Democratic and People's Republic of Algeria
Ministry of Higher Education and Scientific Research
University of 8th May 1945 - Guelma
Faculty of Mathematics, Computer Science, and Material Sciences
Computer Science department

Master's Thesis

Speciality : Computer Science
Option: Information and Communication Science and Technology

Predicting Chronic Kidney Disease: A Machine Learning Approach to Early Detection and Risk Assessment

---

Presented by:
Ilyas GUETTAF

Jury Members :

- **President :** Dr. Adel BENAMIRA
- **Supervisor :** Dr. Hiba ABDELMOUMENE
- **Examiner :** Dr. Karima BENHAMZA
- **Examiner :** Dr. Ghania BERKAT

June 2025

Acknowledgment

First and foremost, all praise and gratitude are due to Allah, the Most Merciful and Compassionate, for granting me the strength, guidance, and perseverance to complete this work.

I extend my deepest appreciation to my supervisor, **Dr. Hiba ABDELMOUMENE**, whose unwavering support shaped my academic journey profoundly. Her guidance began long before my formal supervision, illuminating my educational path with wisdom and patience. Her unique ability to motivate through calm, insightful dialogue, coupled with her exceptional professionalism and dedication, provided me with both reassurance and inspiration at every stage.

I am immensely grateful to **Dr. Lotfi LAMOURI** and **Dr. BELKHAMSA** from Oued Zenati's El Amir Abdelkader Hospital for their invaluable mentorship. Their expertise and generosity in sharing knowledge significantly enhanced my research and optimized my approach to this work.

My sincere thanks go to the jury members **Dr. Adel BENAMIRA**, **Dr. Karima BENHAMZA**, and **Dr. Ghania BERKAT** for their time, thoughtful feedback, and constructive insights.

I also wish to acknowledge all professors in the Computer Science department for their collective support and wisdom. A special mention goes to **Dr. Zakaria ZENDAOUI**, whose encouragement and advice were pivotal to my growth.

Dedication

To my father, **Dr. Mohamed GUETTAF**, my lifelong idol, whose wisdom in research lights my path and whose unwavering presence anchors my spirit.

To my mother, my pillar of strength, for believing in me fiercely—especially when I doubted myself—and for turning my failures into lessons with her grace.

To my aunts, my second mothers, for filling my world with love, guidance, and the warmth of a home.

To my three brothers: Abderrahmane, Adnane and Taha. your loyalty is my fortress.

To my beloved nephews, the twins: Nouh & Youcef, may your futures shine brighter than mine.

To my cousin, colleague, and brother : Ayoub MESSAHEL for walking every step of this journey beside me.

To my friends groupe : thank you for lightening the load, absorbing pressure, and consistently shifting negativity into positivity—your steady support kept me going.

To my classmate and colleague: your camaraderie made mountains feel like hills.

To my entire constellation of family—you are my why, my how, and my always.

Abstract

Chronic Kidney Disease (CKD) is a major global health concern, often developing silently until reaching advanced stages, which makes early prediction vital for timely medical intervention. This research addresses the challenge of predicting CKD onset six months in advance by leveraging both laboratory and clinical data sources. While most existing models either rely on clinical datasets lacking biological markers or laboratory datasets with limited size and availability, our work proposes a hybrid approach to combine the strengths of both.

We first trained a Deep Neural Network (DNN) on the UCI laboratory-oriented dataset to detect CKD using biological parameters. This model was then used as a feature extractor in a transfer learning strategy applied to the NHIRD clinical dataset, which contains extensive claims data but lacks laboratory indicators. Our goal was to assess the impact of incorporating learned biological patterns into clinical prediction tasks.

The proposed transfer learning-based model demonstrated strong performance, particularly in terms of recall, achieving a true positive rate of 92% for predicting CKD six months before clinical onset. These results confirm the added value of integrating laboratory-derived knowledge into large-scale clinical prediction systems, and highlight the feasibility of using such models for real-world healthcare applications, especially in contexts where lab data is scarce.

Key words : Chronic Kidney Disease, Early prediction, Clinical data, Laboratory parameters, Transfer Learning

Résumé

La maladie rénale chronique (MRC) constitue un problème majeur de santé publique à l'échelle mondiale, évoluant souvent de manière silencieuse jusqu'à des stades avancés, rendant ainsi la prédiction précoce essentielle pour permettre une intervention médicale efficace. Ce travail de recherche s'attaque à la problématique de la prédiction de l'apparition de la MRC six mois à l'avance en exploitant à la fois des données biologiques et cliniques. Alors que la plupart des approches existantes se basent soit sur des données cliniques dépourvues de marqueurs biologiques, soit sur des jeux de données biologiques de taille limitée, notre étude propose une approche hybride qui combine les avantages des deux types de données.

Nous avons d'abord entraîné un réseau de neurones profond (DNN) sur le jeu de données UCI, orienté laboratoire, afin de détecter la MRC à partir de paramètres biologiques. Ce modèle a ensuite été utilisé comme extracteur de caractéristiques dans une stratégie d'apprentissage par transfert appliquée au jeu de données NHIRD, basé sur des données cliniques et administratives mais dépourvu de marqueurs biologiques. L'objectif était d'évaluer l'impact de l'intégration des connaissances issues des données biologiques dans des tâches de prédiction cliniques.

Le modèle proposé, basé sur l'apprentissage par transfert, a montré de très bonnes performances, notamment en termes de rappel, avec un taux de vrais positifs de 92% pour la prédiction de la MRC six mois avant sa manifestation clinique. Ces résultats confirment la valeur ajoutée de l'intégration des connaissances biologiques dans les systèmes de prédiction clinique à grande échelle, et soulignent la faisabilité de leur utilisation dans des contextes réels, en particulier lorsque les données biologiques sont rares.

Mots clés : Maladie rénale chronique, Prédiction précoce, Données cliniques, Paramètres de laboratoire, Apprentissage par transfert

Abstract

يُمثل مرض الكلى المزمن (CKD) عبئًا صحيًا عالميًا كبيرًا، وغالبًا ما يتطور بصمت حتى مراحل متقدمة، مما يجعل التنبؤ المبكر أمرًا بالغ الأهمية للتدخل. وقد أظهرت الأبحاث التي تهدف إلى التنبؤ بظهور مرض الكلى المزمن قبل ستة أشهر نتائج واعدة باستخدام بيانات سريرية واسعة النطاق وبيانات المطالبات (البيانات الديموغرافية، والأمراض المصاحبة، والأدوية)، محققةً نتائج قوية، على الرغم من أن هذه الأساليب تفتقر عادةً إلى معايير مختبرية أساسية. في المقابل، أظهرت الدراسات التي تستخدم مجموعات بيانات غنية بالقيم المختبرية دقة تشخيصية عالية لأمراض الكلى المزمنة الحالية، ولكنها غالبًا ما تواجه تحديات مثل البيانات المفقودة الكبيرة ونقص البيانات. ولسد هذه الفجوة بين أنواع البيانات وتعزيز القدرة التنبؤية للكشف المبكر، طبقنا تقنية التعلم بالنقل. وتحديدًا، تم نقل المعرفة من نموذج مُدرّب على بيانات مختبرية لتحسين نموذج شبكة عصبية عميقة (DNN) مُدرّب على بيانات سريرية. وقد عزز هذا النهج المبتكر الأداء بشكل كبير لمهمة التنبؤ لمدة ستة أشهر، محققًا درجة تذكر عالية بلغت 92%، مما يُظهر قيمة دمج الرؤى من التشخيصات المختبرية لتحسين التنبؤ بالمخاطر المستقبلية باستخدام البيانات السريرية الواقعية.

الكلمات المفتاحية: مرض الكلى المزمن، التنبؤ المبكر، البيانات السريرية، معايير المختبر، نقل التعلم

Contents

List of Figures

List of Tables

General Introduction

General context

Chronic Kidney Disease (CKD) is a growing global public health concern, affecting approximately 10% of the world's population [27]. In 2017, an estimated 843.6 million individuals were living with CKD worldwide. The disease poses severe clinical, social, and economic burdens due to its progressive and irreversible nature, often culminating in End-Stage Kidney Disease (ESKD) that requires dialysis or kidney transplantation.

