

ناصر قنابطه

Nasser_CKD_Final.docx

- 1- Final Thesis (Word Format)
- Thesis_(2)_AI901_Dr_Wessam_Ahmed_Third Trimester 2025 (Moodle PP)
- Midocean University

Document Details

Submission ID

trn:oid::1:3450537231

Submission Date

Dec 24, 2025, 9:55 PM GMT+4

Download Date

Dec 25, 2025, 10:46 PM GMT+4

File Name

xcod.473670004_101621_laniF_DKC_ressaN_ناصر قنابطه_45203

File Size

1.2 MB

70 Pages

17,451 Words

102,488 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.





Optimized Hybrid Machine Learning Approach for CKD Diagnosis Using Genetic Algorithms

Prepared By

Name: Nasser Naim Qanabta

ID No.: 251003169

Submitted in fulfillment of the
requirements for the degree Master of
Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

UAE

Supervisor Name: Wesam Ahmed

Collage of Informatics, Midocean
University, Comoros

Email: wesamahmed@midocean.edu.km

Acknowledgements

I would like to thank my supervisor Dr.Wesam Ahmed for her inspiration, patience, and encouragement, your insight and mastery aided me at every point of this project.

Abstract

This thesis proposes an interpretable hybrid machine learning framework toward early and accurate classification of Chronic Kidney Disease (CKD) with intricate clinical dataset. First, we propose a unified pipeline with learning compact, highly discriminative latent feature representations (Z) by applying a CNN encoder. These features are introduced into multiple classifiers, which are optimized for hyperparameters using a Genetic Algorithm (GA) that guarantees fairness and optimal performance for base learners such as Random Forest, Decision Tree, and an Artificial Neural Network. Based on existing works, the primary novelties are the Hybrid (DT_RF) model which adopts a structural integration method of Decision Tree and Random Forest that is specifically engineered to enhance predictive accuracy as well as the traceability and interpretability of the obtained results. Experimental work on the CKD dataset supports the approach and indicates that CNN extracted latent features allow all GA optimised models to achieve near-perfect performance on every evaluation (e.g., $AUC \leq 0.9992$), with a consistently high Recall which is very strong and therefore has minimum clinical false negative burden. Moreover, the tailored Hybrid A model brought the best balance between high-level function with vital transparency. In order to address the relationship between abstract latent features and clinical importance, a SHAP framework and an adaptive Reverse SHAP Mapping approach were applied. This analysis effectively traced the model's best predictive ability back to the most basic clinical markers and, more significantly, Red Blood Cells and Pus Cells emerged as the pivotal factors in CKD classification. This work provides a robust, high-performance and modular solution that satisfies a gap in the existing state of the art for transparency in medical AI, but also provides a reliable theoretical base for adopting traceable Explainable AI techniques into future CKD decision support systems.

Contents

OPTIMIZED HYBRID MACHINE LEARNING APPROACH FOR CKD DIAGNOSIS USING GENETIC ALGORITHMS	1
ACKNOWLEDGEMENTS	2
ABSTRACT	3
LIST OF ABBREVIATIONS	6
LIST OF FIGURES	7
CHAPTER 1 INTRODUCTION	9
1.1 PROBLEM DEFINITION	10
1.2 THESIS OBJECTIVES	11
1.3 THESIS CONTRIBUTION	12
1.4 THESIS ORGANIZATION	13
CHAPTER 2 OVERVIEW	14
2.1 INTRODUCTION	14
2.2 THE ROLE OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE	15
2.3 MACHINE LEARNING TECHNIQUES	17
2.3.1 Support Vector Machine (SVM)	19
2.3.2 Random Forest (RF)	20
2.3.3 XGBoost (Extreme Gradient Boosting)	21
2.3.4 Artificial Neural Network (ANN)	21
2.3.5 k-Nearest Neighbors (k-NN)	22
2.3.6 Decision Tree (DT)	22
2.3.7 Genetic Algorithm for Hyperparameter Tuning	23
2.3.8 Model Selection Justification	24
2.4 DEEP LEARNING TECHNIQUES	24
2.4.1 Convolutional Neural Networks (CNN) for Feature Extraction	25
2.4.2 DL vs. ML Trade-offs in the Hybrid Approach	26
2.5 EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)	27
2.6 EVALUATION METRICS AND CONFUSION MATRIX	28
2.7 SUMMARY	29
CHAPTER 3: RELATED WORK	30
3.1 INTRODUCTION	30
3.2 MACHINE LEARNING STUDIES FOR CKD CLASSIFICATION	31
3.2.1 Summary Table of Machine Learning-Based Studies	32
3.3 DEEP LEARNING STUDIES FOR CKD CLASSIFICATION	34
3.3.1 Summary Table of Deep Learning -Based Studies	36
3.4 SUMMARY	37
CHAPTER 4: PROPOSED APPROACH	38
4.1 INTRODUCTION	38
4.2 DATASET DESCRIPTION	39
4.3 PREPROCESSING DATASET	41
4.4 FEATURE EXACTION PHASE	42
4.4.1 CNN Architecture and Training Details	42
4.4.2 Generating the Latent Embeddings (Z)	43
4.5 GENETIC ALGORITHM (GA) OPTIMIZATION FOR BASE CLASSIFIERS	44
4.5.1 GA Implementation and Operational Flow	44
4.5.2 Search Spaces and Optimized Models	45
4.6 MACHINE LEARNING MODELS FOR CLASSIFICATION	46
4.6.1 Traditional Machine Learning Classifiers	46
4.6.2 Advanced and Ensemble Classifiers	47
4.7 HYBRID MODEL	48
4.7.1 The Two-Stage Hybrid A Workflow	48
4.7.2 Rationale for Model Selection (Hybrid A vs. Performance Leaders)	49

4.8 EXPLAINABILITY AI: SHAP METHODOLOGY	49
4.8.1 <i>The Foundation of SHAP Values</i>	50
4.8.2 <i>Interpretation Tools and Visualization</i>	50
4.9 SUMMARY	51
CHAPTER 5.....	52
EXPERIMENTAL RESULTS AND DISCUSSION	52
5.1 INTRODUCTION	52
5.2 RESULTS OF GA-OPTIMIZED BASE CLASSIFIERS	52
5.3 RESULTS OF THE PROPOSED HYBRID MODELS	54
5.4 RESULTS OF EXPLAINABLE AI (SHAP ANALYSIS).....	56
5.4.1 <i>Global Feature Importance</i>	56
5.4.2 <i>Directional Feature Impact</i>	57
5.4.3 <i>Reverse SHAP Mapping and Clinical Insight</i>	58
5.5 DISCUSSION	60
<i>Validation of the Hybrid A Approach</i>	60
<i>Clinical Traceability through Reverse SHAP Mapping</i>	61
<i>Comparison with Other Hybrid Approaches</i>	62
5.6 LIMITATIONS OF THE PROPOSED WORK.....	62
5.7 SUMMARY	63
CHAPTER 6: CONCLUSION AND FUTURE WORK.....	64
6.1 IMPORTANT FEATURES.....	ERROR! BOOKMARK NOT DEFINED.
6.2 FUTURE WORK.....	64
<i>Enhancing Generalizability and Data Scope</i>	Error! Bookmark not defined.
<i>Methodological and Explainability Advancements</i>	Error! Bookmark not defined.
<i>Clinical Deployment</i>	Error! Bookmark not defined.
REFERENCES	66

List of Abbreviations

- **AI:** Artificial Intelligence
- **ANN:** Artificial Neural Networks
- **AUC:** Area Under the Curve
- **CKD:** Chronic Kidney Disease
- **DL:** Deep Learning
- **ML:** Machine Learning
- **RF:** Random Forest
- **ROC:** Receiver Operating Characteristic
- **SHAP:** SHapley Additive explanations
- **XAI:** Explainable Artificial Intelligence
- **RF:** Random Forest
- **SVM:** Support Vector Machine
- **XGBoost:** Extreme Gradient Boosting
- **TP:** True Positive
- **TN:** True Negative
- **FP:** False Positive
- **FN:** False Negative
- **ANN:** artificial neural networks
- **K-NN:** k-nearest neighbors
- **CNN:** Convolutional Neural Networks
- **DT:** Decision Tree
- **GA:** Genetic Algorithm

List of Figures

2.1 Worldwide age-standardized CKD prevalence per 100,000 populations	6
2.2 Core AI technologies between ML, DL and their role in CKD classification	8
2.3 Hierarchical taxonomy of Machine Learning.....	9
2.3.4 ANN Diagram	13
2.4.1 Basic convolutional neural network (CNN) architecture.....	17
2.3.6 DT Diagram.....	22
4.1 Proposed System Architecture for CKD Classification.....	39
4.2 Distribution of CKD and Non-CKD samples in the dataset	41
4.7.1 Top clinical features influencing CKD classification in the HybridA.....	49
4.7.2 SHAP beeswarm plot illustrating	51
5.3 Performance comparison of the proposed hybrid models	55
5.4 <i>Global Feature Importance Ranking for Hybrid (SHAP Bar Plot)</i>	56
5.5 <i>SHAP Beeswarm Plot Showing Directional Feature Impact for Hybrid A</i>	57
5.6 Top Clinical Features (Hybrid via CNN + RF + DT) Ranked by Reverse SHAP Impact	59

List of Tables

2.3	Comparison of ML Models for CKD Classification	15
2.4	DL Model Comparison	17
3.3.1	Summary Table of Machine Learning-Based Studies.....	23
3.4.1	Summary Table of Deep Learning-Based Studies.....	27
4.2	Description of the CKD Dataset Features	41
5.1	Performance of GA-Optimized Base Models on the Test Dataset	52
5.2	Performance of the Proposed Hybrid Models on the Test Dataset	54
5.3	Top Clinical Features Ranked by Reverse SHAP Impact	58

Chapter 1 Introduction

Chronic Kidney Disease (CKD) is a long-term condition in which the kidneys gradually lose their ability to perform essential functions removing waste, balancing electrolytes, regulating fluids, and producing hormones that control blood pressure and red blood cell production [1]. It affects more than 850 million people worldwide, a number that already surpasses the combined prevalence of diabetes and cardiovascular diseases. The impact is most severe in low- and middle-income countries, where healthcare resources are scarce and risk factors such as hypertension, diabetes, and obesity are on the rise.

CKD presents with a subtle pathophysiology, not least its insidious evolution. It typically starts when there are no manifest signs at the very beginning of the life-course (G1-G3), when the disease remains undiagnosed. Most of the patients are in advanced stages (G4-G5) or have ESRD or End-Stage Renal Disease, a stage that often necessitates dialysis or transplantation and can lead to cardiovascular complications.

As well as the medical cost, the human and the social costs are enormous family members and health care systems are left with elevated costs of hospital admission, expensive medical therapies, and diminished quality of life. This fact allows early detection to become not only a medical necessity, but, more importantly, a worldwide health one. Earlier diagnoses of CKD were based on a few biomarkers, based on serum creatinine and estimated glomerular filtration rate (eGFR). Sure, it's important but not good enough for detecting disease early. They neglect the intricate interaction of multiple clinical parameters such as hemoglobin, blood pressure, albumin levels and comorbidities such as diabetes which are population-specific [2].

This limits sensitivity and specificity, and a number of cases are too late in being detected. Differences between regions make matters even worse when it comes to the difficulties in the healthcare setting. With the use of electronic health records (EHRs) being expanded, a new computational technology has been developed to analyse hidden patterns in patient documents, making early and accurate diagnosis more likely.

Over the last few years, the advance of Artificial Intelligence (AI) specifically deep learning (DL), has shaken up in healthcare with remarkable impact. DL models such as CNNs and transformer-based architectures have achieved phenomenal success in classifying complex,

non-linear relationships in large datasets, which were particularly more successful than traditional ML (support vector machines, SVM, random forests, RF, and XGBoost) models when capturing the complex, non-linear relationships among all three categories [3].

Building on this, the present work adopts a hybrid approach: convolutional neural networks (CNNs) are first used to extract informative features from clinical data, which are then fed into six ML models SVM, RF, XGBoost, artificial neural networks (ANN), k-nearest neighbors (k-NN), and DT. Each model is fine-tuned using a genetic algorithm to optimize hyperparameters and maximize performance.

Then, the best performing model will be selected for potential clinical implementation. To avoid a ‘black box’ of this system, SHAP (SHapley Additive exPlanations) is applied to elucidate how each feature relates to the prediction [4]. For instance, the system may find that elevated creatinine and reduced hemoglobin are the strongest indicators in labeling a patient as high risk information so clinicians can interpret and trust the model’s recommendations.

Hence, an integrated AI pipeline synthesising the representational strength of CNN-based feature extraction with the predictive flexibility of multiple ML classifiers that have been optimized with evolutionary algorithms is presented in this thesis. The aim is clear: assist with early, reliable CKD detection from structured clinical data, contribute to timely medical intervention, and ultimately facilitate better patient outcomes, especially where resources are scarce.

1.1 Problem Definition

Chronic Kidney Disease (CKD) is a major public health concern affecting more than 850 million people worldwide, especially in low- and middle-income countries with limited access to healthcare facilities and increasing prevalence of risk factors such as hypertension and diabetes. The disease’s initial stages (G1–G3) tend to be asymptomatic, causing mass underdiagnosis. Such patients are typically only recognizable once they have progressed to G4–G5 or End-Stage Renal Disease (ESRD), and treatments such as dialysis or kidney transplantation become costly and less effective. Traditional diagnosis requires fixed cutpoints in biomarkers, such as serum creatinine and estimated glomerular filtration rate (eGFR), to attain the detection.

Although useful, clinical assessment techniques are unsuitable to capture the early signs of CKD and are unable to capture the intricate links between the myriad interplay of clinical indicators like hemoglobin, blood pressure, albumin, and comorbidities such as diabetes and hypertension. These interactions vary between populations and ethnicities, resulting in diminishing diagnostic sensitivity and generalization.

The increasing numbers of electronic health records (EHRs) made accessible represent an opportunity to develop a more accurate method for the early detection of CKD that meets the conditions at every threshold. However, both classic statistical and rule-based methods are insufficient to model subtle nonlinear trends in heterogeneous datasets like the UCI CKD database and its related laboratory results, vital signs, and demographic information. Furthermore, lack of transparency in these systems often acts as a barrier to clinical adoption since both the doctors and decision makers need data that is interpretable to gain the trust of the patient.

In this study, we fill this diagnostic void by building a binary classification model (CKD or non-CKD) based on publicly available clinical data. The strategy utilizes Deep Learning (DL) models, particularly Convolutional Neural Networks (CNN) and Transformer-based architectures, as well as SHapley Additive exPlanations (SHAP), for explainability. The aim is to make precise predictions that are clear and transparent, and to show that all features are informative (e.g., the contribution of serum creatinine to CKD risk), which enables to facilitate timely intervention and also assist in improvement of the patient's outcome in all kinds of healthcare settings.

1.2 Thesis Objectives

The purpose of this research is to propose a hybrid artificial intelligence framework to perform early CKD diagnosis based on structured clinical data. Indeed, the framework will:

- **Utilize a Convolutional Neural Network (CNN)** for automated feature extraction, transforming raw clinical variables into high-quality feature representations without manual feature engineering.

- **Integrate multiple machine learning classifiers** (e.g., Random Forest, XGBoost, Support Vector Machine) to perform CKD classification based on CNN-derived features.
- **Apply a Genetic Algorithm (GA)** for hyperparameter tuning of the ML classifiers to maximize predictive performance and generalization capability.
- **Evaluate and benchmark** the hybrid CNN–ML approach against standalone ML models to quantify performance improvements.

We have proposed the proposed framework by training the model within a publicly available clinical dataset (e.g., UCI CKD dataset) to demonstrate its reproducibility and applicability within practical healthcare environments.

1.3 Thesis Contribution

This study adds to the area of early Chronic Kidney Disease (CKD) detection by suggesting a predictive system based on clinical data that is understandable and well-built.

