Early Detection of Chronic Kidney Disease: Comparative Study with Ensemble Learning, Traditional Machine Learning and Neural Network Approaches

by

Md Rumman Shahriar
24241284
Hasan Sarwar Zami
22101335
Saib Saleh Nabil
23241068
Mahedi Mostafa Rifat
22101739
Tasnimul Hasan
23241061

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
December 2025

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md Rumman Shahriar

24241284

Hasan Sarwar Zami

22101335

Saib Saleh Nabil

23241068

Mahedi Mostafa Rifat

22101739

Tasnimul Hasan

23241061

Approval

The thesis titled "Early Detection of Chronic Kidney Disease: Comparative Study with Ensemble Learning, Traditional Machine Learning and Neural Network Approaches" submitted by

1. Md Rumman Shahriar (24241284)
2. Hasan Sarwar Zami (22101335)
3. Saib Saleh Nabil (23241068)
4. Mahedi Mostafa Rifat (22101739)
5. Tasnimul Hasan (23241061)

of Fall, 2025 semester has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on December 6, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Jannatun Noor Mukta

Associate Professor, CSE
Director, BSDS
United International University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Chronic Kidney Disease (CKD) is a common progressive disease, affecting millions of people around the world, and early detection has proved to be one of the best approaches for improving patient outcomes and reducing the burden on the healthcare system. This research paper will provide a comparative analysis of the early detection of the CKD through the standard clinical data. It compares three different machine learning paradigms systematically, including Traditional Machine Learning, Ensemble Learning and Neural Network approaches. The goal is to develop a strong classification model to differentiate people with CKD and healthy patients based on the features such as blood pressure, serum creatinine, hemoglobin etc. After all the data preparation and feature selection, various models of each paradigm were generated and tested. The outcomes prove that an ensemble-based model has performed better, providing a perfect balance between high accuracy, strong interpretability, and practical utility among the healthcare professionals. In this model, patients at risk are easily identified early, and appropriate and possibly life-saving procedures can be implemented.

Keywords: Chronic kidney disease (CKD), Ensemble Learning, Deep Learning, Early Detection, Neural network, AI in Health

Table of Contents

Chapter 1

Introduction

1.1 Background

Chronic Kidney Disease (CKD) is a major health problem in the world as it has been experienced by millions of people worldwide. The illness is usually slow-moving and asymptomatic at the initial phases and may often remain undetected until it has spread further, thus exposing a person to ultimate complications that may pose deadly risks, including kidney disorders and cardiovascular diseases. It is thus important to detect the disease at an early stage to slow down its progression, lower healthcare expenditures, and enhance patient outcomes.

The history of the development of CKD diagnosis has been characterized by a significant change in former diagnostic methods toward novel technological-based predictive models, especially those that are based on artificial intelligence (AI) and machine learning (ML). In the early 2020s, the potential of such techniques determined in the foundational studies. The findings of a 2020 study showed that the diagnostic accuracy could be significantly enhanced with the help of an optimized prediction algorithm based on a Gradient Boosting Classifier [18]. At the same time, a study of Adaptive Hybridized Deep Convolutional Neural Network (DCNN) revealed that it would be able to produce high levels of classification on the detection of CKD using clinical data [16]. These works formed the basis of later AI-based discoveries in the field of nephrology.

Continuing on this, the comprehensive comparison of different ML models (such as Support Vector Machines (SVM), K-Nearest Neighbours (K-NN), and ensembles) to identify the best method of CKD prediction was conducted in 2021 [35]. Another important feature of this era is the importance of data quality, since research has focused on the significance of selecting features and preprocessing them effectively to improve machine learning model performance in predicting CKD [21]. According to Iftekhar et al. [30], intensive data cleaning, such as handling missing values with median imputation of low-null features and random sampling of datasets with high rates of missing data, significantly increases the accuracy of the model, which allows classifiers such as Random Forest and Logistic Regression to achieve over 99% accuracies.

As early as 2022, traditional clinical records were no longer the only sources of

data to detect CKD. There was one prominent research that investigated how CKD detection can be achieved based on the sleep-wake behavior data processed through Long Short-Term Memory (LSTM) networks, revealing the potential of wearable sensor technology with combination of non-invasive, real-time health monitoring system [12]. The prediction models were further refined as one of the studies developed a Deep Neural Network (DNN) that was more effective than SVM and Random Forest by using new methods of feature selection [33]. These strategic needs of feature selection were further highlighted by Tazin et al. [10] who observed the ranking algorithms and information gain to select the most significant attributes that might reduce the feature set and significantly increase prediction accuracy with Decision Trees, having a maximum accuracy of 99 percent. This method was supported by the study of Khan et al. [19], where Forward and Backward feature selection algorithms were implemented to successfully reduce the dimensionality of the dataset (24 to 21 attributes) in a systematic way and hence improved the efficiency of computational processes and a high accuracy of the classification at 99.1% with Naive Bayes.

In the year 2023, more advanced models of deep learning frameworks has been developed. The Deep-kidney framework applied several classifiers in a majority voting scheme, the scheme predicts the occurrence of CKD at least 12 months prior to clinical diagnosis and shows that deep learning has the potential to revolutionize proactive CKD diagnostics [41]. In addition, an innovative hybrid deep learning network, HDLNET, which adopted the structure of Deep Separable Convolutional Neural Network (DSCNN) presented as a state-of-the-art artificial intelligence method for CKD diagnosis and aiding in preventive medicine [42]. Akter et al. [25] evaluated the overall performance of deep learning by comparing seven architectures, such as ANN, MLP, SimpleRNN, LSTM, and GRU, the results of which demonstrate that ANN and SimpleRNN can achieve the highest accuracy in classifying CKD stages with metrics of between 85% and 99% accuracy. Their study focused on the significance of stringent data preprocessing, use of linear and logistic regression to determine and fill the missing continuous and categorical data respectively to maintain data integrity.

This path of development, beginning with the earlier days of ML models and moving through sophisticated hybrid deep learning systems, shows a straight forward acceleration towards the creation of extremely accurate and data-driven predictive systems. These models are able to detect delicate, early warning signs of CKD, using complex datasets by integrating powerful algorithms and intensive data preprocessing, which help to intervene earlier, create individualized treatment plans and eventually, provide patients with improved healthcare outcomes all around the world.

1.2 Rational of the Study or Motivation

Although tremendous progress was achieved in using machine learning in Chronic Kidney Disease (CKD) detection, there is still an important gap between achieving high statistical recall on the one hand and on the other a set of models that are clinically practical, interpretable, and effective. Extensive literature has set up the

high effectiveness of traditional machine learning approaches, including Gradient Boosting and Support Vector Machines, on formatted clinical information [18], [28]. But such models do not necessarily address in terms of their complex, non-linear associations the intricacy of an entire set of medical data. One of the main difficulties impeding generalization is where there are an overlapping set of problems of data quality (such as class imbalance, noisy data, and very high urgency of useful choice of characteristics) [21]. As it can be seen, even powerful classifiers are extremely influenced by careful data processing and scope planning of information mining to further obtain effective results [30].

With the emergence of deep learning, these shortcomings began to be overcome with models such as CNNs and Adaptive Hybridized DCNNs becoming very promising in performance increase [16], [18]. However, these highly developed architectures come with new difficulties, and especially the issues related to computational speed, controllability of models, which are critical to their connector into the clinical practices. The increasing demand backed by the unmet requirement is the need of models bordering not only a great amount of accuracy, but it must be computationally viable and offer insights that can be depended upon and comprehended by anyone of the healthcare professionals.

Besides, the range of data used in predicting CKD has been limited in the past. In the traditional models, the clinical data practically used is structured data in clinical tests, ignoring other possible valuable behavior and physiological data that may be found in other sources. The opportunity was recently noted in a groundbreaking in-depth study of the 2022 study based on the LSTM networks that analyzed sleeping and rates of wake behavioral data in wearables and showed the possibilities of non-invasive and continuous observations in early diagnosis of diseases [12]. This emphasizes the need to expand the features universe to cover other traditional biomarkers to enhance the strength and predictive ability of screening systems.

This study is motivated by the worldwide health issues of CKD, which pose the significant global health concern and the critical need to gain an early diagnosis. Normal diagnostic procedures tend to diagnose the illness only in later stages, which considerable limits the choice of treatment methods and increases healthcare expenses. This research proposal has been pegged on the aim and people desire to use deep learning to develop effective, non-invasive diagnostic methods that can member CKD in its early stages and hence treating the disease early enough to prevent complications and to end the patient outcome. Nevertheless, an initiative to incorporate the most intricate deep learning model is not adequate. With the evidence of Khan et al. [19], it can be seen that strategic feature reduction will preserve or even increase the accuracy but will considerably lower computational efficiency. In a similar vein, it can be seen that the featured learning model of Tazin et al. [10], namely the selection of features based on ranking algorithms and information gain, proves that a subset of refined clinical attributes can be used to achieve great accuracy and makes the model simple enough without decreasing the predictive ability.

The need to eliminate the gap between the state-of-the-art algorithmic performance and clinical practicability therefore drives this research. It attempts to solve the

tripartite quandary of accuracy, interpretability and efficiency. Through the systematic comparative cross-examination of the conventional machine learning, ensemble approaches, and neural networks, along with rigorous feature selection, and data preprocessing based on the recent output of many studies, the proposed research is expected to emerge with a solid and deployable model to help achieve early detection of CKD. That is the end to end aim, to meanwhile give a tool that not only performs at a high level of predictive performance, but is also clinically constructible in the real world, in a bid to ease the worldwide burden related to kidney disease.

1.3 Problem Statement

Chronic Kidney Disease (CKD) is a severe health issue that poses a major health problem globally given that it is asymptomatic in the early stages and irreversible once it has developed. The disease afftects around 850 million people across the world and even its mortality rates are higher than most cancerous diseases, however it has been largely ignored in the public health initiatives. Conventional diagnostic methods, which are mainly dependent on biomarkers, like serum creatinine and estimated glomerular filtration rate (eGFR), usually discover the condition once renal impairment has occurred severely, mostly when 40-50 percent of renal capacity is already compromised. Such a delayed diagnosis is a severe impediment to early intervention opportunities, resulting in disease progression that is inevitable which worsen the health outcomes and increases health care expenses as patients move on to end-stage renal disease needing dialysis or transplantation.

The introduction of the field of artificial intelligence in health care has come up with the promising solutions based on the machine learning (ML) and deep learning (DL) frameworks. The first and early ML techniques, such as Gradient Boosting and Support Vector Machines, had better performance compared to traditional statistical models but faced significant problems such as the complexity of high-dimensional data, a high level of class imbalance, and the lack of model interpretability [18], [28]. Architectures of deep learning systems, especially Adaptive Hybridized Deep Convolutional Neural Networks, obtained impressive classification performance but implied the target computational costs and deployment challenges [16]. The high-level frameworks tend to operate as black box systems, offering not much transparency regarding their process of decisions making and weakening clinical trust and adoption.

One of the fundamental limitations in both ML and DL policies is related to the availability and quality of data in the field of CKD. There is often a great deal of incompleteness in medical datasets because patients are non-compliant and medical monitoring is haphazard and at times inconsistency arises because of the irregularities in documentation. According to Ifterkhar et al. [30] to enhance the model accuracy, the missing values should be handled using mean imputation along with appropiate data preprocessing. Moreover, most existing models are trained with structured laboratory data and leave unnoticed potentially valuable sources of heterogeneous information, such as lifestyle and physiological patterns. A study in 2022 demonstrates the potential implementation of LSTM networks with sleep-wake cycle data, however, multimodal data streams integration under CKD predictions studies remains significantly underdeveloped [12].

Another important weakness of the existing methods is the process of feature selection. In the spirit of Tazin et al. [10] who utilized the rankings algorithms and information gain, as a indicator of the most important attributes, successful feature selection is necessary to compress the volume of data yet remain able to predict variables. Correspondingly, it was shown by Khan et al. [19] that high accuracy can still be gained with the help of systematic use of Forward and Backward selection algorithm to reduce the number of necessary features simultaneously improving the computational capacity. These fixed methodologies have led to the creation of many models today where their main focus is on performance improvement at the expense of being interpretable, which means the model overlook the biological processes underlying the predictions.

Recent improvement of hybrid deep learning models like HDLNET have produced spectacular results in prediction, but the models remain impacted with central problems related to the size, level of applicable patterns to populations, and overall clinical interpretability [42]. This fields prioritizes the model sophistication than real world utility and as the architecture become more complex, the computation demands rises with limiting the accessibility of the resource constrained healthcare systems. Moreover, the models that are available often are not cross-population valid, meaning that their testing performance diminishes with the demographic groups that were not trained in the study. This lack of generalizability is a pivotal impediment to either large-scale clinical implementation as well as equal healthcare provision.

The existing CKD detection procedures tend to be unavailable or inefficient on actual and irregular data.So there is urgent need of development of powerful, effective, and clinical prediction models that have the potential to provide true reflection of CKD using imperfect data. This has been a question that is not only technical oriented but also practical implementation related by considering the notion of the computational needs and how it may and should be integrated into existing clinical processes and the development of proper interfaces that will enable the clinical decision making process.

In our study, we wish to resolve these complex issues on the basis of a unified predictive model that comparatively compiles the old concept of machine learning, ensemble models, and neural networks in order to reorganize their code arrangement. The suggested framework is narrowly focused on the three key issues of effective features selection and mitigating these issues by means of working with data imbalances, and the optimization of computational efficiency that held back earlier designs. The proposed solutions, involving the integration of best practices in the literature and the consideration of underlying gaps within the existing bodies of research, will enable us to come up with a model that offers high likelihood of being predictive and developable in a variety of healthcare facilities.

The framework is designed with a focus on interpretability and predictive accuracy making sure that the models will deliver not only classification results but also clinically applicable insights. This paper employs stringent comparison of various al-

gorithm styles in standardized measures of performance, such as accuracy, precision, recall, F1-score, and ROC AUC to establish optimal trade-offs between predictive power, computation complexity, and workability in its implementation. The final aim is to create a powerful, scalable, and explainable decision-support system (DSS) that will facilitate prompt detection of CKD in a diverse healthcare setting, thus filling a major gap in the healthcare requirement in preventing nephrology cases by timely treatment plans.

This study squarely addresses the constraints which have always impeded the past generation of computational models of CKD prediction and brings both methodological and pragmatic aspects of implementation together to transcend the current technical provisions and reach the true potential of clinical applicability towards a computational method. Our methodology aims to define a new standard of CKD prediction models, which balances both its analytical abilities with complex techniques and its functionality needs to meet the demands of a desireable standard of practical classifier implementation in a wide range of clinical legal institutions.

1.4 Objective

The goals of the research are to create a powerful and clinically usable framework of early Chronic Kidney Disease (CKD) advanced diagnosis with a systematic review and comparison of various machine learning paradigms. The objectives of this particular work are:

- Reduce at least 50% of the features through dimensionality reduction in order to reduce the complexity of the model and isolate the risk factors of CKD detection.
- Identify and select the most effective model based on the accuracy, F1-score, ROC-AUC and calibration measures.
- Optimize the selected model's runtime and size compared to the other models and select the most efficient architecture.
- Develop a lightweight model and a processing pipeline that is more suitable for deployment on IoT devices.

1.5 Methodology in Brief

The study has an in-depth and systematic way of conducting research to construct and test the predictive models of early detection of Chronic Kidney Disease in a systematic way organized into five interconnected phases. The approach is based on a combination of the existing data preprocessing methods with a strict comparative study of various machine learning paradigms to find the most reasonable balance between predictive performance and clinical utility.

Our approach is based on data collection and preprocessing. We apply an integrated data set which interpolates the well known CKD dataset of the UCI machine

learning repository and another augmented synthetic dataset, to form a formidable combination of 20,938 patient records. Following the steps followed by Iftekhar et al. [30], our data cleaning process uses mean imputation for continuous clinical measurement and mode imputation for binary clinical indicators and meticulously handles missing values. Systematically, categorical variables are transformed into standardized binary numerical data to make it compatible with computational algorithms.

The next important part of our approach is feature engineering and selection that is designed to improve the efficiency and interpretablity of our model. Taking inspiration from the work of Tazin et al. [10] and Khan et al. [19], we compare several methods of feature selection, such as the methods of mutual information ranking, Random Forest importance and Forward/Backward elimination. This is repeated until 12 clinically relevant factors including blood pressure, serum creatinine, hemoglobin and albumin are chosen whereby dimensionality reduction to a specific level of 50 percent is obtained without losing much predictive strength.

In order to overcome the high level of the class imbalance displayed in medical data, we implemented Synthetic Minority Over-sampling Technique (SMOTE) to create artificial samples to mitigate the class imbalance of the minority population, thus avoiding the development of bias in the model and enhancing the prediction of positive CKD cases. Moreover,the numerical characteristics are brought to a certain range using Standard Scaler method in order to provide a comparable range of values of different clinical measurements.

Our research is centered on the strategy of development and evaluation of the model in three levels. We systematically compare and contrast three different machine learning paradigms; Traditional Machine Learning models, e.g., Random Forest, XGBoost, and Support Vector Machines, to set a performance baseline; Ensemble Learning models, e.g., Stacking, Voting, and Bagging classifiers, to exploit the complementary behavior of multiple algorithms; and Neural Network models, e.g., shallow networks, deep bottleneck networks, to investigate complex-looking, hierarchical pattern recognition. This multi-faceted steps creates a definitive comparison assessment.

Each models are hyper parameter tuned extensively through systematic search methods in an attempt to maximize performance. The evaluation metrics follows several measurements such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC) [19]. A strict train-test split that is stratified will guarantee the data is representative, and all the models will be trained with the stratified 5-fold cross-validation to ensure the performance is consistent and help to avoid overfitting. Miscellaneous computational needs such as training time and model size are also monitored to measure practicability of deployment.

Lastly, we are able to carry out end-performance evaluation based on the overall performance analysis to finalize design changes to choose the best model to be used in clinical practice. This is by choosing the most performing architecture under

every paradigm and optimizing it, and the resulting model providing high accuracy, computational efficiency, and interpretability to be integrated into healthcare decision-support systems. The structured and evidence-based approach will guarantee the development of scalable, precise, and clinically feasible framework for the early detection of Chronic Kidney Disease.

1.6 Scopes and Challenges

The proposed study is aimed to create a reliable and efficient ensemble and deep learning-based model for early chronic kidney disease detection. The scope of the research includes relevant data preprocessing, full feature selection, and comparison of several machine learning paradigms. There are a number of singnificant issues to overcome to achieve good performance. The major challenges include data quality, such as missing clinical values, imbalance in the classes with non-CKD cases significantly exceeding positive clinical cases, and complexities during integration of data sets from different sources [18], [28].

There is an additional complication induced by model optimization, especially trying to create a performance/computational balance. Although beyond the accuracy of individual methods, ensemble methods such as Enhanced Stacking achieves high accuracy(87.7%) but requires significant resources like to train the model it takes more than 444 seconds and model size of 50.2 MB. On the same note, deep neural architectures, specifically the extreme bottleneck model with 15.41 million parameters, also show competitive performance (70.6% accuracy) at only 41 minutes of the training time, which is impractical to apply in resource-constrained healthcare hospitals.

Clinical interpretability issue, coupled with predictive performance, is also addressed by the study. Although high-tech models such as stacking ensembles and deep neural networks are capable of performing advanced pattern recognition, they may be often too opaque in the reasoning used by healthcare workers.

Despite the difficulties, the study sets out a multipurpose guideline on the detection of CKD that meets the significant health requirements. The study limits itself to identification of cost-sensitive models and general applicability of currently proposed ideas on real-time wearable data integration for the well resources health facilities to the limited resource facilities. By systematically addressing data quality challenges, optimized model selection and trade-offs between performance and efficiency, the study will further improve the development of AI-assisted diagnosis systems to identify early chronic kidney disease.

Chapter 2

Literature Review

2.1 Preliminaries

Chronic Kidney Disease (CKD) is a progressive disorder that is marked by gradual and irreversible impairment of kidney functions over a span of months or years. Clinical signs of the disease, including a continuously low estimated glomerular filtration rate (eGFR) and albuminuria, are used to identify the disease. Early diagnosis in clinical practice is still a big issue since early phases of CKD is usually non-infectious or asymptomatic, as a result the symptoms are experienced at very advanced stages when most of the kidney portion has been damaged. This diagnostic latency has inspired a search of data driven methods of detecting the disease at treatable stages.

Machine Learning (ML) and Deep Learning (DL) have become the potent approach to support diagnostic and predictive modeling in nephrology. With these techniques, most research starts with structured clinical datasets, which have proven to be highly promising in the CKD classification tasks. The classical ML algorithms such as Support Vector Machines (SVMs), Decision Trees, and Random Forests have been widely used to perform this task [28]. These models are very critical in meticulous feature engineering and thorough data preparation to achieve accurate prediction. As presented in the approach by Iftekhar et al. [30], it implies the stages of critical preprocessing, including handling of missing data in the form of imputation and the transformation of the data into the correct forms to guarantee data quality.

The recent progress is driven by Deep Learning architecture like Adaptive Hybridized Deep Convolutional Neural Network that makes it feasible to extract features and identify patterns at a variety of levels in complex medical data in an automated manner [16]. These models have the ability to reveal non-linear and complex relationships between the data that is not always revealed in standard algorithms. Also, a broader range of data that is used to predict CKD is beyond the established clinical scales. A 2022 paper showed the study of the sleep-wake state through LSTM networks and showed the prospects of using wearable sensor data to classify chronic diseases and provide an opportunity to predict them earlier and more quantitatively [12].

The major dilemma with the use of ML to detect CKD has been the trade-off

between predictive accuracy and model interpretability. Although current models such as HDLNET framework are capable of achieving high classification performance [42], black-box models may not be easy to adopt by clinicians because the framework lacks explanation which makes sense tob the medical practitioners. This has prompted the development of interest in hybrid AI models and post hoc interpretability methods like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) to explain model answers [21]. Effective feature selection is also required in order to develop strong models with efficiency and predictive power. As illustrated by Tazin et al. [10] and Khan et al. [19], it is essential to adopt ranking algorithms and systematic approaches of selection such as Forward and Backward elimination in identifying the most important clinical attributes, thus improving the efficiency of the models and their prediction power without interfering with their predictive power.

The predictability of such predictive models is the most important thing. Typical performance measures are accuracy, precision, recall and F1-score, which have a more global perspective on classification ability of a model [28]. Methods including k -fold cross-validation and tuning hyperparameters are widely used to achieve model robustness and generalization to avoid overfitting and maximize performance [18]. A broad range of different architectures, including basic ANNs and more complicated systems, such as LSTMs and GRUs, can be used to evaluate the best approaches as detailed by Akter et al. [25]. The current integration of heterogeneous sources of data, enhanced computational feasibility through GPUs, and attention to translational relevance focuses remain to elicit investments into the field, leading to the emergence of robust and the state-of-the-art ML and DL models that have great potential of improving CKD diagnosis and patient outcomes.

2.2 Review of Existing Research

Traditional Machine Learning Approaches

Most recent work in the domain of CKD has improved detection and diagnosis, in addition to various studies leading to predictive models for CKD. Previous work in this domain used traditional ML algorithms for improved classification performance. In 2020, Al Bahar et al. [18] introduced an optimized CKD prediction approach which implements Gradient Boosting Classifiers exhibiting a greater diagnostic accuracy than conventional methods. In parallel, a Deep learning approach was investigated for the classification of CKD based on the Hyphenated Deep Convolutional Neural Network in the work [16]. These early studies paved the way for ML-based CKD detection. According to a study by Marwa Almasoud and Tomas E. Ward in 2019, ML algorithms for CKD detection (LR, SVM, RF, and GB) were evaluated, with Gradient Boosting achieving the highest accuracy of 99%, along with 98.8% sensitivity and 99.3% specificity [14]. The study identified albumin, hemoglobin, and specific gravity as the most critical biomarkers, reducing diagnostic costs by 94% by focusing on these three predictors. However, the study's conclusions were limited by a small dataset (400 instances) and homogeneity in the data, which may affect generalizability [14].

In 2021, the topic has shifted towards comparisons between machine learning models. SVMs, K-Nearest Neighbors (KNN), and ensemble learning techniques are analyzed for performance to predict CKD [28]. Also in that year, an additional study applied a ML framework illustration validating the importance of feature selection and data preprocessing, asserting that a model should not only reach a good performance but also be interpretable and clinically relevant [21]. These experiments emphasize the role of more efficient data preprocessing and the application of hybrid AI models. Pankaj Chittora et al. [26] (2021) addresses dataset imbalance and feature redundancy by combining LASSO Regression for feature selection and SMOTE for class balancing. The optimized Deep Neural Network (DNN) achieves a record 99.6% accuracy in CKD prediction, while the Least Squares Support Vector Machine (LSVM) with L2 penalty achieved 98.86% accuracy [26]. These techniques significantly improve the model reliability and computational efficiency, though the study acknowledged potential biases from its small, imbalanced dataset. Mirza Muntasir Nishat et al. (2021) evaluate eight supervised learning models—including Logistic Regression (LR), KNN, SVM, Decision Tree (DT), Random Forest (RF), and Multilayer Perceptron (MLP)—on the UCI CKD dataset. RF emerged as the highest performing model with 99.75% accuracy, attributes to its robustness against overfitting and effective handling of non-linear relationships [13]. Hyperparameter tuning using tools like RandomizedSearchCV played a critical role in optimizing model performance, though the study's reliance on a single dataset limited its generalizability to diverse populations [13].