In Algeria, the situation is particularly concerning. In the southeastern regions, the crude incidence of treated stage 5 CKD reached 75 cases per million inhabitants in 2017, while between 2015 and 2018, the prevalence of ESKD in Sidi Bel Abbes was reported at 805.57 cases per million inhabitants [8]. The increasing demand for renal replacement therapies puts additional strain on already limited healthcare resources, especially in low- and middle-income countries.

Recent advances in Artificial Intelligence (AI)—and more specifically, deep learning—have demonstrated considerable promise in supporting early detection and risk prediction of chronic diseases. In the context of CKD, AI-based tools offer a valuable opportunity to identify high-risk patients earlier, enabling preventive care and better clinical outcomes.

Problem statement and motivation

Chronic Kidney Disease (CKD) is a progressive and often asymptomatic condition that can remain undetected until it reaches advanced stages, where treatment options are limited and costly. This late diagnosis reduces the chances for early intervention, increases the burden on healthcare systems, and leads to poorer patient outcomes.

Given this silent progression, there is a growing need to predict CKD early, assess individual risk levels, and enable preventive strategies. Predictive models based on patient data—whether biological or clinical—can support healthcare professionals in identifying at-risk individuals before symptoms appear, thus enabling more timely care and improved quality of life.

However, the development of such models faces a critical challenge: healthcare data comes from diverse sources. In many cases, only clinical or administrative data (e.g., medical history, prescriptions, comorbidities) is available. In other contexts, biological data from laboratory tests (e.g., creatinine, albumin, eGFR) may be accessible. Each type of data brings different insights, and their integration is essential to improve the reliability and precision of prediction tools.

This issue was also observed during a professional internship conducted in the nephrology department of Oued Zenati, Guelma hospital, where discussions with experts and specialists

highlighted the importance of using both clinical and biological data in CKD risk evaluation. Experts stressed that focusing on just one type of data limits the ability to accurately detect or predict the disease, especially in early stages.
This motivates the need to explore AI-based methods that are capable of leveraging multi-source medical data for reliable and scalable CKD prediction, even in contexts with incomplete or non-standardized information.

Objectives

The primary objective of this thesis is to explore the use of deep learning techniques to enhance the early prediction of Chronic Kidney Disease (CKD). Specifically, the research seeks to investigate how different types of medical data—biological (laboratory) versus clinical (administrative/claims-based)—influence the performance and applicability of predictive models. This work aims to:

- Assess the effectiveness of deep learning methods in predicting CKD at early stages.
- Explore how data-driven models can be designed to predict CKD onset or progression based on patient-level information.
- Compare and analyze the predictive value of biological features (e.g., creatinine, eGFR, albumin) versus clinical features (e.g., comorbidities, medications).
- Explore the feasibility of transferring knowledge learned from laboratory-based models to models operating on clinical data, in contexts where lab data are scarce.
- Assess the impact of data diversity on prediction performance—namely, how different types of data (clinical, biological, or combined) contribute to model accuracy and applicability.
- Contribute to the development of intelligent tools capable of supporting healthcare professionals in decision-making and risk assessment for CKD.
- Support healthcare providers with an AI-powered tool that enables early risk assessment, even in data-limited environments.

Main contributions

The main contributions of this work are:
1. Development of a transfer learning strategy that bridges laboratory-based and clinical data sources to improve CKD prediction performance.
2. Demonstration of the feasibility of leveraging administrative health data for CKD forecasting in the absence of lab tests.
3. Provision of a hybrid deep learning model combining feature extraction and classification capabilities to deliver scalable and personalized prediction.
4. Providing an AI-powered tool aimed at supporting doctors in early detection and management of CKD.

Thesis outline

This thesis is organized into the following chapters:

Chapter 1 - State of the art : Provides an overview of Chronic Kidney Disease, including its stages, diagnostic parameters, and epidemiology. It also reviews key machine learning and deep learning methods used in CKD prediction, and summarizes existing work and datasets.

Chapter 2 - Methodology, Materials and Implementation : Describes the proposed approach, including system architecture, data sources, preprocessing steps, model design, and transfer learning process. The results of each step are presented and analyzed to assess the performance of the predictive framework.

Chapter 1

State of the art

1.1 Introduction

Chronic kidney disease (CKD) is a major contributor to global morbidity and mortality from non-communicable diseases (NCDs). Addressing CKD is critical to achieve the United Nations' Sustainable Development Goal (SDG) Target 3.4, which aims to reduce premature mortality from NCDs by one-third by 2030 through prevention and treatment [W1].

In Algeria, CKD poses a significant public health challenge. According to Professor Hind Arzour, a nephrology expert at Mustapha Pacha Hospital (Algiers), an estimated 2 to 3 million Algerian adults are at risk of developing CKD. While national prevalence data remains limited, studies from high-income countries (e.g., the USA, Canada, the UK, and France) suggest an average CKD prevalence of 1–3 % in the general population, highlighting the need for improved surveillance and intervention strategies [W2].

In this chapter, we present an overview of Chronic Kidney Disease, including its risk factors and clinical characteristics. We will also explore machine learning techniques used for CKD prediction, describe the datasets commonly used in this context, outline evaluation metrics for predictive models, and provide a synthesis of key related research studies.

1.2 Presentation of CKD

Chronic Kidney Disease (CKD) figure 1.1 [W3] is defined as a progressive loss of kidney function that persists for three months or more, regardless of the underlying cause. It is characterized by structural or functional abnormalities of the kidneys, with implications for health. CKD often leads, in its most advanced stages, to the need for renal replacement therapy such as dialysis or kidney transplantation [W4].

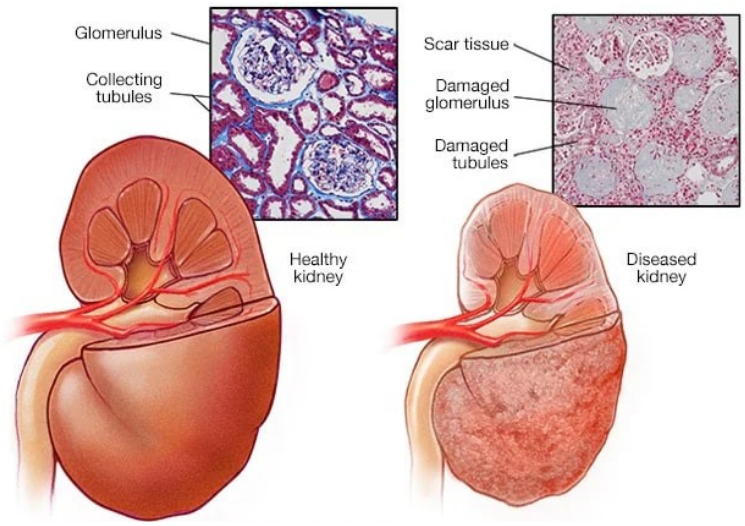


Figure 1.1: Healthy kidney vs Diseased kidney [W3].

The kidneys play a vital role in maintaining overall health. They are responsible for filtering waste products, toxins, and excess fluids from the blood, regulating electrolyte balance, controlling blood pressure, producing hormones that affect red blood cell production, and maintaining bone and mineral health. When kidney function declines, these essential processes are disrupted. In the early stages of CKD, most individuals remain asymptomatic, but as the disease progresses, waste products accumulate in the blood, leading to symptoms such as fatigue, nausea, swelling, and poor appetite. CKD is also associated with complications such as hypertension, anemia, bone disorders, cardiovascular disease, and neurological impairments. These complications often progress silently and may culminate in end-stage renal disease (ESRD), which can occur suddenly and without prior warning.

The diagnosis and staging of CKD primarily rely on two key biomarkers: the estimated glomerular filtration rate (eGFR) and albuminuria.

- **Glomerular Filtration Rate (eGFR)** : eGFR estimates how effectively the kidneys filter blood. A persistently low eGFR ($< 60\ mL/min/1.73m^2$ for $>= 3\ months$) indicates impaired kidney function.

