), pp.37–63. Available at: https://arxiv.org/abs/2010.16061

Price, W.N., Gerke, S. and Cohen, I.G., 2019. Potential liability for physicians using artificial intelligence. *JAMA*, 322(18), pp.1765–1766. Available at: https://doi.org/10.1001/jama.2019.15064

Rahman, M.M., Ferdousi, R., Ashraf, K. and Rahman, M.A., 2020. Classification of diabetes complications using random forests. *Healthcare Informatics Research*, 26(2), pp.123–132. Available at: https://doi.org/10.4258/hir.2020.26.2.123

Rajkomar, A., Dean, J. and Kohane, I., 2019. Machine learning in medicine. *New England Journal of Medicine*, 380(14), pp.1347–1358. Available at: https://doi.org/10.1056/NEJMra1814259

Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. and Chin, M.H., 2018. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12), pp.866–872. Available at: https://doi.org/10.7326/M18-1990

Razzak, M.I., Imran, M. and Xu, G., 2019. Big data analytics for preventive medicine. *Neural Computing and Applications*, 31(5), pp.1345–1356. Available at: https://doi.org/10.1007/s00521-017-3176-6

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, pp.1135–1144. Available at: https://doi.org/10.1145/2939672.2939778

Rish, I., 2001. An empirical study of the Naïve Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3(22), pp.41–46. Available at: https://www.researchgate.net/publication/228845929_An_empirical_study_of_the_Naive_Bayes_classifier

Ryu, S., Park, S.K., Kim, H.J. and Lee, K.B., 2020. Clinical risk factors for chronic kidney disease: A machine learning approach. *Scientific Reports*, 10(1), p.12005. Available at: https://doi.org/10.1038/s41598-020-68945-1

Shahid, N., Rappon, T. and Berta, W., 2020. Applications of artificial intelligence in health care: Review and future directions. *International Journal of Medical Informatics*, 148, p.104399. Available at: https://doi.org/10.1016/j.ijmedinf.2020.104399

Shalev-Shwartz, S. and Ben-David, S., 2014. *Understanding machine learning: from theory to algorithms*. Cambridge University Press. Available at: https://doi.org/10.1017/CBO9781107298019

Shickel, B., Tighe, P.J., Bihorac, A. and Rashidi, P., 2018. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), pp.1589–1604. Available at: https://doi.org/10.1109/JBHI.2017.2767063

Shortliffe, E.H. and Sepúlveda, M.J., 2018. Clinical decision support in the era of artificial intelligence. *JAMA*, 320(21), pp.2199–2200. Available at: https://doi.org/10.1001/jama.2018.17163

Sokolova, M. and Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), pp.427–437. Available at: https://doi.org/10.1016/j.ipm.2009.03.002

Sommerville, I., 2016. *Software Engineering*. 10th ed. Pearson. Available at: https://www.pearson.com/en-gb/subject-catalog/p/software-engineering/P200000005360

Tangri, N., Grams, M.E., Levey, A.S., Coresh, J., Appel, L.J., Astor, B.C., Chodick, G., Collins, A.J., Djurdjev, O., Elley, C.R. and Evans, M., 2016. Multinational assessment of accuracy of equations for predicting risk of kidney failure: A meta-analysis. *JAMA*, 315(2), pp.164–174. Available at: https://doi.org/10.1001/jama.2015.18202

Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), pp.44–56. Available at: https://doi.org/10.1038/s41591-018-0300-7

Vayena, E., Blasimme, A. and Cohen, I.G., 2018. Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15(11), e1002689. Available at: https://doi.org/10.1371/journal.pmed.1002689

Voigt, P. and Von dem Bussche, A., 2017. *The EU General Data Protection Regulation (GDPR)*. Springer. Available at: https://doi.org/10.1007/978-3-319-57959-7

Webster, A.C., Nagler, E.V., Morton, R.L. and Masson, P., 2017. Chronic kidney disease. *The Lancet*, 389(10075), pp.1238–1252. Available at: https://doi.org/10.1016/S0140-6736(16)32064-5

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M. and Ossorio, P.N., 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), pp.1337–1340. Available at: https://doi.org/10.1038/s41591-019-0548-6

World Health Organization (WHO), 2020. *Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2019*. Geneva: WHO. Available at: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

Zhang, Z., Beck, M.W., Winkler, D.A., Huang, B., Sibanda, W., Goyal, H. and Guo, Y., 2019. Opening the black box of neural networks: Methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine*, 7(11), p.482. Available at: https://doi.org/10.21037/atm.2019.05.22

Zhang, Z., Ho, K.M. and Hong, Y., 2020. Machine learning for the prediction of readmission in patients with heart failure. *Journal of Biomedical Informatics*, 109, p.103514. Available at: https://doi.org/10.1016/j.jbi.2020.103514