The main points of this research are:

First, this paper shows a combined learning system that mixes deep feature learning with common machine learning classification. A one-dimensional Convolutional Neural Network (CNN) works as a way to pull out features, turning basic clinical data into useful representations. Unlike full deep learning classifiers, this choice makes the model less hard to understand, but still able to see how different clinical features relate to each other.

Second, the study uses a Genetic Algorithm (GA) to adjust the settings of several machine learning classifiers like Decision Tree, Random Forest, Support Vector Machine, k-Nearest Neighbors, XGBoost, and Artificial Neural Networks. By using the same GA process for all models, the system makes sure that they are compared fairly, work in different situations, and are not changed by hand or personal choice.

Third, a combined hybrid model is suggested, bringing together Decision Tree and Random Forest classifiers. This setup uses the rule-based style of decision trees with the strength of random forests. The hybrid model that comes out of this does well at predicting while still being easy to understand, which is important for systems that help with clinical decisions.

Fourth, the suggested system has a complete testing and comparison process. All models and hybrid designs are checked using performance measures that matter in the clinic, like Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC-ROC). This testing gives a clear idea of how well each model works for CKD prediction.

Finally, the research adds Explainable Artificial Intelligence (XAI) using SHapley Additive exPlanations (SHAP) to make things clearer. SHAP analysis is used to find and measure how much each clinical data, like serum creatinine, blood urea, hemoglobin, urine albumin, and urine sediment indicators, adds to model predictions. This part helps close the gap between good prediction and clinical understanding, allowing the system to help doctors make informed decisions.

In general, this study gives a hybrid machine learning system that is useful in the clinic and easy to understand, balancing prediction, strength, and clarity. The method shows how using CNN-based feature extraction, evolutionary learning, and explainable AI can make early CKD detection better and help with early treatments in different healthcare places.

1.4 Thesis Organization

The remainder of the thesis is structured as follows:

- **Chapter 2** provides an overview of foundational concepts and techniques related to this research.
- **Chapter 3** reviews related work and existing literature on CKD classification.
- **Chapter 4** describes the proposed methodology and hybrid modeling framework.
- **Chapter 5** presents the experimental results, analysis, and discussion.
- **Chapter 6** concludes the thesis and outlines directions for future work.

Chapter 2 Overview

2.1 Introduction

Chronic Kidney Disease (CKD) is a progressive and irreversible disorder characterized by gradual decay of kidney function over the time span of months or years with impaired kidney function on the removal of foreign substances and electrolyte and acid–base homeostasis, control of blood pressure, and synthesis of hormones necessary for the production of red blood cells and bone health [5]. In the face of deteriorating renal functionality, a steady accumulation of impurities in the bloodstream, dysfunctional fluid regulation, and the subsequent onset of anemia, metabolic bone disorders, and a significantly increased risk for cardiovascular complications, which are the leading cause of mortality of CKD patients, could emerge. Clinically CKD is classified in five stages (G1–G5) based on eGFR. Despite this, with Stage G1 representing normal or high ($\text{GFR} > 90 \text{ mL/min/1.73 m}^2$), Stage G5 End-Stage Renal Disease (ESRD) signifies severe kidney failure ($\text{eGFR} < 15 \text{ mL/min/1.73 m}^2$) that typically necessitates dialysis or kidney transplantation for the majority of cases [6].

Worldwide, CKD affects more than 850 million individuals, more than those with diabetes and CVD combined. Its beginning stages (G1–G3) are typically quiet, with few overt signs and symptoms, and the majority of patients are without definitive diagnosis till advanced stages, when therapy is invasive, expensive, and less effective. Delayed diagnosis not only reduces patient outcomes but also increases the chance that life-threatening complications such as cardiovascular diseases will follow. In low-resource settings the economic burden of dialysis is considerable. Annual costs that may outstrip \$100,000 per patient. This results in a tremendous strain on health systems and the health care system itself and on those in their families [7]. Without significant improvements in prevention and early detection, CKD is expected to become the fifth leading cause of death globally by 2040.

The economic and social impacts of CKD go beyond health expenditures. The major costs of medical care in low- and middle-income countries hospitalization, dialysis, and transplants may drive family members into financial hardship. Patients experience a poorer quality of life, face fewer opportunities to participate in the workforce, and are more prone to social and health disparities especially in underprivileged populations with limited access to preventive care [8].

The unequal burden of CKD as a result of geographic variation emphasizes the stark health disparity highlighted by geographic differences in CKD prevalence. As shown in Figure 2.1, sub-Saharan Africa, Central America, and South Asia have some of the highest age-standardized prevalence rates per 100,000 population. The burden of high-risk factors and compromised healthcare systems is also dual in these regions while other high-income areas in the world e.g North America and in Western Europe have better screening programs, early intervention strategies, and disease management [9]. These disparities highlight the absolute necessity for scalable, AI-assisted diagnostic tools capable of utilizing well-structured clinical datasets (e.g. UCI CKD dataset), which can assist in early diagnosis, leading to improved prognoses and decreasing the global burden of CKD.

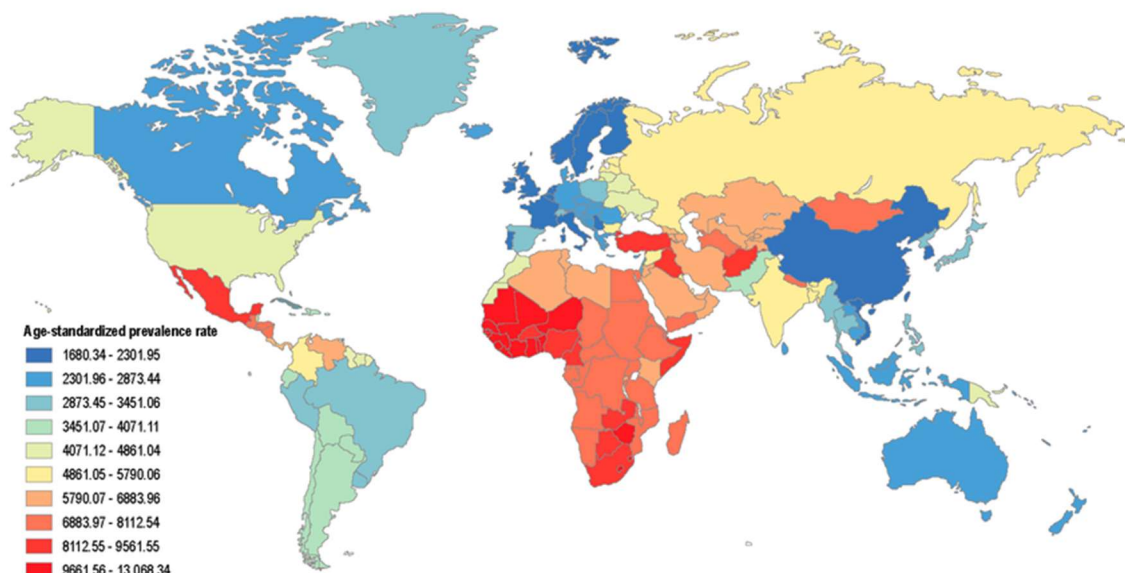


Figure 2.1: Worldwide age-standardized CKD prevalence per 100,000 populations[10].

2.2 The Role of Artificial Intelligence in Healthcare

Artificial Intelligence (AI) is the name given to systems that can compute complex and difficult things as well as complex and efficient processes. It involves many artificial intelligence techniques to reproduce cognitive functions, and with them to process large amounts of data on a large scale and identify patterns or make informed decisions with

minimal human input. As healthcare digitalization from electronic health records (EHRs), medical imaging, lab results, and genomic data continue to grow [12], there is an unparalleled opportunity for AI to further improve diagnostic accuracy, treatment planning, and patient outcomes.

Traditional statistical algorithms often have problems treating heterogeneous and high-dimensional data (whereby, AI is able to deal with complex interactions from various data sources). In healthcare, it facilitates the inclusion of multimodal data like lab test results, imaging scans, and patient demographics into integrated predictive models. One of the major branches of AI known as Machine Learning (ML) trains algorithms that leverage past information to recognize patterns and predict results without being specifically designed to do so [13]. Deep Learning (DL), a higher-order form of ML, employs multi-layered neural networks to model very complex non-linear relationships, which makes it ideal for applications such as medical image analysis, natural language processing, and disease risk prediction [14].

AI technology applications have demonstrated great promise when applying across medical areas from early disease diagnosis and treatment recommendation to support in clinical decision making and patient monitoring [15]. With Chronic Kidney Disease (CKD), AI can analyze structured clinical variables such as blood pressure, hemoglobin, serum creatinine, and comorbidities to identify patients at risk before the onset of the symptoms. AI-based models can also demonstrate better results than classic diagnostic methods due to their ability to detect subtle cues that human observability would overlook and provide earlier interventions, thus leading to decreased health care costs [16].

The concept of Explainable Artificial Intelligence (XAI) is critical in modern healthcare-based AI; it aims at bringing model predictions to interpretable levels for clinicians. That is particularly important in healthcare settings because trust and accountability matter a lot for doctors and nurses alike. In CKD classification, for example, XAI approaches like SHapley Additive exPlanations (SHAP) are able to highlight which characteristics—like elevated creatinine levels or low hemoglobin are most influential on a prediction.

To show relationships between the technologies, a hierarchical representation with the evolution of these solutions from AI to ML and DL with Convolutional Neural Networks (CNNs) and Transformer-based architectures as the most commonly used DL techniques are

shown in Figure 2.2. Throughout, it presents XAI as a cross-cutting layer that permeates through the spectrum, highlighting how it enhances the interpretability and trust for AI-based healthcare applications.

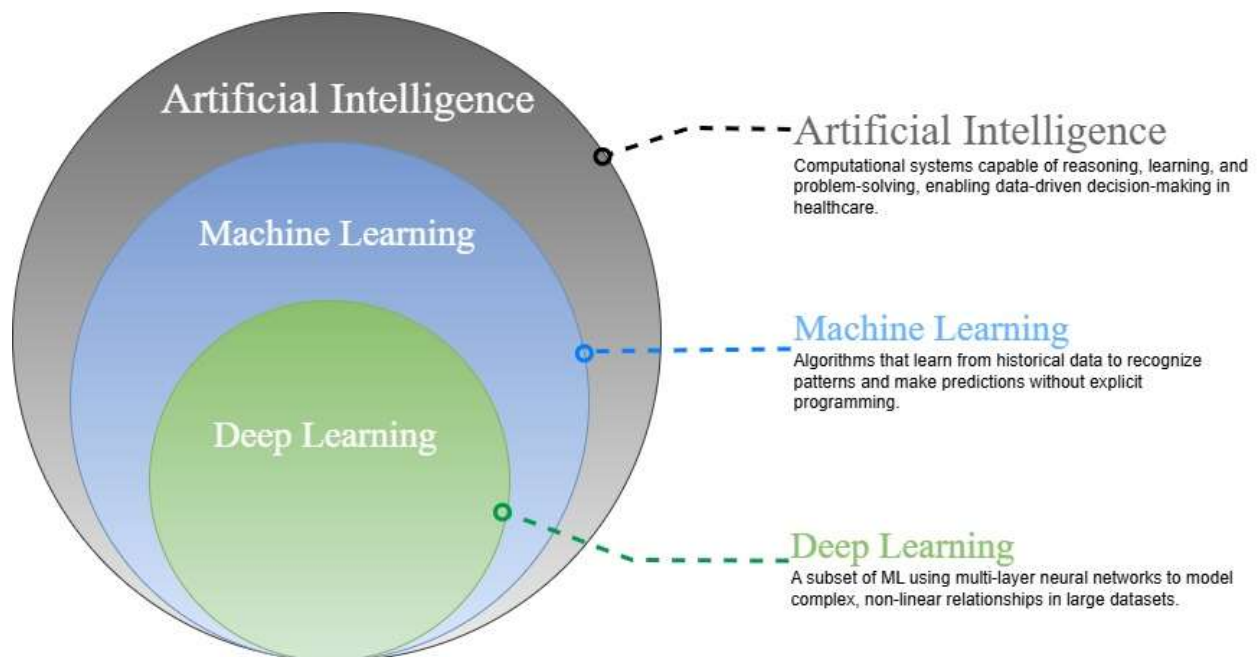


Figure 2.2: Core AI technologies between ML and DL role in classification.

2.3 Machine Learning Techniques

Machine Learning (ML) is a subfield of artificial intelligence that enables computer systems to automatically learn patterns from data and make predictions without explicit programming for each task [17]. In this research workflow, structured clinical data undergoes **feature extraction using a Convolutional Neural Network (CNN)**. The CNN transforms raw input into compact, informative feature representations that enhance downstream classification performance. These extracted features are then used to train **six different ML classifiers**:

- Support Vector Machine (SVM)
- Random Forest (RF)
- XGBoost
- Artificial Neural Network (ANN)
- k-Nearest Neighbors (k-NN)
- Decision Tree (DT)

Each classifier is **fine-tuned using a Genetic Algorithm (GA)** to identify optimal hyperparameters. After tuning, the six models are evaluated and compared, with the highest-performing model selected as the final CKD classifier. The hierarchical taxonomy in Figure 2.3 illustrates the placement of these six classifiers within the broader machine learning landscape, where all belong to the supervised learning category focused on classification tasks

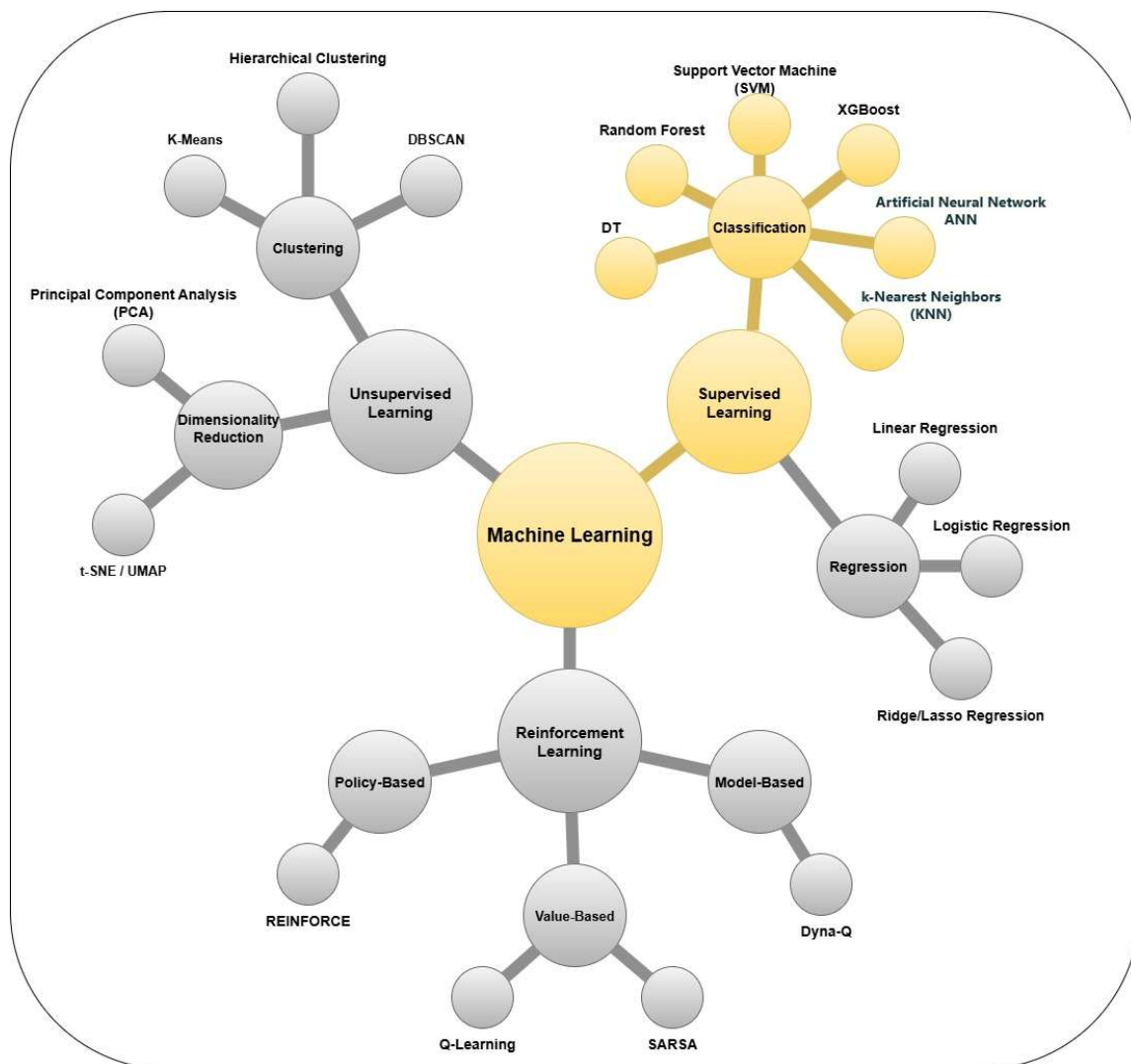


Figure 2.3: Hierarchical taxonomy of Machine Learning

At the highest level, ML can be divided into three main types:

- **Supervised Learning:** In this approach, models are trained on labeled datasets where the correct output is known. Supervised learning itself consists mainly of:

- **Classification:** Assigning inputs to discrete categories or classes, such as differentiating between types of objects, diseases, or customer profiles. Widely used algorithms include SVM, RF, XGBoost, ANN, k-NN, and DT.
- **Regression:** Predicting continuous numerical values, such as prices, temperatures, or measurement results. Common methods include Linear Regression, Logistic Regression, and Ridge/Lasso Regression.
- **Unsupervised Learning:** Models in this category learn patterns or structures from data that lack explicit labels. Typical tasks include:
 - **Clustering:** Grouping similar data points together based on inherent structure in the data. Examples include K-Means, DBSCAN, and Hierarchical Clustering.
 - **Dimensionality Reduction:** Simplifying datasets by reducing the number of features while preserving essential relationships, using techniques such as Principal Component Analysis (PCA) and t-SNE/UMAP.