- **Albumin-to-Creatinine Ratio (ACR)** : Healthy kidneys excrete minimal protein. Elevated urinary albumin (albuminuria) signals kidney damage. ACR, measured in a spot urine sample, quantifies albumin (mg) relative to creatinine (g) and is the preferred screening method for CKD [W5].

The KDIGO (Kidney Disease Improving globcal outcomes) guidelines classify CKD progression into: 6 stages based on eGFR (from G1: normal/high eGFR to G5: kidney failure) (figure 1.2 [W6]), and 3 stages based on proteinuria (A1–A3) to reflect albuminuria severity [60].

We present below the classification based on eGFR :

Stages	GFR value	Classification
I	> 90	Normal or High
II	60-89	Slightly decreased
III A	45-59	Mild to moderately decreased
III B	30-44	Moderately to severely dereased
IV	15-29	Severely decreased
V	< 15	Kidney failure

Table 1.1: CKD progression stages

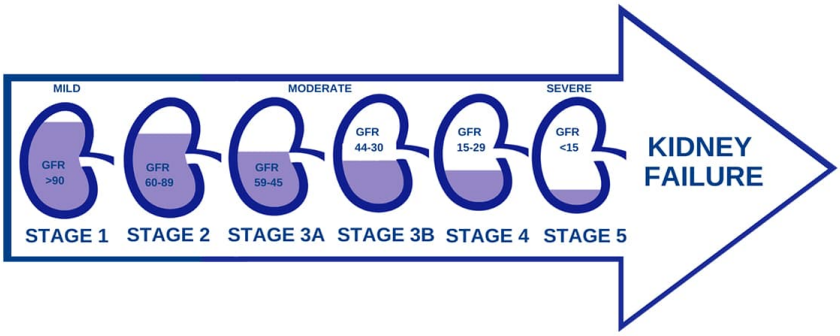


Figure 1.2: CKD Classifiaction based on eGFR [W6].

The following table 1.2 present the classification based on albuminuria.

Category	24-Hour Albuminuria mg/24 h	ACR value	Classification
A1	< 30	< 30	Normal to discrete
A2	30-300	30-300	Moderate
A3	> 300	> 300	Severe

Table 1.2: CKD albuminuria stages

1.2.1 The importance of prediction in CKD management

The effective management of Chronic Kidney Disease (CKD) aims not only to prevent or delay its progression to end-stage renal disease (ESRD), but also to control complications associated with declining kidney function, preserve the patient's quality of life, and ensure timely preparation for renal replacement therapies such as dialysis or kidney transplantation [4].

CKD often develops silently and remains asymptomatic in its early stages, with clinical symptoms typically appearing only in stages 4 or 5. This asymptomatic nature makes the disease difficult to detect early, and contributes to the underestimation of its true incidence and prevalence. Early recognition is therefore crucial, it serves as the first and most essential step toward effective treatment. Identifying the early signs of CKD enables prompt diagnosis and timely intervention, which can significantly slow disease progression and reduce long-term complications [W4].

Because kidney injury is irreversible and disease progression can vary significantly between individuals, early prediction and risk stratification are essential. Predictive models allow clinicians to identify high-risk patients before severe damage occurs, and to personalize management strategies accordingly [31].

Key Benefits of Early Prediction in CKD Management

- **Prevention of disease progression:** Enables timely lifestyle and clinical interventions (e.g., blood pressure and glucose control) to preserve kidney function and delay the onset of ESRD.
- **Reduction in morbidity and mortality:** Allows for earlier detection and management of complications such as cardiovascular disease, one of the main causes of death in CKD patients.
- **Improved Quality of Life:** Helps patients avoid severe symptoms of advanced CKD, such as fatigue, fluid overload, and cognitive decline, thereby maintaining daily functioning and well-being.
- **Optimized Clinical Planning:** Facilitates early nephrologist referral and planning for renal replacement therapy (e.g., timely vascular access or transplant listing).
- **Safer Medication Management:** Enables dose adjustment of renally cleared medications and reduces exposure to nephrotoxic agents.
- **Cost-Effectiveness:** Reduces the economic burden on healthcare systems by avoiding the high costs associated with late-stage dialysis and transplantation.
- **Support for Risk Stratification:** Helps identify and monitor patients with a poor prognosis, allowing clinicians to prioritize interventions for those at greatest risk.

Predictive approaches are vital for transforming CKD care from reactive to proactive, offering clinical, economic, and societal advantages that make early detection and risk assessment central to modern kidney disease management.

1.2.2 Key parameters for CKD prediction

Accurate prediction of Chronic Kidney Disease (CKD) progression requires a comprehensive evaluation of various health indicators, including clinical assessments, biological markers, and associated risk factors. These parameters support early detection, risk stratification, and timely intervention. Depending on their nature, the indicators may be derived from clinical evaluations or laboratory measurements [60].

In the literature, some studies further classify risk factors into two subgroups: non-modifiable risk factors, such as age, gender, and ethnicity, which cannot be changed but provide important background risk context; and modifiable risk factors, including systolic and diastolic blood pressure, proteinuria, and glycemic control, which can be influenced by lifestyle changes or medical interventions [55]. This distinction is essential, as it helps clinicians focus preventive efforts on factors that can be altered to slow the progression of CKD.

In this work, we adopt a classification into three main categories: biological parameters, clinical parameters, and associated risk factors. This organization facilitates a structured assessment of kidney function, the identification of early warning signs, and the estimation of the risk of disease progression.

1.2.3 Clinical Parameters

Clinical parameters are derived from patient history, physical examinations, and non-invasive tests. They provide immediate insights into a patient's health status and potential CKD progression [29].

1. *Blood Pressure (Hypertension)*: Elevated blood pressure is both a cause and consequence of CKD. Persistent hypertension accelerates kidney damage by increasing glomerular pressure.

2. *Diabetes Mellitus*: A leading cause of CKD, diabetes induced hyperglycemia damages nephrons, leading to decreased filtration efficiency.

3. *Age*: Advancing age is associated with a natural decline in glomerular filtration rate (GFR), increasing CKD risk.

4. *Body Mass Index (BMI)*: obesity is a clinical condition often linked to hypertension and diabetes, both of which increase CKD risk.

5. *Anemia*: Common in CKD due to decreased erythropoietin production, resulting in fatigue and reduced oxygen transport.

Other clinical factors may include edema/fluid retention, cardivascular disease, ...

1.2.4 Biological Parameters

Biological parameters are obtained through laboratory tests and provide quantitative measures of kidney function and damage [60], [W4].

1. *Serum Creatinine*: A waste product filtered by the kidneys; elevated levels indicate impaired kidney function.

2. *Estimated Glomerular Filtration Rate (eGFR)*: Calculated using serum creatinine, age, sex, and race; eGFR is a key indicator of kidney function.
3. *Albuminuria (Albumin-to-Creatinine Ratio - ACR)*: The presence of albumin in urine signifies glomerular damage and is a predictor of CKD progression.
4. *Hemoglobin Levels*: Anemia is common in CKD due to decreased erythropoietin production; low hemoglobin levels can indicate disease severity.
5. *Specific Gravity of Urine*: Reflects urine concentration ability; abnormalities may suggest tubular dysfunction.
6. *Electrolytes*: Abnormal levels of potassium, phosphorus, and calcium are frequent in advanced CKD.
7. *Blood Urea Nitrogen (BUN)*: Elevated BUN levels are often observed in patients with reduced kidney clearance capacity.

This list of biological markers is non-exhaustive; these parameters represent the most impactful indicators of kidney function.

1.2.5 Associated Risk Factors

These are conditions or behaviors that increase the likelihood of developing or accelerating CKD [60]), [W4].