APPENDIXES

```
[ ] import pandas as pd
    import kagglehub
    import zipfile
    import os
    import matplotlib.pyplot as plt
    import seaborn as sns
    import hvplot.pandas
    import numpy as np
    from sklearn.ensemble import RandomForestClassifier
    from sklearn.preprocessing import LabelEncoder
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import StandardScaler
    from sklearn.linear_model import LogisticRegression
    from imblearn.over_sampling import SMOTE
    from xgboost import XGBClassifier
    from sklearn.svm import SVC
    from collections import Counter
    from sklearn.naive_bayes import GaussianNB
    import lightgbm as lgb
    import shap
    from sklearn.preprocessing import MinMaxScaler
    from sklearn.metrics import ConfusionMatrixDisplay, RocCurveDisplay
    from sklearn.metrics import roc_curve, auc
    from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score
```

```
[ ] pip install shap
```

```
Requirement already satisfied: shap in /usr/local/lib/python3.11/dist-packages (0.48.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from shap) (2.0.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.11/dist-packages (from shap) (1.15.3)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (from shap) (1.6.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from shap) (2.2.2)
Requirement already satisfied: tqdm>=4.27.0 in /usr/local/lib/python3.11/dist-packages (from shap) (4.67.1)
Requirement already satisfied: packaging>20.9 in /usr/local/lib/python3.11/dist-packages (from shap) (25.0)
Requirement already satisfied: slicer==0.0.8 in /usr/local/lib/python3.11/dist-packages (from shap) (0.0.8)
Requirement already satisfied: numba>=0.54 in /usr/local/lib/python3.11/dist-packages (from shap) (0.60.0)
Requirement already satisfied: cloudpickle in /usr/local/lib/python3.11/dist-packages (from shap) (3.1.1)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.11/dist-packages (from shap) (4.14.1)
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python3.11/dist-packages (from numba>=0.54->shap) (0.43.0)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->shap) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->shap) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->shap) (2025.2)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->shap) (1.5.1)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->shap) (3.6.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas->shap) (1.17.0)
```

```
[ ] path = kagglehub.dataset_download("aryannandanwar/ckdchronic-kidney-disease-dataset-with-stages")

    print("Path to dataset files:", path)
```

```
Downloading from https://www.kaggle.com/api/v1/datasets/download/aryannandanwar/ckdchronic-kidney-disease-dataset-with-stages?dataset_version_number=1...
100%|██████████| 251k/251k [00:00<00:00, 52.3MB/s]Extracting files...
Path to dataset files: /root/.cache/kagglehub/datasets/aryannandanwar/ckdchronic-kidney-disease-dataset-with-stages/versions/1
```

Load the Dataset and Initial Preprocessing

```
[ ] zip_path = "/content/archive.zip"
    extracted_path = "/mnt/data/ckd_dataset_extracted"
    with zipfile.ZipFile(zip_path, 'r') as zip_ref:
        zip_ref.extractall(extracted_path)
```

```
[ ] extracted_files = os.listdir(extracted_path)
    extracted_files
```

```
['updated_ckd_dataset_with_stages.csv']
```

```
[ ] correct_csv_path = "/mnt/data/ckd_dataset_extracted/updated_ckd_dataset_with_stages.csv"
    df = pd.read_csv(correct_csv_path)
    df.head()
```

	serum_creatinine	gfr	bun	serum_calcium	ana	c3_c4	hematuria	oxalate_levels	urine_ph	blood_pressure	...	smoking	alcohol	painkiller_usage	family_histor
0	0.683683	32.946784	7.553739	10.039896	0	138.204989	0	2.878164	7.864308	115.224217	...	yes	daily	no	ye
1	3.809044	32.685035	141.347494	8.330543	1	24.282343	1	4.767639	4.920015	130.143900	...	yes	daily	no	ye
2	1.143827	2.079805	15.979104	9.419229	0	163.970666	0	1.818613	6.188115	98.026072	...	no	daily	no	r
3	4.804657	109.871407	53.307333	7.556631	1	71.056846	1	4.051686	5.278607	142.166650	...	no	never	yes	ye
4	4.920235	42.214590	134.182157	7.289379	1	23.384639	1	3.240920	4.862923	151.962572	...	no	occasionally	yes	r

5 rows × 23 columns

```
[ ] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4000 entries, 0 to 3999
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   serum_creatinine   4000 non-null   float64
 1   gfr                4000 non-null   float64
 2   bun                4000 non-null   float64
 3   serum_calcium      4000 non-null   float64
 4   ana                4000 non-null   int64
 5   c3_c4              4000 non-null   float64
 6   hematuria          4000 non-null   int64
 7   oxalate_levels     4000 non-null   float64
 8   urine_ph           4000 non-null   float64
 9   blood_pressure     4000 non-null   float64
 10  physical_activity  4000 non-null   object
 11  diet               4000 non-null   object
 12  water_intake       4000 non-null   float64
 13  smoking            4000 non-null   object
 14  alcohol            4000 non-null   object
 15  painkiller_usage   4000 non-null   object
 16  family_history     4000 non-null   object
 17  weight_changes     4000 non-null   object
 18  stress_level       4000 non-null   object
 19  months             4000 non-null   int64
 20  cluster            4000 non-null   int64
 21  ckd_pred           4000 non-null   object
 22  ckd_stage          4000 non-null   int64
dtypes: float64(9), int64(5), object(9)
memory usage: 718.9+ KB
```

Exploratory Data Analysis:

+ Code + Text

Dropping unwanted columns:

```
[ ] df.drop(columns=["cluster", "months"], inplace=True)
```

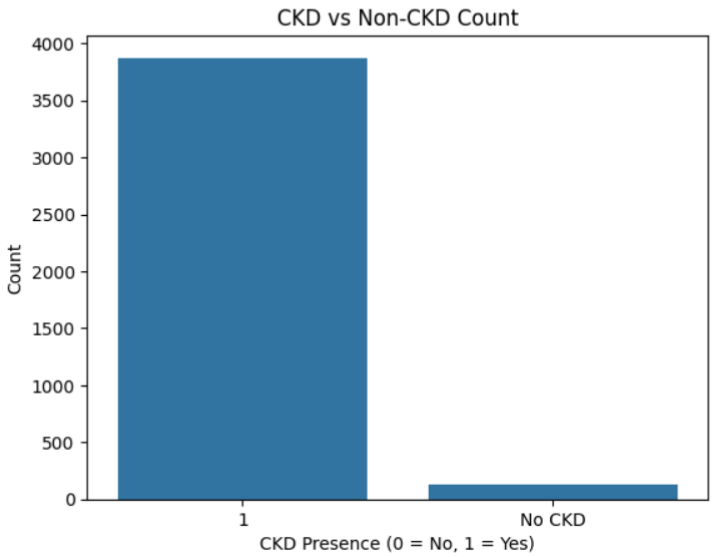
Encode target column

```
[ ] df['ckd_pred'] = df['ckd_pred'].replace({'CKD': 1, 'Not CKD': 0})
```