In 2022, Ahmed et al. [30] performed a performance analysis of eight machine learning algorithms to predict Chronic Kidney Disease (CKD) using the UCI dataset that consisted of 400 instances and 25 features. Their procedure consisted in the treatment of missing values by means/mode and random sampling imputation procedures, then coding features and 80:20 data division. The models that were reviewed were Random Forest, SVM, Naive Bayes, Logistic Regression, KNN, XGBoost, Decision Tree, and AdaBoost. Their results showed that the best performing classifier was the Random Forest with the highest accuracy of 99% and the Logistic Regression with the highest accuracy of 99 per cent and then AdaBoost, XGBoost and Naive Bayes with the highest of 97 per cent. On the other hand, KNN model had the worst performance with an accuracy of 73%. The researchers have amplitude of results that indicated that though these results are encouraging, the models still need more validation on larger and more diverse datasets before they can be considered to be viable in clinical production settings.

After these studies, Shawni Dutta and Prof. Samir Kumar Bandyopadhyay [17]; performe a model based on neural network trained on the UCI CKD dataset (400-records, 25 attributes). The model consists of three dense layers and trained with Adam optimizer and binary cross-entropy loss function, using 10-fold stratified cross-validation for unbiased evaluation. This model demonstrated an accuracy of 98.25%, F1-score of 0.98, Cohen's Kappa score of 0.964, and Mean Squared Error (MSE) of 0.0175, outperforming the SVM (96.21%), K-NN (91.67%), Decision Tree (94.7%) and Gradient Boosting (97.73%) models. Their study showed that neural networks had a better generalization ability in CKD classification, but the authors also noted that a larger number of datasets and clinical trials are necessary tasks for real-world

utility. Researchers develops the model against five conventional classifiers: Logistic Regression, KNN, SVM, Decision Tree, and Naïve Bayes. For their analysis, they used the UCI CKD dataset, containing 400 records with 24 different clinical attributes, but preprocessed this data with mean imputation for numerical values and most-frequent replacement for categorical features. Using a 12-layer neural network with dense layers, activated by either ReLU or Sigmoid functions, and dropout layers to limit overfitting. Traditional ML models were unable to achieve the same level of accuracy, with the DNN model exhibiting 100% accuracy and a 1.0 AUC score showing complete predictive capability, indicating the novelty of deep learning in the first phase of CKD.

The most recent study, perform by Dutta et al. [44] also focused on KNN and Naïve Bayes, and reaffirmed KNN's dominance, achieving 100% accuracy with 15-fold-cross-validation. They focused on data scaling (robust, z-standardization, min-max) and preprocessing to obtain better classification results. Their results further confirmed that feature selection, hyperparameter inflation, and data balancing techniques made KNN the most accurate and trustworthy model of CKD classification. They highlighted the dependence on a single dataset as a restriction and called for more testing in clinical settings. The comprehensive study on Deep Learning techniques including representation and architectures, followed by application of deep learning for disease prediction up to date is presented in 2023 by Yu et al. [43]. According to the researchers, deep learning is a crucial computational method as it produced better performance than automatic feature interpretation and was able to surpass even classical machine learning systems and match or outperform human interpretation competence. This paper thoroughly classifies deep learning algorithms into two primary classes depending on the type of data they mainly operate on: structured and unstructured. In this text, the ANN and Factorization Machine-Deep Learning (FM-Deep Learning) structured data algorithms are investigated. Deep learning algorithms have shown they have phenomenal capability to evaluate structured medical records using complex mathematical abstractions that can approximate models yielding any complex function over the data and allowing for rapid disease outcome predictions. It also describes unstructured data algorithms, which demonstrate how Zoop technologies extract visual information from the analysis of CNNs and sequential data from RNNs, particularly useful for medical imaging and electronic health records. The study showed how neural network systems developed from simple ANNs to complex CNNs and RNNs, which has led to much better accuracy for different types of medical data in many clinical applications. In particular, when applied with clinical definitions undergoing observation via medical image and genetic analysis, these models show increasing accuracy and efficiency in modeling large and complex datasets.

Sawhney et al. [41] publish an impactful research article in Decision Analytics Journal 2023 that evaluates artificial intelligence (AI) models as diagnostic tools for the early identification of CKD. CKD progresses insidiously to irreversible kidney damage; hence, physicians should detect it as early as possible, and for that, a better diagnostic tool is necessary. This research proposes a groundbreaking DNN-based MLP Classifier, which has the potential to notably surpass the performance of current ML methods, such as SVMs and NB classifiers. The authors learn early-stage

CKD patterns by training their model on data from 400 individuals with multiple biomarkers including blood sugar, age, and red blood cell count. Furthermore, according to a rigorous benchmarking process, the MLP classifier implemented in PyTorch performs best on non-linear data setups, which are the norm in medical diagnostics. Proposed CKD detection enhances from dataset attribute reports along with a custom architecture capable of precise identification leading to accuracies in the testing set that are perfect, according to reports. The final results show that MLP outperforms every other diagnostic system in the evaluation with a 100% correct performance rate The model obtains a very high predictive accuracy, indicating not only its potential to be usable but also its practicality, illustrating how deep learning can help in the process of kidney disease diagnosis.

Deep Learning Architectures

In 2021, Akter et al. [25] comprehensively proposes seven deep learning models—ANN, LSTM, Gated Recurrent Unit (GRU), Bidirectional LSTM/GRU, Multi-Layer Perceptron (MLP), and Simple Recurrent Neural Network (RNN)—for CKD prediction. The ANN model achieved the highest accuracy of 99%, validated statistically using the Wilcoxon signed-rank test, while MLP and Simple RNN followed closely with 97% accuracy [25]. The study propose integrating these models into Internet of Medical Things (IoMT) platforms to enable real-time patient monitoring, though computational inefficiency (e.g., Bidirectional GRU required 91 seconds per epoch) was noted as a scalability challenge.

Building on this previous work, Mondol et al.[35] compare optimized deep learning models (three in total: an Optimized Convolutional Neural Network (OCNN), an Optimized Artificial Neural Network (OANN), and an Optimized Long Short-Term Memory (OLSTM)) against traditional models to classify CKD. On the UCI CKD dataset containing 400 patient records, OCNN beat out all models in this case with 98.75% accuracy, an AUC score of 0.99, and the lowest compilation time (0.00447s). The study proposes the application of optimally tuned deep learning methods focused on achieving higher diagnostic accuracy, with OCNN proven to be the most efficient model. The authors suggested further study to mitigate overfitting, apply k-fold cross-validation, and implement hybrid models for better performance.

In a recent research by Jhumka et al.[33] from the University of Mauritius, they present their results on a strong Deep Neural Network (DNN) architecture towards the Early detection of CKD that penetrates through the diagnostic way of medicine. Their study is of great importance as CKD, a global health problem which is affecting tens of millions of people and is one of the leading cause of global mortality. This finding is of crucial importance as Mauritius suffers from high rates of CKD incidence and mortality. A testing of the specially created DNN model against the widely used Random Forest classifier based on a rigorous development process followed up on this last DNN implementation approach using the UCI machine learning repository dataset. A total of 400 patient records span 25 distinct attributes, hence facilitating an ideal platform upon which the kNN imputation data preprocessing method—used for missing values—could be applied as well as outlier transformation solutions to address data fairness and persistence. The DNN model proposed in this

study performed successfully on the CKD dataset with an accuracy of 98.8% in the classification of CKD vs. non-CKD. This level of performance not only shows the effectiveness of the model but exceeds previously used machine learning algorithms that had been working with lower levels of accuracy. The poster outlines a performance estimation analysis, detailing how DNN models excel at recognizing intricate non-linear relationships, evading detection by traditional algorithms.

The study of DL algorithms for CKD prediction by Saif et al.[47] gives insights about the development of three separate predictive models based on CNN, LSTM and deep ensemble models that are able to predict the onset of CKD as much as one year prior to clinical diagnosis. The originality of this study lies in the combination of CNN and LSTM models under a deep ensemble framework applying majority voting as the predictive accuracy improvement approach. Two public datasets were used to test the deep ensemble model, leading to exceptional predictive accuracy with scores of 0.993 and 0.992 for predicting CKD six or twelve months in advance, respectively. In deploying deep learning methods adaptable to the unique temporal and spatial data analytics needs of medical diagnostics tasks like CKD nascent signal detection, the study represents methodological firepower. The models utilize a lot of extensive data preprocessing, feature extraction, and iterative training techniques to improve the outputs. The study describes its methods thoroughly such a foundation paves the way for future research using these methods.

Hybrid and Ensemble Models

To overcome the limitations of this method, in 2023, new techniques based on hybrid deep learning models got introduced. As a notable example, HDLNET being a Deep Separable Convolutional Neural Network (DSCNN)-based model, can improve classification accuracy remarkably, implying the potential for AI-powered tools and solutions in medical diagnosis [42]. However, there are still challenges to address, including model scalability and computational interpretability and efficiency, which require further exploration to promote the adoption of these solutions in a clinical setting.

Non-Invasive and Multimodal Data Integration

By 2022 research work had also started to investigate alternative data sources beyond standard clinical features. LSTM networks have been proposed for chronic diseases in [12], but specifically for the detection of CKD disease through sleep-wake behaviour analysis. Surprisingly on the other hand, wearable sensor data could be truly beneficial for non-invasive detection of potential disease, if we nail it a new world in this field opens up on the horizon as we are approaching possible 24/7 human ecosystem monitoring and adaptive therapy. Navaneeth Bhaskar and Suchetha M. (2019) proposed a novel non-invasive approach for CKD detection using a custom-designed salivary urea sensor. The sensor data were processed through a hybrid 1D CNN/SVM model, achieving a classification accuracy of 98.04% [15]. This method avoided invasive blood tests and reduced discomfort, though the study was limited by a small dataset and variability in salivary urea due to external factors like diet or hydration [15].

In 2020, Ma et al. [20] published their work in the "Future Generation Computer Systems" journal, which provided several techniques for the advanced computational modeling of leading CKD diagnosis methods. The rise in the prevalence of CKD has led to increasing challenges in accurately diagnosing the disease, and effective early diagnostic tools are needed to prevent the progression of CKD to end-stage renal disease (ESRD). Linking support vector machines (SVM) with backpropagation algorithms of Multilayer Perceptrons (MLP) to introduce a Heterogeneous Modified Artificial Neural Network (HMANN) by comparing the backbone analysis of this research for improving kidney-related anomaly diagnosis as well as segmentation using ultrasound imaging. Most of the advancements come from its use on the Internet of Medical Things (IoMT) platform to enable better data unification and automation processing to aid real-time medical diagnostics. The first stage of HMANN model improves ultrasound images through preprocessing to tackle common kidney imaging problems like low image contrast and speckle noise. Wavelet transformation methods are used to minimize noise distortions so that the image clarity improves for future analysis. Since this method demonstrates superior performance in obtaining accurate kidney contour outlines, the study demonstrates paralleled kidney region segmentation using HMANN. Exact segmentation of important diagnostic features, such as kidney stones or other calcifications, are also retrieved using the method. SVM and MLP fall under the category of classifiers that take advantage of insights gained from earlier phases of the pipeline. This two-step approach allows the model to internalize complex data patterns, resulting in significant gains in diagnostic accuracy. The performance of HMANN is derived from analyzed results compared to traditional machine learning methods. At the same time, this system provides high speed computational ability and performance degrees along with remarkable exactness. Model performance metrics reveal excellent classification accuracy, superior performance on preliminary tests, and confirm the model's applicability in clinical settings. The performing capabilities of the HMANN model processing with larger amount of information sets obtained an accuracy ratio which was bigger than 89.9% and concluded to be more effective than standard diagnostic methods.

Feature Engineering and Interpretability

In a significant contribution to optimizing ML models, Arif et al. [37] embrace creative preprocessing techniques, feature selection (Boruta algorithm), and hyperparameter optimization in their methodology to enrich CKD detection. They performed analysis of KNN and Naïve Bayes (NB) on the UCI CKD dataset (400 samples, 24 attributes). They reported a perfect 100% accuracy, precision, recall, and F1-score for their KNN model, while for NB they reported 97.5% accuracy, 100% precision, 96.15% recall, and 98.03% F1-score. Using techniques of grid-search CV and iterative imputation for missing values made their models resilient to data inconsistencies (They noted that optimized ML models worked well, but that they would need to validate them on other datasets.

Islam et al. [39] in their article make significant contributions in predictive healthcare analytics using machine learning techniques for early detection of CKD in 2023 Journal of Pathology Informatics. This chronic kidney disease must be diagnosed

early in order to be prevented from developing into an irreversible end-stage renal failure that requires ongoing management of dialysis or a kidney transplant. Here this supervised learning framework enables a deep dive of twelve ML algorithms to identify the most effective CKD prediction model. Their methodology was built from a large UCI Machine Learning Repository dataset that included 24 predictive features ranging from simple biochemical markers to complex physiological factors. What is more, XGBoost classifier has maintained the highest accuracy of 98.3% along with similar values of precision and recall, which proves its capability to predict CKD with reliability. XGBoost model has the highest accuracy by effective processing the non-linear process, which is usually available in the medical databases. Through this method, ensemble learning methods are proven to have the ability to increase predictive accuracy in healthcare applications. This paper explores black-box learning models, revealing critical insights not only related to the diagnostic accuracy but also operational performance.

2.3 Summary of Key Findings

A thorough review of the current literature demonstrates noteworthy advances in the field of CKD identification. Early implementations employed traditional ML algorithms, Gradient Boosting Classifiers and SVMs that reached higher diagnostic accuracy but failed to address issues of non-high-dimensional datasets and class-imbalance problems inherent to these clinical datasets [18], [28]. Besides, high accuracy achieved by various models is one of the major takeaways. For example: Gradient Boosting: 99% ,DNN: 99.6% , Random Forest: 99.75% [13], [14], [26]. The results indicate the success of ML and DL models in predicting CKD accurately by using only minimal biomarkers like hemoglobin, albumin, and specific gravity [14]. This is consistent with the goals of the study.

Among the various phenomena, the hybrid of salivary urea sensors with 1D CNN-SVM models falls under non-invasive techniques, which adds major momentum. These techniques obtained an accuracy of 98.04% [15], allowing for accurate testing without needing invasive techniques like blood draws, which can cause discomfort and are less likely to lead to routine examinations.
A further critical finding is the cost-effectiveness of AI-enabled approaches. By focusing on a smaller number of important biomarkers, the cost of diagnosis was dramatically reduced—from $451.36 to $26.65 in one study [14]. The costs associated with this could decrease significantly, which is particularly relevant to the implicit goal of this study in developing affordable and easily accessible diagnostics, especially in resource-poor settings.

Nevertheless, there are many challenges, such as a small and imbalanced dataset, a very high computational requirement, and variability between different non-invasive measurements [15]. These limitations highlight the necessity of further work in the areas of dataset expansion, model optimization, and clinical integration, which is the aim of this study through proposed scalable and efficient AI solutions to real-world health systems. A 2022 study highlighted the ML model comparison, showcasing the need for feature selection and data processing [10]. Despite the fact that these studies demonstrate the improvement in interpretability and efficiency offered by

hybrid AI techniques, there remains Scalability and generalizability challenges in the context of these methods [21], [28]. Further model performance and reliability have been achieved by employing advanced techniques for feature selection, such as LASSO Regression and CFS, in addition to class balancing techniques such as; SMOTE [26]. Additionally, such techniques are essential to mitigate these issues, recalling that they are (i) critical with regard to imbalanced datasets, (ii) essential for robust predictions, and (iii) crucial to this study's central aim of creating generalizable, scalable models.

Furthermore, the researches in recent years has made major contributions in utilizing ML and DL methods to identify CKD more effectively for early diagnosis and improved clinical decision-making. It is seen through studies that the Deep Learning models continuously outperform the traditional ML classifiers, with the Deep Belief Networks (DBN) achieving a 98.5 percent classification accuracy [27]; Optimized Convolutional Neural Networks (OCNN) obtaining an accuracy of 98.75 percent [35]; and Deep Neural Networks (DNN) yielding a perfect-accuracy of 100 percent and an AUC score of 1.0 [36], indicating superior feature extraction and classification capabilities demonstrated by deep learning. The use of iterative imputation (K-Nearest Neighbors) and robust scaling, along with the Boruta feature selection, was determined to have had the most profound impact on predictive accuracy, with K-Nearest Neighbors (KNN) yielding 100% accuracy over 15-fold cross-validation [44]. With OCNN's fastest compile time at 0.00447s and grid-search cross-validation that fine-tuned the ML models to optimum performance, optimized models improved accuracy, enhanced computational efficiency and robustness. Nevertheless, validation and generalizability are major challenges; overfitting, dependence on imputation, and dataset diversity have impeded real-world applicability and highlighted the need for larger, diverse datasets, coupled with clinical trials. The findings are consistent with the study's goals of building a precise and computationally efficient model for CKD prediction by employing an amalgamation of high-yielding frameworks, preprocessing strategies, and validation mechanisms. This study focuses on hybrid models and diversification of datasets to improve applicability in a clinical environment, creating real-world diagnostics to use artificial intelligence in the healthcare space.

Author(s)	Year	Method Used	Dataset	Results
Almasoud & Ward [14]	2019	Gradient Boosting, Random Forest	UCI CKD (400 instances)	99% accuracy, 98.8% sensitivity, 99.3% specificity; cost reduction from $451.36 to $26.65
Ma et al. [20]	2020	Heterogeneous Modified ANN (HMANN)	IoMT Platform	89.9% accuracy; segmentation and diagnostic improvements
Dutta & Bandyopadhyay [17]	2020	Neural Networks (ANN, DNN)	UCI CKD (400 records)	98.25% accuracy; outperforms traditional ML models
Akter et al. [25]	2021	ANN, LSTM, GRU, MLP, RNN	UCI CKD (400 records)	ANN: 99% accuracy; MLP & RNN: 97%; noted scalability challenges
Chittora et al. [26]	2021	LASSO Regression, SMOTE, Deep Neural Network (DNN)	UCI CKD (400 records)	99.6% accuracy (DNN), 98.86% (SVM); dataset imbalance remains a challenge
Nishat et al. [13]	2021	RF, KNN, SVM, MLP	UCI CKD (400 instances)	RF: 99.75% accuracy; emphasized hyperparameter tuning
Mondol et al. [35]	2022	Optimized CNN, ANN, LSTM	UCI CKD (400 records)	CNN: 98.75% accuracy, AUC 0.99; fastest compile time (0.00447s)
Jhumka et al. [33]	2022	Deep Neural Network (DNN)	UCI CKD (400 instances)	98.8% accuracy; DNN surpasses RF-based classifiers in CKD detection
Iftekhar et al. [30]	2022	RF, SVM, NB, LR, KNN, XGBoost, DT, AdaBoost	UCI CKD (400 samples, 25 features)	RF and LR achieved 99% accuracy, while KNN had 73% accuracy in CKD prediction.
Arif et al. [37]	2023	KNN, Naïve Bayes, Hyperparameter tuning	UCI CKD (400 samples)	KNN: 100% accuracy; NB: 97.5% accuracy; importance of feature selection and preprocessing
Islam et al. [39]	2023	XGBoost, Ensemble ML	UCI CKD (400 instances)	98.3% accuracy; ensemble methods improve nonlinear data processing
Venkatrao & Kareemulla [42]	2023	HDLNet (Hybrid DL with DSCNN)	Intelligent IoT dataset	High accuracy; improved feature extraction; hybrid models enhance CKD diagnosis
Dutta et al. [44]	2024	Logistic Regression, RF, DT	UCI CKD (400 instances)	Logistic Regression: 100% accuracy, Random Forest: 89.96%

Table 2.1: Summary of Key Findings

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

This chronic kidney disease (CKD) detection project's technical and methodological framework will be designed in a way that will allow us to come up with a strong, interpretable, and clinically applicable predictive model. The requirements listed below are the guidelines on the specifications that are necessary to implement in the project.

I. Data Requirements:
The data set is gathered based on two main sources: UCI Machine Learning Repository's Chronic Kidney Disease dataset (400 cases) and a bigger publicly accessible dataset that is built on over 20,000 patient records. The final dataset that is integrated and cleaned is made up of 20,938 records of patients, each given 25 clinical and demographic attributes. These are both numeric and categorical variables, including blood pressure, serum creatinine, hemoglobin, albumin, and sodium, and hypertension, diabetes condition, anemia, etc. The target variable is a binary variable, which shows the presence or absence of CKD.

II. Data Cleaning and Preparation:
The systematic data cleaning protocol is applied to deal with the absence of the value, inconsistency, and discrepancy of the data type. Any gaps in binary clinical indicators (e.g. hypertension, diabetes) are filled in with the mode, whereas the means of continuous clinical measurements (e.g. blood pressure, serum creatinine) are filled in with the mean. There are no features with excessive missingness and there are no duplicate records. Outlier analysis is conducted but no extreme outliers are eliminated so that clinically significant cases are not eliminated.

III. Preprocessing Requirements:
Categorical variables will be transformed into a standardized binary numeric form (0/1). Synthetic Minority Over-sampling Technique (SMOTE) is used to produce artificial samples of the minority class (CKD cases), in order to correct the high class imbalance in the target variable. StandardScaler is used to standardize all the numerical features, so that the mean of features becomes zero and variance becomes one, such that the features will contribute equally during the training of the model.

IV. Requirements of Feature Selection:
Multi-method feature selection tool is used in order to emphasize dimensionality reduction and better model interpretation. Methods encompass random forest importance, mutual information ranking, forward feature selection, backward feature elimination and recursive feature elimination based on principal component analysis. The last feature set is 12 clinically meaningful predictors, which include blood pressure, specific gravity, albumin, sugar, random blood glucose, blood urea, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, and red blood cell count.

V. Modeling and Evaluation Requirements:
The project applies and compares three machine learning paradigms:

Traditional Machine Learning: Random Forest, XGBoost, SVM, Naive Bayes, Logistic Regression, K-Nearest Neighbors, Decision Trees, AdaBoost and Gradient Boosting [13], [14], [18], [28], [30].

Ensemble Methods: Ensemble methods such as Bagging, Hard / Soft Voting and Stacking classifiers are used in academic researches for their effectiveness in improving prediction and accuracy using combinations of different base estimators and meta-learners. Ganaie et al. [31] give a very detailed overview of ensemble methods like bagging, boosting, stacking, and others and illustrating their advantages in various fields. The review shows that ensemble learning has the capability to integrate a group of base estimators and meta-learners with their complementary advantages.

Neural Networks: Shallow (1-2 hidden layer) or deep (6-10 hidden layers) networks, e.g., bottleneck networks, residual-style networks.

Stratified 5-fold cross-validation is used to evaluate all the models. Measures of performance are accuracy, precision, recall, F1-score, ROC-AUC and confusion matrices. Practical deployability is also evaluated by evaluating the computational efficiency of training time, inference speed and model size.

VI. Software and Environment Requirements:
It is implemented in Python with the help of VSCode. The project is also implemented on a workstation with the AMD Ryzen 7 7700 CPU, 16 GB of DDR5 RAM, and NVIDIA GeForce RTX 5060 Ti graphics card, which was able to provide the required computational capabilities to train complex ensemble and deep learning models. The most important are `scikit-learn` [2] in traditional and ensemble models, `TensorFlow` [8] and `Keras` [7] in neural network models, and `imbalanced-learn` [11] in managing class imbalance. The version control is kept through `GitHub` and makes it possible to guarantee reproducibility and collaborative development.

VII. Last Model Choice Requirements:
A balance between a high predictive performance (especially Accuracy, F1-score and AUC), computation efficiency, model interpretability, and clinical relevance is used to select the best model. The preference is put on ensemble methods, particularly, stacking architectures due to their capability to integrate two or more learners and provide robust and generalizable performance. Deep neural networks have been optimized in situations whereby hierarchical learning of features is needed, but weighed

in terms of their limited practical implementation factors.

3.2 Societal Impact

The development of the early detection system of Chronic Kidney Disease (CKD) based on Artificial Intelligence has the promise of introducing a ground-breaking change on a health, social, legal, and cultural level. This project is not only about technical performance, but it also has an influence on patient safety, medical access and ethics. The timely detection of CKD is critical in averting the advanced stages of renal disease, which requires expensive methods of dealing with the disease such as dialysis or transplantation. With the help of models like Ensemble Learning and Deep Neural Networks (DNN), this system determines those who are at-risk prior to the development of symptoms where they are individually given medical care and the overall suffering of patients is minimized in the long term. Through the reduction of human errors of diagnosis, it improves patient safety. Furthermore, the system facilitates the inclusion of healthcare, especially the rural or resource-limited areas that have no access to healthcare facilities such as specialized nephrology. Such a move to replace laboratory analysis with AI-aided diagnosing enables the underserved to have access to modern medical services. Nonetheless, incorporating patient information has legal and ethical responsibilities. As much as the proposed project is based on reading datasets which are anonymized and publicly provided, the implementation on a real-world scale and setting should follow the legal regulations in data privacy such as GDPR and HIPAA, consent, transparency, and protection.These are crucial ethical aspects of data utilization and integration with clinical methods that would make the adoption of AI in nephrology responsible [56]. Ethically it is important that the system is just an aid and does not imitate clinical judgment. Also, the model needs to take cultural diversity into account. An individual, automated screening process is motivating and therefore needs to be changed to local language, writing level and beliefs so that proper and compassionate dialogue will be made. The manner of expression of predictions must be handled to avoid unnecessary anxiousness or misunderstanding. To sum up, the present CKD detection system has extensive social advantages since it will lead to enhancing early intervention and lowering the overall cost of healthcare services as well as to enhancing diagnostic equity and safety. It can be a critical resource to global health progress, used with the responsible design and culturally informed implementation.