The following sections introduce key Machine Learning models that are common in Chronic Kidney Disease classification. Each model has its own section to make things clearer.

2.3.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a useful model for binary classification, which works well with complex clinical info. SVM creates a decision boundary (hyperplane) that separates two classes as much as possible. If the data can't be separated with a straight line, SVM uses kernel functions like the Radial Basis Function (RBF) to change the info into a higher dimension, according to [18].

The decision function for SVM can be written as:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x_i, x) \quad (1)$$

Where:

- β_0 = Bias
- S = Set of observations
- α = Model parameters that have to be learned
- K: Kernel (e.g., RBF)

The margin (M) is the distance between the hyperplane and the nearest data point from any class. SVMs are trained to make this margin as large as possible, which helps it apply what it learns to new data. The hinge loss function is often used when training SVMs.

$$M = \frac{2}{||\omega||} \quad (2)$$

2.3.2 Random Forest (RF)

Random Forest (RF) is a method that makes many decision trees when learning from data. Each tree is built using a random sample taken from the original data. RF combines the predictions from all the trees by choosing the most common answer for classification or calculating the average for regression. This leads to better prediction and less overfitting than using just one tree [19].

The prediction of Random Forests works as follows:

$$f' = \frac{1}{B} \sum_{b=1}^B f_b(x) \quad (3)$$

Where:

- B: Total number of trees
- $f_b(x)$: Output of the b-th decision tree

RF improves how different and stable the model is by picking random groups of features for each split [25].

2.3.3 XGBoost (Extreme Gradient Boosting)

XGBoost is a gradient boosting method for decision trees, it constructs trees in order. Each new tree attempts to compensate for the mistakes of the previous trees. While Random Forest employs averaging across many trees, XGBoost uses gradient descent, using a second-order Taylor expansion. This means updating the system, both the gradient and Hessian of its loss function is used [20].

The objective function in XGBoost is:

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (4)$$

Where:

- l : Loss function (e.g., logistic)
- $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2$: Regularization to reduce overfitting

This method helps XGBoost model complex nonlinear links in structured data, making it useful for things such as CKD detection.

2.3.4 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) have artificial “neurons. All of them receive input signals, work with them, and send output to the next layer. This allows ANNs to learn patterns from the data. Their algorithms include ones like backpropagation, which alters the internal weighting of the model to lower prediction errors. These models may demonstrate non-linear relationships in clinical data [21].

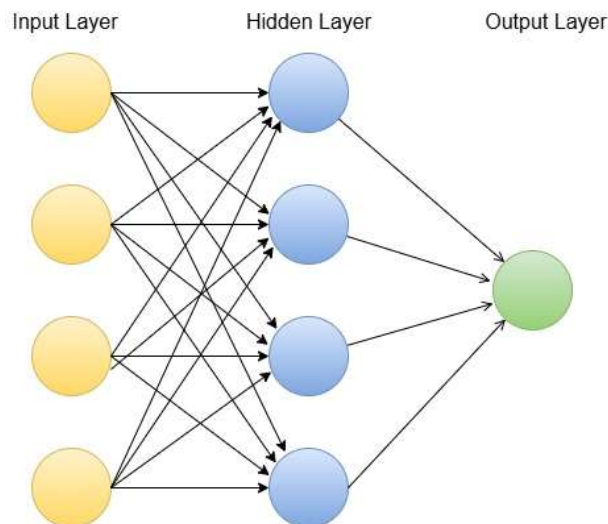


Figure 2.3.4: ANN Diagram

Figure 2.3.4 shows an ANN which has three layers (input layer: patient features, hidden layer: neurons that are connected, output layer: CKD/non-CKD predictions).

2.3.5 k-Nearest Neighbors (k-NN)

The k-Nearest Neighbors (k-NN) algorithm is a very simple classification method used for classification. Different from that approach, which creates a specific model, k-NN saves all training data and predicts the new data according to the most common class among the k closest examples. Distance calculations such as Euclidean distance are commonly used to consider similarity. To predict CKD, when a patient is new, k-NN finds the patients that are closest to each other from the old data and predicts the diagnosis based on their results.

2.3.6 Decision Tree (DT)

Decision Trees (DTs) are classification and regression supervised learning models. Decisions made in a DT are shown as a tree, with a root, decision nodes, and leaves. A rule based on a feature is defined at each decision node, which is followed by a class predicted at each leaf. We see that this allows decision trees to model intricate associations simply [23]. The impurity measures, i.e., Gini Index, Information Gain and others, are used in building a decision tree. These facilitate picking top-quality feature splits at each node. The model repeatedly breaks up the feature space again and again to generate clear decision rules linking the contribution of an input feature to prediction. The greatest advantage of Decision Trees for medical support is that they are easy to comprehend. Doctors can see and understand the

rules which establishes the trust in the model's predictions. Nevertheless, a single decision tree can overfit complex data if not well restricted.

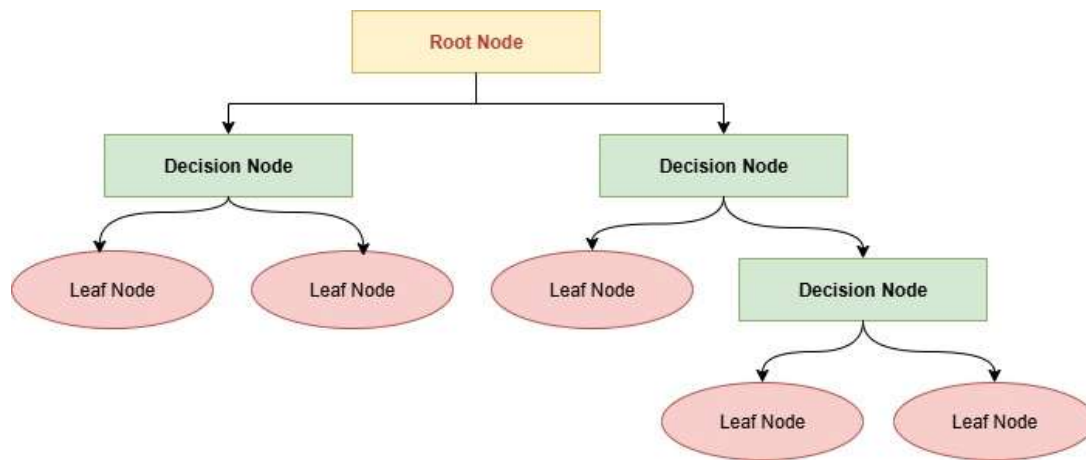


Figure 2.3.6: DT Diagram

Figure 2.3.6 illustrates a decision tree's structure by observing that the root, decision nodes, and leaves are closely associated.

2.3.7 Genetic Algorithm for Hyperparameter Tuning

Genetic algorithms (GAs) draw inspiration from natural selection and use a population-based approach to optimization [24]. In this study, GAs are employed to explore the hyperparameter space for each classifier, following these steps:

1. **Initialization:** Hyperparameter sets are created randomly.
2. **Evaluation:** The performance of each set is scored based on fitness.
3. **Selection:** The top-performing candidates are selected.
4. **Crossover & Mutation:** Parameters are recombined and changed to produce new candidates.
5. **Termination:** The process repeats until improvements stall.

The use of GAs allows efficient exploration within complex parameter spaces. For model tuning, it does better than grid search and random search methods.

2.3.8 Model Selection Justification

SVM was chosen because it works well with many variables, and its margins and kernels are easy to understand.

RF is also good because it can handle electronic health record data that is incomplete or has errors, which is common in actual clinical settings.

XGBoost was selected for its outstanding accuracy and computational speed on tabular data, and for its scalability with large datasets.

ANN Included for its capability to model complex nonlinear clinical patterns.

k-NN Simple yet effective for similarity-based classification in small feature spaces.

LightGBM Selected for its speed and efficiency with large-scale clinical data.

Table 2.3: Comparison of ML Models for CKD Classification

Model	Strengths	Weaknesses	Clinical Relevance
SVM	Clear margin interpretation, kernel tricks	Sensitive to hyperparameters	Interpretable decision rules
RF	Handles missing data, feature importance	Slower inference	Identifies key biomarkers
XGBoost	High accuracy, handles imbalanced data	Requires tuning	Scalable for large datasets
ANN	Captures nonlinear relationships	Requires large datasets, risk of overfitting	Models complex feature interactions
k-NN	Simple, interpretable	Sensitive to irrelevant features, slow with large datasets	Useful for small datasets
DT	Simple structure, transparent decision paths	Prone to overfitting, sensitive to noise	Provides explicit rule-based decisions suitable for clinicians

2.4 Deep Learning Techniques

Deep Learning (DL) is a family of machine learning algorithms that utilize multi-layered artificial neural networks to automatically capture hierarchical feature representations present in the input data. In contrast with the traditional ML method that focuses largely on manual feature generation, DL systems can learn complex, non-linear, nuanced features, by transforming raw inputs into high-level abstract features in the model through multiple layers

of computation. For example, in healthcare machine learning technology such as DL has achieved impressive performance in image analysis, signal processing, and structured data modeling due to its suitability for dealing with high-dimensional and diverse data sets.

In the study, DL is not presented as the end product, but rather as a feature extraction phase before the machine learning classification. More precisely, CNN is used to generate rich, compact, and discriminative features in the raw clinical data [25]. These learned features are subsequently passed to 6 ML models (SVM, RF, XGBoost, ANN, k-NN, and DT) where the hyperparameters are tuned by GA algorithm respectively. This hybrid technique features the representational strength of DL with the interpretability, flexibility, and generally better data efficiency of ML classifiers.

2.4.1 Convolutional Neural Networks (CNN) for Feature Extraction

Convolutional Neural Networks are a specialized type of deep neural network originally developed for image recognition, but which have been adapted successfully for structured and tabular data. CNNs automatically learn spatial or feature correlations by applying convolutional filters to the input, enabling them to detect patterns without manual feature engineering.

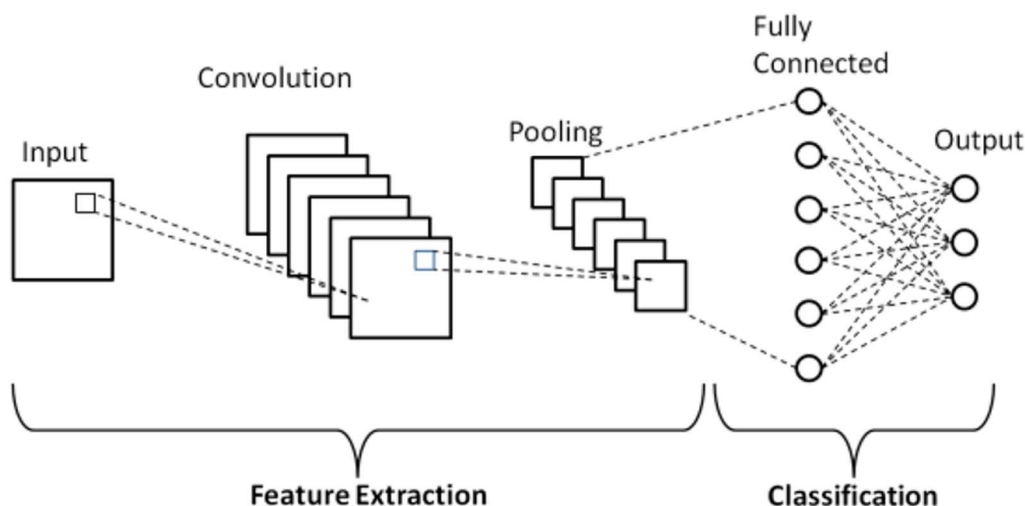


Figure 2.4.1: Basic convolutional neural network (CNN) architecture

Figure 2.4.1 shows the architecture of a Convolutional Neural Network (CNN) as implemented in this research for feature extraction from structured clinical data [26]. The architecture consists of the following components:

- **Input Layer:** Receives the preprocessed clinical dataset, where each variable (e.g., laboratory measurement, demographic attribute) is treated as an input feature.
- **Convolutional Layers:** Apply learnable filters to detect local patterns and correlations between clinical variables. In image-based applications, these filters capture edges and shapes, whereas in structured/tabular data, they identify co-occurring clinical markers and feature combinations.
- **Pooling Layers:** Downsample the feature maps generated by the convolutional layers, reducing dimensionality and computational complexity while retaining the most informative patterns.
- **Fully Connected Layers:** Take the flattened feature vector from the pooling stage and map it to higher-level abstractions, enabling the model to form a richer representation of the data.
- **Output Layer:** Produces the final feature representation in this study, which is then passed to six different machine learning classifiers—Support Vector Machine (SVM), Random Forest (RF), XGBoost, Artificial Neural Network (ANN), k-Nearest Neighbors (k-NN), and DT each optimized via Genetic Algorithm (GA) for CKD prediction.

This structure lets the CNN change raw clinical data into useful features that boost classifier performance without needing manual feature creation, which fits with current trends in clinical prediction models.

2.4.2 DL vs. ML Trade-offs in the Hybrid Approach

Deep networks such as CNNs can find expressive features, but they usually need a lot of computing power and big datasets to avoid overfitting. Regular ML models, especially those using trees or kernels, can do well with smaller datasets and are easier to understand.

By using CNNs only to pull out features, and then using ML classifiers to make predictions, this study gains these advantages:

- CNN's ability to automatically learn and compress complex feature relationships.
- ML models' strengths in handling structured data efficiently and producing explainable decisions (especially when combined with XAI methods like SHAP).
- Optimization via GA, ensuring each classifier is tailored to make the best use of CNN-extracted features.

Table 2.4.2: DL Model Comparison in This Study

Model	Advantages	Limitations	When to Use
CNN (as feature extractor)	Learns complex feature interactions automatically; reduces need for manual feature engineering	Requires careful preprocessing of tabular data; computationally heavier than basic ML	Pre-classification stage to produce high-quality features from structured data

2.5 Explainable Artificial Intelligence (XAI)

Emphasizes making the decision-making process of ML models transparent and interpretable to humans. In this work, SHAP (SHapley Additive exPlanations) is used for analyzing the final Chronic Kidney Disease (CKD) classification model derived by first extracting the features by means of a Convolutional Neural Network (CNN) and then selecting the best-performing machine learning classifier through Genetic Algorithm (GA) optimization [28].