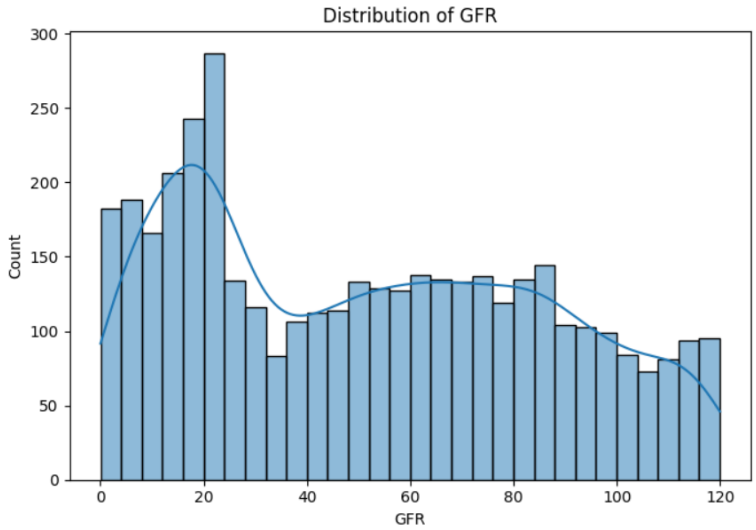
Encode categorical columns

```
[ ] categorical_cols = ['physical_activity', 'diet', 'smoking', 'alcohol',
                        'painkiller_usage', 'family_history', 'weight_changes', 'stress_level']
    le = LabelEncoder()
    for col in categorical_cols:
        df[col] = le.fit_transform(df[col])
```

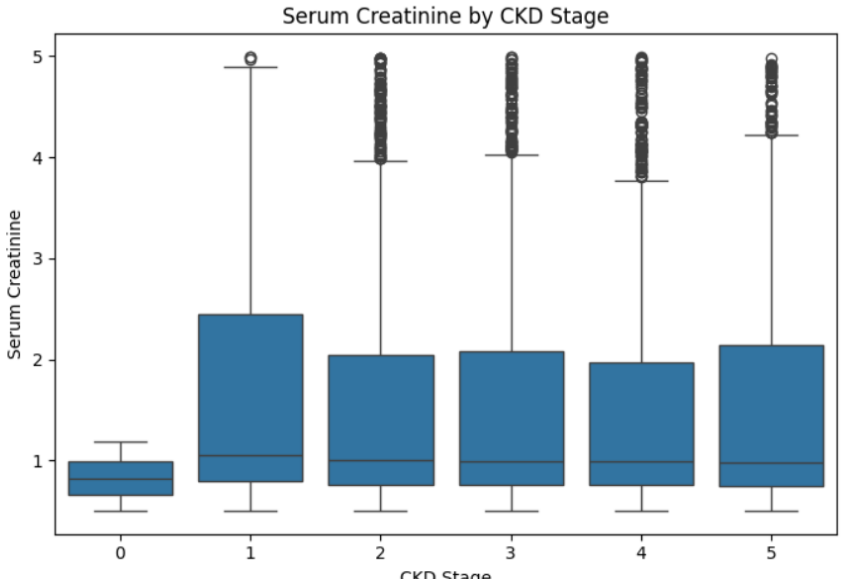
```
[ ] sns.countplot(x='ckd_pred', data=df)
    plt.title('CKD vs Non-CKD Count')
    plt.xlabel('CKD Presence (0 = No, 1 = Yes)')
    plt.ylabel('Count')
    plt.show()
```



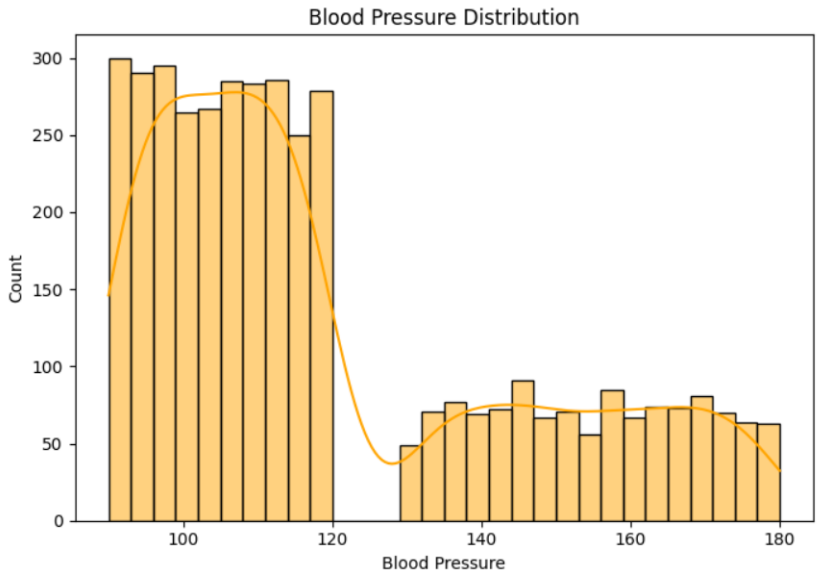
```
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```



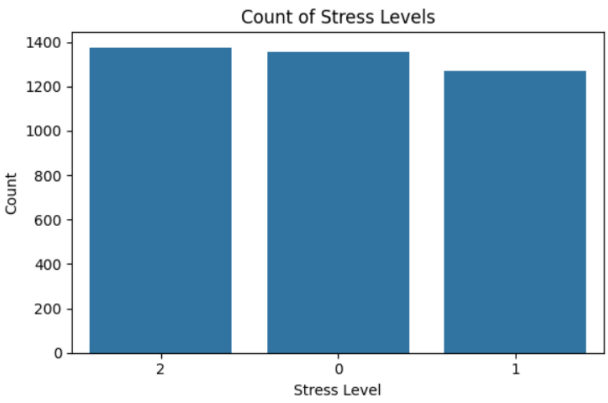