1. *Family History of Kidney Disease*: Genetic predisposition plays a role in CKD susceptibility.
2. *Use of Nephrotoxic Medications*: Prolonged use of certain drugs (e.g., NSAIDs) can harm kidney function.
3. *Physical Inactivity*: Sedentary lifestyle is linked to obesity, diabetes, and hypertension, all of which are CKD risk factors.
4. *Smoking*: Tobacco use contributes to vascular damage and worsens kidney outcomes.
5. *Race/Ethnicity*: Certain populations (e.g., African descent) have higher CKD prevalence due to genetic and socio-economic factors.

1.3 ML techniques used in prediction

Traditional Machine Learning (ML) methods continue to be widely employed in medical prediction studies, often demonstrating robust predictive performance, particularly when dealing with structured datasets [30]. The Support Vector Machines (SVM) [24] [36], Random Forest (RF) [51] [43], Decision Trees (DT) [35] [40], Logistic Regression [42] [20], K-Nearest Neighbors (KNN) [5] [39], Naïve Bayes [25] [48], and Gradient Boosting (GB) [50] / XGBoost [32] / AdaBoost [61] are broadly applied across various medical domains, including the prediction and diagnosis of leptospirosis [56], CKD, heart disease [30], Alzheimer's disease, diabetes, hypertension, melanoma, stroke [1], oncology, neurology, and COVID-19 [26].

Deep Learning (DL) models represent a significant advancement over traditional ML methods, consistently demonstrating superior performance, particularly in the analysis of complex medical images and time-series data (temporal data) [30]. Artificial Neural Networks (ANN) [46] [34] [19], Convolutional Neural Networks (CNN) [53] [58] [41], Recurrent Neural Networks (RNN) [14] [11], Long Short-Term Memory (LSTM) [38] [7] / Gated Recurrent Unit (GRU) [57] [44], Transformers [33] [28] [47], Deep Belief Networks (DBN) [3] [45], and Multilayer Perceptron (MLP) [9] [13], most of them were used with time-series data.

For CKD, both traditional ML algorithms and deep learning architectures have been applied, demonstrating significant potential for early detection and prognosis. The following table shows different algorithms used for different fields in medical prediction.

Model class	**Model**	**Fields**
Neural Networks	Feedforward NN/multilayer perceptron (MLP), Convolutional NN (CNN), Recurrent NN and long short-term memory NN (RNN), Auto-encoder, Extreme learning machine	**CKD**, Heart Disease, Occupational pneumoconiosis, Colorectal cancer, diabetic blood glucose prediction, Covid-19, Diabetes mellitus, incident heart failure,
Tree algorithms	Random Forest, Extreme gradient boosting (XGBoost), Decision tree, Gradient boosting machine, Bagged decision trees, Extremely randomized trees, Light gradient boosting machine, Adaptive boosting machine, Categorical boost	Heart Disease, breast cancer, Liver disease, diabetes mellitus, **CKD**, Lung cancer
Support vector machines	Support vector machines, Genetic algorithm based on SVM, Particle swarm optimization SVM, Simulated annealing particle swarm optimization SVM	Heart Disease, Strok, lung cancer, **CKD**
Logistic Regression	Logistic regression, LASSO logistic regression, Ridge logistic regression, Elastic net logistic regression	Heart Disease, **CKD**

Others	k-Nearest neighbors (kNN), Gaussian Naïve Bayes, Ensemble model, Linear regression, (Adaptive) Neuro-fuzzy Inference System, Partial Least Square Regression, Hidden Markov Model (HMM), k-Means, Cox regression, Hierarchical clustering, Genetic programming, Linear discriminant analysis (LDA), Markov decision process (MDP), Hierarchical clustering	Diabetes mellitus, Heart Disease, Lung cancer, **CKD**

Table 1.3: Machine Learning techniques for medical prediction.

1.4 Available datasets

The development of predictive models for Chronic Kidney Disease (CKD) relies heavily on the availability of quality datasets. However, publicly accessible datasets specific to CKD are limited, particularly those encompassing comprehensive laboratory and imaging data. The datasets commonly utilized in CKD research can be broadly categorized into numerical (structured) data and imaging data. Numerical datasets include clinical and laboratory measurements, while imaging datasets comprise modalities like ultrasound, MRI and CT scans [23] [2] [W7]. This section focuses on prominent numerical datasets employed in CKD prediction studies.

1.4.1 The UCI dataset

The UCI (UC Irvine Machine Learning Repository) Chronic Kidney Disease dataset [W8] is the most widely used resources in CKD prediction research [6] [10] [22]. It comprises 400 instances with 24 features, including demographic, clinical, and laboratory variables. The dataset contains 250 instances labeled as CKD and 150 as non-CKD. Its accessibility and inclusion of laboratory-oriented data make it a popular choice for developing and benchmarking machine learning models.

1.4.2 The NHIRD dataset

Another commonly used dataset is the Taiwan National Health Insurance Research Database (NHIRD) [W9] is a comprehensive claims-based database encompassing health records of over 23 million individuals in Taiwan. It includes extensive clinical data such as diagnoses, prescriptions, and procedures. Although access to NHIRD is restricted, it has been employed in several studies to develop predictive models for CKD, particularly focusing on forecasting disease occurrence 6 to 12 months in advance. For instance, a study utilized a cohort of 18,000 CKD patients and 72,000 non-CKD individuals to train models incorporating demographic, comorbidity, and medication data [52] [29].

1.4.3 UAE Hostpital dataset

Another dataset used in CKD prediction research was collected from 544 patients admitted to Tawam Hospital in Al-Ain City, Abu Dhabi, United Arab Emirates, between January and December 2008 [W10]. This dataset includes various clinical and laboratory parameters. It was employed in a study investigating explainable machine learning models for CKD prediction, emphasizing the importance of model interpretability in clinical settings [17].

Table 1.4 gives a summary of the three datasets' usage. While several datasets are available for CKD prediction [54], the UCI Chronic Kidney Disease dataset remains the most utilized due to its accessibility and comprehensive laboratory data. The NHIRD offers a vast repository of clinical information, albeit with access restrictions, and the UAE hospital dataset provides valuable insights into CKD prediction in a specific regional context. The limited number of publicly available datasets underscores the need for more open-access resources to advance research in CKD prediction.

Dataset	**Datatype**	**Strengths**	**Limitation**
UCI	Laboratory	Standardized Biomarkers	Small sample size
NHIRD	Clinical	Longitudinal, large scale	No lab data, access restricted
UAE	Laboratory&Clinical	Regional diversity	Limited laboratory features (8)

Table 1.4: Available datasets details

1.5 Evaluation metrics

The goal of internal validation is to evaluate the predictive performance of an AI-based model using data that were not involved in training but originate from the same population and setting. This process ensures that the model is not simply overfitting the training data. Performance metrics used during internal validation are specifically designed to assess how reliably the model can predict future events within that context.

The performance of an AI-based prediction model should be evaluated through two key aspects: discrimination and calibration. Discrimination measures the model's ability to distinguish between individuals with and without the outcome. Calibration assesses how closely the predicted probabilities align with the actual outcomes [12].

True positives (TP, the cases predicted 1 and the actual output was also 1), false positives (FP, the cases predicted 1 and the actual output was 0), True negatives (TN, the cases predicted 0 and the actual output was 0) and false negatives (FN, the cases predicted 0 and the actual output was 1) are used to calculate several useful metrics for evaluating models. Which evaluation metrics are most meaningful depends on the specific model and the specific task, the cost of different misclassifications, and whether the dataset is balanced or imbalanced.

In this section, we discuss the most widely used evaluation metrics for assessing discrimination in AI-based prediction models in healthcare [12].

Accuracy

Accuracy is the proportion of all classifications that were correct, whether positive or negative. It is mathematically defined as:

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

Even it's simple and very popular, it's not always the best metric to use, because accuracy simplifies things too much, that's why we need to look at other metrics more detailed like precision and recall.

Recall

The true positive rate (TPR), or the proportion of all actual positives that were classified correctly as positives, is also known as recall. Recall is mathematically defined as:

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

Out of everything that was positive, how many of them the model was able to capture.

Precision

Precision is the proportion of all the model's positive classifications that are actually positive. It is mathematically defined as:

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

Out of everything the model labeled as positive, how many of them were actually positive.