Role of SHAP in This Study:

- **Model-agnostic:** While it can be applied for other types of predictive models, in our workflow, SHAP is specifically used to explain the predictions of the selected ML classifier after CNN-based feature extraction.

- **Local Interpretability:** SHAP picks the clinical characteristics (serum creatinine, hemoglobin, blood pressure, etc.) of each patient that contributed the most to the CKD prediction.

- **Global Interpretability:** By aggregating SHAP values across all patients, SHAP helps identify which features are most consistently important and helps clinicians interpret risk factors.

We integrate SHAP into the last stage of our CNN–ML pipeline so that predictive performance is supplemented by interpretability, which allows healthcare professionals to validate and trust the model’s recommendations [29].

2.6 Evaluation Metrics and Confusion Matrix

To measure how well the ML/DL models categorize CKD, we look at these measures from the confusion matrix:

- **True Positive (TP):** Patient has CKD and the model correctly predicts CKD.
- **True Negative (TN):** Patient does not have CKD and the model correctly predicts non-CKD.
- **False Positive (FP):** Patient does not have CKD but the model predicts CKD.
- **False Negative (FN):** Patient has CKD but the model predicts non-CKD.

A confusion matrix lets us calculate different measurements that give a fuller idea of how the model is doing [30]. Some key ones are below:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Evaluation Metrics:

- **Accuracy:** Ratio of how right the model was overall.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

- **Precision:** This is the proportion of the positive predictions made that were actually correct, so this is an important metric when we want to decrease false positives like how many predicted cases of CKD actually had CKD.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

- **Recall (Sensitivity):** How useful it is for a model to cover all relevant positive cases. It is very important in cases where the costs of missing and not detecting a positive case are high, for example in disease diagnosis.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

- **F1-Score:** A harmonic mean between precision and recall, striking a balance between both metrics.

$$\mathbf{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

- **Specificity:** How well a model identifies negative cases. In medical diagnosis, where we really must prevent false positives, it's especially relevant. Like correctly identifying non-CKD cases.

$$\mathbf{Specificity} = \frac{TN}{TN + FP} \quad (9)$$

- **AUC-ROC:** Measures the ability of the model to discriminate between classes across all probability thresholds.

The metrics give a comprehensive view of the model performance, which is especially vital in dealing with class imbalance as it happens in medical datasets [31].

2.7 Summary

A comprehensive background overview on Chronic Kidney Disease (CKD) and its clinical relevance was given in this chapter, followed by detailed analysis of machine learning & deep learning approaches employed within this study. The hybrid strategy used in this study was specified regarding automatic feature extraction by means of a Convolutional Neural Network (CNN) system for structured clinical data and the subsequent classification by six optimized machine learning models (SVM, RF, XGBoost, ANN, k-NN, and DT), tuned through a Genetic Algorithm (GA). The chapter presented the Explainable AI framework SHAP, which helps model predictions be more transparent and interpretable in clinical settings. Furthermore, the evaluation metrics and confusion matrix also explained how performance measurement and interpretation of the model are handled when using the model in healthcare usage. In the subsequent chapter, we provide the full procedures for data preprocessing, feature extraction using CNN, model training, hyperparameter optimization, and validation.

Chapter 3: Related Work

In this chapter, we aim to synthesize previous work in CKD classification with artificial intelligence by reviewing that work done in CKD classification using technology. The two main sections of the review organize the review as:

1. Machine Learning-Based Approaches, including classical algorithms applied to structured clinical datasets such as laboratory results and patient demographics.
2. Deep Learning-Based Approaches, which leverage automated feature extraction methods, including Convolutional Neural Networks (CNNs), for improved pattern recognition in clinical data.

A contrast of merits, demerits, and methodological norms can then be made in the two groups. However, the chapter also highlights the limitations like the limited dataset diversity, class imbalance, and the underutilization of hybrid frameworks, which the proposed research would address by implementing CNN-based feature extraction along with machine learning classifiers optimized by tuning Genetic Algorithm (GA).

3.1 Introduction

During the past two decades, CKD classification research went from rule-based classification methods to machine learning-driven models, using systems such as Support Vector Machines (SVM), decision trees, and so on using data from hospitals. From 2020 onwards, wearable and home-monitoring devices made it possible to connect together consumption and lifestyle information to support non-invasive early detection. In recent years, deep learning schemes have attracted considerable notice concerning the automatic feature representation learning capacity, while their larger dataset and substantial computational demand still present a challenge. Through consolidation of these works, this review will present the advancement of methods in this context, limitations, and background to the hybrid CNN–ML implementation in this study. The following sections present an extensive survey of recent ML- and DL-based CKD classification studies, and a comparison of their performance and gaps in applicability.

3.2 Machine Learning Studies for CKD Classification

Phuong et al. [32] This work focused on six different ML algorithms applied to UCI CKD dataset, which evaluated SVM, RF, KNN, LR, NB and DT. RF obtained highest precision, followed closely by SVM. Creatinine and hemoglobin were the key CKD predictors identified. But the small scale of the dataset also raised concerns for generalization.

Elshewey et al. [33] The authors proposed Extra Trees incorporated with Biologically-Based Feature Selection (BBFS) to detect CKD classification issues. This method achieved accuracy (over 99%) and provided interpretable feature importance rankings. However, the results pointed out the importance of feature selection but, due to small number of samples, risked overfitting.

Liu et al. [34] Random Forest was applied to predict CKD risk using more than 10,000 clinical EHR records. High AUC values (Excellent handling of imbalanced data). The results highlighted creatinine, eGFR, and age as primary predictors but could not be compared to other ML methods.

Metherall et al. [35] In this work, we investigated at-home creatinine monitoring data to train XGBoost, RF, and SVM models. XGBoost scored highest with accuracy of 94% and was applied for remote CKD screening. However, the dataset only included a brief follow-up period and thus there was little insight into long-term progression of disease.

Ganie et al. [36] AdaBoost and Gradient Boosting were used to identify CKD based on clinical data using Boosting algorithms. Both models reached above 96% accuracy and beat traditional single-learner performances. Although the results were encouraging, the sensitivity of the models to outliers was one of the limitations.

Debal et al. [37] RF, SVM, LR were compared for CKD classification with the UCI dataset. RF showed the highest accuracy rate with 98% indicating its stability in medical prediction tasks. The size and limited advanced preparation of the dataset limited the work.

Islam et al. [38] The authors trained RF, DT and KNN models on CKD data by hospitalized patients, and showed that RF achieved the best accuracy at 93%. Clinical relevance was helped by real patient records. However, single-source data reduced the model's generalizability

Khan et al. [39] used LightGBM, CatBoost, and RF on data from several medical centers to predict CKD. LightGBM had the best accuracy at 95%, due to gradient boosting's skill with complicated feature combinations. The research pointed out that this came at the cost of more computing power.

Ghosh et al. [40] Evaluated multiple ML algorithms on the UCI CKD dataset, achieving top performance with Random Forest (99.2% accuracy, AUC 0.995). The study used SHAP and LIME to interpret key features, finding serum creatinine and hemoglobin as the most influential. While results were strong, reliance on a single public dataset limited real-world generalization.

Chowdhury et al. [41] Compared classical ML algorithms for CKD detection in T1DM patients, finding Random Forest most accurate (95.6%). The work offered a baseline for diabetic CKD screening and emphasized model comparison. However, its scope was limited to a specific population without external validation.

3.2.1 Summary Table of Machine Learning-Based Studies

Ref	Year	Dataset	Models	Results	Advantages	Limitations
Phuong et al. [32]	2025	UCI CKD dataset (400 samples, 24 features)	SVM, RF, KNN, LR, NB, DT	RF highest accuracy (~99%), SVM second (~98%), others 90–96%	- Comprehensive comparison of six popular ML algorithms on the same dataset. - Identified creatinine, hemoglobin, and specific gravity as top predictors. - Straightforward methodology for replication in similar datasets.	- Small dataset limits generalization. - No external validation performed. - Only binary classification; no CKD stage prediction
Elshewey et al. [33]	2025	UCI CKD dataset	Extra Trees + BBFS	Accuracy 99.2%, F1-score 0.992	- Hybrid feature selection improved model interpretability. - Outperformed baseline Extra Trees without BBFS. - Identified clinically relevant features while reducing redundancy.	- Risk of overfitting due to limited sample size. - Tested on a single dataset without external validation. - Computational cost of BBFS may be high for large datasets.
Liu et al.	2024	>10,000 EHR	Random	Accuracy	-world dataset	- No comparison

[34]		patient records	Forest	~92%	improved robustness. - RF handled imbalanced data effectively. - Identified top CKD risk factors with high interpretability.	with other ML models. - Missing data imputation approach not deeply discussed. - Performance on rare CKD cases less certain.
Metherall et al. [35]	2025	At-home monitoring dataset (short-term)	XGBoost, RF, SVM	XGBoost: ~96% accuracy; RF ~94%; SVM ~92%	- Incorporated at-home measurements (creatinine, BP) for early detection. - Compared multiple algorithms on identical features. - Useful in remote patient monitoring contexts.	- Short monitoring period limits ability to predict long-term progression. - Potential sensor inaccuracies in home data. - Limited demographic diversity in dataset.
Ganie et al. [36]	2023	Clinical dataset	AdaBoost, Gradient Boosting	AdaBoost ~94%, Gradient Boosting ~95%	- Boosting improved handling of complex feature interactions. - Models robust to small feature noise. - Provided feature importance rankings for clinical insight.	- Sensitive to outliers and noisy data. - No testing on external datasets. - High computational requirements for large-scale application.
Debal et al. [37]	2022	UCI CKD dataset	RF, SVM, LR	RF ~98%, SVM: 96%, LR: 92%	- Straightforward preprocessing steps for reproducibility. - RF performed consistently across folds. - Demonstrated stability with minimal parameter tuning.	- Limited dataset size and diversity. - Absence of advanced feature engineering. - No inclusion of temporal or longitudinal data.
Islam et al. [38]	2023	Hospital CKD records	RF, DT, KNN	RF: 97%, DT: 93%, KNN: 92%	- Easy-to-implement ML methods.- RF robust against overfitting in small datasets. - Highlighted top predictors for CKD risk.	- Data from a single hospital limits external applicability. - No cross-validation beyond hold-out split.

						- Lacked integration of lab trends over time.
Khan et al. [39]	2024	Multi-center patient data	LightGBM, CatBoost, RF	LightGBM: 98.5%, CatBoost: 97.9%, RF: 97%	- Large, diverse dataset improved model generalization. - Gradient boosting models handled categorical and numerical data well. - Fast training times compared to deep learning.	- Higher training costs than simpler ML models. - Potential overfitting on rare feature combinations. - Interpretability requires additional tools.
Ghosh et al. [40]	2024	Clinical dataset	XGBoost, RF, SVM	XGBoost: AUC ~96%, RF: AUC ~95%, SVM: AUC ~93%	- Focused on explainable AI, providing SHAP-based insights. - High predictive performance with interpretable results. - Useful for clinician trust and adoption.	- Requires domain knowledge for interpreting SHAP results. - Performance may drop if features are highly correlated. - No multi-class classification tested.
Chowdhury et al. [41]	2021	Type 1 diabetes CKD dataset	RF, SVM, LR, KNN, NB	RF: ~95%, SVM: ~94%, LR: ~92%, KNN: ~91%, NB: ~90%	- Compared performance of multiple classic ML models. - Dataset focused on diabetic CKD detection. - Identified early markers in diabetic patients.	- Narrow patient population limits generalization. - No advanced ensemble or hybrid methods tested. - Limited to static data without time-series analysis.

3.3 Deep Learning Studies for CKD Classification

Al-Rasheed et al., [42] This study trained a dense fully-connected deep neural network to classify CKD using the UCI CKD tabular datasets (400 samples, 24 features). The optimized 3-layer network (20–15–10 neurons) with ReLU and Adamax achieved nearly perfect performance. It outperformed conventional classifiers on the same data.

Chowdhury et al. [43] proposed a TabNet model for CKD stage prediction in diabetes patients using 10-year CRIC cohort data. Two variants (31-feature "complete" and 15-feature "simplified") outperformed XGBoost/RF (94.1% accuracy, AUC $\approx$ 0.95), with SHAP/LIME identifying creatinine, HbA1c, age, and sex as key predictors. While robust, performance dipped ($\sim$ 88%) in external validation, suggesting limited generalization beyond diabetic populations.

Rao et al. [44] developed a graph-tabular fusion model (GNN + MLP) for CKD prediction, with patients acting as nodes in a similarity-based graph. Using inter-patient relationships this ensemble obtained 95.1% accuracy over the UCI dataset as compared with just individual MLP and GNN models (91% and 93% respectively). Despite its strength, weaknesses are such as reliance on manual construction of graphs and lack of validation on larger EHR datasets.

Alsekait et al. [45] introduced OptEnsemble, a stacked deep learning model (CNN + LSTM + SVM meta-learner) using mutual-information-based feature selection, achieving 99.7% accuracy on the UCI CKD datasets. Results indicate that although performance is above single models (DNN: 95%) and robust, only near-perfect results indicate overfitting and that external validation was not performed to determine generalization. The study describes trade-offs between complexity and small data performance.

Tian et al. [46] created a deep radiometric model that aggregated CNN and texture features derived from kidney ultrasounds that outperformed radiologists in early detection of CKD (AUC 0.918 compared with 0.869) particularly in stages G1-G3 of CKD with low visual changes. Although matching expert performance for severe CKD (G4-G5), the model has limited clinical utility due to its black box nature and the need for multi-center validation. This cutting-edge research shows AI's ability to pick up on diagnostic features from everyday ultrasounds that may go unrecognized by human visual assessment.

Priyadharshini et al. [47] have implemented "OptiNet-CKD", a population-based algorithm (POA) optimizing deep neural network that has obtained 100% on the UCI CKD datasets by evading local minima via evolutionary weight perturbations. Although the model shows true accuracy metrics (precision, recall, F1, AUC), their unblemished performance on these datasets have led to concerns about overfitting, despite generalized claims. Advanced optimization techniques can optimize with limited data, as the study shows.

3.3.1 Summary Table of Deep Learning -Based Studies

Ref	Year	Dataset	Models	Results	Advantages	Limitations
Al-Rasheed et al., [42]	2025	UCI CKD dataset (400 patients, 24 features)	3-layer DNN (20–15–10)	Accuracy: 99%, AUC: 0.9821, F1 = 0.99	- Simple to implement, - well-suited for tabular data, - high performance on standard dataset, - aligned well with clinical risk indicators.	- Small dataset, - risk of overfitting (100% train accuracy), - no external validation, binary-only classification.
Chowdhury et al. [43]	2025	CRIC cohort (57,304 labs from 3,844 diabetic patients)	TabNet (full 31-feature and simplified 15-feature models)	Accuracy: 91–94%, AUC $\approx$ 0.95	- Interpretable attention mechanism, - effective on longitudinal lab data, - feasible simplified model.	- Limited to diabetic cohort, - lower performance on external data, - depends on multiple biomarkers not always available.
Rao et al. [44]	2023	UCI CKD (400 patients)	GNN + DNN + late-fusion ensemble with SVM	Fusion Accuracy: 95.1%, F1 = 0.95, AUC $\approx$ 0.95	- Incorporated patient similarity networks, - improved generalization using ensemble stacking.	- Requires expert-defined patient graphs, - limited evaluation on real-world data, - complexity reduces interpretability.
Alsekait et al. [45]	2023	UCI CKD dataset (400 samples)	Stacked DNNs (CNN, RNN) + SVM meta-classifier	Accuracy: 99.69%, F1 = 0.997, AUC: 0.999	- Leveraged multiple model strengths, - population based optimization, - improved learning via feature selection.	- Possible overfitting, - ensemble complexity hinders real-time use, - untested on real-world EHRs.
Tian et al. [46]	2024	Renal ultrasound (1,049 patients, 4,365 images)	ResNet-34 CNN + radiomic features + fusion ensemble	AUC: 0.918 overall, 0.867–0.931 (G4–G5)	- Outperformed physicians in early detection, - Hybrid of CNN and	- Single-center training, - Low interpretability, - Lacked integration with

				stages)	expert texture features, - Highly relevant to clinical practice.	clinical data.
Priyadharshini et al.[47]	2025	UCI CKD dataset (400 instances)	4-layer DNN with Population Optimization Algorithm (POA)	Accuracy: 100%, F1 = 1.0, AUC = 1.000	- Avoided local minima, robust optimizer, - extensible to other domains.	- No external validation, overfitting risk, - complex and computationally expensive training.