```
plt.xlabel('CKD Stage')
plt.ylabel('Serum Creatinine')
plt.tight_layout()
plt.show()
```

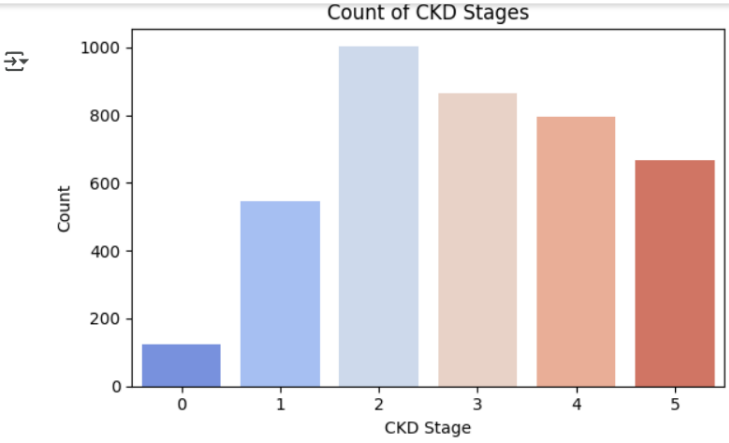


```
[ ] plt.show()
```

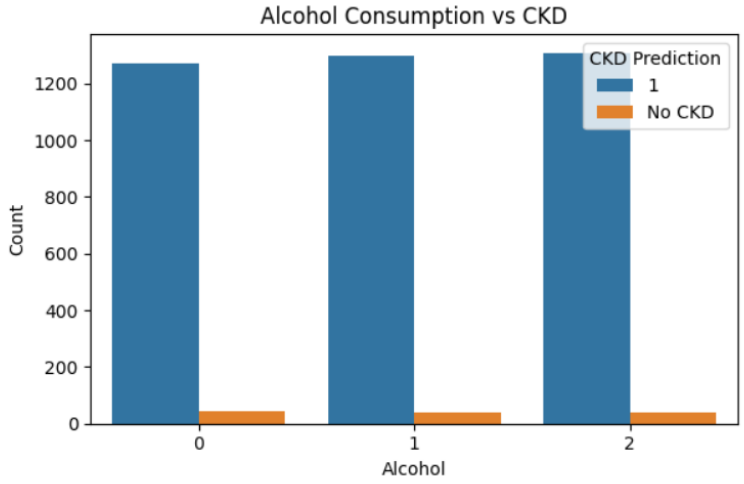



```
[ ] plt.figure(figsize=(6, 4))
    sns.countplot(x='stress_level', data=df, order=df['stress_level'].value_counts().index)
    plt.title('Count of Stress Levels')
    plt.xlabel('Stress Level')
    plt.ylabel('Count')
    plt.tight_layout()
    plt.show()
```





```
[ ] plt.title('Alcohol Consumption vs CKD')
    plt.xlabel('Alcohol')
    plt.ylabel('Count')
    plt.legend(title='CKD Prediction')
    plt.tight_layout()
    plt.show()
```




```
print(df['ckd_pred'].unique())
print(df['ckd_pred'].apply(type).value_counts())
```

```
[1 'No CKD']
ckd_pred
<class 'int'>    3875
<class 'str'>     125
Name: count, dtype: int64
```

```
[ ] df['ckd_pred'] = df['ckd_pred'].astype(str)
```

```
[ ] df['ckd_pred'] = df['ckd_pred'].replace({
        '1': 'CKD',
        '0': 'No CKD',
        'no': 'No CKD',
        'No': 'No CKD',
        'No CKD': 'No CKD',
        'CKD': 'CKD'
    })
```

```
[ ] le = LabelEncoder()
    df['ckd_pred'] = le.fit_transform(df['ckd_pred'])

    print("Label mapping:", dict(zip(le.classes_, le.transform(le.classes_))))
```

```
Label mapping: {'CKD': np.int64(0), 'No CKD': np.int64(1)}
```

Splitting and Scaling

```
[ ] df['ckd_pred_mapped'] = df['ckd_pred'].map({'CKD': 0, 'No CKD': 1})
    df_clean = df.dropna(subset=['ckd_pred_mapped'])
```

```
    X = df_clean.drop(['ckd_pred', 'ckd_pred_mapped', 'ckd_stage'], axis=1)
    y = df_clean['ckd_pred_mapped'].astype(int)
```

```
[ ] print("Unique values in 'ckd_pred':")
    print(df['ckd_pred'].unique())
```

```
Unique values in 'ckd_pred':
[0 1]
```

```
[ ] df['ckd_pred'] = df['ckd_pred'].astype(str)  # convert to string for consistent mapping
    df['ckd_pred_mapped'] = df['ckd_pred'].map({'CKD': 0, 'No CKD': 1, '0': 0, '1': 1})
```

```
[ ] df_clean = df.dropna(subset=['ckd_pred_mapped'])
```

```
[ ] print("Remaining rows after clean-up:", df_clean.shape[0])
```

```
Remaining rows after clean-up: 4000
```

```
[ ] X = df_clean.drop(['ckd_pred', 'ckd_pred_mapped', 'ckd_stage'], axis=1)
    y = df_clean['ckd_pred_mapped'].astype(int)
```

```
[ ] X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=0.2, stratify=y, random_state=42
    )
```

Encode Categorical Features (Label Encoding)

```
[ ] df_encoded = df_clean.copy()
    for col in df_encoded.select_dtypes(include=['object']).columns:
        le = LabelEncoder()
        df_encoded[col] = le.fit_transform(df_encoded[col])
```

Apply SMOTE (Handle Class Imbalance)