F1-Score

The F1 score can be interpreted as a harmonic mean of the precision and recall. It is mathematically defined as:

$$\text{F1} = \frac{2 * TP}{2 * TP + FP + FN}$$

An F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal.

AUC-ROC

The previous set of model metrics, all calculated at a single classification threshold value. But if we want to evaluate a model's quality across all possible thresholds, we need the ROC curve.

The ROC (Receiver-operating characteristic) graph summarizes all the confusion matrices produced by each threshold, by comparing True Positive rate with False Positive rate.

The AUC (Area Under Curve) makes it easy to compare one ROC curve to another. The following figure represent ROC and AUC of two hypothetical models. The curve on the right, with a greater AUC, represents the better of the two models.

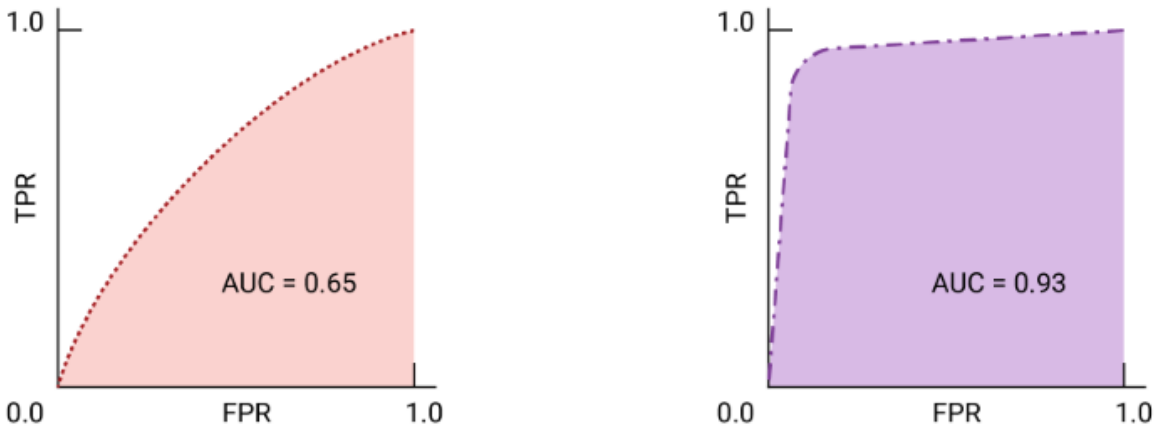


Figure 1.3: ROC and AUC of two hypothetical models.

ROC curves make it easy to identify the best threshold for making a decision, and the AUC can help to decide which categorization method is better.

Metrics discucssion

Precision and Recall are fundamentally different from AUC-ROC; they offer complementary perspectives on evaluating model performance. While AUC-ROC summarizes the trade-off between the true positive rate and false positive rate across thresholds, Precision and Recall provide more focused insight, especially in imbalanced medical datasets like those involving Chronic Kidney Disease (CKD).

When dealing with CKD cases, Recall is often considered the most critical metric. It ensures that the model successfully identifies as many true CKD cases as possible, minimizing false negatives. This is especially important in medical applications, where missing even a single positive case can have serious health consequences. In contrast, Precision focuses on how many of the predicted CKD cases are actually correct, helping to reduce false positives.

In healthcare, and particularly for CKD detection, Recall is typically prioritized over Precision, as early detection and intervention are vital. However, Precision also remains important to avoid overburdening medical professionals with too many false alarms.

In situations where the dataset is imbalanced—with significantly more non-CKD than CKD samples—AUC-ROC becomes a more robust evaluation metric than accuracy, as accuracy may be misleading. In such cases, Precision is also more reliable than the false positive rate, as it is not affected by the large number of true negatives. This is particularly relevant in our study, where the dataset reflects a real-world population in which CKD cases are relatively rare.

In our work, accurately identifying all CKD-positive samples is a critical goal. To achieve this, we may choose to lower the classification threshold to favor higher Recall, even at the

cost of generating more false positives. This strategy ensures that at-risk individuals are not missed, which is essential for early intervention and public health management.

1.6 Summary of related work

To conduct a comprehensive review of the use of machine learning (ML) and deep learning (DL) techniques in the prediction of Chronic Kidney Disease (CKD), we carried out a structured search on the Mendeley database using the query: "Chronic Kidney Disease prediction based ML."

This initial search returned 2,377 research papers. We first filtered the results to retain only scientific journal articles, reducing the number to 1,691. Then, we applied a publication date filter to include only articles published between 2020 and 2024, yielding 1,617 articles. To ensure accessibility, we further narrowed the selection to open access publications, resulting in 152 articles.

These remaining papers were then sorted by citation count, and the top 50 most cited articles were shortlisted. After reviewing the abstracts, introductions, and conclusions, we selected 20 studies that were most relevant to our research.

Following a thorough reading and critical evaluation of these selected works, we identified a core subset that is highly relevant to our specific area of interest in CKD prediction using ML and DL techniques. This subset is summarized in Table 1.5.

Work	**FS technique**	**Algorithms**	**Dataset**	**Dataset issues (solution)**	**Metrics**
[22]	filter feature selection approach	ANN, AdaBoost, DT, XGBoost, CatBoost, KNN, RF, GB, Stcoh GB, LGBM, Extra Tree, SVM, HML	UCI dataset	-Missing Values (KNN) -Unbalanced (Stratified folds)	ACC = 0.983 PREC = 0.98 REC = 0.98 F1 = 0.98 For XGboost

[6]	chi-square test (Chi2), recursive feature elimination (RFE), and mutual information.	pretrained DL models with SVM as the metalearner model	UCI dataset	None	ACC = 0.996
[17]	fruit fly optimization algorithm (FFOA), improved teacher-learner-based optimization (ITLBO), correlation-based feature selection (CFS), and the Apriori algorithm.	Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), and Extreme Gradient Boosting (XGBoost)	UAE dataset	irrelevant or redundant features (FS)	AUC of 0.9689 and an accuracy of 93.29% for XGBoost
[16]	None	logistic regression, decision tree, XGBoost, RF, SVM, AdaBoost, CS AdaBoost	UCI Dataset	-Missing values (mean imputation technique) -Unbalanced (biasing the weighting technique)	ACC = 0.993 for 6m 0.992 for 12m For ensemble model
[10]	Correlation-based, Wrapper method, LASSO regression	ANN, C5.0, CHAMID, logistic regression, LSVM with penalty L1 and with penalty L2, Random Tree	UCI Dataset	-Missing values () -Unbalanced (SMOTE)	ACC = 98.86 REC = 100% PRE = 96.67% AUC = 100% F-ME = 98.3% GINI = 0.99 For LSVM L2

[29]	LightGBM	Logistic regression, decision tree, RF, CNN, BLSTM, LightGBM	90,000 instances of the NHIRD	None	AUROC = 0.957 for 6m 0.954 for 12m For CNN
[52]	none	CNN, LSTM, Deep ensemble model (CNN + LSTM + LSTM-BLSTM)	90,000 instances of the NHIRD	none	ACC = 0.993 for 6m 0.992 for 12m For ensemble model
[59]	None	Gradient Boosting Machine (GBM)	De-identified electronic health records (EHRs) of 14,039 adult patients with type 2 diabetes	None	AUC : At year 2 since diabetes onset: 0.83 , At year 3: 0.78 , At year 4: 0.82
[15]	None	Logistic Regression, XGBoost, Stochastic Gradient Descent classifier	23,948 instances from NHIRD dataset	None	AUC = 0.77 for LR

Table 1.5: Related works

Since the datasets are divided into clinical and laboratory data, we can classify the previous studies into clinical-oriented studies and biological-oriented studies.

Laboratory-oriented studies

Several studies used laboratory-oriented datasets to develop CKD predictive models. The work in [22], Islam et al. applicate a collection of 12 prediction models on "UCI dataset", with the greatest performance were for the XgBoost classifier. The features selection was discovered that hemoglobin, albumin, and specific gravity had the biggest impact when it comes to predicting CKD. In order to increase model's generalization performance, a significant amount of a more sophisticated data will be used for training the model in the future.