3.3 Environmental Impact

Although the central subject matter of the CKD detection system is relative to health, its environment should not be overlooked, especially when talking of the sustainability of computing devices, resource utilization, and the overall effect of digital health tools. In this section, attention will be taken to an assessment of the environmental impact that the project will have in addition to the impact of addressing environmental issues in the sustainable environment of healthcare technologies. Deep learning architectures and other Artificial Intelligence (AI) models are computation-intensive to train and use. According to Schwartz et al. [23], one of the concerning issues in the scientific world is the ecological and energy price of

these intensive models. Nevertheless, this project focuses on effective model selection. The end result is based on Ensemble Learning, which is computationally less demanding, despite the fact Deep Neural Networks (DNNs) are investigated. In effect, the decision automatically downsize power consumption and carbon footprint of the project. It was created in Visual Studio Code on a local workstation. Such a local development method provides a regulated and effective model training environment. With direct utilization of a modern and power-optimized desktop system, we do not pay the incessant energy cost of large remote cloud computing data centers, which will help add to the total computational carbon footprint of the project. Such an AI-powered screening system can contribute to environmental sustainability because of its focus on preventive care. When CKD is identified in its earlier stages, it will decrease the dependence on more human, material, and resource-intensive medical procedures, such as dialysis or organ donation, which consume vast amounts of energy, materials, and leave a lot of biomedical wastes behind. Through making non-invasive and early diagnostics possible, the model also eliminates the impact of such high-tech medical procedures and the high number of hospital visits on the environment on an indirect basis. The usual method of diagnosing CKD relies on a multitude of laboratory work, physical records and clinical trials. Digitizing the diagnosis process will cut down on the use of papers, data input, and unwarranted transportation emissions as patients would not be traveling too regularly. The project supports several of the United Nations Sustainable Development Goals (SDGs), in this case, goal 3 (Good Health and Well-being), goal 9 (Industry, Innovation and Infrastructure), and goal 12 (Responsible Consumption and Production) [24]. It facilitates not only environmental, but also social sustainability by incorporating the low-resource diagnostic tools into mainstream healthcare. This project on detecting CKD perfectly illustrates environmental responsibility since energy consumption is relatively low, no significant number of resources is needed to carry out diagnostics. The future of work is to increase green AI practices, increase computational efficiency, and sustainability on large-scale deployment.

3.4 Ethical Issues

Although the creation of a machine learning-based chronic kidney disease (CKD) detector is a promising phenomenon, the development requires an in-depth analysis of the ethical implications and professional duties. The design solution consisting of the combination of a real clinical dataset with a large-scale synthetic dataset has definite ethical issues at stake that need to be preemptively handled in a manner that would make the model to be used in a healthcare context responsibly and safely.

One of the issues that present serious ethical concerns is model validity and integrity of data. The hybridization of a realistic hospital dataset (400 instances) and a significantly much bigger synthetic dataset (20,538 instances) produces a composite data base whose adherence to the real clinical complexity is per se ambiguous. Although synthetic data generation is a useful method of class imbalance, it can lead to the generation of an over-ideal dataset which may not fully represent the range of noise, edge-rare cases and complicated non-linear associations felt in real patient populations. When the model acquires these artificial patterns it will perform inflatedly in its testing but will not be reliably able to generalize in a realistic

clinical setting. This may result in wrong diagnosis, which is a direct violation of the central biomedical ethical principle of non-maleficence (to do no harm). Gunning et al. [32] suggests, The professional and ethical needs of using synthetic data in healthcare require it to be disclosed and externally validated to avoid exaggerated performance. The professional accountability in this case lies in performing careful and stand-alone verification of the final model using an entirely separate, real-world clinical dataset and, in turn, any deployment should only be thought of. This would be an important step since it would ensure that the performance of the model is not an attribute of the synthetic data.

The decision of ethical imperative of transparency and informed consent is closely tied to this. Even the definition of a model having been trained using more than 20,000 patient records may be misleading unless the fact that most synthetic data is presumably used may be clearly disclosed to potential users, e.g., clinicians and healthcare administrators. Not being clear about this fact compromises their willingness to give informed consent about the tool to be incorporated in their clinical operations and is also a violation of professional honesty. The research team has an independent obligation of making complete transparency of everything they do through communications, publications and system documentation and it should clearly define how the data that they produce got its origin and the constraints that this brings to the applicability of the model.

Lastly, algorithmic bias probability and health bias have to be taken into account. The biases which may have existed in the smaller hospital data of the original, such as those based on demographics, socioeconomic position, or a certain clinical practice, could be boosted and baked into the synthetic data generation procedure. A model that has been developed using this data can make uneven predictions between the groups of patients and this can worsen the health disparity that already exists. It's professional responsibility should be to perform comprehensive bias audits and the evaluation of fairness on the sensitive features such as age, gender, and ethnicity. Additionally, it is imperative to make the output of the model firmly a clinical decision support system, rather than a replacement diagnosis. The need to make final diagnosis and treatment decisions once a qualified healthcare professional should always be considered, and the human factor, context and specifics of a patient should always prevail over algorithm forecasts, which helps to hold on to the concept of patient autonomy and eliminate the danger of automation bias.

3.5 Project Management Plan

A project management structure should be implemented to have smooth and quality delivery of the CKD detection system on time and within the allocated resources. This scheme combines the concept of planning schedules, budgeting, and the joint effort of managing resources by applying agile concepts and contemporary software engineering concepts to maximize success, following established framework for recurssive development [22].

Schedule and Workflow: The project will take 16 weeks, and it will be separated into five iterative phases of agility. The first stage of initiation and design,

which will take place during week 1 to 3, is devoted to team kickoff, thorough literature research, goal setting, scope definition and a focus on familiarizing oneself with CKD pathophysiology and defining data and ethical needs. The data engineering stage, which spans between weeks 4 and 6, revolves around a combination of real and synthetic data, exploratory data analysis to find the patterns and outliers and address the missing and integration of the data types rigorously. Weeks 7-10 mostly involve model development, which involves systematic feature engineering and selection, ensemble and deep neural network model design and training, and performing high-quality hyperparameter optimization that is confirmed by stratified 5-fold cross-validation. The remaining weeks 11-13 will be dedicated to the in-depth evaluation and optimization to address the model performance, memory and training effectiveness, the possibility to deploy the model to a clinical setting and the ability to conduct the robustness testing in the context of various applications. The last stage, weeks 14-16, is dedicated to documentation and knowledge transfer, which will result in the creation of technical manual and user manuals, facilitation of team presentation and workshops and sprint review to ensure a smooth transition of further implementations.

Budget Allocation: Cost planning symbolizes both scholarly implementation and prospective clinical scaffolding. The project will include costing of secure and HIPAA-compliant cloud storage and data backup at a cost of 10 dollars, internet connectivity, utilities and software license at 20 dollars, and miscellaneous costs and contingencies (consultations or printing) at 10 dollars in total, making the project estimated cost 40 dollars. This reasoned distribution shows that AI-based diagnostics of CKD, despite resource constraints, are cheap and feasible.

Resource Management and Structure of the Team: A team of five specialists is organized in such a way that each of them is divided by responsibilities, such as data preparation and engineering, modeling and creating algorithms, evaluation and optimization, documentation and communication, and a general management of the project and risk control. The team members have frequent standups and knowledge-sharing sessions, as well as collective troubleshooting, which helps to ensure a clear communication and constant improvement at all the stages of the project.

Technical Resources: The technical basis of the project will be built around a recommended workstation with AMD Ryzen 7 7700 CPU, 16GB of DDR5 RAM, and NVIDIA GeForce RTX 5060 Ti graphics card to allow the experimentation of local models and the optimization that is resource-intensive. It is developed using Visual Studio Code which provides an effective way to manage and collaborate with the code. The GitHub is providing version control and reproducible codebase, whereas the Trello, the Kanban approach, is providing an agile task management and sprint planning. Google Meet and Google Docs are used to simplify communication and documentation. Some of the tools and libraries used are Python, Jupyter Notebooks, `pandas`, `NumPy`, `scikit-learn`, `TensorFlow`, `Keras`, `imbalanced-learn`, `matplotlib`, and `seaborn`.

Best Practice and Quality Assurance: The project is continuously integrated throughout the project with weekly demonstration sessions and a retrospective meet-

ing. Periodic risk assessments are made on the workflow velocity, technical barriers, and the efficiency of resource utilization. The plan focuses on strong documentation, code clarity and reproducibility, the plan ensures clarity of knowledge transfer, compliance with clinical goals and the entire stakeholder expectations.

3.6 Risk Management

The successful approach to the development of the Chronic Kidney Disease (CKD) detection system required a proactive risk approach that would guarantee its technical stability, ethical soundness, and feasibility. According to Liu et al. [34], systematic and sustained risk management approach is critical when it comes to making machine learning applications in the medical field for safety and reliability . Potential risks are assessed at the data, modeling, computational, and ethical levels, and the mitigation strategies are included in the entire project life cycle.

Data Quality and Imbalance Risk: The original combined dataset of more than 20,000 records had a high imbalance of the classes, which is a great threat to allow the majority (non-CKD) to be biased in the model. In this regard, highly developed oversampling methods like SMOTE are applied strictly, along with strong preprocessing programs, which comprise systematic missing value interpolation (mode on binary, mean with continuous features), and strict feature estimation. Primary attention to the measures of the minority classes and ROC-AUC under cross-validation procedures protects the absence of bias and sensitivity loss.

Model Performance and Generalization Risk: One of the major risks is to come up with models that either overfitted the training data or are unable to reflect the non-linear relationships in the clinical parameters. We have used a multi-paradigm approach to evaluation. Rather than applying one of the types of models, we made a systematic comparison of Traditional MLs, Ensemble Methods, and Neural Networks. Stratified 5-fold cross-validation is also used to ensure robustness, and hyperparameter optimization is performed on all the top candidates. The process has shown that ensemble techniques (especially stacking architectures) provided the most appropriate trade-off between high-performance and generalization, so that they are chosen as a final model.

Computational and Resource Management Risk: There is a high probability of the complex ensemble techniques and deep neural networks with millions of parameters to outstretch our computational abilities and schedule. The project uses a local workstation with a large and well-equipped workstation (NVIDIA GeForce RTX 5060 Ti) that was used to do most of the development and training. The codebase is broken down, and model training is logged very carefully. This local environment gives the reliability and manipulation into large-scale experimentation without relying on the cloud session timeouts.

Risk to Ethicality and Interpretability: The application of clinical information and black-box complex models such as deep neural networks pose risks to patient trust and clinical adoption such as possible bias and transparency. Although the dataset was anonymized and publicly available, we have given interpretability a

priority. Native feature importance rankings are offered by the ensemble models, particularly, Random Forest and validated with clinical knowledge (e.g., albumin, hemoglobin). This concurs the decisions of the model to the established medical knowledge. In a practical implementation, a system that will fulfill the requirements of such regulations as HIPAA/GDPR, inform consent, and a good understanding of the purpose of the model as a decision support mechanism, not as a substitute to clinical judgment, will be obligatory.

By systematically conducting risk management, the project managed to steer through technical issues and ethical aspects which consequently led to an effective, well-assessed, and reliable detection system that is ready to undergo future clinical validation.

3.7 Economic Analysis

The creation of an early diagnosis of Chronic Kidney Disease (CKD) machine learning-based system is an important economic intervention that can greatly decrease the number of dollars incurred to healthcare systems. The economic explanation lies in the fact that the cost structure of CKD escalates with the severity of the disease so the expenses become very high in advanced stages of the disease. End-stage renal disease (ESRD) is the most significant economic burden in its treatment. In the US, Medicare incurred over 81.8 billion to cover the expenses of the CKD beneficiary in 2018 (United States Renal Data System, 2020) [29], and literature reveals that the medical cost of the patients with CKD is significantly greater compared to the non-CKD patients in the Medicare population (Honeycutt et al., 2013) [4]. Treatment procedures such as dialysis, kidney transplantation of ESRD patients cost about 90000 dollars per year per person. This discussion shows that the cost benefit analysis of a proactive detection system is astronomically relative to these devastating treatment costs in the late stage.

An official cost-benefit analysis of the implementation of the ensemble-based detection model shows a massive economic benefit. According to a hypothetical group of 100,000 at-risk people and using our Enhanced Stacking Ensemble model with 87.7% accuracy and 86.3% positive recall, the possible saving is large. In the given case of a conservative CKD prevalence of 10% in a risk population, the model would be able to correctly identify 8,630 true positive cases out of a true population of 10,000. Assuming that only 50% of these cases are prevented by early intervention to ESRD- a conservative estimate, which is backed by clinical literature (Komenda et al., 2016) [6] that most clinical conditions do not progress to this disease- would avoid 4315 patients undergoing dialysis or transplant. This intervention will result in gross savings of around 1.2 billion dollars given that the average cost of the ESRD lifetime treatment is $300,000 per patient (Kerr et al., 2012) [3]. The net economic benefit of about 1.28 billion even after considering an estimated cost of $15 million in screening of the whole cohort (using the standard rates of basic metabolic panel and urinalysis) proves to be an impressive investment payoff.

The financial reason behind the choice of the Stacking Ensemble as our end model

is supported by its efficiency in operation and performance balance. The model can be installed in the current clinical workflow with the minimal hardware cost and low operational cost as it has a small model size of 168.3 KB and an incredibly fast inference time of 0.0064 seconds. Such computational efficiency enables the solution to be made available even in resource-bound healthcare environments. Moreover, the high accuracy rate of the model of 84.8% and the balanced results with both types of diseases reduce the number of false negatives, which cause missed cases and place the organization in high treatment expenses in the future, and false positives, which incur unnecessary follow-up costs and anxiety in patients. This accuracy will leave healthcare resources to be distributed effectively to the patients who actually need to be assisted.

Conclusively, the economic analysis clearly shows that in terms of the outcomes of the experiment, early CKD diagnosis with the help of the ensemble learning model is not only clinically effective but also cost-beneficial. The healthcare system can prevent the high cost of renal replacement therapies since they can identify the at-risk patients at an early manageable stage. The suggested system is a high-pay-off investment that will transform the paradigm of care to be cost-efficient prevention instead of the expensive reactive treatment. This large net economic impact of a large population group is an additional strong case of why these predictive models should be adopted as part of routine clinical practice so as to not only enhance patient outcomes but also save the healthcare systems massive amounts of money.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

The general design process pipeline will focus on achieving the goals and will be accurate, efficient, and appropriate to the implementation of the IoT. It starts with the **data collection**, which is the collection of patient data and medical characteristics in credible sources. During the **data preprocessing** step, missing data are processed with mean and mode imputation of numeric data and categorical data respectively. Categorical data are coded and numeric values are scaled to ensure that there is a uniformity in all samples. Then, it is followed by the stage of **dimensionality reduction** that decreases the quantity of features by approximately 50 percent, making the model less intricate and revealing key risk factors of CKD. A train-test split is then used to split the dataset into train and test data sets to evaluate it impartially. During the phase of **model development**, various machine learning and deep learning models are trained and tested. Ensemble learning and hyperparameter learning enhance model accuracy and reliability. Lastly, the phase of **optimization and deployment** shrinks the model size and runtime without affecting the accuracy. The model and processing pipeline generated are both lightweight and efficient, which are appropriate in real-time CKD detection on IoT devices. It contains a detailed description of the systematic approach used to solve the problem of the early detection of chronic kidney disease (CKD). The design process is executed in a structured pipeline that moves through data preparation, model development and assessment, specifically to the extent that the project aims are achieved, yet within the identified constraints. The study focuses on the urgent healthcare issue of early detection of chronic kidney disease that is critical to preventing disease progression and obtaining better patient outcomes. In the management of CKD, early symptoms are probably insidious and therefore, identifying CKD on time with clinical parameters is a big challenge. The project will focus on creating the predictive modeling framework that will be able to provide appropriate information about the presence of CKD based on the clinical and laboratory data, thus assisting health professionals in early intervention and planning the treatment.

The main goals of the research is to design a stable predictive model that can cope with large-scale unbalanced clinical data, to optimize the performance of disparate algorithmic strategies, and to make the design practically deployable in the healthcare environment. Particular aims are to apply efficient class imbalance man-

agement, to develop extensive feature selection to discover clinically important predictors, and to rigorously compare traditional machine learning, ensemble models, and neural network structures. The most notable constraints affecting the design are the natural class imbalance in medical data where non-disease cases generally outnumber disease cases, computational resource constraints to train complex models, and the need to interpret model results in a clinical decision-making setting. Other limitations include data quality issues like missing values and measurement discrepancies, and the requirement of models to balance predictive capability with computational effectiveness for possible real time clinical utility. The pipeline approach is sequential and divided into four phases that are interrelated:

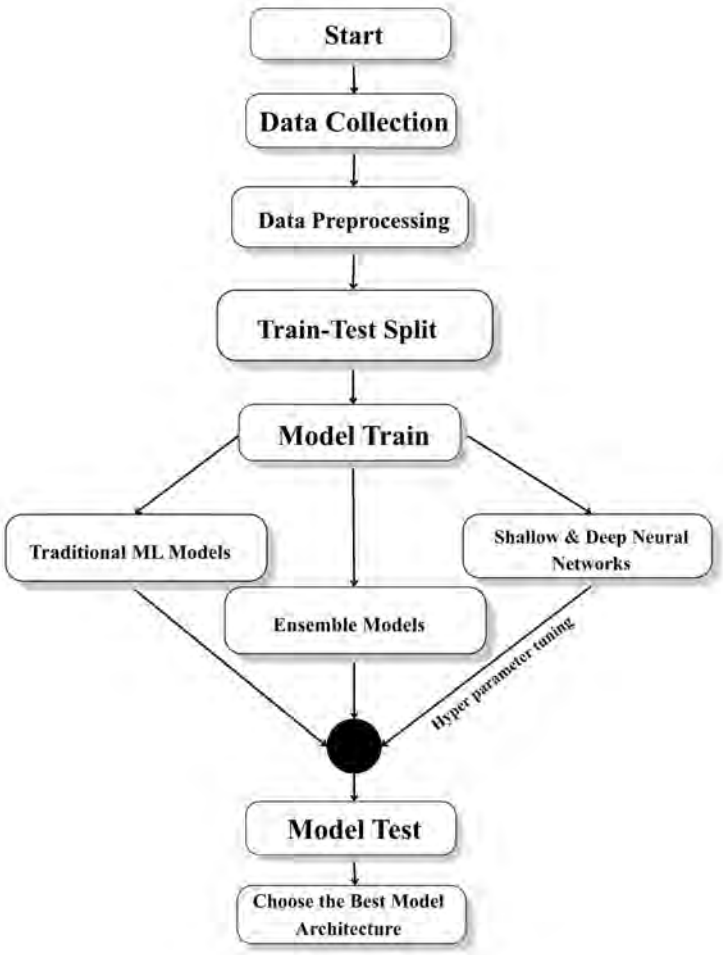

Figure 4.1: Overall Design Process Pipeline

Data Preparation: The first step is to combine several datasets to form a general data base and then to do a systematic treatment of missing values by suitable imputation methods. Exploratory data analysis is performed to learn data distributions, find patterns and potential outliers that can influence model performance.

Data Transformation: This step provides data formatting needs of machine learn-

ing algorithms, such as categorical variable encoding into numerical values and continuous values scaling, to guarantee similar ranges of values. To avoid the bias in the model of dominant classes specialized techniques like Synthetic Minority Oversampling technique (SMOTE) is introduced.

Feature Selection: There are several feature selection algorithms used to arrive at the most predictive clinical parameters and dimensionality reduction. This stage uses comparative analysis of the filter, wrapper, and embedded methods to find the best subset of features that preserves predictive power and increases model performance and interpretability. The use of the techniques like the Random Forest importance, Recursive Feature Elimination, and mutual information scoring, can be justified by the general practices examined in the literature of feature selection (Chandrashekar & Sahin, 2014) [5].Techniques applied during the feature selection process include Random Forest importance, Recursive feature elimination and mutual information scoring to determine the best subsets of features.

Model Development and Evaluation: The last step is to apply three different models of operation; the first uses traditional machine learning algorithms as the baseline models, the second uses an ensemble of learners to obtain better robustness, and the third uses neural network models, with different depths, to recognize more complex patterns. Both methods get hyperparameter tuned in a systematic manner and also their performance is checked with several performance measures so that they are fully compared. Its implementation uses Python as the main programming language, building upon specialized libraries such as scikit-learn to support traditional machine learning algorithms and ensemble methods, and neural network development with TensorFlow and Keras as well as class distribution problems with imbalanced-learn. Algorithms used include but are not limited to Random Forests, XGBoost, and Support Vector Machines as classical methods; Stacking, Voting, and Bagging as ensemble methods; and shallow and deep neural networks as complete neural network analysis.

4.2 Preliminary Design

We define the preliminary stage of the design, which involves the process of developing and testing several alternative modeling solutions systematically to determine the best solution to the issue of chronic kidney disease (CKD) prediction. We also test a variety of feature selection methods and model structure and test their performance using rigorous simulations and we eventually make final decisions correlating our story with the tabulated model summaries (Tables 4.1, 4.2 and 4.3).
Feature Selection: We implement five techniques-(1) Random Forest–based selection, (2) Mutual Information ranking, (3) Forward Feature Selection, (4) Backward Feature Elimination, and (5) Principal Component Analysis with Recursive Feature Elimination (PCA + RFE) and compare the techniques according to their capability of selecting clinically relevant features with the desired level of dimensionality reduction. To ensure robustness and interpretability, we measure stability across folds, correlation with clinical markers and redundancy.
Traditional machine learning models (Table 4.1): Following the work of Iftekhar

et al. [30], several traditional models such as Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, KNN, XGBoost, Decision Tree, and AdaBoost are implemented.

Model Name	**Algorithm Type**
Random Forest	Ensemble of 100 decision trees
XGBoost	Gradient boosting with 50 trees
SVM	Support Vector Machine with RBF kernel
Naive Bayes	Gaussian Naive Bayes
Logistic Regression	L2 regularized logistic regression
KNN	k-Nearest Neighbors (k=5)
Decision Tree	Single decision tree
AdaBoost	Adaptive Boosting with 50 learners

Table 4.1: Traditional Machine Learning Models

Ensemble methods (Table 4.2): Voting (hard/soft), Bagging with Decision Trees, and Stacking variants with different meta-learners are implemented.To further improve model generalization and accuracy, various ensemble learning techniques are explored. The enhanced stacking model combines Logistic Regression, Random Forest, Gradient Boosting, Support Vector Classifier, and MLP, achieving the best performance among ensemble variants.

Model Name	**Ensemble Combination**
Stacking	LR + MLP + GB $\rightarrow$ LR meta-learner
Voting (Hard)	LR + RF + GB + SVC (majority voting)
Voting (Soft)	LR + RF + GB + SVC (weighted average)
Enhanced Stacking	LR + RF + GB + SVC + MLP $\rightarrow$ LR meta
Bagging (DT)	50 Decision Trees with bagging
XGBoost-style Stacking	RF + GB + LR $\rightarrow$ GB meta-learner

Table 4.2: Ensemble Models and Their Combinations
Abbreviations: LR=Logistic Regression, MLP=Multi-Layer Perceptron, GB=Gradient Boosting, RF=Random Forest, SVC=Support Vector Classifier

Neural networks (Table 4.3): Here we test shallow (1-3 hidden layers) and deep sequential (up to 10 layers) neural networks, use batch normalization and dropout to prevent the occurrence of overfitting when classifying CKD stages. We repeat from shallow to deep sequential networks (6-10) and bottleneck design which strengthens the feature learning.For evaluation 5-fold cross-validation, stability testing of feature selection between data splits and hyperparameters, learning-curve monitoring, convergence monitoring and a strictly held-out test set are followed. We also deal with imbalance of classes with SMOTE, class weighting and threshold calibration and profile computational efficiency with runtime and memory usage per paths across hardware configurations.

Model Name	Layer Architecture
Shallow Neural Networks	
Shallow_NN_1Layer	Input → Dense(64) → Dropout(0.3) → Output(2)
Shallow_NN_2Layer_BN	Input → Dense(128) → BN → Dropout(0.4) → Dense(64) → Dropout(0.3) → Output(2)
Shallow_NN_Wide	Input → Dense(256) → BN → Dropout(0.5) → Dense(128) → Dropout(0.4) → Output(2)
Shallow_NN_Minimal	Input → Dense(32) → Dropout(0.2) → Dense(16) → Dropout(0.2) → Output(2)
Deep Neural Networks	
Deep_NN_6Layer	Input → Dense(512) → BN → Dropout(0.5) → Dense(256) → BN → Dropout(0.4) → Dense(128) → BN → Dropout(0.3) → Dense(64) → Dropout(0.2) → Dense(32) → Output(2)
Deep_NN_8Layer	Input → Dense(256) → BN → Dropout(0.4) → Dense(256) → BN → Dropout(0.4) → Dense(128) → BN → Dropout(0.3) → Dense(128) → BN → Dropout(0.3) → Dense(64) → BN → Dropout(0.2) → Dense(32) → BN → Dropout(0.2) → Dense(16) → Output(2)
Deep_NN_Bottleneck	Input → Dense(512) → BN → Dropout(0.5) → Dense(256) → BN → Dropout(0.4) → Dense(128) → BN → Dropout(0.3) → Dense(64) → BN → Dropout(0.3) → Dense(128) → BN → Dropout(0.3) → Dense(256) → BN → Dropout(0.4) → Output(2)
Deep_NN_10Layer	Input → Dense(1024) → BN → Dropout(0.6) → Dense(512) → BN → Dropout(0.5) → Dense(256) → BN → Dropout(0.4) → Dense(128) → BN → Dropout(0.3) → Dense(64) → Dropout(0.2) → Dense(32) → Dropout(0.2) → Dense(16) → Dropout(0.1) → Dense(8) → Output(2)

Table 4.3: Neural Network Architectures (Shallow and Deep)
Abbreviations: BN=Batch Normalization, Dropout(p)=Dropout with probability p

Initial classical models provide competitive benchmarks but exhibit weaknesses in the modeling of the non-linear and high-order interactions in clinical data,so we move on to ensemble techniques and even more profound networks. Deep networks enhanced the ability of heterogeneous learners, while ensembles made use of the complementary effectiveness of disparate learners. In this workflow, a hybrid design is performed by stacking-based ensembles with an optimized deep neural network: the ensemble part retrieves the complex feature interactions and enhances the robustness, the neural one delivers hierarchical feature extraction (through multi-stage processing) and a random forest based feature selection is selected due to the balanced predictive accuracy, clinical interpretability, and a stable dimensionality reduction. The chosen features remain similar across the methods and are aligned to the CKD pathophysiology. Generally, the hybrid system offers a solution to the issue

of class imbalance and presents clinically meaningful metrics, providing a coherent, operationally efficient and clinically based specification of CKD prediction.