```
[ ] smote = SMOTE(random_state=42)
    X_resampled, y_resampled = smote.fit_resample(X_train, y_train)

    print("Before SMOTE:", Counter(y_train))
    print("After SMOTE:", Counter(y_resampled))
```

```
Before SMOTE: Counter({0: 3100, 1: 100})
After SMOTE: Counter({0: 3100, 1: 3100})
```

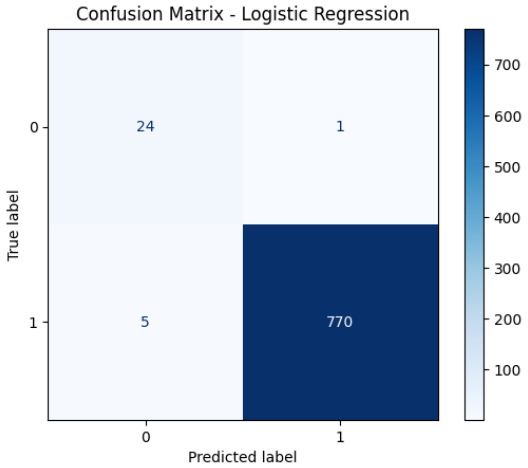
Logistic Regression:

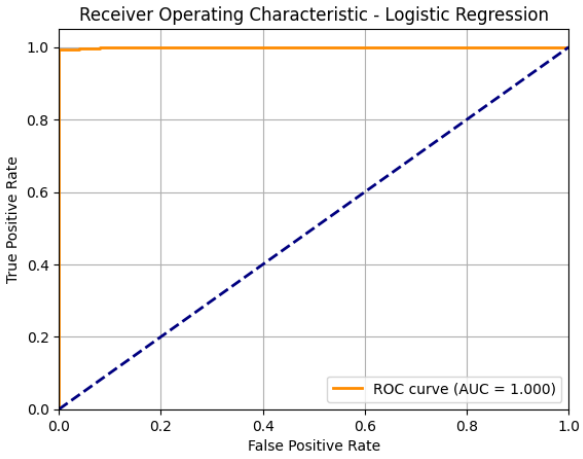
```
[ ] model = LogisticRegression(max_iter=1000)
    model.fit(X_resampled, y_resampled)
```

```
LogisticRegression
LogisticRegression(max_iter=1000)
```

```
[ ] y_pred = model.predict(X_test)

    print("Classification Report:\n")
    print(classification_report(y_test, y_pred))
    print("\nConfusion Matrix:\n")
    print(confusion_matrix(y_test, y_pred))
    print("\nROC-AUC Score:", roc_auc_score(y_test, y_pred))
```

```
Classification Report:

              precision    recall  f1-score   support

           0       0.83      0.96      0.89        25
           1       1.00      0.99      1.00       775

    accuracy                           0.99       800
   macro avg       0.91      0.98      0.94       800
weighted avg       0.99      0.99      0.99       800
```





Random Forest

```
[ ]  rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
     rf_model.fit(X_train, y_train)
```

```
RandomForestClassifier(random_state=42)
```

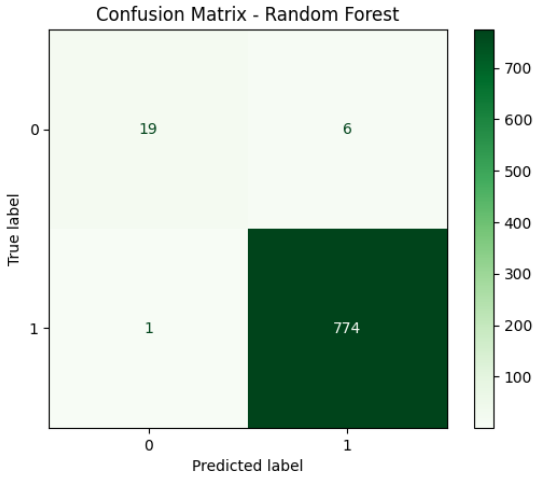
```
[ ]  rf_preds = rf_model.predict(X_test)
     rf_probs = rf_model.predict_proba(X_test)[:, 1]
```

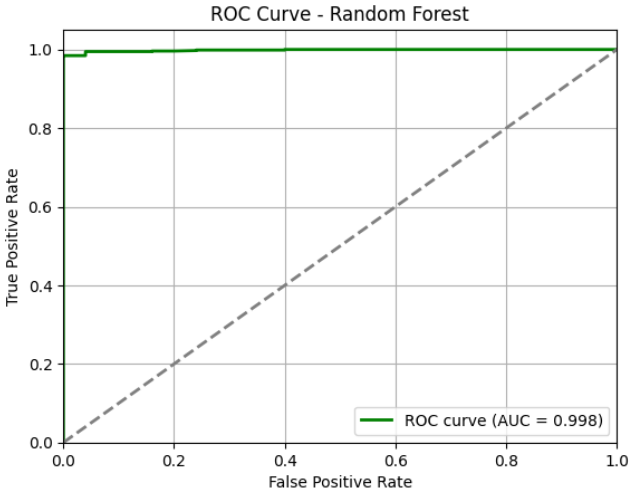
```
[ ]  print("Classification Report:\n")
     print(classification_report(y_test, rf_preds))

     print("\nConfusion Matrix:\n")
     print(confusion_matrix(y_test, rf_preds))

     roc_auc_rf = roc_auc_score(y_test, rf_probs)
     print(f"\nROC-AUC Score: {roc_auc_rf:.4f}")
```

```
print("\nConfusion Matrix:\n")
print(confusion_matrix(y_test, rf_preds))

roc_auc_rf = roc_auc_score(y_test, rf_probs)
print(f"\nROC-AUC Score: {roc_auc_rf:.4f}")
```