The feature selection of the UCI dataset positive impact was improved on the performance of the various classifiers on [16]. CKD screening time and cost was saved thanks to the few clinical test attributes identified (18 out of 24) and needed for the diagnosis. The IG technique

ranked albumin, hemoglobin, packed cell volume, red blood cell count, and serum creatinine as the most informative features. Giving more attention to the minority class, AdaBoost trained with the reduced feature set achieved the best classification performance. This predictive model could be applied to more imbalanced medical datasets. The use of reduced feature set decreases the computational cost of training the models, which performed better than those trained with the complete feature set.

Seven machine learning models and one deep learning model were combined with three feature selection technique and the UCI dataset in [10]. The LASSO regression was the best FS technique, where the six most important features were rbc, pc, al, ba, su and pcc. But it has been shown that LASSO regression with SMOTE outperform the LASSO regression without SMOTE. Among classic ML technics, the LSVM with penalty L2 using SMOTE with full features gave the best performance, the LVSM also outperform all the 6 others when it comes to SMOTE with LASSO's selected features. It has been noted that deep neural network achieved the highest accuracy of 99.6, the DNN gave strong result and important features were extracted by itself.

The study [17] utilized feature selection techniques such as FFOA, CFS, and ITLBO to enhance model performance by identifying relevant clinical attributes. Multiple algorithms including LR, RF, DT, NB, and notably XGBoost were evaluated, with XGBoost achieving the highest accuracy (93.29%) and AUC (0.9689). The dataset consisted of clinical and demographic data from 544 patients at Tawam Hospital in UAE, with feature selection addressing issues like redundancy and overfitting. Model performance was assessed using accuracy and AUC metrics, highlighting XGBoost as a highly effective and interpretable model for CKD prediction.

Clinically-oriented studies

On another hand, some studies tried to predict CKD using clinical-oriented datasets.In [29], the tree-based LightGBM model is used because of his ability to capture complex relationship on large features space, to see that the most prominent features was diabetes mellitus, age, gout, and use of sulfonamides and angiotensin. Wherein the aim was predicting CKD 6–12 months in advance using 90,000 instances of "the NHIRD dataset". Among the machine learning models, deep neural networks (CNN and BLSTM) chosen because they took advantage of temporal information, and outperformed the classical models. The CNN model performed best for the 6-month and 12-month predictions. In term of computing, these models could be efficiently used in resource management because they were not very large and complex. For the application of such models into clinical practice dealing with individual patients, the feature set would have to be expanded to include laboratory measurementsand possibly lifestyle information, which falls within the scope of future work.

CNN, LSTM and a deep ensemble model are the three predictive models proposed in the research [52] for CKD prediction within 6 or 12 months earlier based on medication, demographic, and comorbidity data of two different public benchmark datasets obtained from Taiwan's NHIRD, where one of this method advantages is that it does not need laboratory data as related studies in this field. The Ensemble model fuses three base deep learning classifiers (CNN, LSTM, and LSTM-BLSTM) using the majority voting technique. The authors choose the Ensemble learning algorithms, because ML research has shown that combining the output of several individual classifiers can reduce generalization errors and perform better in many

applications than individual deep learning classifiers, and are able to extract features without human's intervention. This study used the research [29] as a comparative paper, and the proposed model performed better. This ensemble model needs more memory storage and longer learning time than deep learning models which requires more memory and learning time than traditional machine learning techniques. Laboratory data is needed for clinical validation, they plan to test the robustness of their developed models against datasets based on patient laboratory data collected from various sources. There are a lack of previously unknown features in the dataset, where the risk factors for this disease, such as a family history of kidney failure, hypertension, and diabetes, were not determined.

The study [59] evaluated the performance of the proposed Landmark-Boosting model for predicting 1-year diabetic kidney disease (DKD) risk in patients with type 2 diabetes, achieving high discrimination and calibration across multiple landmark times. Specifically, the model reached an AUROC of 0.83 at year 2, indicating excellent predictive ability, and maintained strong performance at years 3 and 4 with AUROCs of 0.78 and 0.82, respectively. It also demonstrated superior sensitivity (83%) and specificity (78%) compared to other temporal models, while maintaining good calibration as reflected by favorable observed-to-expected risk ratios. The model effectively integrated longitudinal electronic health record data, adapting dynamically over time to improve risk stratification, and outperformed other approaches in both discrimination and calibration metrics across the study period.

Dovgan et Al. [15] appliyed machine learning algorithms such as Logistic Regression, XGBoost, and SGD to predict the need for renal replacement therapy within 12 months in CKD patients, based solely on diagnoses and comorbidities from Taiwan's NHIRD dataset. The dataset included over 19,000 patients diagnosed between 1998 and 2011, but lacked laboratory and personal characteristic data, which posed limitations. The researchers applied different feature extraction methods but did not perform feature selection or dimensionality reduction in the best models, achieving an AUC of approximately 0.77. Despite dataset limitations and potential biases, the models demonstrated promising predictive performance, supporting their potential use for healthcare planning in resource-limited settings.

While numerous studies have explored CKD prediction using either laboratory-oriented or clinical-oriented datasets, few have addressed the challenge of combining insights from both sources to enhance early diagnosis and risk assessment in resource-constrained settings. Most laboratory-based studies leverage rich biological parameters to train highly accurate models but suffer from limited sample sizes and generalizability. Conversely, studies utilizing large-scale clinical datasets like Taiwan's NHIRD demonstrate high performance at the population level but lack critical laboratory measurements required for individualized decision-making.

Our proposed work addresses this gap through a novel transfer learning approach. We aim to bridge the divide between laboratory and clinical data by first training a robust deep learning model on the UCI CKD dataset, which contains rich biological markers such as serum creatinine, hemoglobin, and albumin. This model is then used as a feature extractor, transferring its learned representations to a second model trained on the NHIRD dataset, which includes large-scale clinical records (demographic, comorbidity, and medication data) but lacks laboratory data.

1.7 Conclusion

Chronic Kidney Disease (CKD) represents a growing global health concern, characterized by its asymptomatic onset and progressive nature. Early detection and risk assessment are crucial for slowing disease progression, optimizing treatment strategies, and reducing the economic and societal burden associated with end-stage renal failure. In this context, artificial intelligence—particularly machine learning (ML) and deep learning (DL) approaches—offers promising avenues for enhancing CKD prediction.

This chapter presented a comprehensive overview of CKD, including its clinical definition, stages, diagnostic parameters, and key risk factors. We then explored the types of data used in predictive modeling, distinguishing between biological/laboratory data and clinical/claims-based records. The availability and characteristics of major CKD datasets were discussed, highlighting their strengths and limitations in both research and real-world applications.

We also reviewed a broad range of recent studies applying ML and DL techniques to CKD prediction. These works revealed valuable insights into effective algorithms, feature selection strategies, and performance metrics. While many studies achieve high accuracy using laboratory data, others have demonstrated scalable solutions using clinical data alone. However, a persistent gap remains: integrating the predictive power of laboratory data with the accessibility and scale of clinical datasets.

This observation motivates the approach proposed in this thesis: to investigate the effectiveness of transfer learning from a laboratory-based dataset (UCI) to a large clinical dataset (NHIRD), aiming to build a hybrid model that supports early and accurate CKD prediction across diverse healthcare environments.

In the following chapter, we present the methodology, materials, and implementation details of our proposed solution, including system architecture, preprocessing steps, model design, and evaluation procedures.

Chapter 2

Methodology, Materials and Implementation

2.1 Introduction

After reviewing the theoretical foundations and related studies on CKD prediction, this chapter presents the proposed approach, which relies on a transfer learning strategy. The goal is to leverage a deep learning model trained on a laboratory-oriented dataset (UCI) to enhance prediction capabilities on a large-scale clinical dataset (NHIRD) that lacks biological markers. This approach aims to bridge the gap between detailed but limited lab data and scalable clinical data, enabling more robust and generalizable CKD prediction.