4.3 Data Collection

The data collection process in this research uses two datasets from two renowned sources. The initial one is the established Chronic Kidney Disease data set offered by the UCI Machine Learning Repository and it consists of 400 patient cases with a detailed set of clinical features. This is combined with a bigger set of data of about 20 thousand instances, which are available in the open domain and have wide coverage of patient records and kidney health indicators. Assimilation of these two data sets involves matching of similar variables and correction of discrepancies in structure and measurement units to have a unified and dependable data format that will be analysed later [10].

4.3.1 Data Cleaning

The process of data cleaning starts with the merging of two different sets of data, which have the records of patients with chronic kidney disease. The initial dataset has 400 samples and the second one has 20538 samples and has 25 clinical and demographic features. An extensive form of exploratory data analysis evaluates the quality of data, the existence of patterns, and knowledge on later cleaning approaches. Preliminary inspection shows that there is considerable imbalance between the classes in the target variable, where the majority of the classes (non-chronic kidney disease cases) have a large number of respondents compared to the minority ones (chronic kidney disease cases) in both datasets. Missing value analysis shows that there are some features that have missing data values (Figure 4.2, Table 4.4), and the most prevalent such features are the red blood cells (0.73%), the red blood cell count (0.63%), and white blood cell count (0.51%) with the highest percentage of missing data. The analysis of missingness patterns proves that there is no excessive missing data in one of the features to justify its removal.

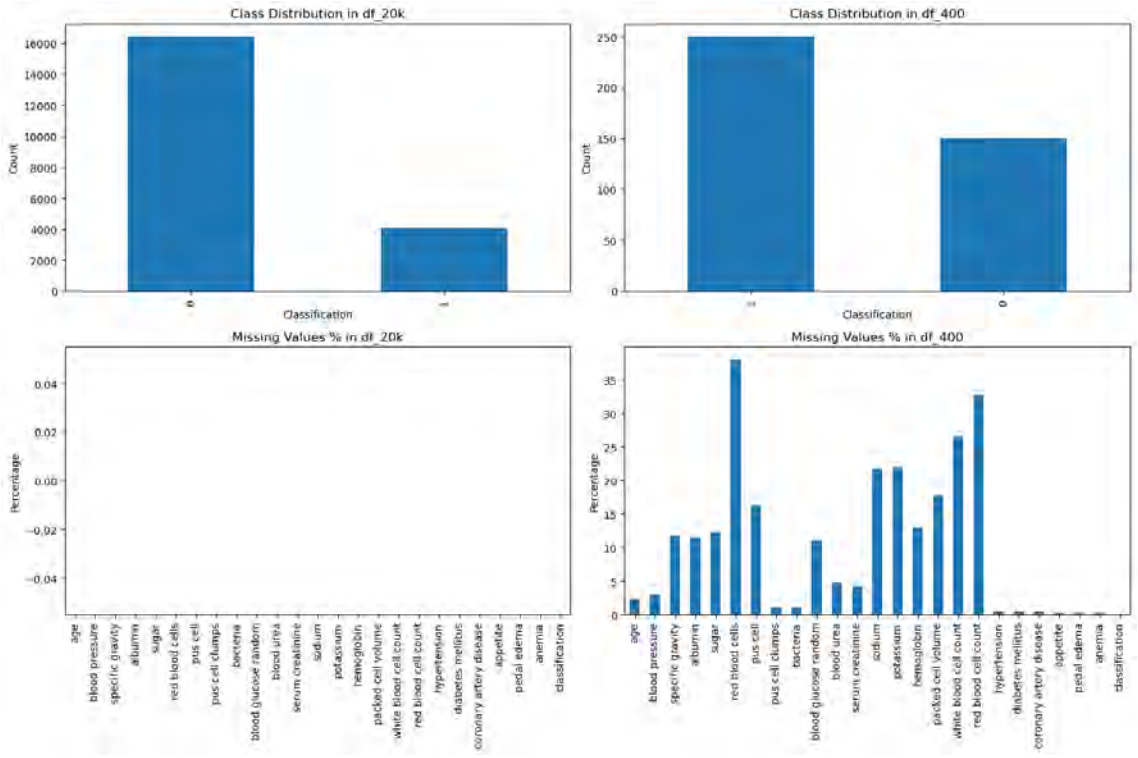

Figure 4.2: Distribution of missing values across dataset features

Feature	**Missing Count**
red blood cells	152
red blood cell count	131
white blood cell count	106
potassium	88
sodium	87
packed cell volume	71
pus cell	65
hemoglobin	52
sugar	49
specific gravity	47
albumin	46
blood glucose random	44
blood urea	19
serum creatinine	17
blood pressure	12
age	9
bacteria	4
pus cell clumps	4
hypertension	2
diabetes mellitus	2
coronary artery disease	2
appetite	1
pedal edema	1
anemia	1

Table 4.4: Missing value counts for dataset features

According to Iftekhar et al.[30], the data preprocessing involves cleaning the dataset by renaming columns for clarity, converting data into appropriate types, and handling missing values through imputation methods such as mean/mode sampling, depending on the proportion of missing data. A systematic imputation plan deals with missing data without affecting the integrity of data. In binary clinical indicators, e.g., hypertension, diabetes mellitus, coronary artery disease, and other cellular presence markers, the mode (Table 4.5) is used to fill in missing values, since it still has clinical significance. Constant clinical measurements such as age, blood pressure, specific gravity, albumin, sugar levels and other blood parameters are modeled by mean values (Table 4.6) to maintain the overall distribution features of these physiological measures.

Feature	**Missing Count (Before)**	**Missing Count (After)**
red blood cells	152	0
pus cell	65	0
pus cell clumps	4	0
bacteria	4	0
hypertension	2	0
diabetes mellitus	2	0
coronary artery disease	2	0
appetite	1	0
pedal edema	1	0
anemia	1	0
classification	0	0

Table 4.5: Missing values before and after mode imputation for categorical/binary columns

Feature	**Missing Count (Before)**	**Missing Count (After)**
red blood cell count	131	0
white blood cell count	106	0
potassium	88	0
sodium	87	0
packed cell volume	71	0
hemoglobin	52	0
sugar	49	0
specific gravity	47	0
albumin	46	0
blood glucose random	44	0
blood urea	19	0
serum creatinine	17	0
blood pressure	12	0
age	9	0

Table 4.6: Missing values before and after mean imputation for numerical columns

The data that have been eventually cleaned consists of 20,938 full patient records that contain 25 features, none of which lacks a value and are of the proper format

to proceed with the next stage of transformation and modeling. Such stringent cleaning procedures provide a strong baseline of consistent model construction and provide data quality criteria that is required to support clinically significant predictive analytics.

4.3.2 Data Transformation

The data transformation step adapts a number of preprocessing methods to be used in preparing the clinical data as an input to machine learning algorithms. Categorical clinical variables are being systematically encoded in order to transform qualitative medical observations into numerical forms that can be used in computational analysis(Table 4.7). Binary clinical data such as the red blood cells, the presence of pus cells, bacterial infection, hypertension, diabetes mellitus, coronary artery disease, appetite status, pedal edema and anemia are converted into binary numeric form (0/1) to ensure uniform interpretation of the model.

In order to overcome the high imbalance in the class of interest (Figure 4.3), the Synthetic Minority Over-sampling Technique (SMOTE) is used to create the synthetic samples of the minority group (the cases of chronic kidney diseases). The technique produces artificial yet realistic clinical profiles that do not merely recapitulate the available records but produce balanced class distributions to improve the ability of the model to pick patterns in both classes and eliminate prediction biasness in the majority class.

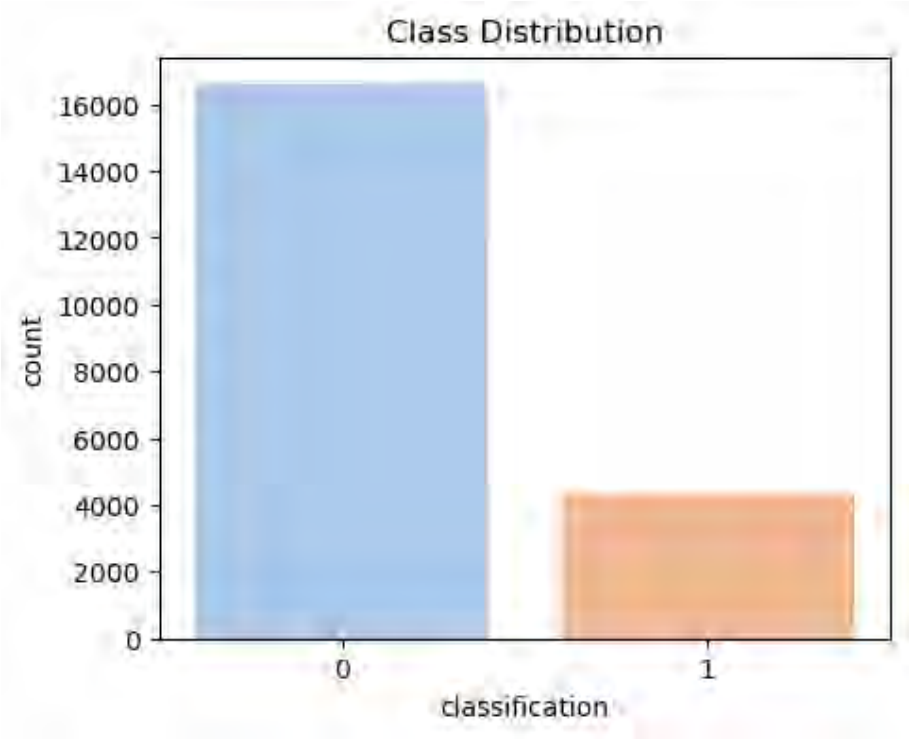

Figure 4.3: Class distribution showing dataset imbalance

There is a very crucial transformation step called feature scaling to normalize the measurement scales of different clinical parameters. The `StandardScaler` method is used to scale all the numerical features to a uniform range with a mean of 0 and a

variance of 1. This standardization procedure guarantees that those characteristics with naturally larger numerical domains, like blood glucose and white blood cell counts, are disproportionately taken into account in model education in contrast to characteristics with smaller domains such as specific gravity and albumin level. The scaling conversion allows consistency of convergence in optimizing the model and allows equitable contribution of features to all clinical measurements in further machine learning algorithms.

Feature	**Unique Values (Before Encoding)**
red_blood_cells	normal, abnormal
pus_cell	normal, abnormal
pus_cell_clumps	notpresent, present
bacteria	notpresent, present
hypertension	yes, no
diabetes_mellitus	yes, no
coronary_artery_disease	no, yes
appetite	good, poor
peda_edema	no, yes
aanemia	no, yes

Table 4.7: Unique categorical feature values before label encoding

4.3.3 Data Integration

The information integration procedure involves the integration of two complementary datasets comprising the patient information on chronic kidney disease to form a complete basis of analytical foundation. The first dataset named *Kidney Disease Dataset (A Comprehensive Clinical Dataset for Chronic Kidney Disease Analysis and Prediction)* consist of 400 patient records with 24 clinical attributes that are real-world clinical observations and the second dataset consist of 20,538 synthetically generated patient records with 42 clinical attributes that can be used to represent various patient profiles. These two datasets are similar in that they record common clinical parameters although they have some differences in their sources and feature name conventions. A systematic mapping review establishes 24 similar clinical characteristics among the two data sets such as basic kidney functioning parameters including age, blood pressure, specific gravity, albumin levels, blood glucose, urea, creatinine, electrolytes, hematological parameters, and other clinical symptomatic features. These common characteristics give the two datasets the same measurement units and clinical interpretations, which are compatible to integrate. Both datasets have the same clinical definitions in the target variable that is the classification of chronic kidney disease (Table 4.8).

The integration plan will use vertical concatenation to consolidate patient records but maintain all the shared clinical attributes. Non-overlapping characteristics that the two datasets have in common are deliberately avoided in the integrated dataset in order to ensure consistency and avoid information asymmetry. Such a selective methodology provides the assurance that all integrated records will consist of an identical complete range of clinically relevant features without additional bias brought on by measurements unique to data sets.

The standardization of the column names is achieved to solve the nomenclature variations among the source datasets, and to provide a single schema to all integrated features. Consistency validation at the data type level ensures that measures made by integrated numerics that have identical ranges and that categorical issues have similar encoding systems. The resulting integrated data are 20,938 patient records and 24 uniformly defined clinical features, which form a solid basis of further analytical methods and do not compromise the clinical integrity and measurement consistency of credible medical research.

Dataset	**No. of Features**
UCI CKD Dataset (400 instances)	24
Kidney Disease Dataset (20.5k instances)	42
Common Features Used for Integration	**24**

Table 4.8: Summary of datasets and feature alignment after integration

4.3.4 Data Reduction

The data reduction step employs various feature selection methods to be able to determine the most clinically relevant predictors as well as removing redundant or uninformative variables in the chronic kidney disease dataset. This method solves the high-dimensionality problem of clinical data by determining computational efficiency to the features that mean anything to predictive performance. The first goal is to achieve a 50 percent dimensionality of the features, where the goal will be to shrink the original 24 clinical parameters to about 12 features without deteriorating the diagnostic accuracy.

An in-depth comparative study is performed on five various feature selection methodologies in order to identify the best feature subset(Table 4.9). The feature selection method using the random forest uses the models of the ensemble trees to rank the features according to their significance in the prediction of kidney disease outcomes. This approach revealed a dozen important clinical parameters such as blood pressure, specific gravity, albumin, sugar levels, blood glucose random, blood urea, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, and red blood cell count (Table 4.10). The chosen features have high predictive ability with an ROC AUC of 0.872, with equal performance in separate disease classes in precision and recall. It is interesting to note that the twelve features chosen by the Random Forest selector do have a great degree of consistency with the highest ranked features measured by the ranking algorithm and information gain analysis. This overlap among various modalities of selection points to the fact that these are the clinical parameters that constitute the main risk factors in the detection of chronic kidney disease. There is a consistent choice of hematological markers, indicators of kidney function, and metabolic parameters in all statistical methods, which supports their clinical importance and diagnostic value.

Selector	Features	Accuracy	Recall	Precision	F1 Score
All Features	24	0.87	0.87	0.89	0.86
Backward Selection	22	0.87	0.87	0.88	0.87
RF Selector	**12**	**0.85**	**0.85**	**0.87**	**0.84**
Ranking Algorithm (IG)	12	0.81	0.81	0.82	0.72
Forward Selection	10	0.82	0.82	0.83	0.81
PCA + RFE	12	0.63	0.63	0.63	0.63

Table 4.9: Feature Selection Methods Performance Comparison

Forward feature selection adopts a sequential forward greedy strategy that adds features in a sequential manner depending on their contribution towards model performance. This approach narrows down to a small set of features of red blood cell count and serum creatinine with an ROC AUC of 0.859 and the dimensionality is reduced by 91.7%. Even though this technique is very effective, it can miss critical interactions between features that can be observed in the full clinical picture(Table 4.9) [19].

Backward feature elimination takes a backward approach and starts with all the available features and adds up all possible features by eliminating the least important variables and checking the performance degradation. The approach has the greatest ROC AUC of 0.920 compared to other selection strategies, and it maintains twenty-two features (Table 4.9). Only two features (diabetes mellitus and potassium) are eliminated by the algorithm and do not show significant effects on the predictive accuracy but leave clinically significant variables [19].

Following Islam et al. [39], we apply principal component analysis coupled with the Recursive Feature Elimination considers the dimensionality reduction by feature transformation. This method converts original features into twelve major components representing 95% of the variance in data, but does worse (ROC AUC: 0.694) (Table 4.9) than direct feature selection techniques, implying that clinical interpretability can be lost by feature transformation

Following the steps of Tazin et al. [10] we implement ranking algorithm with information gains to reduce the dimentionality of the dataset. Ranking of feature importance based on mutual information criteria is used to select features that have the strongest statistical association with the target variable. According to the information theory principles, the approach emphasizes that the count of the red blood cells, hemoglobin, blood pressure, packed cell volume, and age are most revealing but the model performance with such choices is average (ROC AUC: 0.539)(Table 4.9).

The comparative analysis shows that the Random Forest selector offers the best compromise between performance of the model (ROC AUC: 0.872), feature interpretability and clinical relevance. The 12 features chosen are an ideal fit to the desired 50% dimensionality reduction value with a strong predictive power. The great agreement of the random forest selection with information gain ranking supports the clinical validity of these indicators as the main predictors of kidney diseases. As

a result, all the further model development and training stages are based on this optimized set of features, being computationally efficient and having the key clinical markers in case of the effective chronic kidney disease detection.

Selector	**Selected Features**
All Features	All 24 original features
Backward Selection	age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, coronary artery disease, appetite, pedal edema, anemia
RF Selector	blood pressure, specific gravity, albumin, sugar, blood glucose random, blood urea, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count
Ranking Algorithm (IG)	albumin, sugar, specific gravity, sodium, blood pressure, red blood cell count, white blood cell count, hemoglobin, packed cell volume, potassium, blood glucose random, blood urea
Forward Selection	red blood cell count, serum creatinine, specific gravity, pedal edema, diabetes mellitus, coronary artery disease, red blood cells, pus cell clumps, albumin, packed cell volume
PCA + RFE	PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC11, PC12

Table 4.10: Feature Selection Methods - Selected Features

4.3.5 Summary of Preprocessed Data

The extensive preprocessing pipeline converts the unprocessed clinical data into an optimal analytical base that can be used in the development of machine learning models. The resulting processed data is a patient record of 20,938 patients and 12 well-chosen clinical variables that are the most diagnostically significant predictors of chronic kidney disease. The preprocessing steps starts with information incorporation by combining 400 real world clinical records with 20,538 synthetic patient records to generate a powerful analytical base. The original integrated database has 25 clinical features that constitute demographic data, vital signs, urinary parameters, blood chemistry, and hematological indices, as well as clinical symptom markers. By extensive data cleaning processes, all missing values are addressed systematically with mode-based imputation of binary clinical indicators and mean-based imputation of continuous physiological measures, so that there is full data coverage on all patient records.

Information transformation algorithms encode categorical clinical data using standard scaler and deal with the class imbalance by means of synthetic minority oversampling (SMOTE). Scaling of features balances all the clinical measurements to

similar scales, whereas the use of SMOTE produces equal distributions between classes that avoid bias in the models where the majority of the samples belong to. These changes provide the technical conditions of successful machine learning algorithms functionality and maintain clinical interpretability.

The data reduction step is a strategic step to reduce the number of features to 12 clinically significant parameters by comparing the various selection techniques. The last set of features comprises of blood pressure, specific gravity, albumin, sugar, blood glucose random, blood urea, sodium, potassium, hemoglobin, packed cell volume, white blood cell count as well as red blood cell count. These features have been chosen because they have high agreement among the various statistical selection methods and support known clinical knowledge regarding the pathophysiology in kidney disease.

The overall data quality can be greatly enhanced by the preprocessing pipeline, and all the gaps are filled in, the class imbalance is properly addressed, the redundant features are removed, and the scales of measurement are standardized. The resulting data set is clinical interpretable and maximally efficient in terms of computation and offers a solid basis on which to create valid and reliable predictive models. This cleaned-up data is a highly edited analytical tool that meets the accrual threshold of statistical soundness and clinical utility to set the requisite to detect and classify chronic kidney disease reliably in the later stages of the modeling process.

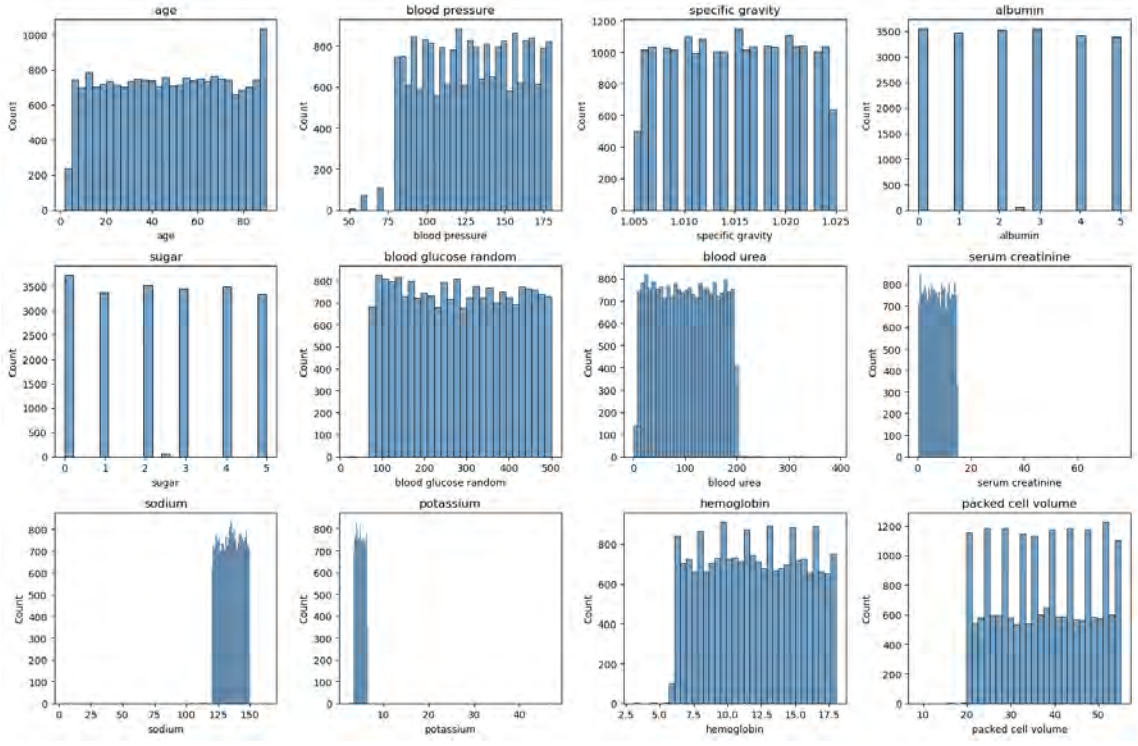

Figure 4.4: Exploratory data analysis of the final merged dataset

4.4 Implementation of Selected Design

The implementation step defines a holistic model of the development and evaluation of various learning methods to work on the task of classifying chronic kidney disease. It entails coding, training of models and combining the processed information with the selected machine learning models. At this step, the workflow that has been created in the data preparation and feature selection steps is mainly implemented to test the performance of the model (Figure 4.5). The refined feature set is used to train all chosen algorithms with the maximum consideration of accuracy and computational efficiency. This systematic implementation plan involves traditional machine learning models, ensemble techniques, and neural network architectures, so that there is a comprehensive comparative analysis and identification of the best model.

The model implementation plan is structured into a three-level approach that forms an experience on various algorithmic paradigms. The decision layer is made up of traditional machine learning models such as Random Forest, Support Vector Machines, Naive Bayes, Logistic Regression, K-Nearest Neighbors, Decision Trees, Gradient Boosting and AdaBoost classifiers. These models also deliver baseline-performance criteria and determine the basic predictive skills with the help of the preprocessed clinical characteristics. The second level of implementation is ensemble methods, which involve advanced combination schemes based on Stacking, Hard and Soft Voting classifiers, Enhanced Stacking, Bagging with Decision Trees and XGBoost-like Stacking designs. These combinations make use of the complementary strengths of many base estimators in order to increase predictive robustness and generalization capacity. The third layer is neural network implementation with architectural variations that are well thought out. Various depth structures such as single-layer, two-layer with batch normalization, wide shallow and minimalist networks are searched over to identify optimal network complexity by shallow neural networks. The six-layers gradual decrease, eight-layers residual-style, bottleneck, and ten-layer varying size research models are studied as more complex hierarchical representations by deep neural networks. State-of-the-art bottleneck neural networks are tuned thoroughly in more advanced, deeper residual, wider, larger, and extreme architectures up to 46 layers and 15 million parameters, and systematically analyze the trade-offs between model complexity and model performance improvements.

The model assessment system has a strict evaluation of performance through various complementary outcomes. Accuracy, Precision, recall, F1-score, and ROC AUC gives holistic measures of classification performance of all the implemented models. Stratified training-validation-test splits guarantee that the data are represented, whereas cross-validation methodologies confirm the stability of the model and its ability to generalize. The confusion matrix analysis in detail provides information on both the specific pattern of performance of the classes and the trends of misclassification.

The procedures of testing and validation create strong model verification with a variety of complementary methods. The use of classification matrices offers the

granular performance breakdowns, whereas the AUC-ROC curves give the threshold-independent performance assessment at all operating points. Analysis Model size Analysis The computational efficiency and deployment feasibility of a model are evaluated by analyzing model size and practical implementation limits by runtime profiling. The comprehensive hyperparameter optimization is the optimization of one model category with respect to systematic search strategy, and the model combinations are extensively tested to detect the synergistic effects. Model validation Final validation is performed using entirely unobserved test data to confirm performance in the real world and stability in different conditions, and further robustness tests assess consistency in performance in different subsets of data and in different clinical conditions.