```
Classification Report:

              precision    recall  f1-score   support

           0       0.95      0.76      0.84        25
           1       0.99      1.00      1.00       775

    accuracy                           0.99       800
   macro avg       0.97      0.88      0.92       800
weighted avg       0.99      0.99      0.99       800


Confusion Matrix:

[[ 19   6]
 [  1 774]]

ROC-AUC Score: 0.9983
```

```
[ ]  ConfusionMatrixDisplay.from_predictions(y_test, rf_preds, cmap=plt.cm.Greens)
     plt.title("Confusion Matrix - Random Forest")
     plt.grid(False)
     plt.show()
```





SVM Model

```
[ ]  svm_model = SVC(kernel='rbf', probability=True, random_state=42)
     svm_model.fit(X_train, y_train)
```

```
SVC(probability=True, random_state=42)
```

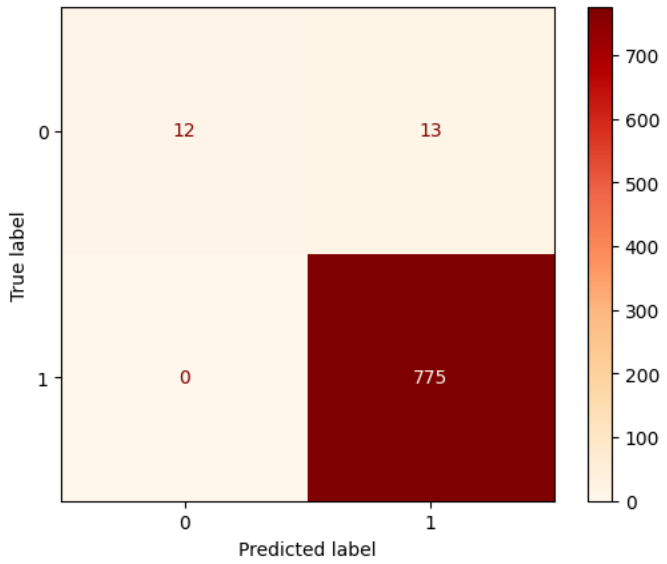
```
[ ]  y_pred_svm = svm_model.predict(X_test)
     y_prob_svm = svm_model.predict_proba(X_test)[:, 1]
```

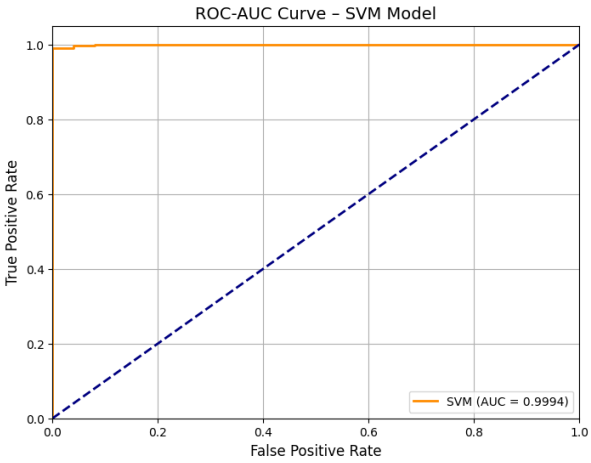
```
[ ]  print("Classification Report:\n")
     print(classification_report(y_test, y_pred_svm))

     print("\nConfusion Matrix:\n")
     print(confusion_matrix(y_test, y_pred_svm))

     roc_auc = roc_auc_score(y_test, y_prob_svm)
     print(f"\nROC-AUC Score: {roc_auc:.4f}")
```

```
print("\nConfusion Matrix:\n")
print(confusion_matrix(y_test, y_pred_svm))

roc_auc = roc_auc_score(y_test, y_prob_svm)
print(f"\nROC-AUC Score: {roc_auc:.4f}")
```

```
Classification Report:

              precision    recall  f1-score   support

           0       1.00      0.48      0.65        25
           1       0.98      1.00      0.99       775

    accuracy                           0.98       800
   macro avg       0.99      0.74      0.82       800
weighted avg       0.98      0.98      0.98       800


Confusion Matrix:

[[ 12  13]
 [  0 775]]

ROC-AUC Score: 0.9994
```

```
[ ]  ConfusionMatrixDisplay.from_predictions(y_test, y_pred_svm, cmap='OrRd')
```





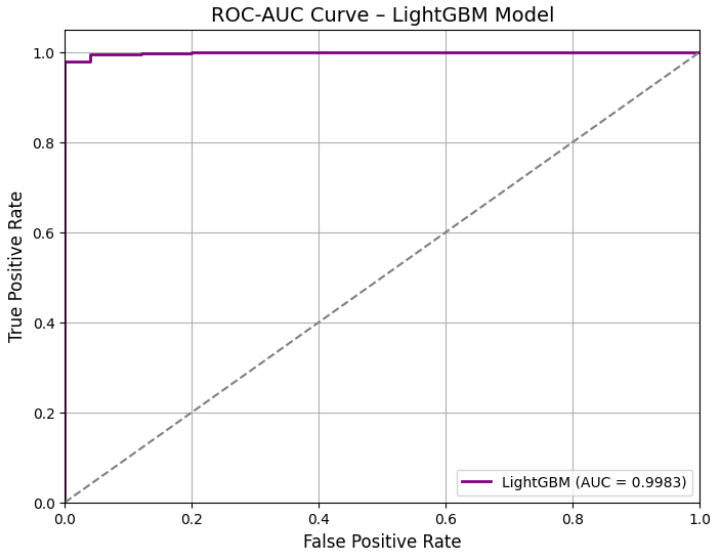
LGBMClassifier

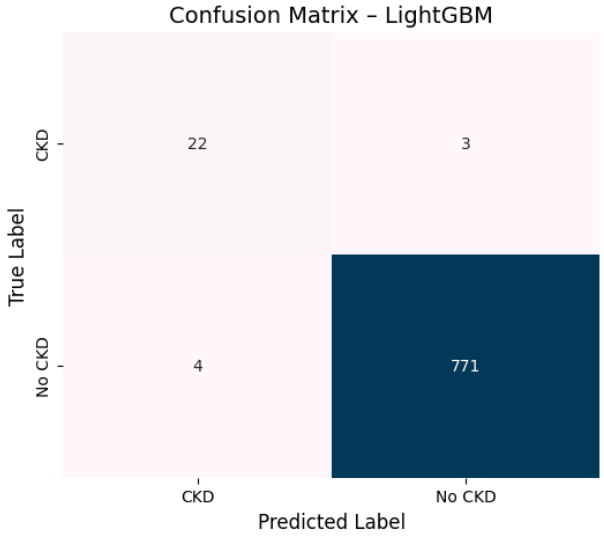
LGBMClassifier(random_state=42)