This chapter details the methodology adopted throughout the study, including system design, data preparation, model development, and implementation. It is organized to first introduce the overall system architecture, followed by descriptions of the data sources, preprocessing steps, feature selection, base model training, and the transfer learning process. Each step is explained alongside its implementation outcomes.

2.2 Problematic and motivation

Chronic Kidney Disease (CKD) is a "silent" condition, often remaining asymptomatic until irreversible damage occurs [37]. Early detection is critical to slow its progression and prevent complications such as transplantation, dialysis, or death. However, traditional reliance on biological laboratory tests—such as estimated glomerular filtration rate (eGFR) and albuminuria—poses several challenges:

- Limited accessibility: Laboratory data are often unavailable in large-scale administrative health databases limiting their use for population-wide screening and early detection strategies.
- Delayed diagnosis: In clinical settings, patients may not undergo routine laboratory testing until symptoms manifest, resulting in missed opportunities for early intervention.

In response to these limitations, claims-based datasets like NHIRD have been increasingly used to train predictive models for CKD onset using features such as comorbidities and med-

ication histories [29]. These models have demonstrated strong performance at the population level, confirming the potential of clinical data for large-scale forecasting. However, these works [29] acknowledge that claims data alone are sufficient for epidemiological forecasting, but that integrating laboratory markers remains essential for clinical decision-making and patient-level management. This emphasizes the complementary value of biological markers for fine-grained, individual-level assessment.

This context underscores a key research motivation: to explore how combining clinical and biological data may enhance the clinical utility of CKD prediction models. Specifically, there is a need to investigate whether leveraging both types of data—via approaches such as transfer learning—can enable scalable, risk-sensitive tools that serve both broad health monitoring and individualized care. By evaluating this hybrid strategy, the aim is not only to build adaptable predictive systems but also to assess the added value of each data source in different clinical scenarios.

2.3 Objectives

The main objective of this study is to develop a deep learning–based approach for the early detection and risk assessment of Chronic Kidney Disease (CKD) by leveraging both biological laboratory data and clinical data from multiple sources.

Specifically, this work aims to:

- Build a predictive model using biological parameters to identify CKD cases from laboratory-based data,
- Apply transfer learning to adapt the biological model for use with clinical and claims-based data.
- Assess the ability of clinical data to support CKD prediction in the absence of laboratory results.
- Investigate the combined impact of integrating both data types on CKD risk assessment and prediction.
- Provide a framework that supports scalable and individualized prediction of CKD in different healthcare contexts.

2.4 Proposed approach

To address the challenge of early detection and prediction of Chronic Kidney Disease (CKD), we propose a deep learning–based approach that leverages both biological and clinical data sources. The method combines two distinct datasets through a transfer learning framework: a base model is first trained on biological parameters from the UCI CKD dataset, then its learned representations are transferred to a new model built on clinical and administrative data from the NHIRD dataset. This hybrid approach enables the use of laboratory-specific insights even in environments where lab data are unavailable, thus improving the adaptability of the prediction system across various healthcare settings.

2.4.1 System architecture

The overall system architecture is illustrated in 2.1. It outlines the two-stage training process and data flow used in our study. The pipeline begins with the preprocessing and feature selection steps applied separately to the UCI and NHIRD datasets. A deep learning model is first trained on the UCI dataset using laboratory data, and the resulting base model is then used as a feature extractor. In the second phase, this extracted representation is transferred to initialize the NHIRD model, which is subsequently trained on clinical and medication-based features to perform CKD risk prediction. This diagram presents a modular view of each component in the pipeline, from data preprocessing to final prediction.

- Step 1: Preprocess and expand features in the UCI dataset
- Step 2: Train a DNN (base model) on UCI for CKD detection
- Step 3: Save and reuse the UCI model by extracting learned representations
- Step 4: Apply preprocessing and feature selection on NHIRD
- Step 5: Build a new DNN model on NHIRD, initialized with transferred features
- Step 6: Predict CKD risk based on NHIRD data

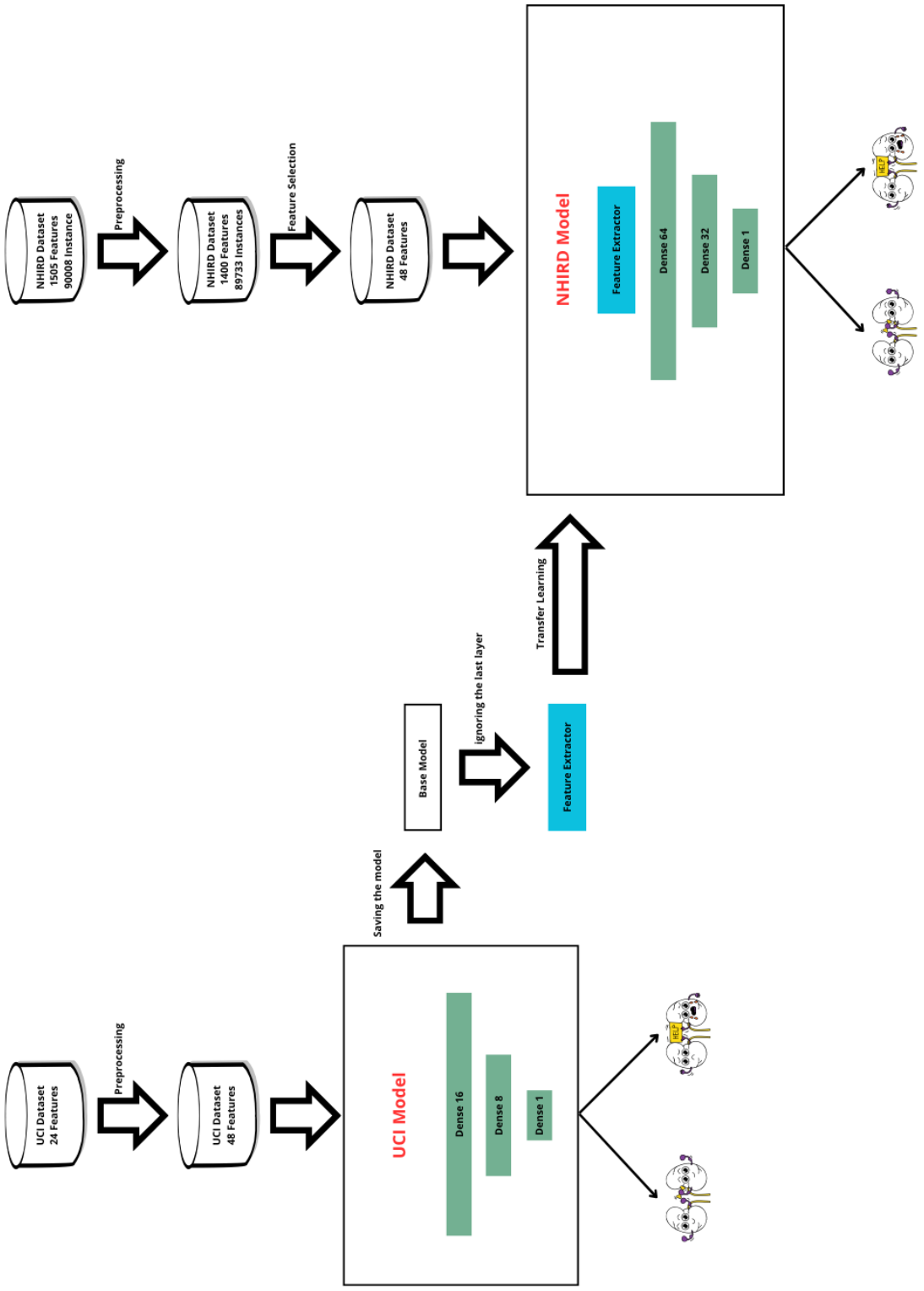


Figure 2.1: System architecture

2.4.2 Data sources description

This study is based on two distinct datasets that provide complementary types of information for Chronic Kidney Disease (CKD) prediction. The first dataset, derived from laboratory measurements, enables the development of a baseline model based on biological markers. The second dataset consists of large-scale administrative and clinical data, allowing for extended prediction via transfer learning. These datasets are described below.