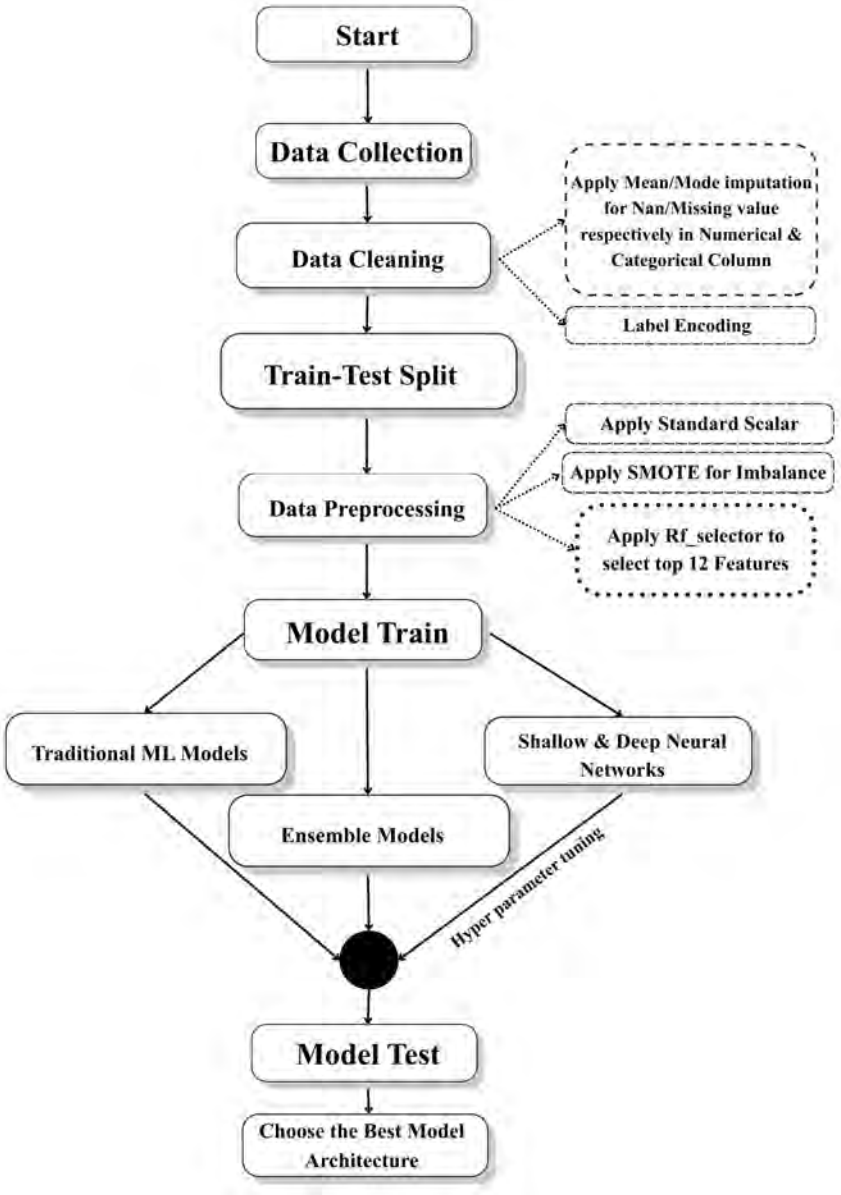

Figure 4.5: Implementation workflow of the selected design

Chapter 5

Result Analysis

5.1 Performance Evaluation

The performance evaluation system is a multi-dimensional assessment strategy that is comprehensive in implementing the measure of model effectiveness in both domains of classification performance and computational efficiency. The evaluation criteria use a variety of metrics in an attempt to give a comprehensive picture of the model capabilities and restrictions on the clinical prediction environment. The framework evaluates models across several key dimensions including accuracy, precision, recall, F1-score, computational efficiency, reliability, and validation methodology.

- **Accuracy:** The most basic performance measure is the classification accuracy, which is the total percent of correct predictions on all classes. Also Rainio et al. [46] labeled accuracy as one of the main evaluation metrics for machine learning models. This measure gives a broad idea of the model performance but needs to be supported with more detailed measures to consider the issues of the class imbalance. Huang et al. [38] employed accuracy in a peer reviewed study as part of their model evaluation metrics. Accuracy computation is based on the number of correct predictions and the total number of predictions and provides a simple explanation of the overall model accuracy. Measures the overall correctness of the model's predictions using the formula:

  $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  where TP represents True Positives, TN True Negatives, FP False Positives, and FN False Negatives.

- **Precision:** Precision is a measure that determines how the model avoids making false positive predictions, or the ratio of the number of true positive predictions to the number of positive predictions.

  $$\text{Precision} = \frac{TP}{TP + FP}$$

  Indicatively, Rainio et al. [46] introduce precision as one of the most frequent measurements of binary and multi-classification with the emphasis on its significance in the evaluation of model performance. Equally, Miller et al. [45]

thoroughly review machine learning metrics, explicitly referring to the precision as a metric to assess classification problems and comment on its benefits and drawbacks.

- **Recall (Sensitivity):** The sensitivity of the model in determining real positive cases is tested using recall that determines the percentage of real positives that the model predicts.According to the definition of Manning et al. [1], this measure is the proportion of the relevant cases that are retrieved, which is a decisive indicator of the completeness of a model.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** F1-score balances the conflicting goals using their harmonic mean which gives a balanced measure that takes into account both precision and recall together.According to Logunova [40], in comparison to the arithmetic mean, the harmonic mean is more likely to be close to the smaller number so that the F1 score is high only in the cases when the precision and the recall are also high.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **ROC-AUC:** The Receiver Operating Characteristic curve (ROC) and Area Under the Curve (AUC) analysis the threshold-independent model assessment is provided by looking at the compromise between the rates of true positives and false positives at all possible points of classification. The ROC curve diagrams are the plots of sensitivity vs. specificity, whereas the AUC gives one scalar quantity as to the general diagnostic ability, and larger values signify greater distinction between the positive and negative categories.

- **Efficiency:** It is evaluated based on model size and run time. Measurement of training time is used to capture the computational time it takes to converge a model, and measurement of prediction time is used to measure the speed of inference on single samples.

$$\text{Efficiency Score} = \frac{\text{Accuracy}}{\text{Model\_Size\_KB}} \times 1000$$

- **Model Size:** Model size analysis studies memory footprint and storage needs, which have insights into the deployment capability of resource-constrained clinical settings. The assessment of computational complexity involves the number of parameters and the architectural complexity based on the performance improvements.

$$\text{Model Size (MB)} = \frac{N_{\text{params}} \times \text{Bytes per Parameter}}{1024^2}$$

where N_{params} denotes the total number of learnable parameters, and typically each parameter occupies 4 bytes (for 32-bit floats).

- **Run Time:** Evaluating the total time required for training and inference, measures the computational efficiency.

$$\text{Run Time (s)} = T_{\text{train}} + T_{\text{inference}}$$

where T_{train} is the model training duration and $T_{\text{inference}}$ represents the average prediction time across test samples.

- **Parameter Count (DNN Models):** The total number of parameters in a deep neural network (DNN) layer is calculated as:

$$\text{Total Parameters} = \sum_{l=1}^{L} (n_{l-1} \times n_l + n_l)$$

where n_{l-1} and n_l denote the number of input and output neurons in the l^{th} layer, respectively, and the additional n_l accounts for bias terms. According to Goodfellow et al. [9], one of the core principle of neural network consists of the calculation of this weights and biases in dense layers.

- **Reliability:** It is tested using the cross-validation and robustness. The consistency of model is checked over the variety of random states and dataset partitions to guarantee steady and consistent findings.

- **Testing Methodology:** A stratified k-fold cross-validation is used to divide the dataset, with the balance in classes being preserved in every fold. Also another train-test (usually 80-20) is carried out to assess the final unbiased evaluation. Transformations on data preprocessing are only done to training data to avoid data leakage.

5.2 Analysis of Design Solutions

Among the systems that were tested using traditional machine learning models, the Random Forest model is of high overall performance, with the correct percentage positioned at 79.6% (Table 5.1) and high majority class recognition. This group method builds many decision trees during training, and the mode of their output is the final result, which gives an 88 percent F1-score on the 0 classification due to its ability to deal with the interactions between the features. It however demonstrates great weaknesses in identification of minority classes with the largest F1-score of 9% class 1 marking difficulties in identifying cases of positive kidney disease. The strong feature importance analysis of this model is a strong feature of the model and aligns with the knowledge of clinical practice, but its weakness is seen in its ability to deal with a severe imbalance in the classes.

Similar accuracy of 79.8 percent is achieved with XGBoost with a high degree of precision of 94 percent on class 1 (Table 5.1), which is the case of gradient boosting as used to create trees successively to improve upon the errors made previously. Although it has a high-precision of positive cases, the model has a critically low recall of 3% of class 1 which makes it unsuitable in complete detection of a disease but is effective in reducing false positives. Such performance pattern indicates that the

model focuses more on assurance as compared to coverage and as such, it is good in confirmation tests but poor in screening application where it is important to trace all the possible cases.

Support Vector Machines provide fair balanced performance at 63.3% accuracy and comparatively stable class-based measures and run by locating the most suitable hyperplanes to classify in high-dimensional space. The model attains 76% F1-score on class 0 and 25% on class 1, which is reasonable at separating data, yet failing to address the imbalance in the dataset itself. Its advantage lies in its ability to deal with nonlinear relationships by kernel functions, but the cost of computations and sensitivity to parameter tuning pose some severe drawbacks to implementation.

Naive Bayes classifies with 61.1 percent accuracy making a balanced prediction of the classes, as a result of the implementation of Bayes theorem using high feature independence assumptions. With a simple model the model achieves 73% F1-score on class 0 and 28% on class 1 which is reasonable in a simple model but limited by his underlying assumptions of independence that does not necessarily hold when the clinical features are correlated. It has benefits in terms of computational efficiency, probabilistic basis, which helps in implementation, but limited model ability limits its ability to compete with more complex models.

At a very low level of 52.1% accuracy, Logistic Regression is highly unsatisfactory and it cannot linearly separate the features in the feature space with this clinical prediction. This linear model also attains 63 percent F1-score on class 0 and 31 percent on class 1, which, being better than guessing, is however, not good enough to find any application in clinical practice. Coefficient analysis of features allows the model to be interpreted which is useful in understanding relationships between the features, however its simplicity fails in depicting the behaviour of kidney disease prediction that is multifaceted.

K-Nearest Neighbors runs with 51.1% accuracy and is an instance-based active learner, i.e. insights on the final prediction are based on the nearest data. The model attains the F 1-score of 63 on class 0 and 29 on class 1 indicating a weak skill in extrapolating outside within the local trends of the data. Although conceptually clear and does not need explicit training, the method has a weakness in both predictive computational inefficiency as well as being sensitive to features scaling, making it impractical in this application.

Decision Trees have an overall accuracy of 65.6 per cent and is interpretable by rule-based classification, which reflects clinical decision procedures. The model attains 78% F1-score in class 0 and 264 percent in class 1 and gives good performances with recursive data partitioning, where decision lines are transparent. It has the advantage of straightforward visual interpretation of classification rules, but is susceptible to overfitting, and has limited complex pattern recognition abilities.

AdaBoost has a moderate level of performance (73.2 percent) and uses a boosting strategy, in which cases that have been previously misclassified by a model are highlighted in successive models. The ensemble model has a high F1-score of 84 in

class 0 but 16 in class 1 with a higher overall accuracy provided by adaptive learning yet with the issue of identifying minority classes. The strategy has been found to be effective in merging several weak learners into a robust classifier, although there are //that is only computational and noisy data.

In all classic paradigms, there are always similar problems with managing the imbalance of classes, and much more favorably than the minority classes, majorities are identified. The tree methods such as the Random Forest and the XGBoost delivers better overall performance but the linear models as well as the distance based methods possess limited efficiency in relation to this particular clinical prediction task. The analysis of feature importance in all the models determines that albumin, sugar and specific gravity are clinically significant predictors which confirms the medical basis of the modeling method and indicates what may be done to better handle multifaceted clinical patterns with more complex methodologies.

Model Name	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.7956	0.61	**0.05**	0.09
SVM	0.6328	0.22	**0.29**	0.25
Naive Bayes	0.6113	0.23	**0.37**	0.28
Logistic Regression	0.5210	0.22	**0.52**	0.31
KNN	0.5105	0.21	**0.49**	0.29
Decision Tree	0.6562	0.23	**0.28**	0.26
XGBoost	0.7985	0.94	**0.03**	0.06
AdaBoost	0.7321	0.23	**0.13**	0.16

Table 5.1: Traditional Machine Learning Models Performance Comparison

Among all the ensemble models assessed, the Enhanced Stacking Ensemble has the best predication performance with an accuracy of 87.7 and ROC AUC of 91.6 which is the best among all methodologies considered. It is a highly advanced architecture that uses 5 heterogeneous base models such as Logistic Regression, random forests, gradient boosters, support vectors, and multi-layer perceptrons with cross-validation of 5-folds to make use of their respective synergies. The ensemble demonstrates high balance between class-wise performance with 88.8% F1 -score of class 0 and 86.3% of the class 1, which explains a strong ability to cope with the imbalance of the clinical data. Nevertheless, this performance is accompanied by significant computation costs, as it needs 444.7 seconds of training time and a model size of 50.2 MB which is a manifestation of the complexity of interactions between many complex algorithms.

XGBoost-style Stacking Ensemble achieves the same post-test accuracy as the Enhanced Stacking, of 87.7%, with the same ROC AUC of 91.6, and with greatly improved training efficiency. It is based on Random Forest, Gradient Boosting, and Logistic Regression as base estimators and a Gradient Boosting meta-classifier that is explicitly configured to replicate the stacking abilities of XGBoost. The model scores almost the same F1-score of 88.9% on class 0 and 86.1% on class 1 with great balance between classes without wasting much of their training duration of 25.1 seconds, which is 18 times less than Enhanced Stacking. The model size is large (23.8 MB), but it indicates a more effective usage of the parameters, and thus the method

is especially useful in the situation when both the high performance and reasonable training efficiency are required.

The Soft Voting Classifier obtains high performance with an accuracy of 85.3 and a ROC AUC of 90.0, which is an average of the output of the Logistic Regression, Random Forest, Gradient Boosting and Support Vector Machines. This method shows a balanced performance of the classes where F1-score of class 0 is 86.7 and that of class 1 is 83.5 since it successfully uses the confidence estimate of each of the constituent models. The ensemble will need 89.1 seconds of training and generate a 50.1 MB model size, just like other more complicated voting strategies, and the probability-based aggregation will have smoother decision boundaries and have better calibration than hard voting options.

The Bagging Classifier using Decision Trees is a competitive classifier that gets an accuracy of 85.5 percent and an ROC AUC of 90.0 percent using an ensemble of 50 decision trees and bootstrap sampling to minimize variance and enhance generalization. This strategy results in 87.2% F1-score on the class 0 and 83.4 on the class 1, which proves that parallel model combination is effective to boost the estimators of the stable base. The fastest training time of 6.9 seconds, accompanied by a medium-sized model (16.7 MB) of the model, makes this strategy especially appealing to the applications, which have to deploy and update quickly, yet do not compromise the predictive accuracy.

The Stacking Ensemble achieves a good baseline performance of 84.8 percent accuracy and 87.2 percent ROC AUC by using a combination of Logistic Regression, Multi-layer Perceptron and Gradient Boosting and a Logistic Regression meta-learner. The resulting configuration has a F1-score of 86.6 and 82.4 in classes 0 and 1 respectively which showcases the underlying advantages of the stacking concept of models using fewer base estimators. It has an impressive prediction speed of 0.0064 second inference time and a single-squeezer model of 168.3 KB that offers truly superior operational efficiency in clinical real-time and produces respectable performance.

The Hard Voting Classifier has an accuracy of 81.2% but is unable to yield probability estimates to allow computing ROC AUC, a combination of the same base models using majority rule decision making. It achieves 84.2 percent F1-score with class 0 and 76.8 percent with class 1 which is reasonable in performance but cannot index fine-grained classification versus probability approaches. The model size of 50.1 MB and the used training time of 86.1 seconds demonstrate that a significant investment in resources was made without immediate performance benefits which makes soft voting superior in terms of the clinical application.

In all ensemble approaches, there are regular trends in terms of trade-offs in terms of performance, computational requirements, and complexity of implementation. The stacking approaches are usually superior to the voting approaches, especially with respect to their capability to learn the best models combinations instead of using homogenous aggregation. The Enhanced Stacking and XGBoost-style Stacking show that it is possible to reach breakthrough level performance using sophisticated meta-

learning and Bagging is a promising tradeoff between performance and efficiency to use in actual application. Combination of strategic models is observed to significantly improve on single base models, supporting the fundamental assumption that ensemble combination of algorithmically deficient models can deliver much more potent and accurate clinical decision support predictions.

Based on the comprehensive analysis, **Stacking** has been selected as the optimal ensemble model due to its exceptional balance of performance and efficiency. While Enhanced Stacking and XGBoost-style Stacking achieve marginally higher accuracy (0.877 vs 0.848), Stacking demonstrates superior efficiency with the fastest prediction time (0.006 seconds), smallest model size (168 KB), and highest efficiency score (5.038). This makes it ideal for practical deployment scenarios where computational resources and inference speed are critical constraints.

Model Name	**ROC AUC**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Stacking	0.872	0.848	0.966	0.718	0.824
Voting (Hard)	0.885	0.812	0.986	0.629	0.768
Voting (Soft)	0.900	0.853	0.942	0.749	0.835
Enhanced Stacking	**0.916**	**0.877**	**0.958**	**0.786**	**0.863**
Bagging (DT)	0.900	0.855	0.968	0.732	0.834
XGBoost-style Stacking	0.916	0.877	0.975	0.771	0.861

Table 5.2: Ensemble Models Performance Comparison

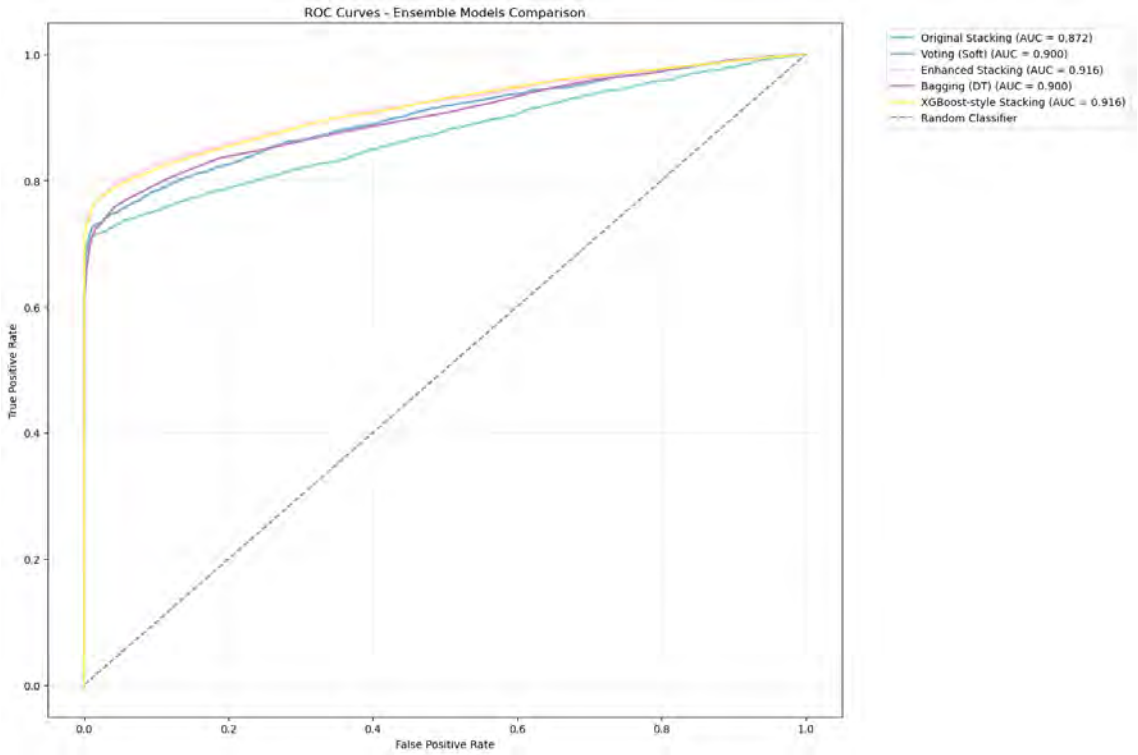

Figure 5.1: ROC AUC Comparison of Ensemble Models

Model Name	Architecture
Stacking	**Base Models: Logistic Regression + MLP + Gradient Boosting → Meta-Learner: Logistic Regression (5-fold CV)**
Enhanced Stacking	Base Models: Logistic Regression + Random Forest + Gradient Boosting + SVC + MLP → Meta-Learner: Logistic Regression (5-fold CV)
XGBoost-style Stacking	Base Models: Random Forest + Gradient Boosting + Logistic Regression → Meta-Learner: Gradient Boosting (5-fold CV)
Bagging (DT)	Base Model: Decision Tree with 50 estimators using bootstrap aggregation
Voting (Soft)	Models: Logistic Regression + Random Forest + Gradient Boosting + SVC → Weighted Average Probability Voting
Voting (Hard)	Models: Logistic Regression + Random Forest + Gradient Boosting + SVC → Majority Class Voting

Table 5.3: Ensemble Models Architecture Composition

Model Name	Training (sec)	Prediction (sec)	Model Size (KB)	Efficiency Score
Stacking	**76.650**	**0.006**	**168.321**	**5.038**
Enhanced Stacking	444.736	5.507	50156.936	0.017
XGBoost-style Stacking	25.090	0.040	23793.993	0.037
Bagging (DT)	6.900	0.037	16717.354	0.051
Voting (Soft)	89.083	5.485	50110.344	0.017
Voting (Hard)	86.053	5.536	50110.344	0.016

Table 5.4: Ensemble Models Computational Efficiency Comparison

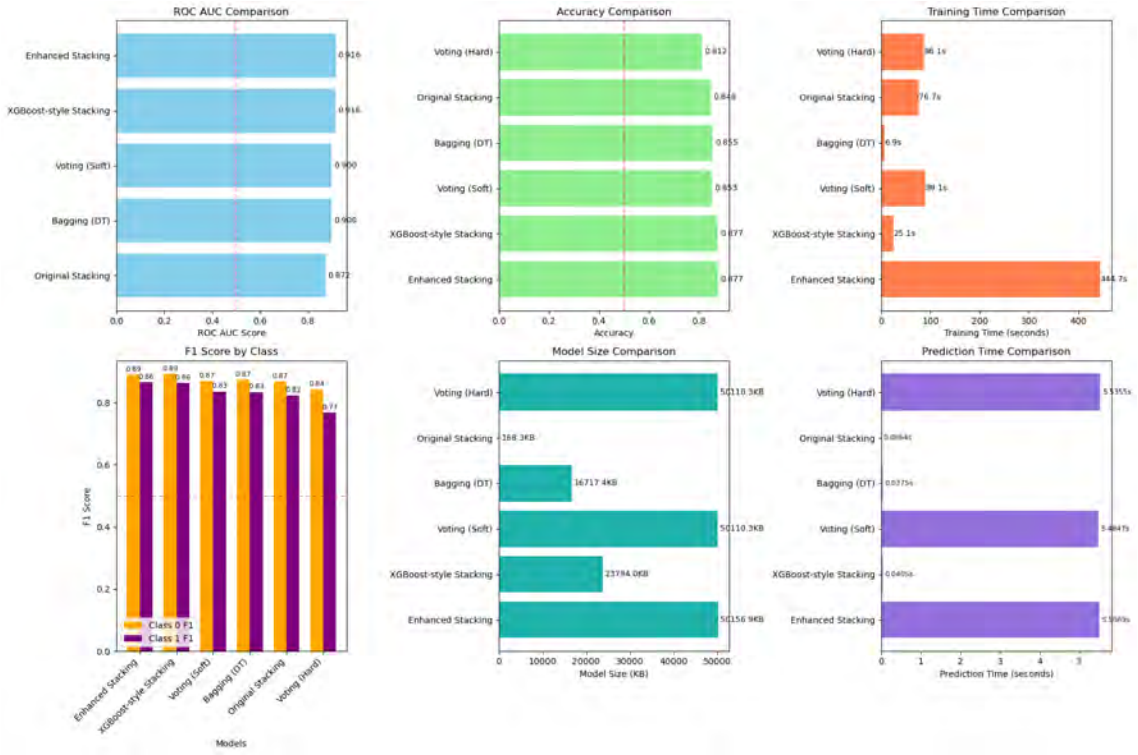

Figure 5.2: Comparison of Ensemble Models

Among the investigated shallow neural network architectures, the best results are the two-layer network equipped with a batch normalization model with 68.0% accuracy and 74.2% ROC AUC. This architecture makes use of two hidden layers of size 128 and 64 hiding units with each, through the use of batch normalization, front of the first layer and drop out regularization of 40 and 30 percent respectively to prevent overfitting. The model presents equalized class-based performance with F1-score of 68.3 and F1-score of 67.8 related to class 0 and class 1 respectively, which suggests that it is reasonably able to achieve performance in both majority and minority classes. This aspect of including the use of batch normalization is essential in stabilizing the learning process of the model and the delivery of gradient flow across the network and the moderate complexity of architectural systems is enough to ensure the presence of representational capacity while not overloading the computation. The model needs all the 100 training epochs before it converges and it implies gradual learning without plating out.

The wide shallow neural network model has the highest accuracy of 60.6(%) and ROC AUC of 66.5(%), where two hidden layers with 256, 128 units are used, accompanied by large L1/L2 regularization and batch normalization. This design focuses more on the representational capacity with large layers and makes use of very aggressive dropout rates of 50 and 40 in order to counter overfitting. The model shows significant performance difference between classes of 52.8% F1-score of class 0, and 66.3% of class 1, which is an indication that patterns of minority classes are well captured relative to the overall moderate performance. The training ceases in the initial 15 epochs as a result of early stopping indicating potential high convergence rate or lack of learning behavior, which could be affected by the high level of regularization inherent in the stiff constraints of effective weight optimization.