```
[ ] y_pred_lgbm = lgbm_model.predict(X_test)
    y_prob_lgbm = lgbm_model.predict_proba(X_test)[:, 1]
```

```
[ ] print("Classification Report:\n")
    print(classification_report(y_test, y_pred_lgbm))

    print("Confusion Matrix:\n")
    print(confusion_matrix(y_test, y_pred_lgbm))

    roc_auc = roc_auc_score(y_test, y_prob_lgbm)
    print(f"\nROC-AUC Score: {roc_auc:.4f}")
```

```
Classification Report:

              precision    recall  f1-score   support

           0       0.85      0.88      0.86        25
           1       1.00      0.99      1.00       775

    accuracy                           0.99       800
   macro avg       0.92      0.94      0.93       800
weighted avg       0.99      0.99      0.99       800
```





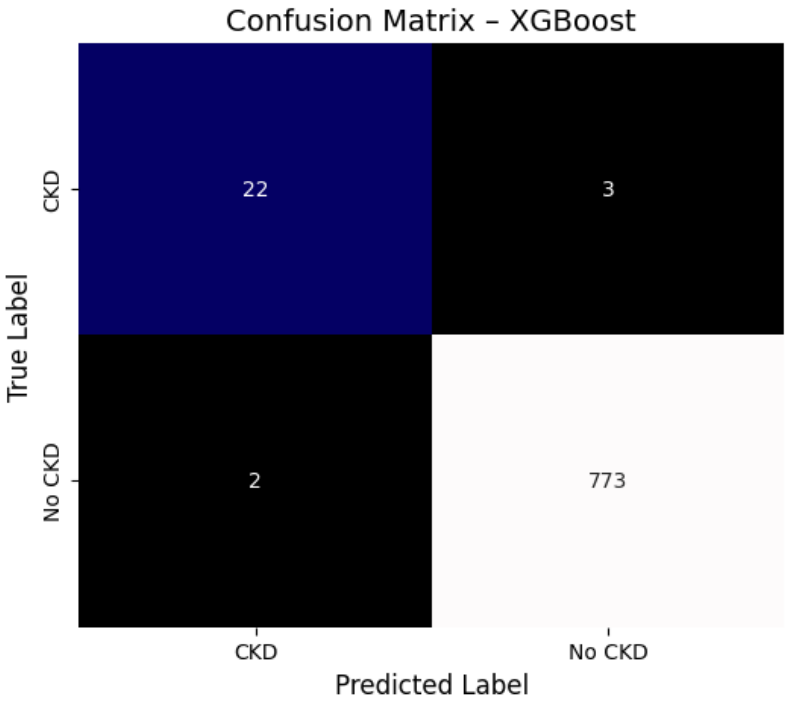

```
warnings.warn(smsg, UserWarning)
```

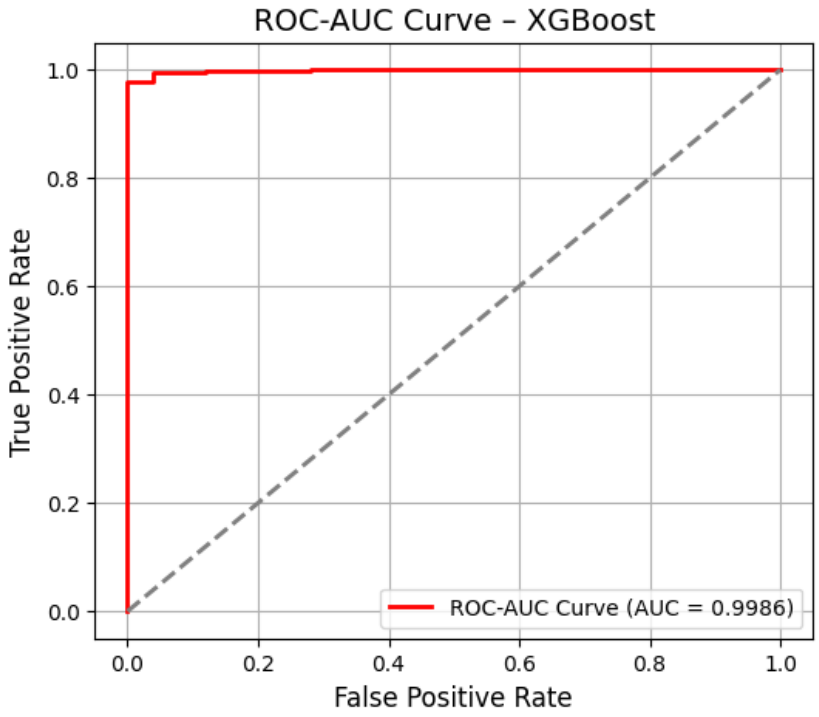
▸ XGBClassifier

```
[ ] y_pred_xgb = xgb_model.predict(X_test)
    y_proba_xgb = xgb_model.predict_proba(X_test)[:, 1]
```