3.4 Summary

This chapter has evaluated how chronic kidney disease (CKD) has been previously described or identified in regards to traditional machine learning (ML) models and advanced deep learning (DL) methods, and explainable AI (XAI) approaches. ML models with high accuracy Random Forest, XGBoost, and Support Vector Machines are also often manually engineered by many decision providers and do not effectively model complex features. While deep learning (DL) models, such as Convolutional Neural Networks (CNNs), can learn rich feature representations automatically, large datasets and low interpretability limit their training. An enduring gap in the literature is the limited inclusion of hybrid frameworks that can leverage the representational strengths of DL while also being interpretable and robust by ML models optimally using evolutionary algorithms such as Genetic Algorithm (GA). In addition, the limited research from different sources is subject to few external verification and their datasets are often of low quality or have been generated by single-source datasets, which tends to limit the generalizability. We propose an approach that utilizes CNN-derived feature extraction, GA-optimized ML classifiers, and SHAP-based explanations to generate a predictive framework that can compromise between accuracy, interpretability, and clinical application. The methodology will follow in the next chapter, covering data preprocessing and process and model training, optimization, and validation strategies.

Chapter 4: Proposed Approach

4.1 Introduction

This chapter presents a novel deep learning and machine learning approach for chronic kidney disease (CKD) classification. This framework was essentially designed to create a powerful prediction pipeline, which enables translating raw clinical data into interpretable diagnoses with the greatest accuracy whilst minimizing data leakage and overfitting. This approach is applied using a defined, multi-step workflow. A cleaning step for the raw CKD dataset is used for standardization rather than transformation.

For this step, you will need to translate feature names and categorical data to English, remove duplicates, and construct a binary target variable representing cases of CKD versus non-CKD. At this stage no data splitting (as well as scaling or imputation) is done, there is no encoding or programming but only to preserve original statistical characteristics. Using clean data, the dataset is divided into training and testing sets by stratified sampling to preserve the balance in the class. All subsequent preprocessing like missing value imputation and the encoding of categorical features and numerical scaling apply only to the training data and after that on the test data using the learned parameters. The above design criteria are vital to prevent data leakage and to enable an unbiased evaluation of model quality.

After preprocessing, a one-dimensional Convolutional Neural Network (1D-CNN) encoder is used to learn a single CNN encoder to obtain compact latent representations of clinical features. CNN is like a feature extractor automatically, extracting features from the input files, in an effort to learn non-linear relationships and patterns between the variables that is hard to capture through manual feature engineering. The generated latent vectors can be used as a coherent feature space to be used by all later classifiers. The CNN-derived embeddings are then employed to train several machine learning models. Genetic Algorithm (GA) is used to optimize hyperparameters of each classifier to systematically search through the parameter space and enhance predictive skills. Besides single classifiers, hybrid modeling techniques are presented for improving robustness, stability, and generalization purposes. At last, the framework

combines an explainability component of SHapley Additive exPlanations (SHAP). This element offers a more explicit explanation of what was really going on in the optimization of a best-performing hybrid model by mapping those predictions back to clinically relevant features. The structure of the proposed CKD classification framework is explained in figure 4.1 where there are significant steps from data cleaning and preprocessing, to CNN-based feature extraction, classification, hybrid model and explainability.

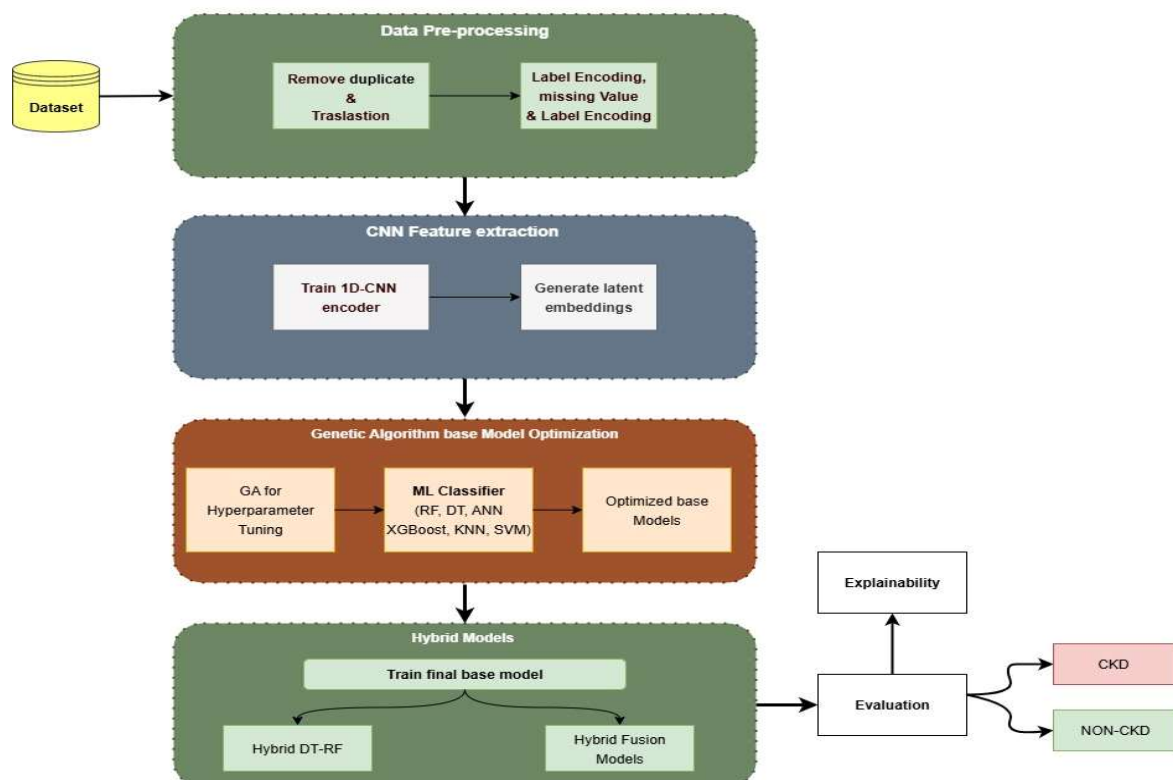


Figure 4.1: Proposed CKD Prediction Framework

4.2 Dataset description

This dataset represents an expanded and cleaned version of the original UCI CKD dataset and has been widely used in recent CKD prediction studies due to its clinical relevance and improved data quality 484848 [48].

The dataset comprises approximately **15,051 patient records**, where each instance is described by a combination of numerical and categorical features derived from routine clinical examinations. These features include blood pressure measurements, urine analysis indicators (such as specific gravity, albumin, sugar, red blood cells, and pus cells), hematological measurements (e.g., hemoglobin and packed cell volume), and biochemical markers including serum creatinine, blood urea, sodium, and potassium levels. All selected variables are clinically meaningful and commonly used in CKD diagnosis and monitoring.

A target variable is given as a binary class label which signifies chronic kidney disease (CKD vs. non-CKD) and a supervised binary classification for the task can be defined. To make this easier and faster, the dataset was examined for feature distribution, any possible class imbalance, and the presence of missing values ahead of model construction. This pre-piloting helped to make the preprocessing and model decisions made later in the experimental pipeline.

In order to delve into the dataset features, the class distribution was interpreted. As illustrated in Figure 4.2, the distribution of CKD to non-CKD classes is fairly evenly distributed. This balance minimizes classification bias risk in order to enable reliable model training and evaluation.

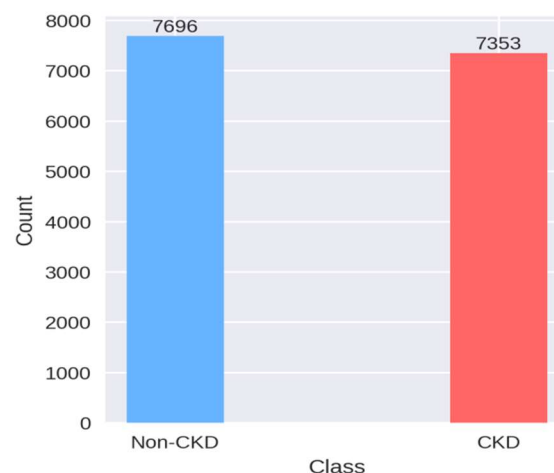


Figure 4.2: Distribution of CKD and Non-CKD samples in the dataset

A detailed description of all dataset attributes, along with their corresponding data types, is provided in **Table 4.1**.

Table 4.1 Description of the CKD Dataset Features

Feature Name	Description	Data Type
Patient_ID	Unique identifier for each patient record	Categorical
Age	Age of the patient (years)	Numerical
Blood_Pressure	Blood pressure measurement	Numerical
Urine_Specific_Gravity	Specific gravity of urine	Numerical
Urine_Albumin	Albumin level in urine	Numerical
Urine_Sugar	Sugar level in urine	Numerical
Red_Blood_Cells	Red blood cell status in urine	Categorical
Pus_Cells	Presence of pus cells in urine	Categorical
Pus_Cell_Clumps	Presence of pus cell clumps	Categorical
Bacteria	Presence of bacteria in urine	Categorical
Random_Blood_Glucose	Random blood glucose level	Numerical
Blood_Urea	Blood urea concentration	Numerical
Serum_Creatinine	Serum creatinine level	Numerical
Blood_Sodium	Sodium concentration in blood	Numerical
Blood_Potassium	Potassium concentration in blood	Numerical
Hemoglobin	Hemoglobin concentration	Numerical
Packed_Cell_Volume	Packed cell volume percentage	Numerical
WBC_Count	White blood cell count	Numerical
RBC_Count	Red blood cell count	Numerical
Appetite	Patient appetite status	Categorical
Chronic_Kidney_Disease	Target class label (CKD / Non-CKD)	Binary

4.3 Preprocessing Dataset

The dataset preprocessing is necessary to ensure the trustworthiness and generalization capability of the machine learning models, particularly in the clinical datasets which have frequent missing values, mixed data, and inconsistencies. In terms of preserving the integrity of the data and reducing the risk of information leakage, the preprocessing was a two-step process in this study.

The CKD raw dataset was first cleaned. With the aid of this step standardization was done, yet the underlying statistical characteristics of the data remained unchanged. In addition, feature names and categorical values were translated into English, duplicate records were removed, and the target variable was defined as a binary label with CKD or non-CKD cases. None of it involved splitting, encoding, scaling or imputation.

In Step Two, the cleaned dataset was prepped for model's training and evaluation. The data were split into training and testing sets applying stratified sampling of 70-30 per cent as the first to keep the original class distribution. No preprocessing was

performed on any other dataset but the training set. For numerical features, these operations included imputing missing values with median for numerical features and mode for categorical features, label encoding categorical variables, and applying standard scaling to numerical features. Similar learned transformations were later applied to the test data to generate meaningful results. This only preprocessing-up-until-trained approach leaves no traces of the test data, so that data leakage can be kept to a minimum and the performance of the model evaluated independently.

For CNN encoder and distance-based classifiers, feature scaling was crucial as it makes training more stable and improves convergence behavior. The result of this preprocessing stage will be the finished, and standardized training and testing datasets, to be used as inputs to the CNN encoder and further machine learning models.

4.4 Feature Exaction Phase

When developing a machine learning framework for a complex clinical condition like Chronic Kidney Disease, traditional approaches often demand that the researcher manually select which features matter most. However, such methods can easily miss the subtle, non-linear interactions hidden within the data. To overcome this limitation, our methodology strategically utilizes a **one-dimensional Convolutional Neural Network (1D-CNN) encoder** to perform an automated, data-driven **Feature Extraction**.

Essentially, we allow the network to learn the most informative ways to represent clinical data a process known as representation learning.

4.4.1 CNN Architecture and Training Details

At the heart of this strategy is a one-dimensional Convolutional Neural Network (1D-CNN) encoder. Instead of relying on manual selection or statistical ranking of the original clinical variables, the 1D-CNN is trained exclusively on the preprocessed training data (70% split) to automatically learn compact and highly informative latent representations, a technique that has seen increasing adoption in medical diagnosis models. For building our CNN architecture, we had a very narrow aim: to efficiently compress the input into a concise, yet

powerful, latent vector. The design consisted of two main convolutional blocks, both utilizing the ReLU activation function and a kernel size of 3, as detailed in Table 4.2.

Table 4.2: Summary of the 1D-CNN Encoder Architecture

Layer Type	Parameters	Output Shape	Feature
Input Layer	19 features	(None, 19, 1)	Input sequence of features
Conv1D (Block 1)	32 filters, Kernel Size=3	(None, 19, 32)	Initial feature pattern detection
Batch Normalization		(None, 19, 32)	Stabilizing network training
Dropout	0.3	(None, 19, 32)	Preventing overfitting
Conv1D (Block 2)	64 filters, Kernel Size=3	(None, 19, 64)	Learning higher-level feature abstraction
Batch Normalization		(None, 19, 64)	Stabilizing network training
Dropout	0.3	0.3	Preventing overfitting
Flatten		(None, 1216)	Preparing for the bottleneck layer
Latent Vector (Dense)	32 neurons	(None, 32)	Final compressed representation (Z)
Output (Dense)	1 neuron (Sigmoid)	(None, 1)	Classification head (used only for training)

We flattened, then forwarded from the CNN layers the obtained feature maps to the dense network segment, in which the encoder bottleneck was caused. The output of this layer is a 32-dimensional latent feature vector, Z (Row 9 in Table 4.2).

We ran this convolutional neural network driven classifier on 100 epochs with Early Stopping based on the validation AUC (Area Under the Curve), ensuring that we received training weights from its highest performance epoch (Epoch 6 in our case) to prevent overfitting and degradation.

4.4.2 Generating the Latent Embeddings (Z)

After ensuring the 1D-CNN was properly trained, we detached the last classification layer (the Sigmoid output) to build the separate encoder model. The encoder was then employed for creating the final feature representations of both the training and test datasets:

- Training Embeddings: Z_{train} (Shape: 10534×32)
- Testing Embeddings: Z_{test} (Shape: 4515×32)

Importantly, the Z vectors for the test set were created by an encoder that did not look at the test data when entering the learning phase. This strong separation is important to ensure that these subsequent machine learning models are tested on truly unseen data and that we achieve impartial, scientifically valid performance measures. The resulting 32-D feature space (Z) was the input for all subsequent GA-optimized classifiers.

4.5 Genetic Algorithm (GA) Optimization for Base Classifiers

The previous phase gave us a strong, compressed feature space (Z). But even the top features need to be well tuned by the classifiers to reach a high level of predictive performance. In this paper, instead of the laborious manual tuning and the complex grid searches, we turned to the Genetic Algorithm (GA), a biologically inspired adaptive search mechanism to systematically optimize the hyperparameters of our six base machine learning models.

The decision to use the GA was rooted in its proven ability to efficiently explore complex, high-dimensional search spaces, providing a robust solution for hyperparameter optimization across diverse machine learning paradigms.