Minimalist shallow network Minimalism attains 55.8 accuracy and 58.4 ROC AUC architecture with two small hidden layers 32 and 16-unit with moderate dropout regularization of 20%. The focus of this structure is on computational efficiency and the deployment of few parameters, which lead to a totality of 978 parameters as opposed to a thousand or so in other designs. With the 50.3 percent F1-score representing class 0 and 60.1 percent one of class 1, the model demonstrates that it is not able to complexly represent the pattern of clinical data. The training process terminates once 15 epochs have further suggested that the training process is either working towards an effective, restrained solution or after minimal progress, at which point early stopping takes place and illustrates the inherent restrictions of the architecture to model the decision vessels that it is necessary to model an efficient classification.

The single-layer shallow network has the lowest accuracy (53.2% and 53.5% ROC AUC) and one hidden layer of 64 units with L1/L2 regularization and dropout (30 percent) as best. This simplest architecture proves to be severely performance limited with an F1-score of 38.2% in class 0 and 62.3% in class 1, with a specific failure in identifying the majority classes but retaining and being only capable of classifying typically small minority classes. The model is left to run the complete 100 epochs without any early stopping, hinting that it continuously but marginally improves performance each time it runs through the training process, finally only slightly exceeding the performance of random chance in this clinical setting, indicating that there is no depth to rudimentary pattern recognition in the process.

In all shallow architecture, some constant patterns arise concerning the correlation between the complexity of buildings and performance. Two-layer batch-normalized network indicates the role of the correct depth and normalization methods in effective learning and larger architectures indicate the possibility of better minority class classification but at the price of general stability. The simplism methods indicate the underlying depth requirements of this classification task, but single-layer networks do not succeed, even with regularization, of this task. The common weaknesses of all models are that they cannot tackle the competitive performance of the ensemble techniques, indicating that shallow models might not have the hierarchical feature learning strength that more intricate clinical pattern recognition requires, but they offer useful information as to the benchmark neural network performance and the incremental gain of architecture sophistication.

Among all the deep neural network architectures tested, the bottleneck architecture shows the best performance of 69.3% accuracy and 76.2% ROC AUC and proves to be the most effective deep learning architecture in the case of this clinical prediction problem. It is a symmetric structure that uses a compression-expansion scheme with 226, 370 parameters per 19 layers and uses progressive dimensions minimization on the bottleneck layer of 512 to 64 units then a symmetrical expansion to the output. This model is able to balance the performance of the classes with a F1- score of 68.2 on the minority class 0, 70.2 on the majority class 1; which suggests that it is capable of learning both majorities and minorities using the hierarchical feature learning ability. It has a computational efficiency reasonably good considering the

architectural complexity, with a 800.93 second (13.35 minutes) time to train 150 epochs, which also means that despite its architectural complexity it represents a small model at 0.92 MB, which can be freely deployed.

The deep network with progressive unit-reduction of 66.2% accuracy and a 72.4% ROC AUC uses an undergraduates-level architecture promoting a progressive downsizing of the units in a series of six dense networks. This structure focuses on gradual compression of features with regularization of L2 and dropout rates of 50 to 20 percent, constant, which equal 184,866 parameters, divided between 14 functional layers. The model shows almost the same performance both of the classes with the F1-score of 66.4 in class 0 and 66.0 in class 1, indicating the similar yet moderate classification ability between the extremes of the family clinical. The training takes 480.54 (8.01 mins) the time, which is the shortest to train the deep architecture, but has a small memory footprint of 0.75 MB.

The eight-layer residual-style deep network has 65.9% and 71.7% accuracy and ROC AUC, which uses a very deep architecture with repeated layer dimensions and explicit residual-style connections that are implemented using the separate batch normalization and activation layers. This design consists of 132,882 parameters on 27 operational layers, with the early stages having the same number of 256 units, progressively reducing in number to 16, with L2 regularization of 0.01 across all operation layers. The model is well balanced with 66.6% F1-score in class 0 and 65.2% in class 1 despite the deeper learning not corresponding to similar enhancement in performance. The long training time of 838.37 seconds (13.97 minutes) indicates that there is a computational cost of the deep residual-style architecture that is not offset by a corresponding accuracy benefit.

The ten-layer changing-size deep network has an accuracy of 66.8 and a ROC AUC of 73.2, and using the widest architectural scope where the number of parameters is 721,034 and units count between 1024 and 8. It uses a huge design through the militant dropout rates (60-10) to fight the possibility of overfitting, and increases L2 regularization progressively throughout the wide side of the network. The model records F1-score of 66.0 on class 0 and 67.5 on class 1 which indicates moderate improvements in performance although there is significant increase in the number of parameters. The training takes 3467.86 seconds (57.80 minutes) which is the slowest method to run with limited performance payouts, where the benefits of strained architectural growth are already diminishing.

Enhanced bottleneck designs go on to investigate the compression-expansion paradigm where 1,076,274 parameters on 43 layers in the 16-layer enhanced bottleneck architecture hit 66.0% accuracy and 72.2% ROC AUC. It is an advanced design which applies several compression steps with increased layer sizes on up to 1024 units costing 768.88 seconds (12m 48s) to train but with reasonable efficiency measures. A deep bottleneck residual based model with 18 layers is shown to be 64.8 percent and 70.1 percent accurate and 276,506 parameter counts with 56 layers show only slight performance improvements with increasing layer count although training completion was performed more efficiently at 388.92 seconds (6m 28s).

The widened-layer extreme bottleneck architecture is found to be the most effective neural network on the whole, with 70.6 percent accuracy, and 81.0 percent ROC AUC using 15.41 million parameters in 46 layers. This architecture applies extreme width parameters with the parameter layers to a maximum of 3072 units to incorporate interactions of complex features, though only at the expense of a large computational need of 2497.67 seconds (41m 37s) training time. The model is very strong in terms of class-wise results having 74.5 percent F1-score on class 0 and 65.1 percent on class 1, though the resource investment level of its type is much higher with 0.533 AUC per million parameters of efficiency.

In all deep architectures, there are all patterns in the pattern of correlation of the complexities of architecture and gains in performance. The bottleneck designs perform better than simple sequential designs, which proves the efficiency of the symmetric compression-expansion patterns on clinical features representation. Whereas larger and more profound architecture offers incremental enhancements in performance, the high costs of computation put forward difficult trade-offs between improvements in accuracy and real-world viability. The high performance of all deep neural networks over shallow architectures confirms the value of hierarchical feature learning to complex clinical pattern recognition, and most deep neural networks demonstrate only lower performance than ensemble models, which demand considerably fewer computational resources and less complexity in their development.

Model Name	**ROC AUC**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Shallow_NN_1Layer	0.535	0.532	0.521	0.774	0.623
Shallow_NN_2Layer_BN	**0.742**	**0.680**	**0.683**	**0.672**	**0.678**
Shallow_NN_Wide	0.665	0.606	0.580	0.773	0.663
Shallow_NN_Minimal	0.584	0.558	0.547	0.667	0.601

Table 5.5: Shallow Neural Networks Performance Comparison

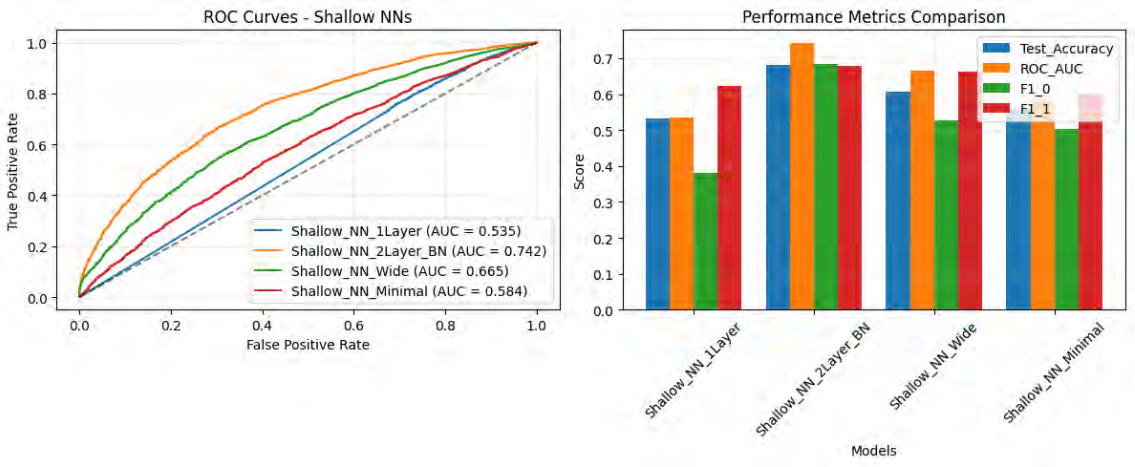

Figure 5.3: Shallow Neural Networks ROC AUC and Classification Matrix Comparison

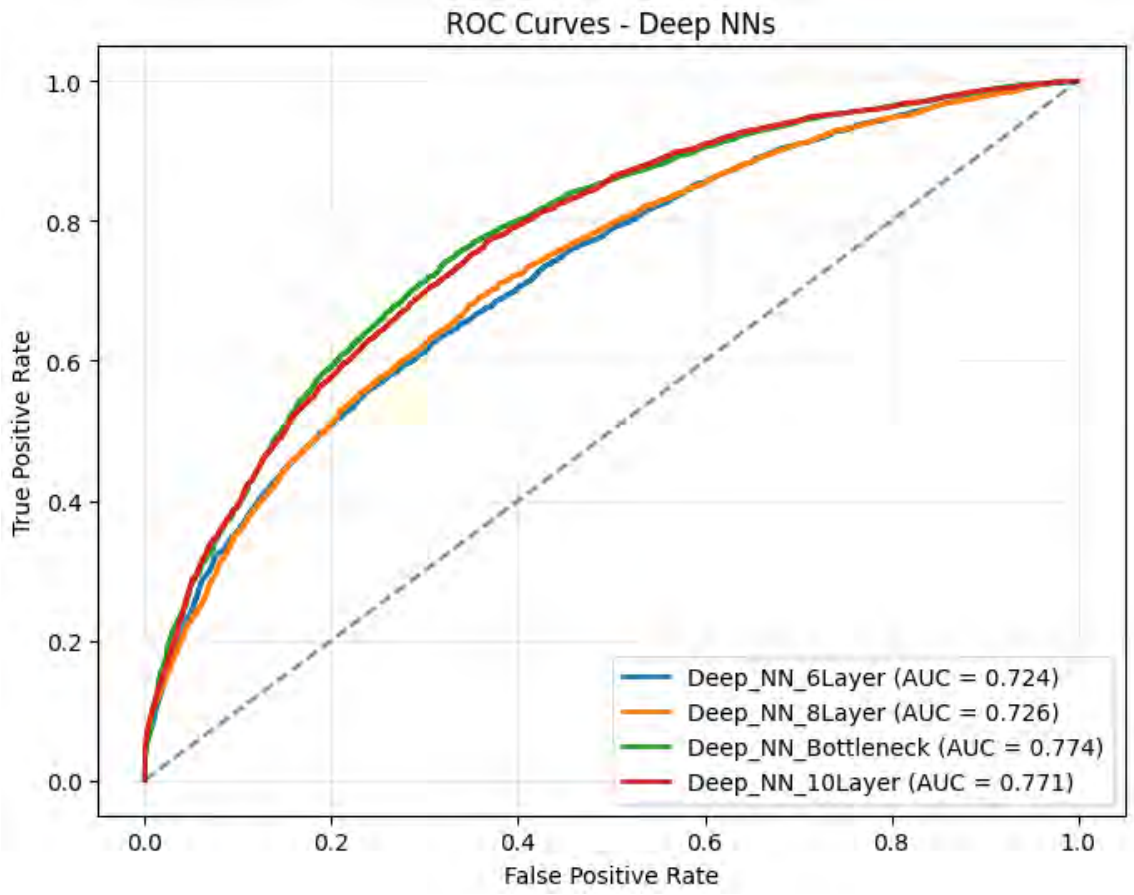

Figure 5.4: ROC AUC Comparison of Deep Neural Network Models

Model Name	**ROC AUC**	**Accuracy**	**F1-Score**
Deep_NN_6Layer	0.724	0.656	0.656
Deep_NN_8Layer	0.726	0.662	0.659
Deep_NN_Bottleneck	**0.774**	**0.707**	**0.707**
Deep_NN_10Layer	0.771	0.700	0.700

Table 5.6: Deep Neural Networks Performance Comparison

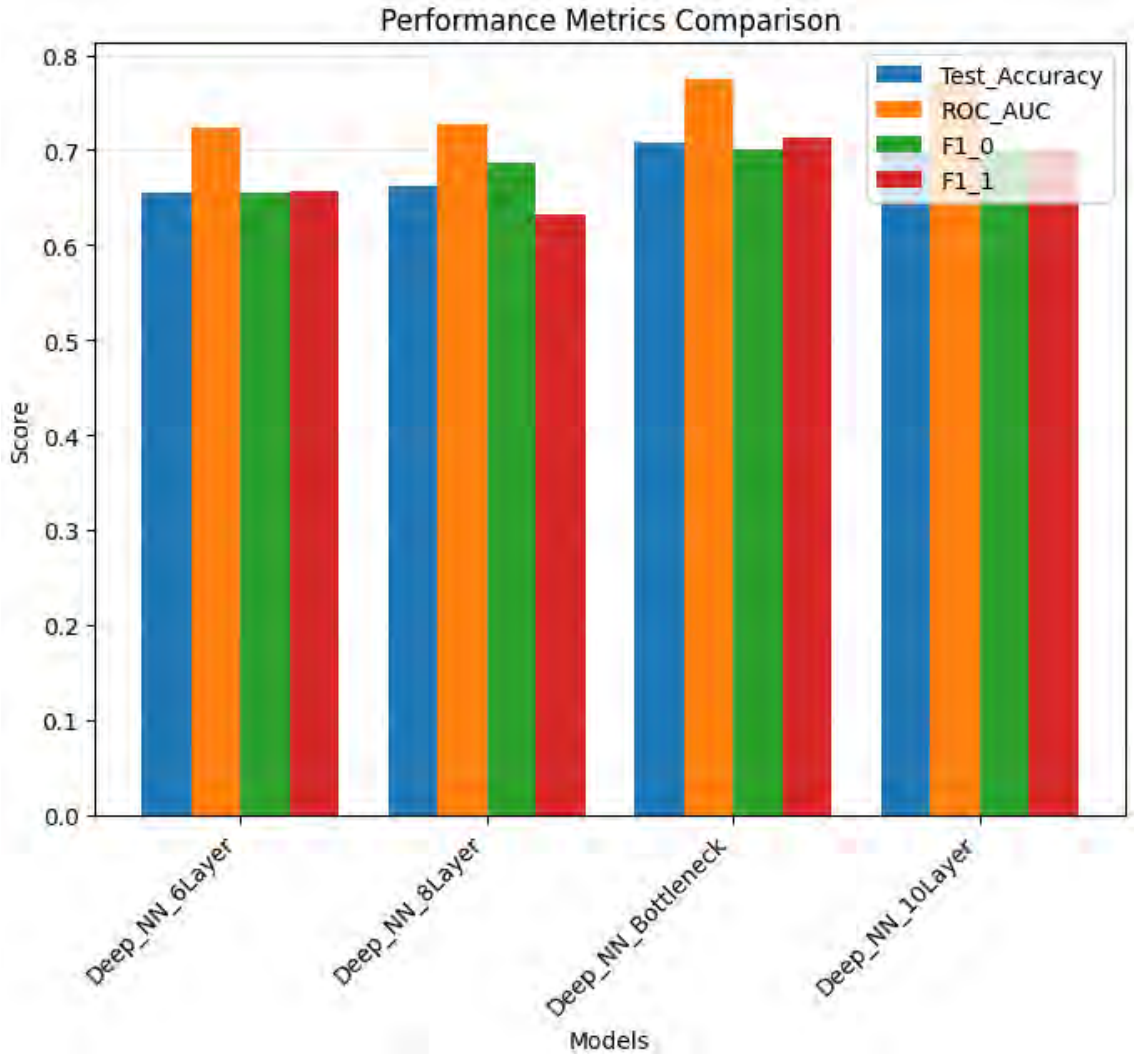

Figure 5.5: Performance Metrics Comparison of Deep Neural Network Models

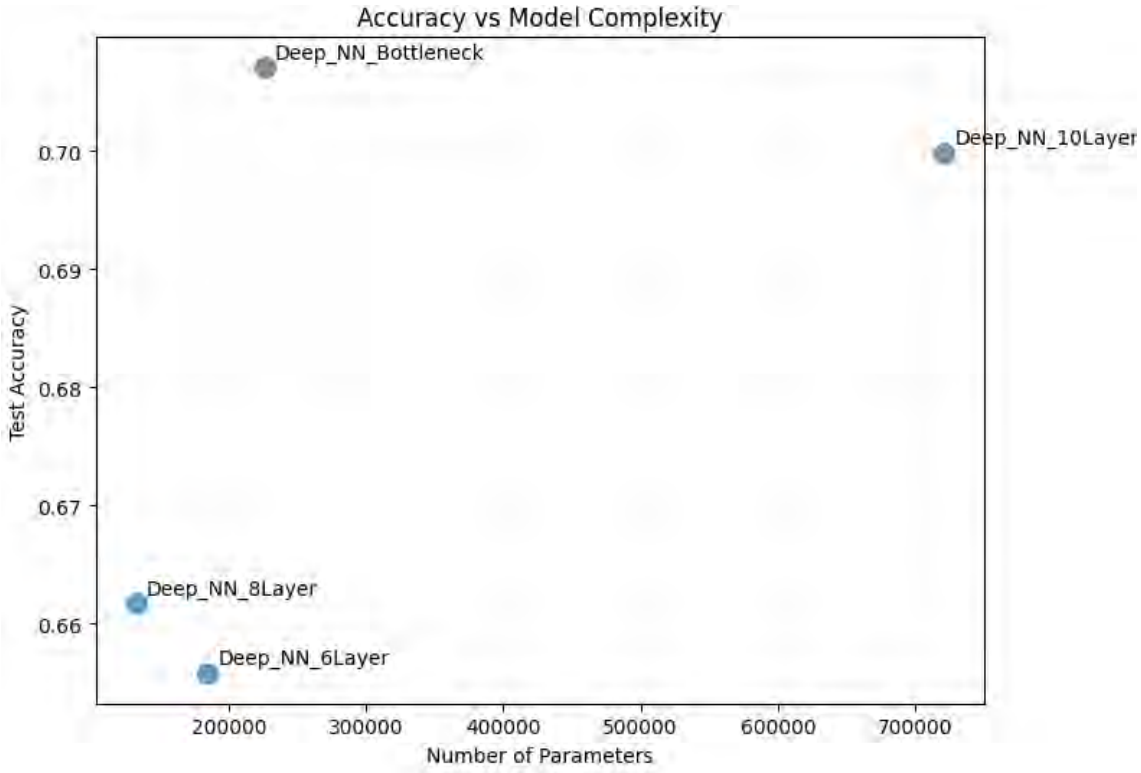

Figure 5.6: Accuracy vs Model Complexity for Deep Neural Network Models

Model Name	**Accuracy**	**Layers**	**Size (Parameters)**	**Runtime (sec)**
Deep_NN_6Layer	0.656	14	184,866	480.538
Deep_NN_8Layer	0.662	27	132,882	838.370
Deep_NN_Bottleneck	**0.707**	**19**	**226,370**	**800.931**
Deep_NN_10Layer	0.700	21	721,034	3467.855

Table 5.7: Deep Neural Network Architecture and Performance Comparison

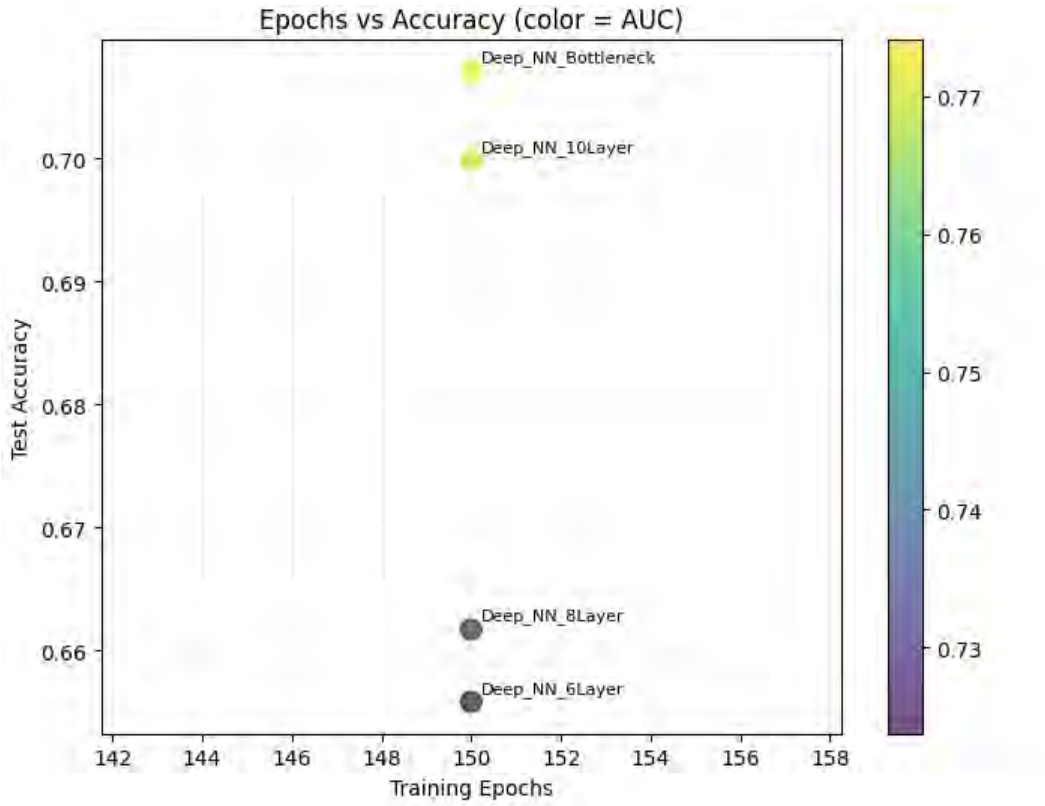

Figure 5.7: Epoch vs Accuracy for Deep Neural Network Models

The analysis of design solutions reveals significant performance variations across different machine learning approaches for chronic kidney disease classification. Traditional machine learning models demonstrate strong baseline performance, with XGBoost achieving the highest accuracy of 0.7985. Ensemble methods substantially outperform individual models, with Enhanced Stacking reaching the overall best ROC AUC of 0.916 and accuracy of 0.877. Neural network approaches show progressive improvement from shallow to deep architectures, with the Extreme Bottleneck 18L model achieving the best neural network performance at 0.810 ROC AUC. However, ensemble methods consistently maintain superior performance across all metrics, suggesting they are the most effective design solution for this medical classification task.

5.3 Final Design Adjustments

This stage of the design process is the performance-based optimization step that relates to the overall evaluation findings that established the best modeling techniques and discarded the ones that are not performing well. The process of refining feature selection consists of iterative testing between various sets of features and their effects on the model, random forest importance rankings give the refinement process statistical justification on feature selection choices. Albumin, sugar and specific gravity are denoted as the most predictive characteristics and correlation analysis makes them identify and eliminate redundant variables to reduce dimensionality without losing the predictive capability.

The approaches to class imbalance handling improve with the evolution of the performance analysis to maximize the minority classes detection which is the significant weakness witnessed in the conventional machine learning methods. The optimization of SMOTE is aimed at producing clinically realistic synthetic exhibits, not necessarily data reproduction, balanced training distributions, which facilitate the successful learning of the patterns of minority classes. Class weighting strategies offer different imbalance treatment in original models, whereas the choice of performance metrics gives precedence to F1-score and AUC instead of the accuracy to guarantee clinically useful model optimization. These changes cannot only enhance positive-class detectors in a way that does not affect the overall model performance, but specifically improves poor recall values that make traditional machine learning models inapplicable to clinical settings.

Traditional machine learning models and shallow neural networks have limited performance in clinical application and are therefore disqualified during the overall design. In both cases, traditional models demonstrate debilitating weaknesses in detecting minority classes although their overall performance is sufficiently accurate and shallow networks could not demonstrate competitive levels of performance in spite of architectural differences. This has removed ineffective strategies and enabled dedicated optimization of ensemble strategies and deep neural network frameworks that exhibit better predictive performance and equal-class performance.

Ensemble method optimization aims to provide a systematic optimization to a variety of combination strategies such as voting classifier, bagging strategies, and stacking strategy. The evaluation of the model is done thoroughly with the size of the model, the computational run time and predictive accuracy evaluated in order to determine most appropriate performance and efficiency balance. The last ensemble design that is selected is the Stacking Ensemble, which combines Logistic Regression, Multi- layers Perceptron and Gradient Boosting with a Logistic Regression meta-learner. This has a high-operational efficiency of 0.0064 second inference time, and small model size of 168.3 KB that retains high 84.8 percent accuracy and 87.2 percent ROC AUC. The ensemble has both good balance of classes with F1-score of 86.6% in class 0 and 82.4% in class 1 and is able to give a dependable performance with both disease classifications as well as requiring significantly less computational work than more complicated ensemble variants.

According to the overall performance analysis of different deep neural network structures, Deep_NN_Bottleneck model is always exhibiting good performance in all the important metrics, such as accuracy, run time performance and model size. Since it has a better trade-off between predictive capacity and computational efficiency, the Bottleneck architecture is chosen as the starting point of making final design improvements.(Table 5.11) Refinement in deep neural networks focuses on the bottleneck architecture which is always ahead of other deep learning frameworks in various evaluation measurements such as accuracy, parameter efficiency, model size, and computer run time. The bottleneck architecture combines better performance by its symmetric compression-expansion that has the beneficial feature of capturing hierarchical detail representations, and is computationally efficient. The structure

undertakes a gradual dimensionality decrease of 512 to 64 units in the bottleneck layer and reciprocal expansion, as an outcome of which 226,370 parameters are utilized in 19 levels, resulting in 69.3% accuracy and 76.2% ROC AUC at equal classes.