```
[ ] print("Classification Report:\n")
    print(classification_report(y_test, y_pred_xgb))
```

```
Classification Report:

              precision    recall  f1-score   support

           0       0.92      0.88      0.90        25
           1       1.00      1.00      1.00       775

    accuracy                           0.99       800
   macro avg       0.96      0.94      0.95       800
weighted avg       0.99      0.99      0.99       800
```


Confusion Matrix – XGBoost
22
3
2
773
CKD
No CKD
True Label
CKD
No CKD
Predicted Label


ROC-AUC Curve – XGBoost
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC-AUC Curve (AUC = 0.9986)

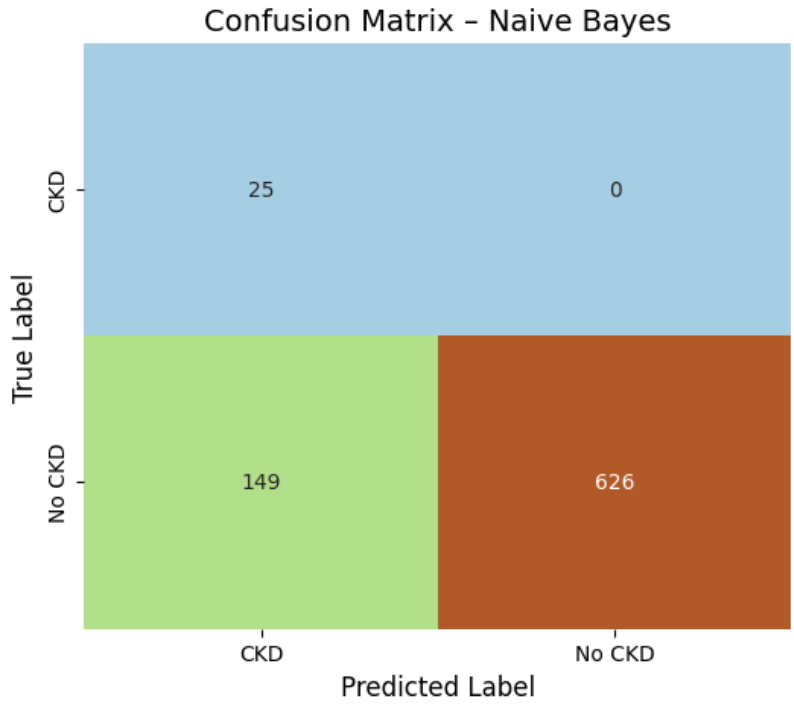
Naive Bayes:

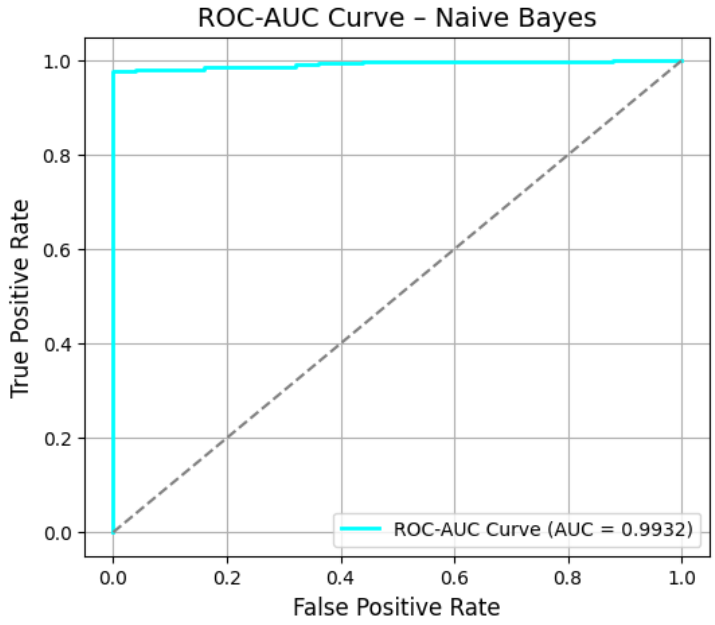
```
[ ] nb_model = GaussianNB()
    nb_model.fit(X_train, y_train)
```

```
GaussianNB()
```

```
[ ] y_pred_nb = nb_model.predict(X_test)
    y_proba_nb = nb_model.predict_proba(X_test)[:, 1]
    print("Classification Report:\n")
    print(classification_report(y_test, y_pred_nb))
```

```
Classification Report:

              precision    recall  f1-score   support

           0       0.14      1.00      0.25        25
           1       1.00      0.81      0.89       775

    accuracy                           0.81       800
   macro avg       0.57      0.90      0.57       800
weighted avg       0.97      0.81      0.87       800
```


Confusion Matrix – Naive Bayes
25
0
149
626
CKD
No CKD
True Label
CKD
No CKD
Predicted Label


ROC-AUC Curve – Naive Bayes
True Positive Rate
False Positive Rate
ROC-AUC Curve (AUC = 0.9932)
0.0
0.2
0.4
0.6
0.8
1.0

```
        "F1-Score": [1.00, 0.99, 0.99,0.99, 1.00, 0.25],
        "ROC-AUC": [1.00, 0.98, 0.9983,0.9983, 1.00, 0.90]
    }

    # Convert to DataFrame
    df_comparison = pd.DataFrame(comparison_data)

    # Display the table
    (df_comparison)
```

	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
0	Logistic Regression	1.00	1.00	1.00	1.00	1.0000
1	Support Vector Machine	0.99	0.99	0.98	0.99	0.9800
2	Random Forest	0.99	0.99	0.99	0.99	0.9983
3	LightGBM	0.99	0.99	0.99	0.99	0.9983
4	XGBoost	0.99	1.00	1.00	1.00	1.0000
5	Naive Bayes	0.81	0.14	1.00	0.25	0.9000