4.5.1 GA Implementation and Operational Flow

The GA is based on the three basic methods of evolution: selection, crossover, and mutation. We assumed the hyperparameter format of each model as a chromosome and we were trying to "evolve" several generations of these chromosomes to find the best performing one. After performing a fair and consistent optimization process, the GA applied a standardized operational flow for all six base models (DT, RF, SVM, XGB, KNN, ANN):

- 1. Initialization:** A population of 18 new random hyperparameter sets (chromosomes) was generated in the predefined search space for each model.
- 2. Fitness Evaluation:** The heart of the GA. We rated each chromosome based on its level of prediction power. In terms of fitness function, we selected the Area Under the ROC Curve (AUC), which was determined on the training set by applying 3-fold Stratified Cross-Validation technique to that. Here, we ensured that the hyperparameters selected would be robust and that they generalized to various subsets of the training data.

- 3. Selection:** We used Tournament Selection ($k=3$), a method in which we select a small and competitive subset of individuals (three chromosomes) from which the best fitness individual (best AUC) can be selected to be parent to the next generation. It is a balance of exploiting the best solutions and exploring new solutions.
- 4. Reproduction:**
 - Crossover: To generate new offspring, selected parents were paired through Uniform Crossover (Probability = 0.5) and the offspring took on some characteristics of both parents.
 - Mutation: A controlled degree of randomness was achieved using Integer Mutation (Rate = 0.2), permitting the GA to outgrow the local optima by changing a hyperparameter's value in an arbitrary manner within its accepted set of values.
- 5. Termination:** The process was done repeatedly for 12 generations, a quantity that was found empirically to strike a compromise between computational cost and convergence to a near-optimal solution.

4.5.2 Search Spaces and Optimized Models

The task faced by the GA was to map its internal integer-based chromosome representation to the actual hyperparameter values of the classification models. The search spaces were carefully defined to cover a wide range of model complexity, ensuring that the optimal settings were not missed. Table 4.3 presents the key optimized models from the GA along with the GA names used in this framework and the important tuning parameters targeted.

Table 4.3: GA Optimized Base Classifiers and Targeted Hyperparameters

Model	GA Name	Key Hyperparameters Tuned
XGBoost	XGB_GA	n_estimators, max_depth, learning_rate
ANN	ANN_GA	hidden_layer_sizes, activation, alpha

KNN	KNN_GA	n_neighbors, weights, p (distance metric)
Random Forest	RF_GA	n_estimators, max_depth, min_samples_split
SVM	SVM_GA	C (Regularization), gamma
Decision Tree	DT_GA	max_depth, min_samples_split, min_samples_leaf

Utilizing this GA-based optimization, we did not have to test only small but arbitrary model configurations and instead ensured all six base classifiers entered the final evaluation phase with their data (Z) from the CNN-derived feature space in place of any parameters of non-standard design.

4.6 Machine Learning Models for Classification

Based on this high-quality, 32-dimensional latent feature representations (Z) extracted through the 1D-CNN encoder, the next critical step was to use a diverse repertoire of different machine learning algorithms. The strategic deployment of six different classifiers (three traditional ML, three advanced), was not only to contrast, but to perform systematic analysis towards understanding how strong the Z feature space is in different learning modes. All six models were trained with the same Z features, and hyperparameters were optimized carefully using the Genetic Algorithm (GA) described in the previous section. This integrated approach allowed the model to account not only for such differences in final efficacy but also to see that any differences were the result of the model's inherent learning bias (its ability to infer about the Z features) and not arbitrary hyperparameter tuning.

4.6.1 Traditional Machine Learning Classifiers

This group also includes well-known algorithms for this task, which are usually stable and interpretable, and provide a variety of decision-making strategies to evaluate the latent features.

A. Decision Tree (DT):

The Decision Tree was designed to be able to model the classification process as a sequence of simple, hierarchical decisions. Training on Z features gave the DT highly comprehensible rules that basically translated the 32-dimensional space back to simple logical conditions. The GA made sure that the complexity of the tree (max_depth) was kept in check so we wouldn't overfit, but still have the predictive potential.

B. K-Nearest Neighbors (KNN):

The KNN investigated the density and spatial coherence of the Z feature space as a distance-based, non-parametric classifier. If the CNN successfully clustered samples by class (CKD/Non-CKD), KNN would perform exceptionally well. The GA was used in this case to determine a relevant neighborhood size (n_neighbors) and the distance formula (p) to be used as a measure of similarity within this normalized embedding space.

C. Support Vector Machine (SVM):

SVM is well-suited for finding our single best hyperplane for classes separation, very well in high-dimensional feature spaces. With the Radial Basis Function (RBF) kernel, the optimization of the SVM required tuning the soft margin parameter (C) and kernel influence (gamma). This enabled the model to build this complex and non-linear boundary very suitable to the CNN-based embeddings.

4.6.2 Advanced and Ensemble Classifiers

This collection of models features high-performance, complex learners known to provide state-of-the-art results by using ensemble techniques or deep non-linear transformation.

D. Random Forest (RF):

This is why the Random Forest is also an ideal ensemble to mitigate the instability of a single DT by averaging predictions across multiple trees. RF trained several trees on diverse subsets of the Z features, resulting in a highly robust, variance-reducing model. The GA struck the perfect balance between the number of trees (n_estimators) and their respective depths to maximize generalization.

E. Extreme Gradient Boosting (XGBoost):

XGBoost is one of the premier gradient boosting algorithms, renowned for its efficiency and ability to sequentially correct the errors of previous weak learners. Its inclusion was essential to test the final predictive ceiling possible with the Z features. The GA focused on balancing the model's aggression (`learning_rate`) with its complexity (`max_depth`) to avoid aggressive overfitting while pushing the predictive metrics to their limit.

F. Artificial Neural Network (ANN):

We proposed an ANN as a Multi-Layer Perceptron (MLP) to utilize another layer of non-linear transformation on the Z embeddings. Because the Z features themselves are a product of a deep network, their incorporation into the ANN tested for the benefit of additional non-linear mixing. The GA was essentially responsible for the search for the best architecture (`hidden_layer_sizes`) and activation functions to capture complex feature interactions and to make sure that the final output was as accurate as it could be.

4.7 Hybrid Model

Although the prior stages determined baseline performances for a few tuned classifiers, the major objective of our framework was to define a model that can provide an optimal trade-off between predictive capability and clinical interpretability, rather than to obtain the best metric. We present Hybrid (DT-RF Integration), our designed structural hybrid model, which was chosen by SHAP specifically for further elaboration.

Hybrid A is a complex two-stage architecture based on integration of explicit rule learning of the Decision Tree (DT) and the high stability and robustness of the Random Forest (RF) ensemble. This guarantees that the model's complicated decisions can be traced back to clear, rule-based features from the original data.

4.7.1 The Two-Stage Hybrid A Workflow

1. Decision Tree Rule Extraction:

At the beginning, the GA-optimized Decision Tree is trained on the CNN-generated latent feature vectors (Z). This is not because of DT directly predicting but because it's a feature encoder itself: each time (training & testing), you are storing in the DT the leaf node index of where the instance lands. This index will essentially represent a discrete, unique rule (or pathway) learned by the model.

2. *Feature augmentation and ensemble training:*

Next, we convert the leaf node indices to a one-hot encoded feature matrix, which explicitly represents the internal logic of DT in a binary format. The original continuous Z latent features are then concatenated with this new feature set, resulting in a very informative augmented feature space.

3. *Final Prediction:* From here, the GA-optimized Random Forest (RF) classifier is trained on this complete, augmented feature space (Z plus One-Hot DT Rules). This strategic injection of DT rules into the RF allows the ensemble to learn how to couple the continuous latent features with the explicit decision paths.

4.7.2 Rationale for Model Selection (Hybrid A vs. Performance Leaders)

Although there are a few other models, like the XGBoost_GA or Fusion models (XGB_GA+SVM_GA, DT_GA+XGB_GA, ANN_GA+RF_GA, ANN_GA+XGB_GA and RF_GA+DT_GA) achieving marginally superior performance metrics (e.g., AUC), Hybrid A was selected as the main model for deeper studies. The reason behind the selection is because of the good fit of the SHAP interpretation framework to the proposed model:

- **Rule-Driven Interpretability:** Since the DT's decision rules are added directly to the feature set, Hybrid A predictions are naturally more traceable (i.e., readable) and clinically interpretable against a black-box model (e.g., as well as a standalone XGBoost or a deep Neural Network)
- **Balancing Act:** Hybrid A gives us the best compromise for the study where performance scores are kept competitive with the best models, but significantly increase the transparency needed for clinical deployments. Thus, such a focus on Explainable AI (XAI) renders it an even more robust candidate for real-world medical application.

4.8 Explainability AI: SHAP Methodology

The high predictive performance pursuit in medical applications as shown in our Hybrid A model, is just half of the equation. At the other equally important half is trust and transparency. Clinicians care not only for the accuracy of the predictions, but also need a transparent and traceable rationale of their decision making. To solve this issue, we implemented Explainable Artificial Intelligence (XAI) into our framework by performing SHAP (SHapley Additive exPlanations) to the Hybrid A classifier. It is essential for clinical

use to utilize SHAP, due to its ability to facilitate understanding of model behavior in sophisticated prediction tasks.

4.8.1 The Foundation of SHAP Values

SHAP is a unified framework of game theory and local explanation, derived from classic Shapley values. Here, it offers a model-agnostic way to determine how much each feature contributed to a given prediction. This approach yields two main results:

1. **Local Explanation:** It traces the way in which prediction model value for each patient, one at a time, moved the prediction from baseline rate towards the final predicted outcome (CKD or Non-CKD).
2. **Global Explanation:** It combines the contributions from all local features of the overall test dataset to the overall decision-making of the model to get a general indication of which features contribute to the model's overall judgement.

In this study, SHAP was properly applied for our “Hybrid” (DT-RF Integration) model using its tree based nature, enabling our models to quickly quantify the precise contribution of the augmented feature set (the 32-dimensional Z features and the generated DT decision rules), which in turn is used for the CKD prediction outputs.

4.8.2 Interpretation Tools and Visualization

The SHAP framework produces several helpful visuals that we use in the results section to understand how the model works. Our work led to two main types of visuals for detailed study:

- **Summary Plot (Bar Plot):** This plot visualizes the overall, or global, importance of the features by taking the mean absolute SHAP value for each feature across all test instances. This clearly ranks the latent features and DT rules by their predictive magnitude.
- **Beeswarm Plot:** This crucial visualization extends the summary plot by showing the distribution of SHAP values for each feature.⁴ It indicates the density of samples and allows us to see the correlation between a feature's value (represented by color) and its directional impact on the prediction (represented by position on the x-axis).⁵

By integrating this SHAP analysis into our framework, we successfully transform the high-performing Hybrid A from a 'black-box' tool into a transparent and clinically verifiable decision-support system.

4.9 Summary

This chapter introduced our proposed hybrid classification framework for Chronic Kidney Disease covering the experimental pipeline starting from data preprocessing and ending with the prediction and explainability. The chapter initiated by presenting the dataset and a preprocessing approach intended to maintain data integrity and prevent the leakage of information. Then they proposed a 1D-CNN-based feature extraction method for learning compact and informative hidden features in the clinical dataset, in which they converted its own features into a highly predictive 32-dimensional feature space.

A range of machine learning classifiers were subsequently optimised and trained using a Genetic Algorithm (GA) to allow for fair comparisons amongst all starting classifiers. After that, the Hybrid A structural integration approach was described, which couples the Decision Tree's rule-based logic with the ensemble stability of Random Forest. Hybrid A was chosen as the core proposed model specifically due to the impressive predictive performance it achieved, along with a high degree of interpretability and transparency.

Lastly, SHAP-based explainability analysis was applied to offer visibility into the decision process of the model. The contribution of the CNN-derived latent features and the detected DT rules to our analysis has been emphasized, which confirms that the behavior of the model was consistent with established medical knowledge.

Chapter 5

Experimental Results and Discussion

5.1 Introduction

This chapter illustrates and provides high-level results for the experimental pipeline described in Chapter 4. Our primary aim is to confirm the performance of the proposed method that incorporates CNN-based feature extraction, GA-optimized classification, and a new hybrid model. We structure the discussion so that the baseline performance is achieved using the base classifiers separately. It then proceeds to analyze the effectiveness of the hybrid strategies, concluded by a model transparency review. It concludes the chapter with an analysis of the decision-making of the chosen model Hybrid DT-RF and a discussion of the SHAP framework, in order to reveal clinical implications required for translational research.

5.2 Results of GA-Optimized Base Classifiers

In this case we explore and describe performance of each of the individual (base) machine learning models trained using the CNN-derived latent feature representations (Z). The aim of this analysis is to establish an effective and reliable baseline of performance to evaluate the performance of the proposed hybrid modeling strategies. All the base models were optimized with the help of a Genetic Algorithm (GA) and were evaluated on the unseen test dataset using standard classification metrics such as Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC). Table 5.1 summarizes the performance of the GA-optimized base models on the test dataset.

Table 5.1: Performance of GA-Optimized Base Models on the Test Dataset

Model	Accuracy	Precision	Recall	F1-score	AUC
ANN_GA	0.9971	0.9946	0.9995	0.9971	0.9998
XGB_GA	0.9978	0.9968	0.9986	0.9977	0.9997
RF_GA	0.9978	0.9964	0.9991	0.9977	0.9994
SVM_GA	0.9980	0.9968	0.9991	0.9980	0.9989
DT_GA	0.9978	0.9964	0.9991	0.9977	0.9984
KNN_GA	0.9982	0.9968	0.9995	0.9982	0.9982

Table 5.1 summarizes that all the base models evaluated had strong predictive performance in all metrics. This shows that the CNN encoding worked, and learned both an informative and

discriminative latent representation from the clinical dataset, making the available learning algorithms efficient for the same feature space.

In particular, K-Nearest Neighbors (KNN_GA) performed the best with respect to its accuracy and F1-score (≈ 0.9982), closely followed by its sibling SVM_GA. On the other hand, in total discriminative accuracy, the Artificial Neural Network (ANN_GA) was the model with highest AUC score (≈ 0.9998), followed by XGBoost (XGB_GA) and Random Forest (RF_GA).

Indeed, these results indicate neural-based and ensemble-based classifiers especially can leverage CNN-derived latent features. These numerical differences may appear small, but they are crucial for important medical classification tasks to produce consistent generalization across domains in terms of stability. The majority of base models have recall values nearing 0.999, and this finding is important, especially when predicting CKD, as high levels of recall help to avoid false negatives and minimize the critical risk of missing cases due to disease.

Simultaneously, the very high precision of each of the models indicates a very small proportion of false positives. Notably, the Decision Tree model (DT_GA) showed a good performance as well, but it had somewhat lower AUC compared to ensemble-based and margin-based classifiers. This phenomenon might be predictable, because the same decision tree can be influenced more by the variability of the data. Yet DT is of critical importance as a structural component for the present study because of a fundamental interpretable nature, and the fact that it extracts explicit decision rules, which eventually helps generate the Hybrid DT-RF model.

Generally, the performance trends of the base models lead to two important conclusions. Firstly, CNN-based feature extraction significantly enhances the quality of the input feature space, permitting various machine learning approaches to obtain close to optimal discrimination performance. Second, the performance similarity between multiple base models indicates the need for hybridization methods, not only to achieve small improvements in predictive accuracy, but also to facilitate robustness, stability, and interpretability, which are of utmost importance in clinical decision-support systems.

5.3 Results of the Proposed Hybrid Models

In this section, the performance of the proposed hybrid modeling methodologies developed to improve the robustness and generalization capability of the CKD classification framework is presented and discussed. Based on the strong baseline established by the individual models, hybrid approaches were applied to exploit complementary strengths of different classifiers through structural integration or probability-level fusion.

The performance of hybrid models was assessed using the same performance metrics applied to the base models, namely Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC). This ensures a fair comparison and allows direct assessment of the impact of hybridization on classification performance.

The performance of the proposed hybrid models on the test dataset is summarized in **Table 5.2**.