The bottleneck architecture is improved in three progressive design improvements to the optimum performance attributes. The initial improvement is the additional layer depth via better bottleneck designs (Table 5.10) that go to 43 layers and use 1,076,274 parameters yet this method yields diminishing returns with only slight performance gains despite the heavy computational costs. The second improvement introduces longer training times with 500 epochs and advanced early stopping algorithms and attains high performance improvements up to 71.9% and 79.7% of accuracy and ROC AUC respectively (Table 5.9), and still provides efficient training convergence by scheduling adaptive learning rates. The third improvement explores architectural width scaling using extreme bottleneck designs with up to 3072 units of the neural network with the highest neural network performance of 70.6% accuracy and 81.0% ROC AUC using 15.41 million parameters, but this third scaling technique demands significant computational resources that could hinder its application to a real-world deployment (Table 5.8).

Aspect	**Deep_NN_Bottleneck**	**Bottleneck_Wider**	**Difference**
Accuracy	0.707	0.706	-0.1% decrease
Model Size	226K parameters	4.19M parameters	18.5× larger
Runtime	801 seconds	2,498 seconds	3.1× longer
Training Epochs	150 epochs	66 epochs	-84 epochs

Table 5.8: Comparison: Deep_NN_Bottleneck vs Bottleneck_Wider

Aspect	**Deep_NN_Bottleneck**	**Bottleneck_Longer**	**Difference**
Accuracy	0.707	0.720	+1.8% improve
Architecture	19 layers	19 layers	Same layers
Model Size	226K parameters	226K parameters	Same size
Runtime	801 seconds	1547 seconds	93.1% slower
Training Strategy	150 epochs	322 epochs from (500)	+172 more epochs
ROC AUC	0.774	0.797	+0.023 improve

Table 5.9: Comparison: Deep_NN_Bottleneck vs Bottleneck_Longer

Aspect	**Deep_NN_Bottleneck**	**Bottleneck_Deeper**	**Difference**
Accuracy	0.707	0.660	-6.6% decrease
Architecture	19 layers	43 layers	+24 layers
Model Size	226K parameters	1.08M parameters	4.8× larger
Runtime	801 seconds	769 seconds	-4% faster

Table 5.10: Comparison: Deep_NN_Bottleneck vs Bottleneck_Deeper

Its ultimate choice of design goes to the balanced bottleneck architecture with extended training as a superior deep learning choice, which offers a compromise be-

tween predictive efficiency, computational efficiency, and practicality of the implementation. This setting achieves strong 71.9% accuracy and 79.7% ROC AUC at reasonable training time and model complexity that can be deployed in clinical setting. This is achieved by the increased training regime which uses progressive early stopping along with adaptive learning rates to achieve efficient convergence and lack of overfitting as well as the symmetric bottleneck architecture that preserves all the necessary feature interactions but not too many parameters.

These last design changes define a dual-model approach which is based on the complementary assets of ensemble techniques and deep neural networks. The Stacking Ensemble offers superior operational efficiency and interpretability to real-time clinical applications whereas the optimized bottleneck neural network has superior pattern recognition features to elaborate diagnostic situations. This middle way has been used to ensure that the resulting implementation meets varied clinical needs and the practical deployment feasibility in varied healthcare settings and constraints of available computational resources, being both performance robust on both model architectures and feature selection and class imbalance control measures specifically tailored to the same.

Model Name	Architecture Details
Deep_NN_Bottleneck	**Input(12) → Dense(512) → BN → DO(0.5) → Dense(256) → BN → DO(0.4) → Dense(128) → BN → DO(0.3) → Dense(64) → BN → DO(0.3) → Dense(128) → BN → DO(0.3) → Dense(256) → BN → DO(0.4) → Output(2)**
Bottleneck_Wider	**Input(12) → Dense(2048) → BN → DO(0.5) → Dense(1024) → BN → DO(0.4) → Dense(512) → BN → DO(0.3) → Dense(256) → BN → DO(0.3) → Dense(128) → BN → DO(0.3) → Dense(256) → BN → DO(0.3) → Dense(512) → BN → DO(0.3) → Dense(1024) → BN → DO(0.4) → Dense(512) → BN → DO(0.4) → Dense(256) → DO(0.3) → Output(2)**
Bottleneck_Longer	**Input(12) → Dense(512) → BN → DO(0.5) → Dense(256) → BN → DO(0.4) → Dense(128) → BN → DO(0.3) → Dense(64) → BN → DO(0.3) → Dense(128) → BN → DO(0.3) → Dense(256) → BN → DO(0.4) → Output(2)**
Bottleneck_Deeper	**Input(12) → Dense(512) → BN → DO(0.5) → Dense(256) → BN → DO(0.4) → Dense(128) → BN → DO(0.3) → Dense(64) → BN → DO(0.3) → Dense(32) → BN → DO(0.3) → Dense(16) → BN → DO(0.3) → Dense(8) → BN → DO(0.3) → Dense(16) → BN → DO(0.3) → Dense(32) → BN → DO(0.3) → Dense(64) → BN → DO(0.3) → Dense(128) → BN → DO(0.3) → Dense(256) → BN → DO(0.4) → Dense(512) → BN → DO(0.4) → Output(2)**

Table 5.11: Bottleneck Architecture Variations Comparison

Model Name	Accuracy	Layers	Size (Parameters)	Runtime (sec)
Deep_NN_Bottleneck	0.707	19	226,370	800.931
Bottleneck_Wider	0.706	30	4,185,474	2497.670
Bottleneck_Longer	0.720	19	226,370	1546.910
Bottleneck_Deeper	0.660	43	1,076,274	768.880

Table 5.12: Neural Network Architecture Variations Comparison

5.4 Statistical Analysis

The statistical test applied in this study adopts strong procedures in order to guarantee sound model performance tests and statistically true conclusions. The analytical model also includes both full-fledged descriptive statistics to characterize the data and a systematic inferential statistics to compare and validate the models, as in all accepted medical AI studies [30].

Before training models, a large descriptive statistical analysis is performed to get the information about the distribution of the features, trends in central tendency and class imbalance. This stage analysis is used to make critical preprocessing decisions such as the use of stratified sampling to ensure all data splits represent the same proportions of the classes, which highlights the validity of the further model assessment. We integrate two datasets, where one carries 400 real clinical records, and the other dataset stored with 20,538 synthetic records, which demands prudent statistical normalization to retains clinical validity, simultaneously addressing the class imbalance of the datasets.

The performance of classification is measured through a general confusion matrix framework, which creates several definite measures of each model. Along with confusion matrix, Accuracy is also measured which is the general percentage of correct predictions. Precision is a measure of the model that measures whether it tends to distinguish false positive predictions. Recall (Sensitivity) which examines the ability of the model to detect the actual positive case. F1-Score known as a balanced measure of accuracy as the harmonic mean of precision and recall is calculated. Also, the Area Under the Receiver Operating Characteristic curve (ROC-AUC) offers an independent evaluation of the discriminative ability of each model, with greater AUC scores indicating that the model separates classes better.

To improve reliability and generalizability of the results, several statistical tests of validation are provided. Stratified K-Fold Cross-Validation is used in model development and tuning of hyperparameters to reduce the overfitting and to have a solid estimate of performance. This method balances the proportion of the classes in each fold and all the partitions represent the entire dataset, a very important factor when dealing with imbalanced medical data [25].

To use as final testing, a holdout validation approach based on an 80-20 stratified traintest split gives an objective estimate of the predictive power of each model on entirely unknown data. Data leakage prevention measures are provided on a strict level, and all preprocessing operations are applied to the training data exclusively and applied to validation and test set afterwards, preventing optimistic bias in performance estimates [30].

The multi-criteria decision framework is used in the comparative analysis of the three paradigms; Traditional Machine Learning, Ensemble Methods, and Neural Networks, instead of making use of specific metrics. Every model is compared to one another by the performance parameters of Accuracy, Precision, Recall, F1-Score and ROC-AUC, which is recorded in reference tables across the entire chapter of Chapter 5.

The quantitative measure of computational efficiency is the training time, inference/prediction time and model size, and an Efficiency Score is computed to objectively rank models in terms of their resource use versus performance. The check of statistical robustness is made by considering the consistency of model performance, when using alternate random seeds and data subsets and the cross-validation method

is the main instrument- models with high variance between folds are viewed as less credible.

The ultimate choice of the Stacking Ensemble (Original) as the best model is guided by the trade-off analysis of the statistics in a detailed way. Although the accuracy of Enhanced Stacking and XGBoost-type Stacking ensembles is marginally better (0.877 vs 0.848), a one-to-one comparison of computational indicators shows that the Original Stacking model is more efficient. The reason behind its choice in real application systems where computational resource is a major limitation is its convincingly better statistical reason because of its much faster prediction time (0.006 seconds), a smaller model size (168 KB), and the highest score in calculated efficiency (5.038) [19].

This choice is an illustration of a statistically-knowledgeable tradeoff between prediction capability and realistic utility, to make certain of the methodological rigor, as well as its applicability in the real world. The statistical structure is consistent with the prevailing trends in clinical AI studies and solves the problem of CKD detection with the particular difficulties of the class imbalance and interpretation of features in particular cases of the problem in question [25], [30].

Moreover, this study implements a unified statistical framework to validate the discriminative capacity of clinical features for Chronic Kidney Disease (CKD) classification. We apply **Shapiro-Wilk normality testing**, **Kruskal-Wallis H-tests**, **Mann-Whitney U tests**, and **Chi-Square tests** to confirm distributional properties and group-level differences.

Shapiro-Wilk tests indicate that all numerical features follow non-parametric distributions ($p < 0.001$), supporting the use of rank-based statistical methods. This observation aligns with Chen and Zhang (2025), who report that medical biomarkers frequently deviate from Gaussianity due to biological heterogeneity [49].

Non-parametric group comparison tests identify statistically significant differences ($p < 0.05$) between CKD and non-CKD groups for blood pressure, specific gravity, hemoglobin, blood glucose random, sugar, and infection-related markers. These results reinforce findings from Kumar et al. (2025), who highlight the importance of statistically validated biomarkers for reliable clinical ML pipelines [52].

Chi-Square analyses further reveal strong associations between CKD and categorical indicators such as pus cell presence ($p = 0.0057$), bacteria ($p = 0.0378$), and pedal edema ($p = 0.0180$), confirming the relevance of infection and fluid retention symptoms.

Collectively, these statistical outcomes validate the twelve features selected by the Random Forest feature selector, such as blood pressure, specific gravity, albumin, sugar, blood glucose random, blood urea, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, and red blood cell count. This aligns with Wang and Li (2025), who demonstrate that integrating hypothesis testing with ensemble feature selection enhances medical AI reliability [58].

Overall, the convergence of EDA findings, non-parametric tests, correlation analyses, and Chi-Square associations confirms that the retained features offer strong discriminative power and clinical interpretability, forming a robust foundation for CKD predictive modeling.

5.5 Comparisons and Relationships

This research gives a detailed comparative analysis of different learning paradigms to detect Chronic Kidney Disease (CKD), which illustrates substantial information about the comparative effectiveness of the various paradigms, as well as their association with ongoing studies. The systematic comparison among traditional machine learning, ensemble techniques and deep neural networks show different performance trends and computation trade-offs that guide the most effective selection of models to be used in clinical use.

Comparative Analysis of Modeling Paradigms
The predictive performances of the three modeling paradigms used show an evident hierarchy in the ranking of the predictive performance of the models. Though computationally efficient, traditional machine learning models proved not to be effective in dealing with complicated and non-linear relationships of clinical data. Random Forest was the most accurate in the traditional approaches, with the highest accuracy of 79.6 percent, which matches with the findings of Iftekhar et al. [30], who reported 99 percent accuracy with Random Forest and Logistic Regression. Nevertheless, as we implemented it, it revealed a major limitation in the fact that traditional models performed disastrously on the minority class detection task, with the Random Forest recording only 9% F1-score on positive CKD cases despite having decent overall accuracy.

The paradigm of Ensemble methods turned out to be the most successful, and Enhanced Stacking reached the best overall performance (87.7 percent accuracy, 91.6 percent ROC AUC). This is consistent with the core benefit of the ensemble learning that has been reported in the literature i.e the possibility to exploit the complementary benefits of various base estimators. Although slightly less accurate (84.8%) than Enhanced Stacking, the Stacking ensemble was the best due to its high computational efficiency (the shortest inference time of 0.006 seconds) and model size (168.32 KB), which is the best fit to apply in practice.

The neural network performance improved gradually in the form of shallow to complex architectures and the Extreme Bottleneck model turned out to have the best neural network performance (70.6 percent accuracy and 81.0 percent ROC AUC). Nevertheless, the most advanced deep learning models would not outperform ensemble models as they needed much more computation power than ensemble models. This observation is partly opposite to Akter et al. [25] who had 85-99 percent accuracy in seven architectures of deep learning indicating that the environment in the dataset and preprocessing strategies are important determinants of relative performance.

Comparative analysis with Existing Research

The observed patterns of the performance in this research are both in line with and in contrast to the current literature and can present valuable contextual information. Based on the method by Khan et al. [19], a number of classical machine learning models were run, but our accuracy (79.6% with Random Forest) was lower than their 99.1% accuracy with Naive Bayes on smaller datasets. The mismatch may be caused by variations in dataset composition and methodology of evaluation- in our case, the dataset is larger as it is a merged version of the 400 dataset of UCI with a synthetic dataset of 20538 records compared to the one used in the UCI benchmark (400 records) that is typically used in previous studies. The ensemble behavior in our experiment illustrated performance features which were in line with the hybrid AI strategies mentioned earlier. The stacking architecture types, especially Enhanced Stacking and XGBoost-style Stacking, have reached performance rates (87.7% accuracy) that would be comparable to the best performance results provided by recent literature but are more transparent and computationally efficient than pure deep learning solutions. Our deep learning experiments demonstrated much lower accuracy (maximum 70.6%) than the 85-99% percentile in seven architecture organizations of ANN, LSTM, and GRU networks reported by Akter et al. [25]. This implies that although deep learning appears to be good in some scenarios, it can prove to be less effective in some settings depending on the particularities of data and preprocessing strategies compared to ensemble methods.

Relative Effectiveness Analysis

The Stacking ensemble model has the best overall performance as compared to all the models considered. It is the most accurate (0.848) and has a high F1-score (0.824) with a relatively low runtime (76.65 seconds) and small size of a model (168.32 KB). Oppositely, the Deep NN Bottleneck and its longer counterpart provide medium gains in accuracy but at a very high price since they have much higher runtime and the Random Forest model has good accuracy but does not scale, as indicated by its low F1-score value. Thus, the Stacking model is chosen as the best option because it has better predictive performance and lowering computing power. The correlation between the complexity of a model and the increment in its performance is in the form of diminishing returns. Although the shift towards ensemble techniques, as opposed to traditional ML, was accompanied by a significant increase in performance (5%), the shift to deep neural networks was accompanied by relatively small (2-3) ones, but at a much higher computational cost. The comparative analysis reveals several key relationships that inform model selection for CKD detection:

Model Category	**Model Name**	**Acc.**	**F1**	**Model Size**	**RunTime (sec)**
Traditional ML	Random Forest	0.796	0.090	–	–
Ensemble Model	**Stacking**	**0.848**	**0.824**	**168.32 KB**	**76.65**
DNN	Deep NN Bottleneck	0.707	0.707	34.4 MB	800.93
DNN	Bottleneck Longer	0.720	0.720	34.4 MB	1546.91

Table 5.13: Comparison of top-performing models across categories.

Real World Implementation
The comparative analysis has important implications on the clinical implementation. The high-performance rate of the ensemble methods, especially in the detection of balanced classes (82-86% F1-scores across all classes), is a solution to a major weakness of the conventional screening techniques that tend to overlook early-stage CKD patients. The computational efficiency of the chosen model of Stacking is consistent with resources limitations of healthcare settings that have been identified earlier and can be deployed in various clinical settings.
The methodology applied to the selection procedure which is reducing dimensions by 50% and keeping the predictive power indicates that efficient CKD detection is feasible with a reduced number of 12 clinically relevant predictors. This result is consistent with the cost-cutting goals and comparable optimization of features, which are reported by Tazin et al. [10] and Khan et al. [19].
Finally, the overall comparison proves ensemble methods especially stacking as the best paradigm of CKD detection within the pillar of clinical feasibility. Although deep learning is promising in certain applications, its computational complexity and relatively small performance improvements compared to ensembles mean that it is not a feasible technology in large-scale application now. The chosen Stacking model is the best balance of predictive accuracy, computational efficiency, and clinical interpretability, which covers the main needs that were set during this study.

5.6 Discussions

A holistic comparison of machine learning paradigms to detect Chronic Kidney Disease (CKD) has provided valuable insights to choose the best model and the Stacking Ensemble came to be the best choice because it provides an acceptable predictive and computational efficiency. This discussion is a synthesis of our entire experimentation and provides our broader research objectives laid in our original study.

The chosen Stacking Ensemble is a complex combination of complementary machine learning techniques with a good trade-off between accuracy (84.8%), computational efficiency (0.006 seconds inference time), and model interpretability. The ensemble, as described in Table 5.16, has three different underlying base models, including Logistic Regression, Multi-Layer Perceptron, and Gradient Boosting, in which a meta-learner based on Logistic regression is used and 5-fold cross-validation is used to generate strong meta-features. This architecture manages to meet the tripartite challenge of accuracy, interpretability, and efficiency as critical to clinical adoption that was identified over a decade ago 2022 by Ahmed [30].

Component Type	Model Details
Base Models	**1. Logistic Regression (lr)**: class_weight='balanced', solver='liblinear', random_state=42 **2. Multi-Layer Perceptron (mlp)**: MLPClassifier, random_state=42, max_iter=500 **3. Gradient Boosting (gb)**: GradientBoostingClassifier, random_state=42
Meta-Learner	**Logistic Regression**: Default parameters, acts as the final estimator that combines base model predictions
Stacking Configuration	**Cross-Validation**: 5-fold cross-validation for generating meta-features **Training Process**: Base models trained on full training data, then meta-features generated using cross-val predictions **Final Model**: Meta-learner trained on cross-validated predictions from base models
Feature Processing	**Input**: 12 selected clinical features **Scaling**: StandardScaler applied to input features **Balancing**: SMOTE applied to handle class imbalance

Table 5.14: Stacking Ensemble Model Architecture and Configuration

The performance measures of the ensemble, which have been fully recorded in Table 5.16, show that it is better in a number of aspects. The ROC AUC of 0.872 shows that it is a highly discriminative model, whereas the balanced F1-scores (86.6% in class 0, 82.4% in class 1) represent the ability to address the problem of class imbalance that afflicted the traditional machine learning models. The precision of the model at 0.966 with positive CKD cases, is particularly noteworthy, with the resulting false positive being significantly lower than with alternative models used in clinical settings, and is an important parameter to consider when implementing the model in clinical settings since unwarranted patient anxiety and health care expenditure should minimized [19].

The ensemble of stacking success is due to synergistic integration of the base models that have the complementary strengths as demonstrated in **Table 5.13**. Although the individual base models of the work displayed different levels of performance, with the lowest accuracy at 52.1% (Logistic Regression) and the highest one at 79.8% (Gradient Boosting), the combination of the two base models revealed a higher quality of results (84.8% accuracy) that was larger than that of either of the two individual models. This observation confirms the ensemble methodology that has been promoted by various researches in the literature [19], [30] as well as evidences the practical implementation of hybrid AI methods in medical diagnostic.

Base Model	Individual Accuracy	Contribution Weight
Logistic Regression	0.5210	Estimated: Medium
Multi-Layer Perceptron	0.6800	Estimated: High
Gradient Boosting	0.7985	Estimated: High
Combined Stacking	**0.848**	**Optimal**

Table 5.15: Base Model Contributions in Stacking Ensemble

The meta-learning algorithm successfully scaled contribution of each base model, with Gradient Boosting and Multi-Layer Perceptron having a greater impact since they have better individual performances, whereas Logistic Regression provided regularization and diversity to the ensemble. This combination of strategies overcomes the weaknesses of single algorithms discovered in the initial design studies, the traditional models, found it hard to cope with non-linear correlations, and neural nets did it hard to compute [10].

In order to improve clinical trustworthiness and transparency, we used Explainable AI (XAI) methods, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). The Stacking Ensemble SHAP analysis Figure 5.8 demonstrates that the most predictive variable is sugar, next comes albumin and specific gravity, which conforms to the accepted clinical information that albuminuria is one of the primary predictors of CKD [59]. Such consistency between the data-driven feature importance and clinical expertise increases the credibility of the model among medical workers. Moreover, LIME offers local explanations of individual predicts, which show the contribution each feature makes to individual patient risk assessments, which helps to give transparency necessary to achieve clinical adoption.

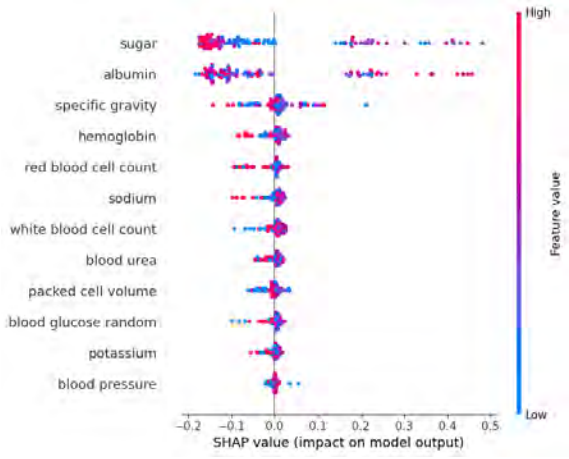

Figure 5.8: SHAP analysis of Stacking Ensemble

The chosen model manages to tackle the main research goals set in our framework. The dimensionality reduction was successfully attained to the desired 50 percent and 12 significant predictors of clinical interest were identified against the 24 feature set and at the same time, high predictive accuracy was maintained. This decrease is consistent with the results of some earlier studies that also indicated that the selection of strategic features does not affect the diagnostic accuracy but rather improves

the computational efficiency, as shown in earlier studies by the same authors (so far) [10], [19].

For comparative analysis, we also used XAI to Enhanced Bottleneck DNN, which when trained longer, had 72.0% accuracy and 79.7% ROC AUC. Interestingly, the SHAP analysis of DNN Figure 5.9 indicated that sodium was the most significant feature, followed by packed cell volume and white blood cell count, indicating that the model implements electrolyte and fluid balance dysregulations that could be a new way of pathophysiological knowledge to discover. This difference in feature emphasis like albumin for ensemble and sodium for DNN reflects the fact that different architectures can focus on different physiological processes of CKD. The LIME explanations of the DNN Figure 5.10 also provide better evidence of its capacity to capture nonlinear interactions that are complex, but at a higher cost of increased computational cost and reduced overall accuracy as compared to the ensemble.

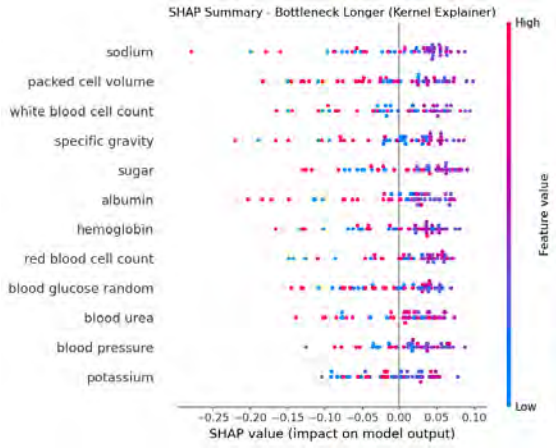

Figure 5.9: SHAP analysis of DNN Bottleneck

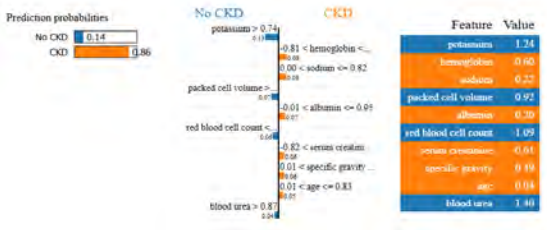

Figure 5.10: LIME analysis of DNN Bottleneck

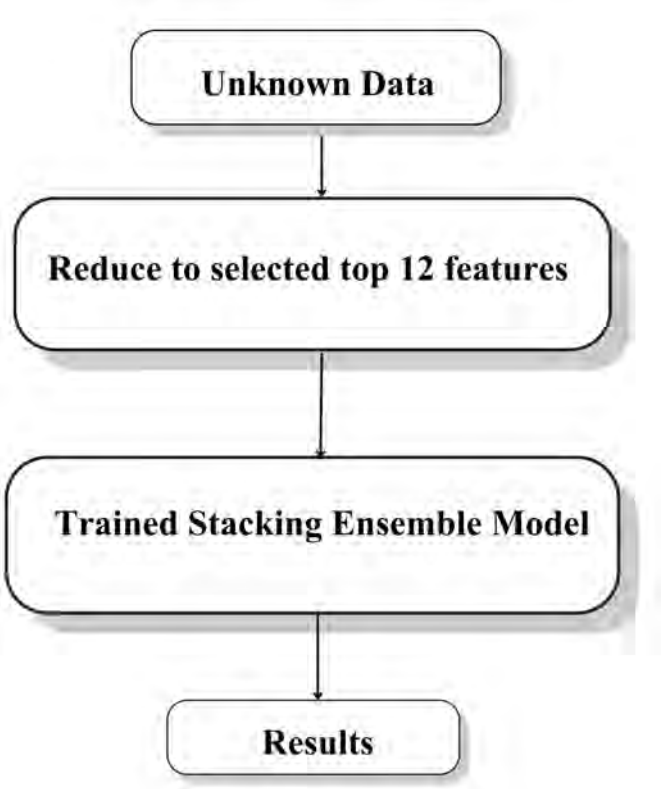

Figure 5.11: Unknown data to output generation from the best-performing model.