Table 5.2: Performance of the Proposed Hybrid Models on the Test Dataset

Model	Model Type	Accuracy	Precision	Recall	F1-score	AUC
Hybrid_DT_RF	Structural Hybrid	0.9980	0.9968	0.9991	0.9980	0.9995
HybridB_Fusion (ANN_GA + XGB_GA)	Probability Fusion	0.9973	0.9950	0.9995	0.9973	0.9998
HybridD_Fusion (ANN_GA + RF_GA)	Probability Fusion	0.9973	0.9950	0.9995	0.9973	0.9998
HybridF_Fusion (DT_GA + XGB_GA)	Probability Fusion	0.9980	0.9968	0.9991	0.9980	0.9995
HybridE_Fusion (RF_GA + DT_GA)	Probability Fusion	0.9980	0.9968	0.9991	0.9980	0.9994
HybridC_Fusion (XGB_GA + SVM_GA)	Probability Fusion	0.9980	0.9968	0.9991	0.9980	0.9992

The results indicate that hybridization is primarily connected to stability and balanced classification behavior improvements, as opposed to large numerical gain in accuracy. Of the tested approaches, Hybrid (DT–RF integration) demonstrated a well-balanced performance profile while maintaining interpretability through its structural design. The hybrid approach (DT–RF) presented a strong, well-balanced performance profile (AUC $\approx$ 0.9995), and it was able to uniquely preserve interpretability through the structural design.

The fusion models also achieved slightly higher AUC values but do not explicitly embed structural decision information, and are therefore not suitable for transparent clinical deployment. Hybrid (DT–RF) was finally selected as the final model, as it provides the optimum trade-off between the predictive performance and the critical explainability; therefore, it is the best candidate for a clinical decision-support system, where transparent decision-making is paramount. To enhance comparison of hybrid model performance under various evaluation indicators, a visual performance comparison is provided.

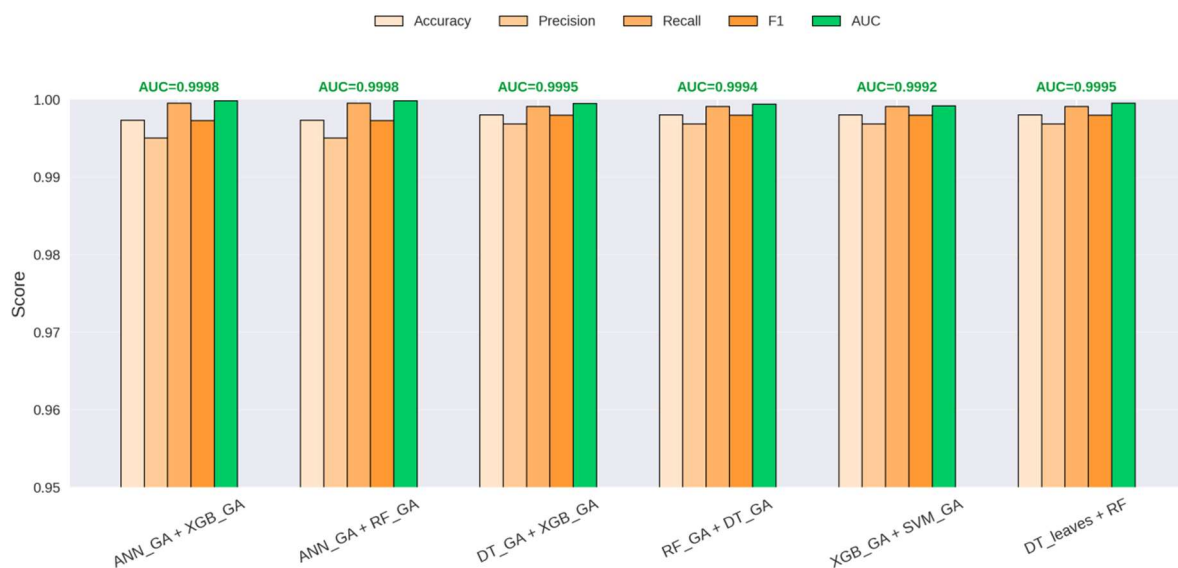


Figure 5.3: Performance comparison of the proposed hybrid

The hybrid models, as shown in Figure 5.3, show uniformly good performance in Accuracy, Precision, Recall, F1-score, and AUC. Similar trends were noted for probability fusion models as well, underscoring the applicability of hybrid modeling approaches for CKD classification. On a macro level the results show that hybrid approaches make concrete improvements beyond individual classifiers, leading to improved robustness and consistency.

In particular, Hybrid (DT–RF) shows a reasonably good trade-off between performance and explainability, which is fitting for clinical decision-support systems where clear decision-making is required.

5.4 Results of Explainable AI (SHAP Analysis)

To meet the need for clear medical decision support, the SHAP framework was used only on the selected model, Hybrid (DT-RF Integration). We used the TreeExplainer to efficiently compute the contribution of each feature, the CNN latent features (Z) and the augmented DT rules, to the unseen test set predictions.

5.4.1 Global Feature Importance

The SHAP Summary Bar Plot (Figure 5.4) gives the mean absolute SHAP value for each feature. It is a global ranking of importance based on the feature's overall magnitude of influence on the model's output across the entire test set.

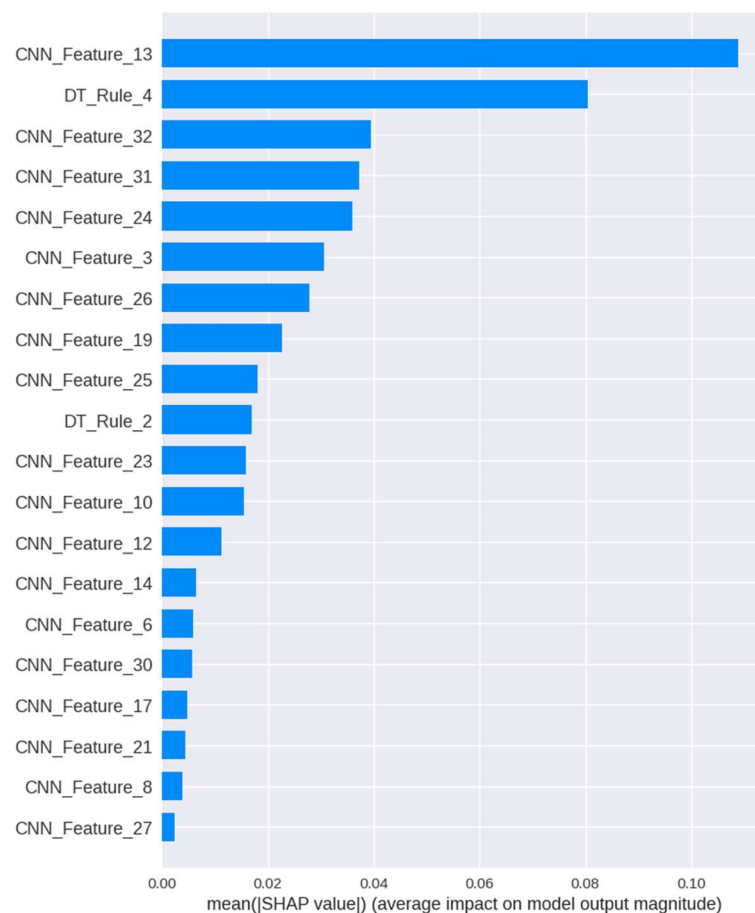


Figure 5.4: Global Feature Importance Ranking for Hybrid A (SHAP Bar Plot)

- Point to note: The graph clearly shows that the main feature of interest is the CNN latent characteristic of CNN_Feature_13 and comes before the structural rule DT_Rule_4. The most striking result that validates the feature extraction step is this most useful test of the

model's prediction results, which suggests that the predictive ability mostly comes from the automatically learned compressed and latent representations.

- Asserting the Structural Integration: the high performance of DT_Rule_4 is indicative of the structural integration: the top feature position such as DT_Rule_4 (second overall) and the dominance of DT_Rule_2 among some key characteristics overall strongly validate the basic structural hybridisation. This shows that the Random Forest ensemble is using abstract continuous latent features but also includes the explicitly made decision using Decision Tree logic.

5.4.2 Directional Feature Impact

The SHAP Beeswarm Plot (Figure 5.5) shows you the degree of this phenomenon more closely, as it captures a deeper, example-focused view on feature impacts which shows the direction shift of the influence of the features on the predicted risk of CKD (positive SHAP values are leading to CKD).

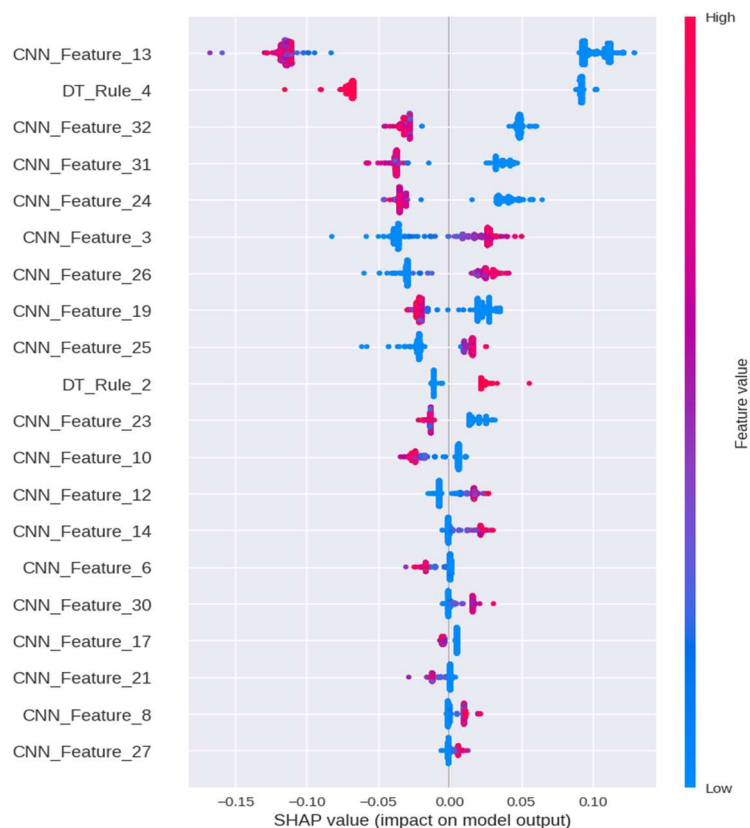


Figure 5.5: SHAP Beeswarm Plot Showing Directional Feature Impact for Hybrid A

- **Interpretation of Latent Features:** For the most important latent feature, **CNN_Feature_13**, the plot shows a clear split: **low** feature values (blue points) strongly push the prediction towards CKD (positive SHAP values, right side), while **high** feature values (red/pink points) strongly push the prediction away from CKD (negative SHAP values, left side). This inverse relationship highlights the complex, non-linear pattern that the CNN has learned.
- **Interpretation of DT Rules:** For the binary rule feature, **DT_Rule_4**, the plot shows that when the rule is active (high feature value, red/pink points), it strongly pushes the prediction towards CKD (positive SHAP values). This allows for direct interpretation of the Decision Tree's contribution: activating the condition represented by **DT_Rule_4** significantly increases the patient's risk profile as determined by the model.

5.4.3 Reverse SHAP Mapping and Clinical Insight

Although the analysis in Sections 5.4.1 and 5.4.2 confirmed the predictive performance of the latent features (CNN_Feature_x), it did not necessarily connect these abstract features back to the original, clinically interpretable parameters. Addressing this drawback and facilitating in delivering significant clinical commentary, a Reverse SHAP Mapping approach was performed. This method uses the Chain Rule to compute the total and ultimate consequences of original clinical features (X) on the final prediction, mediated through the top most significant latent features (Z) in the Hybrid DT_RF model. The most notable clinical features, identified by their Normalized Reverse SHAP Impact, are shown in Table 5.3 and visualized in Figure 5.6.

Table 5.3: Top Clinical Features Ranked by Reverse SHAP Impact

Rank	Feature	Normalized Reverse SHAP Impact
1	Red_Blood_Cells	1.000000
2	Pus_Cells	0.805288
3	Urine_Specific_Gravity	0.212246
4	Urine_Sugar	0.174208
5	Hemoglobin	0.144881
6	WBC_Count	0.138552
7	Appetite	0.131761
8	Blood_Pressure	0.120512
9	Random_Blood_Glucose	0.114070
10	Blood_Potassium	0.113800
11	Blood_Sodium	0.110705
12	Pus_Cell_Clumps	0.104546
13	RBC_Count	0.100098
14	Packed_Cell_Volume	0.097348
15	Blood_Urea	0.096743
16	Urine_Albumin	0.08360044
17	Age	0.080336735
18	Bacteria	0.07390439
19	Serum_Creatinine	0.04218411

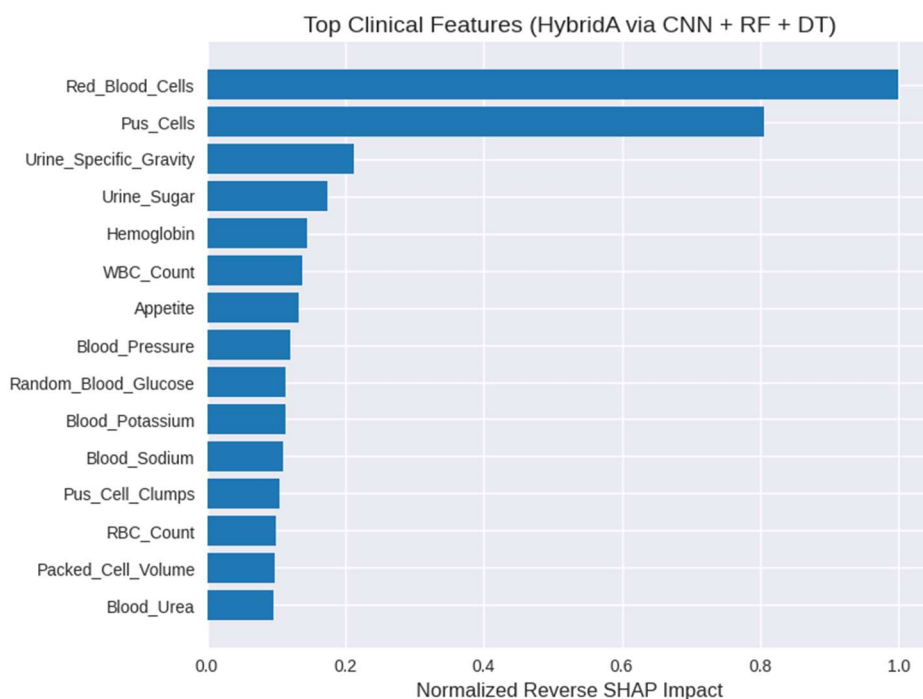


Figure 5.6: Top Clinical Features (Hybrid via CNN + RF + DT) Ranked by Reverse SHAP Impact

Key Clinical Findings:

- **Dominant Predictors:** The study findings reveal that Red_Blood_Cells and Pus_Cells are the most important features, with significantly larger Reverse SHAP Impact (≈ 1.00 and ≈ 0.81 , respectively) than the next feature, Urine_Specific_Gravity (≈ 0.21).
- **Validation of Medical Knowledge:** Prevalence of these features corresponds to the knowledge of clinical pathology, wherein red blood cells and pus cells in urine sediment (Hematuria and Pyuria) are indicators of likely renal or urinary tract damage and likely the development of CKD.
- **Comprehensive Risk Profile:** Although blood count parameters (Hemoglobin, WBC Count, RBC Count) and some basic chemistry parameters (Urine Sugar, Blood Pressure, Blood Glucose) are relevant as well, their importance is limited to a moderate degree. This indicates that the CNN mostly learns patterns concerning microscopic urine analysis and employs them to generate the top-level features that have a notable influence which the final Hybrid A model depends on.

This Reverse SHAP method effectively converts the model from a high-accuracy "black box" to a clinically traceable predictor, in accordance with which the internal mechanism (often abstract) returns systematically to practical clinical markers.

5.5 Discussion

The results will summarize the experimental results in this section on the successful assimilation of the methodologies, validate the core hybrid strategy, and describe the implications of the outcomes for diagnosis of CKD.

Validation of the Hybrid A Approach

Validation of hybrid approaches. The fundamental aim of this study was the proposed and tested high-performance and transparent CKD-prediction framework based on the Hybrid A (DT-RF Structural Integration) model. This objective is confirmed by the results in a multi-faceted analysis:

1. **High-Performance Baseline:** The superior performance of all base models ($AUC \geq 0.9982$), which is induced by CNN feature extractor, confirmed the quality of the

input feature space (Z). This set a very high threshold suggesting that the CNN converted crude clinical data into an extremely discriminative manifold.

2. **Strategic Model Selection:** As other fusion models (Hybrid B, Hybrid D) and top base model (ANN_GA) achieve numerically equivalent (≈ 0.9998) or marginally greater AUC, Hybrid A was selected for the best overall result (AUC ≈ 0.9995). It was a very purposeful and strategic choice. Interpretability and transparency are not simply wished for but are now clinical needs in medical decision support that supersede minimal departures in AUC at this high performance ceiling.
3. **Proof of Structural Integration:** With SHAP analysis, Hybrid A's architectural capability was verified; the high ranking of DT Rule features (e.g., DT_Rule_4) alongside the prominent CNN ones corroborated that Random Forest was successfully using all clear, rule-oriented knowledge harvested from Decision Tree. This win demonstrated that the hybrid design was exactly what it set out to be: achieving very high ensemble performance and yet retaining structural transparency.

Clinical Traceability through Reverse SHAP Mapping

The groundbreaking Reverse SHAP Mapping (Section 5.4.3) was instrumental in connecting abstract-based performance of the model with its clinical relevance, thus solving the main challenge of deep learning feature extraction – identifying and extracting features efficiently.

- **De-Abstracting Z:** The framework was developed in the context of Z, and was able to de-Abstract Z by systematically connecting the internal effect of the latent features (Z) with the input factors (X) in a systematic process leading the system to a model that was converted from being a very successful "black box" (highly reliable) model by transforming into a clinically traceable predictor.
- **Medical Validation:** Red_Blood_Cells and Pus_Cells were confirmed to be the most influencing clinical factors in the accuracy of the model via the mapping. This result is also extremely noteworthy and offers a strong, numerical evidence that the remarkable high performance of the model is based on the well-known clinical pathology characteristic of Renal Disease leading to acceptable clinical practice and confidence.

Comparison with Other Hybrid Methods

The assessment with other hybridization modalities (Probability Fusion Models) proved that:

- **Fusion Models (Hybrid B, D):** These models had best numerical AUC but are inherently opaque. Which means they work from averaging or aggregating the probabilities of individual outputs, without any explanation for the prediction.
- **Structural Hybrid (Hybrid A):** A compromise on a small numerical increase for increased transparency. Hybrid A presents a clean explanation (SHAP values by CNN_Feature and DT_Rule) that the fusion models cannot provide, and it is therefore best suited for high-risk medical problems in which explaining the reason for a diagnosis is a priority.

Overall, the hybrid A (DT-RF Integration) model constitutes the most convincing work of this research in that it offers top-of-the-line predictive power and fulfills the key requirement of transparency via an integrated structural design and a confirmed traceability through the Reverse SHAP method.

5.6 Limitations of the Proposed Work

The developed hybrid framework has been great, but also faced some inherent challenges and limitations which should be addressed in future work:

- **Standardization of the data quality:** The near-perfect performance directly depends on the quality and noise level of the public CKD data being used. The quality of the framework is only as good as real-world clinical data, as the variability of measurement and errors in data entry and the greater quantity of missing values in clinical samples may not allow for direct clinical deployment at the time of treatment.
- **Latent feature abstraction:** Even though SHAP analysis produces importance scores, the strongest features (CNN_Feature_x) are still highly abstract numerical features. To reconstruct CNN_Feature_5 from what was originally calculated (for example albumin levels or creatinine levels) requires a complicated decoding approach that compromises the ability for the original clinical variables to be immediately, intuitively interpreted.

- **Computational Complexity:** The use of the Genetic Algorithm (GA) for hyperparameter optimization in the six reference models led to a substantial increase in computation cost and the training time compared to when these models were trained by standard algorithms. This complexity is justified due to marginal increased stability and performance; however, retraining the model, or training the model in resource-constrained environments are difficult tasks.

5.7 Summary

This chapter provided an experimental evaluation and discussion of the proposed Chronic Kidney Disease (CKD) classification framework. Performance of individual machine learning models was initially studied for establishing a reasonable baseline, which proved the robustness of CNN-derived latent feature representations. It was subsequently followed by a thorough analysis of hybrid modeling approaches for enhancing robustness and interpretability.

The experimental results show that all GA-optimized base models obtained near-perfect prediction performance consistently. Most importantly, the hybrid approaches proposed, in particular Hybrid A (DT-RF Structural Integration), enhanced the robustness of performance by presenting a reasonable trade-off between accuracy ($AUC \approx 0.9995$) and explainability.

In addition, the original Reverse SHAP Mapping identified the clinical indications leading to the model's prediction, identifying Red Blood Cells and Pus Cells which were directly referenced to the original clinical features, confirming their dominance as clinical indicators. The results were contextualized by highlighting the benefits of merging deep feature extraction, hybrid modeling and explainability into a single framework proving the Hybrid A model to be a transparent and trustworthy clinical decision support tool. Furthermore, the chapter presented the essential limitations of the proposed work along with some avenues for further research. In general, the proposed methodology was effective and original, hence this chapter validated the contribution of the proposed methodology and made way for the final chapter where the main contribution is summarized and the potential extensions of the proposed framework are detailed.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

Our proposed model presents an interpretable hybrid machine learning framework for the early and precise categorization of chronic kidney disease (CKD), ensuring a balance between superior prediction accuracy and clinical transparency. The framework gets almost perfect accuracy and extremely high recall by using CNN-based latent feature extraction and GA-optimized classifiers. This cuts down on false negatives by a lot. The main contribution is a Hybrid DT_RF model that combines Decision Trees and Random Forests to provide good results while also being easy to understand. Using genetic algorithms for optimization makes guarantee that all models get fair and strong tuning. To improve explainability, SHAP and Reverse SHAP Mapping were used to connect hidden characteristics to clinically relevant variables. This showed that Red Blood Cells and Pus Cells were the most important markers of CKD. The framework improves explainable medical AI and provides a solid, flexible base for future CKD and other clinical decision support systems.

6.2 Future Work

Despite the fact that the proposed framework was excellent and yielded interpretable results, it appears to suggest several important avenues for further strengthening and extension to extend, and improve the work towards making it relevant in practice. Improving Generalizability and Data Diversification

- **Evaluation on Data Variability:** Another next step is to conduct comprehensive evaluations of the proposed approach using wider and larger data sets collected from different healthcare institutions and geographic communities in the future. This validation could improve the confidence in generalizability and robustness of the model with heterogeneous populations of patients.

- **Integration of Longitudinal Data:** One possible strategy is to incorporate longitudinal patient dataset that would enable better modeling of CKD progression. Predicting the stages of disease or likelihood of progression provides greater opportunity for clinical intervention earlier -- in contrast to mere binary classification.
- **Integration of Multi-Modal Data:** Adding additional multi-modal data — like medical imaging data, laboratory trends, or genetic information — could have significantly enriched the information in the feature space and aided diagnostic and prognostic accuracy.

Development in Methodology and Explainability

- **Alternative Architectures:** Future work can also investigate alternative deep learning architectures specifically addressed tabular data (e.g., TabNet, specific self-attention architectures) along with advanced ensemble strategies for improving model performance and stability.
- **Refining Interpretability:** Extending the interpretability assessment to include additional interpretation techniques (e.g., counterfactual explanations) or clinician-in-the-loop evaluation could help to increase the practical applicability of the system, ensuring that explanations are inherently intuitive for the healthcare workers.
- **Decoding Latent Space:** In the future it will be important to develop more advanced methods for decoding the entire 32-dimensional latent feature space (Z), through something like variational autoencoders or clustering to make sure to generate a better clinical interpretation of the abstract patterns learned by the CNN.

Clinical Deployment

- **Supporting Decision-Making in Real-Time:** Real-time clinical implementation of proposed framework as a tool for decision support is an important objective for the longer run. This use will be sensitive for short-term limitations of ensuring data privacy (federated learning), system interoperability with EHR, and extensive clinical validation in a pilot setup. The new Real-Time Approach will significantly help early detection and therapy of Chronic Kidney Disease.

References

1. Bikbov, B., Purcell, C. A., Levey, A. S., Smith, M., Bikbov, M., et al. (2020). Global, regional, and national burden of chronic kidney disease, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 395(10225), 709–733.
2. Chen, T. K., Knicely, D. H., & Grams, M. E. (2019). Chronic kidney disease diagnosis and management: A review. *JAMA*, 322(13), 1294–1304.
3. Senanayake, S., Karunagoda, K., & Perera, N. (2024). Deep learning for chronic kidney disease prediction: A review. *Artificial Intelligence in Medicine*, 148, 102762.
4. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
5. Webster, A. C., Nagler, E. V., Morton, R. L., & Masson, P. (2017). Chronic kidney disease. *The Lancet*, 389(10075), 1238–1252.
6. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. (2013). KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney International Supplements*, 3(1), 1–150.
7. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group (2024). KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney international*, 105(4S), S117–S314.
8. Luyckx, V. A., Tonelli, M., & Stanifer, J. W. (2018). The global burden of kidney disease and the sustainable development goals. *Bulletin of the World Health Organization*, 96(6), 414–422.
9. Pan, Q., & Tong, M. (2024). Artificial intelligence in predicting chronic kidney disease prognosis. A systematic review and meta-analysis. *Renal failure*, 46(2), 2435483.
10. Rayner, B. L., Jones, E. S. W., Davidson, B., & Wearne, N. (2023). Advances in Chronic Kidney Disease in Africa. *Applied Sciences*, 13(8), 4924.
11. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4), 230–243.
12. Razzak, M. I., Imran, M., & Xu, G. (2019). Big data analytics for preventive medicine. *Neural Computing and Applications*, 32(9), 4417–4451.

13. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.
14. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
15. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.
16. Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in Healthcare* (pp. 25–60). Academic Press.
17. Polat, H. (2017). Diagnosis of chronic kidney disease based on support vector machine. *Journal of Medical Systems*, 41(1), 11.
18. Rodríguez-Pérez, R., & Bajorath, J. (2022). Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *Journal of computer-aided molecular design*, 36(5), 355–362. <https://doi.org/10.1007/s10822-022-00442-9>
19. Chen, Z., Wang, Y., Xu, Y., et al. (2024). Interpretable XGBoost model for distinguishing renal fibrosis severity in CKD patients. *International Journal of Medical Informatics*, 184, 104227.
20. Liu, B., Tan, X., Jin, Y., Yu, W., & Li, C. (2021). Application of RR-XGBoost combined model in data calibration of micro air quality detector. *Scientific reports*, 11(1), 15662.
21. Sattari, M. (2023). Using data mining techniques to predict chronic kidney disease. *International Journal of Occupational Medicine and Environmental Health*, 36, 1896–01809.
22. Abdel-Fattah, M. A., Othman, N. A., & Goher, N. (2022). Predicting Chronic Kidney Disease Using Hybrid Machine Learning Based on Apache Spark. *Computational intelligence and neuroscience*, 2022, 9898831.
23. Mendapara, K., Patel, N., & Shrimali, D. (2024). CKD risk prediction using Random Forest on gene expression data. *Frontiers in Genetics*, 14, 1409755.
24. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
25. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

26. Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 18932–18943.
27. Sharma, A., Vans, E., Shigemizu, D., & Boroevich, K. (2019). DeepInsight: A methodology to transform non-image data to an image for CNN. *Scientific Reports*, 9, 11399.
28. Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science*, 17(11), e70056. <https://doi.org/10.1111/cts.70056>
29. Molnar, C. (2022). *Interpretable machine learning*.
30. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.
31. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6.
32. Phuong, N. D., Tuyen, N. T., Linh, V. T. T., Nguyen, N. N., & Nguyen, T. Q. (2025). Machine learning techniques in chronic kidney diseases: A comparative study of classification model performance. *Bioinformatics and Biology Insights*, 19, 11779322251356563.
33. Elshewey, A. M., Selem, E., & Abed, A. H. (2025). Improved CKD classification based on explainable artificial intelligence with extra trees and BBFS. *Scientific Reports*, 15, 17861.
34. Liu, P., Liu, Y., Liu, H., Xiong, L., Mei, C., & Yuan, L. (2024). A random forest algorithm for assessing risk factors associated with chronic kidney disease: Observational study. *Asian/Pacific Island Nursing Journal*, 8, e48378.
35. Metherall, B., Berryman, A. K., & Brennan, G. S. (2025). Machine learning for classifying chronic kidney disease and predicting creatinine levels using at-home measurements. *Scientific Reports*, 15, 4364.

36. Ganie, S. M., Dutta Pramanik, P. K., Mallik, S., & Zhao, Z. (2023). Chronic kidney disease prediction using boosting techniques based on clinical parameters. *PloS one*, 18(12), e0295234.
37. Debal, D. A., & Sitote, T. M. (2022). Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data*, 9, 1–19.
38. Islam, M. A., Majumder, M. Z. H., & Hussein, M. A. (2023). Chronic kidney disease prediction based on machine learning algorithms. *Journal of pathology informatics*, 14, 100189.
39. Khan, N., Raza, M. A., Mirjat, N. H., Balouch, N., Abbas, G., Yousef, A., & Touti, E. (2024). Unveiling the predictive power: a comprehensive study of machine learning model for anticipating chronic kidney disease. *Frontiers in artificial intelligence*, 6, 1339988.
40. Ghosh, S. K., & Khandoker, A. H. (2024). Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific reports*, 14(1), 3687.
41. Chowdhury, N. H., Hossain, F., Rahman, K., et al. (2021). Performance analysis of conventional machine learning algorithms for identification of chronic kidney disease in type 1 diabetes mellitus patients. *Diagnostics*, 11(12), 2267.
42. Al-Rasheed, A., Saqib, S. M., Asghar, M. Z., Mazhar, T., Osman, A. S. A., Shahid, M., Iqbal, M., & Khan, M. A. (2025). Classifying kidney disease using a dense layers deep learning model. *SLAS technology*, 33, 100324.
43. Chowdhury, M. N. H., Bin Ibne Reaz, M., Ali, S. H. M., Crespo, M. L., Ahmad, S., Salim, G. M., Haque, F., Ordóñez, L. G. G., Islam, M. J., Mahdee, T. M., Zaman, K. S., Hemel, M. S. K., & Bhuiyan, M. A. S. (2025). Deep learning for early detection of chronic kidney disease stages in diabetes patients: A TabNet approach. *Artificial intelligence in medicine*, 166, 103153.
44. Rao, P. K., Chatterjee, S., Nagaraju, K., Khan, S. B., Almusharraf, A., & Alharbi, A. I. (2023). Fusion of Graph and Tabular Deep Learning Models for Predicting Chronic Kidney Disease. *Diagnostics*, 13(12), 1981.
45. Alsekait, D. M., Saleh, H., Gabralla, L. A., Alnowaiser, K., El-Sappagh, S., Sahal, R., & El-Rashidy, N. (2023). Toward Comprehensive Chronic Kidney Disease Prediction Based on Ensemble Deep Learning Models. *Applied Sciences*, 13(6), 3937.

46. Tian, S., Yu, Y., Shi, K., Jiang, Y., Song, H., Wang, Y., Yan, X., Zhong, Y., & Shao, G. (2024). Deep learning radiomics based on ultrasound images for the assisted diagnosis of chronic kidney disease. *Nephrology (Carlton, Vic.)*, 29(11), 748–757.
47. Priyadharshini, M., Muruges, V., Samkumar, G. V., Chowdhury, S., Panigrahi, A., Pati, A., ... Sahu, B. (2025). A population based optimization of convolutional neural networks for chronic kidney disease prediction. *Scientific Reports*, 15, 14500.
48. <https://www.kaggle.com/datasets/marcklugerhart/chronic-kidney-failure-dataset-reboot> Last Access 14/12/2025