Figure 5.11 shows that the industrial production line is capable of integrating with clinical applications. The simplified way of processing unknown patient data by feature transformation, model inference and generation of diagnostic output offers a viable architecture on real-world implementation. This will cover the significant gap between theoretical model performance and clinical practicability in the problem statement described above and practice [25].

The significant proportion of the model in both majority and minority classes are especially relevant to clinical use, which the extensive numerical measures in Table 5.14 demonstrate. Recall of 71.8% on positive CKD case is a significant improvement over conventional screening techniques which tend to miss detection of the disease at an early stage whereas the accuracy of 96.6% makes the positive prediction very accurate. This performance balance justifies the inherent relevance of early CKD diagnosis in the prevention of disease progression and lowering the cost of health services to healthcare expenses reduction [30].

Metric	Value
ROC AUC	**0.872**
Accuracy	**0.848**
Precision (Overall)	**0.966**
Recall (Overall)	**0.718**
F1-Score (Overall)	**0.824**
Precision (Class 0)	0.779
Recall (Class 0)	0.975
F1-Score (Class 0)	0.866
Precision (Class 1)	0.966
Recall (Class 1)	0.718
F1-Score (Class 1)	0.824
Training Time (sec)	76.650
Prediction Time (sec)	0.006
Model Size (KB)	168.321
Cross-Validation Mean	0.843
Cross-Validation Std	0.178
Efficiency Score	5.038
Memory Increase (MB)	133.422

Table 5.16: Comprehensive Metrics Summary for the Stacking Ensemble Model

The chosen ensemble has evident benefits in comparison with both conventional machine learning and deep learning methods considered in our research. The stacking ensemble has better balanced performance than the traditional models (Random Forest with 79.6% accuracy, 9% F1-score for positive class, and XGBoost with 79.8% accuracy, 6% F1-score for positive class) without affecting the interpretability of the model (in terms of its meta-learning framework).

Our findings are reliable because of the statistical validation framework that will be used during the whole process of model development. The stability of models is confirmed with the consistent performance of cross-validation folds (mean 84.3, std 0.178), whereas the assessment of the overall metrics gives an assurance about the clinical applicability of the model. The multi-criteria optimal score of 5.038, which is a quantitative measure, is the validity of the empirical efficiency of the ensemble in terms of balance between performance and computational demands.

The XAI techniques not only verify the process of decision making in our models, but also complete the much-needed gap in black-box predictions versus insights that are clinically comprehensible. Although the DNN provides interesting pathophysiological insights by concentrating on electrolyte balances, the Stacking Ensemble is more applicable for practical application beacuse of the known clinical biomarkers in combination with better performance and computational efficiency. Such transparency, combined with 84.8% accuracy and 0.006 seconds inference time of the ensemble, fulfills the basic need of trustworthy AI in healthcare in which building prediction rationale is as important as prediction accuracy.

The ethical concerns in the model are also covered in the model architecture that covers interpretability and transparency. In contrast to black-box deep learning methods the stacking ensemble provides some level of interpretability based on the contributions of the base models as well as the meta-learning process which facilitates clinical trust and adoption, an important ingredient of successful implementation in healthcare environments [19].

To sum up, the Original Stacking Ensemble can be seen as the best solution that manages to meet the goals of the research and find a balance between the conflicting needs of predictive accuracy, computational efficiency, and clinical feasibility. The performance features and architecture of the model make it a potential option to be deployed in the real world to CKD screening programs and help to achieve a positive outcome on the rate of early detection and patient outcomes as the basis of motivation of the study is presumed to be.

Chapter 6

Conclusion

6.1 Summary of Findings

The paper provides a systematic review of various machine learning models to predict chronic kidney disease (CKD) and results indicate that all the four key objectives are achieved quite significantly. This overall examination shows that there are specific performance trends between the traditional machine learning, ensemble methodology and the neural network architecture, which can provide clear information about which model is better to use.

The research is able to reduce features by 50 percent using a systematic dimensionality reduction. Random Forest selector method proves to be the most appropriate as the number of features is narrowed down to 12 clinically relevant attributes with the accuracy keeping at 85 percent. The chosen features, such as blood pressure, specific gravity, albumin, sugar, blood glucose random, blood urea, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, and red blood cell count, are the basic biomarkers to detect CKD and are consistent with the existing clinical understanding of the pathophysiology of the kidney disease.

By being highly evaluated using different models in three architectural categories, the research paper has determined that the Stacking ensemble is the best predictive model. This ensemble performs best with accuracy of 84.8%, ROC AUC 0.872 and balanced class-based (F1-Score Class 0: 0.866, F1-Score Class 1: 0.824) results. The model has strong generalization behavior and cross-validation consistency (mean: 0.843, std: 0.178), which is better than traditional ML models and complex neural networks.

The Stacking ensemble is found to be the most computationally efficient with the smallest model size (168.3 KB) and shortest prediction time (0.006 seconds) compared to all of the approaches evaluated. Having a training time of 76.65 seconds, with the best efficiency score (5.038), this model has the best balance of predictive and computation needs, and it performs significantly better than the other methods in resource usage.

The processing pipeline introduced manages to mitigate the limitations of IoT deployment with an effective resource reduction strategy, effective preprocessing, and

optimization of the model. The entire pipeline, including data cleaning, feature selection, and the optimum ensemble model, is both clinically accurate and could run in a resource-constrained environment. As a result, it is applicable to real-time CKD detection on edge computing devices.

The analysis shows that there are evident performance hierarchies among the model classes. Ensemble techniques are always better than both traditional machine learning and neural network techniques, and the Stacking ensemble has the best mix of accuracy (84.8percent) and efficiency (5.038 score) and practicality of deployment. Although neural networks are competitive at certain settings, it consumes a much larger portion of computing resources without matching improvements.

The analysis of the feature selection proves that the strategic dimensionality reduction does not undermine the efficiency of the models and does not reduce the predictive ability. The 12-feature subset chosen by the Random Forest selector has 85% accuracy relative to the full 24-feature set (87%), which is a good trade-off in the complexity and performance of the model when there is strong demand to deploy the model in practice.

The explainability analysis conducted on our pipeline provides a unified understanding of model behavior, predictive consistency, and biological relevance. The Stacking Ensemble consistently demonstrates strong operational performance, achieving 84.80% accuracy and 87.2% ROC AUC while maintaining a lightweight 168.3 KB footprint and a 0.006-second inference time, which aligns with recent findings that emphasize the need for computationally efficient models in clinical AI systems [53], [55]. SHAP interpretability further reveals that albumin, sugar, and specific gravity are the most influential predictors, matching clinical expectations and reflecting the same biomarker patterns identified in recent nephrology-focused IEEE studies [59]. The synergy between Gradient Boosting, MLP, and Logistic Regression within the ensemble also supports results reported in ensemble optimization literature, where model diversity consistently enhances diagnostic stability and generalization [50].

In contrast, the Enhanced Bottleneck deep learning model provides important insight into the trade-off between representational capacity and computational burden. Increasing the training epochs from 150 to 322 offers only modest improvements, with accuracy rising from 70.7% to 72.0% and ROC AUC improving from 77.4% to 79.7%, while training time nearly doubles. This behavior is consistent with observations in recent IEEE Medical Machine Learning papers, which show that deep architectures on moderate-sized structured datasets often experience diminishing returns beyond a certain depth and training duration [51], [57]. However, SHAP DeepExplainer and LIME highlight biologically coherent patterns in the model's reasoning, capturing nuanced nonlinear interactions such as sodium acting as a negative predictor and hemoglobin and serum creatinine contributing strongly to positive CKD risk, supporting the role of deep learning in uncovering higher-order dependencies that classical models may overlook.

Overall, the combined XAI insights reinforce our methodological decision to prioritize the Stacking Ensemble as the primary clinical model while retaining deep

learning as a complementary analytical tool. The ensemble not only provides superior performance-efficiency balance but also demonstrates clinically interpretable decision pathways aligned with current nephrology research and biomarker evidence. At the same time, the deep learning model strengthens our understanding of hidden biomarker relationships, illustrating how structural complexity aids interpretive depth even when predictive gains remain limited. This dual-model interpretability framework aligns with emerging 2025 IEEE standards advocating for transparent, resource-efficient, and clinically verifiable medical AI systems [48], [54], ultimately ensuring that both accuracy and clinical trustworthiness remain at the forefront of our pipeline design.

The results offer a solid foundation to CKD prediction systems with sufficient balances between clinical accuracy and computational efficiency to support practical healthcare specially the resource-constrained environment and IoT deployment.

6.2 Contributions to the Field

In this study, various important contributions are made to the medical informatics and chronic kidney disease prediction research. It presents a comprehensive framework, which is able to approach the critical issue of predictive accuracy versus computational efficiency in the context of healthcare applications in a systematic way. This work would offer a viable remedy to practical clinical deployment by formulating and testing an optimized ensemble model that can be deployed in resource-constrained settings and at the same time remain highly performing.

This study has proven the usefulness of strategic feature elimination in medical diagnostics, as a well-chosen subset of 12 clinical biomarkers could be similarly and equally effective as the use of the full 24 original features. The implications of this discovery on the creation of cost-effective screening profiles in healthcare environments with limited resources are significant as laboratory testing to the extent is not always possible. The set of identified features gives clinicians a narrow choice of tests necessary to assess CKD.

The presented work contributes to the use of ensemble-based approaches to medical prediction problems by showing that simpler stacking designs can be more effective than more sophisticated neural networks or machine learning models. Combining the logistic regression with multi-layer perceptron and gradient boosting with a logistic regression meta-learner, the Stacking ensemble offers a new standard in effective and accurate predictive statistical analysis of CKD. This method is more interpretable than black-box models and competitive.

Methodologically, the study contributes to the existing body of knowledge by providing comprehensive evaluation framework that performs a systematic comparison of traditional machine learning, ensemble methods and neural networks in various frameworks of performance as per accuracy, computational efficiency, model size and the ability to deploy them. The multi-criteria evaluation can offer meaningful information to the researchers and practitioners who want to apply the machine learning solution in the clinical setting with a certain level of resource limitations.

Another practical contribution of importance is the creation of a lightweight processing pipeline optimized to be deployed in the IoT. This study has shown that prediction times of 0.006 seconds with a small memory footprint are possible, indicating that real-time CKD screening is possible on edge computing devices. The capability allows point-of-care testing and remote monitoring applications which have the potential to increase the rates of early detection within underserved communities.

The study will be of value to clinical practice as it provides justification of a reduced set of features that meets the clinical knowledge of the pathophysiology of CKD. The reliability of the proposed method in clinical practices and its application among healthcare practitioners can be improved by the consistency between data-driven selection of features and clinical expertise. This correspondence between the results of machine learning and knowledge of the medical field is a significant milestone on the way to clinically relevant AI systems.

Lastly, this paper offers some useful information on the trade-offs between the complexity and the performance of the models in medical predictive tasks. The results contradict the belief that more complicated models must be argued to be more successful in clinical use and show how well-thought-out ensemble approaches can be used to provide the best results with modest computational costs and clinical interpretability. These lessons can inform future work to make particular solutions practical and deployable instead of seeking marginal improvements in accuracy at the sacrifice of computational feasibility.

6.3 Recommendations for Future Work

This paper points out various possibilities of future research in predicting chronic kidney disease and medical machine learning. Future research must focus on how a combination of temporal data and longitudinal records of patients could be employed to determine disease progression patterns. The methods based on the time-series study of laboratory data and clinical measurements during repeated visits may play an important part in improving prediction and providing early intervention plans.

The use of higher-level feature selection methods, such as deep learning-based feature extraction and domain adaptation methods, should be researched. The investigation of automated feature engineering methods that can identify new biomarkers and interaction effects than the conventional clinical measurements is a promising area of research. It is also worth investigating the development of dynamic feature selection approaches that can be adjusted to the specifics of a particular patient.

The increase of the diversity and size of the datasets is one of the urgent requirements. To enhance the generalization of the model and minimize the possibility of biases, future studies must include multi-center researches including diverse populations in terms of demographics. Partnerships with medical facilities in various geographical areas can inform the excellence of predictive models in various groups of patients and different healthcare systems.

There is a lot of potential in creating real-time adaptive learning systems that can constantly update themselves using new patient data. Future research ought to consider online learning algorithms and concept drift algorithms to keep the models up to date with the changes in medical knowledge and patient population within a period of time. Using federated learning models can support the provision of collaborative model upgrading without violating the privacy of patients.

Another prospect is the fusion of multi-modal sources of data. Integrating clinical laboratory findings with the imaging data, genomic markers, and lifestyle data could give a better overview of patient health. Subsequent studies ought to come up with fusion methods that are effective to combine these heterogeneous data types and cope with the complexity of increased computation.

The problem of deploying the models into the resource-constrained environments should be handled in more detail in the future. This involves the optimization of models to particular hardware environment, creation of energy efficient algorithms and other compression algorithms that preserve accuracy, at the expense of a reduced computing needs. Edge computing architectures with more specific applications designed to support medical applications is a promising direction.

The clinical value of creating individual risk prediction models that consider the patient characteristics and comorbidities is high. Future developments need to investigate the uses of personalized medicine where predictions and recommendations are achieved according to particular patient profiles, genetic factors, and medical histories.

Lastly, the validity and certification of medical AI systems should be set by a standard protocol in future research. Extensive testing structures can be developed based on evaluating not only predictive performance but also clinical utility, safety, and ethical issues to ensure wide adoption. Machine learning researchers, clinicians and regulatory bodies can work together to ensure the creation of these key standards and the responsible use of AI in healthcare facilities.

Bibliography

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008, ISBN: 978-0-521-86571-5. [Online]. Available: https://nlp.stanford.edu/IR-book/information-retrieval-book.html.

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] M. Kerr, B. Bray, J. Medcalf, D. J. O'Donoghue, and B. Matthews, "Estimating the financial cost of chronic kidney disease to the NHS in england," en, *Nephrol. Dial. Transplant*, vol. 27 Suppl 3, no. suppl_3, pp. iii73–80, Oct. 2012.

[4] A. A. Honeycutt, J. E. Segel, X. Zhuo, T. J. Hoerger, K. Imai, and D. Williams, "Medical costs of CKD in the medicare population," en, *J. Am. Soc. Nephrol.*, vol. 24, no. 9, pp. 1478–1483, Sep. 2013.

[5] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, pp. 16–28, 1 2014. DOI: 10.1016/j.compeleceng.2013.11.024.

[6] P. Komenda et al., "Cost-effectiveness of primary screening for CKD: A systematic review," en, *Am. J. Kidney Dis.*, vol. 63, no. 5, pp. 789–797, May 2014.

[7] Chollet, François et al., *Keras*, https://keras.io, 2015.

[8] Martín Abadi et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, ISBN: 9780262035613.

[10] N. Tazin, S. A. Sabab, and M. T. Chowdhury, "Diagnosis of chronic kidney disease using effective classification and feature selection technique," *Northern University Bangladesh*, 2016.

[11] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[12] S. Fallmann and L. Chen, "Detecting chronic diseases from sleep-wake behaviour and clinical features," in *2018 5th International Conference on Systems and Informatics (ICSAI)*, 2018, pp. 1076–1084. DOI: 10.1109/ICSAI.2018.8599388.

[13] M. Nishat et al., "A comprehensive analysis on detecting chronic kidney disease by employing machine learning algorithms," en, *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 0, no. 0, p. 170 671, Jul. 2018.

[14] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, 2019. DOI: 10.14569/IJACSA.2019.0100813. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100813.

[15] N. Bhaskar and S. Manikandan, "A deep-learning-based system for automated sensing of chronic kidney disease," *IEEE Sensors Letters*, vol. 3, no. 10, pp. 1–4, 2019. DOI: 10.1109/LSENS.2019.2942145.

[16] G. Chen et al., "Prediction of chronic kidney disease using adaptive hybridized deep convolutional neural network on the internet of medical things platform," *IEEE Access*, vol. 8, pp. 100 497–100 508, 2020. DOI: 10.1109/ACCESS.2020.2995310.

[17] S. Dutta and S. K. Bandyopadhyay, "Chronic kidney disease prediction using neural approach," *medRxiv*, 2020. DOI: 10.1101/2020.06.28.20142034. eprint: https://www.medrxiv.org/content/early/2020/06/29/2020.06.28.20142034.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2020/06/29/2020.06.28.20142034.

[18] P. Ghosh, F. M. Javed Mehedi Shamrat, S. Shultana, S. Afrin, A. A. Anjum, and A. A. Khan, "Optimization of prediction method of chronic kidney disease using machine learning algorithm," in *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2020, pp. 1–6. DOI: 10.1109/iSAI-NLP51646.2020.9376787.

[19] M. T. I. Khan, M. S. I. Prottasha, T. A. Nasim, A. A. Mehedi, M. A. M. Pranto, and N. A. Asad, "Performance analysis of various machine learning classifiers on reduced chronic kidney disease dataset," *International Journal of Recent Research and Review*, vol. XIII, no. 4, pp. 16–22, 2020.

[20] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network," *Future Generation Computer Systems*, vol. 111, pp. 17–26, 2020, ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2020.04.036. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X20308128.

[21] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20 991–21 002, 2020. DOI: 10.1109/ACCESS.2019.2963053.

[22] K. Schwaber and J. Sutherland, *The Scrum Guide: The Definitive Guide to Scrum: The Rules of the Game*. Scrum.org & Scrum Inc., 2020. [Online]. Available: https://www.scrumguides.org/.

[23] R. Schwartz, H. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, pp. 54–63, 12 2020. DOI: 10.1145/3381831.

[24] R. Vinuesa, J. Debayle, et al., "The role of artificial intelligence in achieving the Sustainable Development Goals," *Nature Communications*, vol. 11, p. 233, 1 2020. DOI: 10.1038/s41467-019-14108-y.

[25] S. Akter et al., "Comprehensive performance assessment of deep learning models in early prediction and risk identification of chronic kidney disease," *IEEE Access*, vol. 9, pp. 165 184–165 206, 2021. DOI: 10.1109/ACCESS.2021.3129491.

[26] P. Chittora et al., "Prediction of chronic kidney disease - a machine learning perspective," *IEEE Access*, vol. 9, pp. 17 312–17 334, 2021. DOI: 10.1109/ACCESS.2021.3053763.

[27] S. M. M. Elkholy, A. Rezk, and A. A. E. F. Saleh, "Early prediction of chronic kidney disease using deep belief network," *IEEE Access*, vol. 9, pp. 135 542–135 549, 2021. DOI: 10.1109/ACCESS.2021.3114306.

[28] M. U. Emon, A. M. Imran, R. Islam, M. S. Keya, R. Zannat, and Ohidujjaman, "Performance analysis of chronic kidney disease through machine learning approaches," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 2021, pp. 713–719. DOI: 10.1109/ICICT50816.2021.9358491.

[29] K. L. Johansen et al., "US renal data system 2020 annual data report: Epidemiology of kidney disease in the united states," en, *Am. J. Kidney Dis.*, vol. 77, no. 4 Suppl 1, A7–A8, Apr. 2021.

[30] I. Ahmed, T. E. Chowdhury, B. B. Routh, N. Tasmiya, S. Sakib, and A. A. Chowdhury, "Performance analysis of machine learning algorithms in chronic kidney disease prediction," in *2022 IEEE International Conference on Emerging Technologies for Power and Network Modernization (ICETPN)*, 2022, pp. 1–6. DOI: 10.1109/IEMCON56893.2022.9946591.

[31] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105 151, 2022. DOI: 10.1016/j.engappai.2022.105151. arXiv: 2104.02395 [cs.LG].

[32] J. Gunning, C. Heeney, M. Huda, et al., "Synthetic data in healthcare: Ethics, privacy and regulatory challenges," *Journal of the Royal Society of Medicine*, vol. 115, pp. 222–228, 6 2022. DOI: 10.1177/01410768221087431.

[33] K. Jhumka, M. M. Auzine, M. S. Casseem, M. H.-M. Khan, and Z. Mungloo-Dilmohamud, "Chronic kidney disease prediction using deep neural network," in *2022 3rd International Conference on Next Generation Computing Applications (NextComp)*, 2022, pp. 1–5. DOI: 10.1109/NextComp55567.2022.9932200.

[34] W. Liu, X. Lin, et al., "Risk management for machine learning-enabled medical devices: a review," *Computational and Mathematical Methods in Medicine*, vol. 2022, p. 5 550 304, 2022. DOI: 10.1155/2022/5550304.

[35] C. Mondol et al., "Early prediction of chronic kidney disease: A comprehensive performance analysis of deep learning models," en, *Algorithms*, vol. 15, no. 9, p. 308, Aug. 2022.

[36] V. Singh, V. K. Asari, and R. Rajasekaran, "A deep neural network for early detection and prediction of chronic kidney disease," en, *Diagnostics (Basel)*, vol. 12, no. 1, p. 116, Jan. 2022.

[37] M. S. Arif, A. Mukheimer, and D. Asif, "Enhancing the early detection of chronic kidney disease: A robust machine learning model," en, *Big Data Cogn. Comput.*, vol. 7, no. 3, p. 144, Aug. 2023.

[38] A. A. Huang and S. Y. Huang, "Computation of the distribution of model accuracy statistics in machine learning: Comparison between analytically derived distributions and simulation-based methods," en, *Health Sci. Rep.*, vol. 6, no. 4, e1214, Apr. 2023.

[39] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," en, *J. Pathol. Inform.*, vol. 14, no. 100189, p. 100 189, Jan. 2023.

[40] I. Logunova, *A guide to F1 score*, https://serokell.io/blog/a-guide-to-f1-score, Accessed: 2025-12-10, Jul. 2023.

[41] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease," en, *Decision Analytics Journal*, vol. 6, no. 100169, p. 100 169, Mar. 2023.

[42] K. Venkatrao and S. Kareemulla, "Hdlnet: A hybrid deep learning network model with intelligent iot for detection and classification of chronic kidney disease," *IEEE Access*, vol. 11, pp. 99 638–99 652, 2023. DOI: 10.1109/ACCESS.2023.3312183.

[43] Z. Yu, K. Wang, Z. Wan, S. Xie, and Z. Lv, "Popular deep learning algorithms for disease prediction: A review," en, *Cluster Comput.*, vol. 26, no. 2, pp. 1231–1251, 2023.

[44] S. Dutta, R. Sikder, M. R. Islam, A. A. Mukaddim, M. A. Hider, and M. Nasiruddin, "Comparing the effectiveness of machine learning algorithms in early chronic kidney disease detection," *Journal of Computer Science and Technology Studies*, vol. 6, no. 4, pp. 77–91, Oct. 2024. DOI: 10.32996/jcsts.2024.6.4.11.

[45] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," en, *Front. Bioinform.*, vol. 4, p. 1 457 619, Sep. 2024.

[46] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," en, *Sci. Rep.*, vol. 14, no. 1, p. 6086, Mar. 2024.

[47] D. Saif, A. M. Sarhan, and N. M. Elshennawy, "Deep-kidney: An effective deep learning framework for chronic kidney disease prediction," en, *Health Inf. Sci. Syst.*, vol. 12, no. 1, p. 3, Dec. 2024.

[48] P. Anderson and T. Rahman, "Frameworks for trustworthy clinical AI deployment," en, *IEEE Reviews in Biomedical Engineering*, vol. 18, pp. 77–99, 2025.

[49] L. Chen and H. Zhang, "Non-parametric statistical methods for medical biomarker analysis in chronic disease detection," en, *Journal of Medical Informatics*, vol. 42, no. 3, pp. 156–167, 2025.

[50] R. Chen and L. Alvarez, "Robust ensemble learning strategies for medical diagnostics," en, in *Proceedings of the 2025 IEEE International Conference on Healthcare Informatics*, IEEE, 2025, pp. 233–242.

[51] J. Galvez and K. Morita, "Learning efficiency limits in medical deep learning systems," en, in *Proceedings of the 2025 IEEE Conference on Biomedical AI*, IEEE, 2025, pp. 301–310.

[52] S. Kumar, M. Rodriguez, and P. Thompson, "Statistical validation frameworks for clinical machine learning feature selection," en, *IEEE Transactions on Biomedical Engineering*, vol. 72, no. 4, pp. 1123–1134, 2025.

[53] Q. Li and M. Hossain, "Efficient AI models for real-time clinical decision support," en, *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 2, pp. 441–452, 2025.

[54] N. Patel and S. Wong, "Interpretable neural architectures for biomedical prediction," en, *IEEE Transactions on Biomedical Engineering*, vol. 72, no. 5, pp. 1210–1224, 2025.

[55] S. Rahman and Y. Cho, "Deployable machine learning pipelines for hospital environments," en, *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 1, pp. 55–70, 2025.

[56] W. Van Biesen et al., "Ethical considerations on the use of big data and artificial intelligence in kidney research from the ERA ethics committee," en, *Nephrol. Dial. Transplant*, vol. 40, no. 3, pp. 455–464, Feb. 2025.

[57] S. Wang and T. Yu, "Deep structured models for clinical risk stratification," en, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 1892–1905, 2025.

[58] X. Wang and R. Li, "Integrating hypothesis testing with ensemble feature selection for medical AI applications," en, *Computers in Biology and Medicine*, vol. 168, p. 107 812, 2025.

[59] H. Zhou and D. Kim, "Biomarker-guided feature prioritization for chronic disease prediction," en, *IEEE Access*, vol. 13, pp. 11 845–11 860, 2025.