UCI dataset

The first dataset used in this study was obtained from the University of California Irvine (UCI) Machine Learning Repository [49]. It is commonly used in CKD prediction studies due to its accessibility and the nature of its features, which reflect typical laboratory test results that can be collected within a short clinical observation period (approximately two months).

The dataset contains 400 patient records, of which 250 are labeled as positive for CKD and 150 as negative. It includes 24 features, comprising 13 categorical attributes and 11 numeric ones, along with a binary class label: 1 indicating the presence of CKD; 0 indicating its absence. This dataset serves as the foundation for training the base model focused on biological data.

Table 2.1 provides a detailed breakdown of the categorical and numerical features.

Table 2.1: UCI Dataset Variables Description

Name	**Role**	**Type**	**Description**	**Units**	**Missing Values**
age	Feature	Integer	Age	year	yes
bp	Feature	Integer	blood pressure	mm/Hg	yes
sg	Feature	Categorical	specific gravity		yes
al	Feature	Categorical	albumin		yes
su	Feature	Categorical	sugar		yes
rbc	Feature	Binary	red blood cells		yes
pc	Feature	Binary	pus cell		yes
pcc	Feature	Binary	pus cell clumps		yes
ba	Feature	Binary	bacteria		yes
bgr	Feature	Integer	blood glucose random	mgs/dl	yes
bu	Feature	Integer	blood urea	mgs/dl	yes
sc	Feature	Continuous	serum creatinine	mgs/dl	yes
sod	Feature	Integer	sodium	mEq/L	yes
pot	Feature	Continuous	potassium	mEq/L	yes
hemo	Feature	Continuous	hemoglobin	gms	yes
pcv	Feature	Integer	packed cell volume		yes
wbcc	Feature	Integer	white blood cell count	cells/cmm	yes
rbcc	Feature	Continuous	red blood cell count	millions/cmm	yes
htn	Feature	Binary	hypertension		yes
dm	Feature	Binary	diabetes mellitus		yes
cad	Feature	Binary	coronary artery disease		yes

Continued on next page

Name	Role	Type	Description	Units	Missing Values
appet	Feature	Binary	appetite		yes
pe	Feature	Binary	pedal edema		yes
ane	Feature	Binary	anemia		yes
class	Target	Binary	ckd or not ckd		no

NHIRD dataset

The second dataset used in this study is clinically oriented and was obtained from Taiwan's National Health Insurance Research Database (NHIRD) [W9]. The NHIRD is a large-scale administrative claims database that includes electronic health records (EHRs) of over 99% of Taiwan's population, covering longitudinal data such as patient demographics, diagnoses, prescriptions, and medical procedures from 1997 to 2012.

This rich dataset enables researchers to perform population-level studies on various chronic diseases, including CKD. The NHIRD offers access to multiple data types for approved research, including:

- Sampling datasets (2 million patients),
- Disease-specific databases, and
- Full-population datasets.

In the work of [29], the sampling dataset was used to generate three distinct data representations:
1. Aggregated format (used in our work).
2. Monthly temporal sequences, and
3. Quarterly temporal sequences.

For the aggregated data that we use in this study, the temporal dimension was discarded by summing the total occurrences of each comorbidity and medication code across a predefined observation window. As a result, each patient is represented as a fixed-length feature vector, where each feature corresponds to the frequency of a comorbidity or a prescribed medication, along with demographic information such as age and sex.
The final processed dataset contains 1504 features, including diagnosis codes (ICD-9), medication codes (ATC), and basic demographic attributes. In our work, we use the 6-month aggregated version of this dataset, which includes clinical records of patients over a six-month observation window before CKD diagnosis or a matched index date.
Table 2.2 provides a detailed description of the variables used from the NHIRD dataset and figure 2.2 presents a fragment of this dataset.

Table 2.2: NHIRD Dataset Variables Description

Variable Name	Role	Type	Description	Units	Missing Values
age	Feature	Numerical	Age	year	No
gender	Feature	Categorical	Gender	0/1	No
Diagnosis	Feature	Numerical	ICD-9 based frequencies of visits with a diagnosis	Number of diagnoses	No
Medication	Feature	Numerical	ATC-based frequencies of prescriptions		no
CKD	Target	Categorical	ckd or not ckd		No

	id	ckd	age	sex	1	2	3	4	5	6	7	8
1	720898	1	84	1	0	0	0	0	0	0	0	0
2	32772	1	54	1	0	0	0	0	0	0	0	0
3	622596	1	86	1	0	0	0	0	0	0	0	0
4	786441	1	75	0	0	0	0	0	0	0	0	0
5	9	1	49	1	0	0	0	0	0	0	0	0
6	458764	1	71	1	0	0	0	0	0	0	0	0
7	196622	1	30	0	0	0	0	0	0	0	0	0
8	262159	1	62	1	0	0	0	0	0	0	0	0
9	655375	1	78	1	0	0	0	0	0	0	0	0
10	98319	1	70	1	0	0	0	0	0	0	0	0
11	98322	1	55	0	0	0	0	0	0	0	0	0
12	32787	1	81	1	0	0	0	0	0	0	0	0
13	917524	1	67	1	0	0	0	0	0	0	0	0
14	65557	1	11	0	0	0	0	0	0	0	0	0
15	131090	1	75	0	0	0	0	0	0	0	0	0

Figure 2.2: NHIRD dataset fragment

2.4.3 Preprocessing pipeline

UCI Dataset Preprocessing

The following figure 2.3 illustrates the preprocessing pipeline applied to the UCI dataset. This dataset presents several challenges, notably the presence of missing values. To address this, missing values in numerical features were imputed using the median strategy, while missing values in categorical features were handled using the most frequent value strategy.
The class distribution consists of 150 CKD instances and 250 non-CKD instances, which did not require any balancing techniques due to the relatively small imbalance. Categorical variables were encoded using one-hot encoding, generating 24 additional binary features, resulting in a total of 48 features for model training.

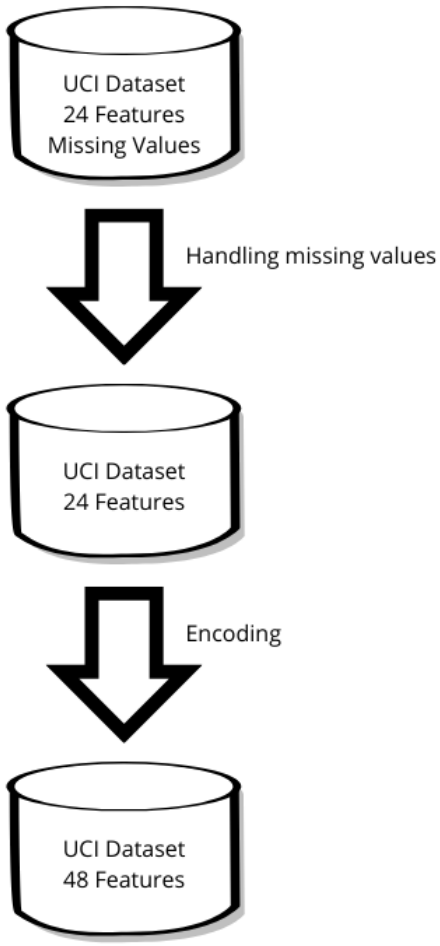


Figure 2.3: UCI dataset preprocessing pipeline

NHIRD Dataset Preprocessing

Unlike the UCI dataset, the NHIRD dataset does not suffer from missing values. However, it presents a significant class imbalance, with a disproportionately high number of non-CKD samples compared to CKD cases. This issue can be addressed using class weighting during model training.

The preprocessing pipeline begins by removing irrelevant or outlier data, such as the ID column and instances where the patient's age exceeds 100. Additionally, the dataset contains numerous zero-variance features—columns in which all observations have the same value, resulting in no variability. Such features do not contribute to model learning and may increase complexity.

To address this, a zero-variance selector was applied, reducing the number of features by approximately 100. The following figure2.4 illustrates the preprocessing steps performed on the NHIRD dataset.

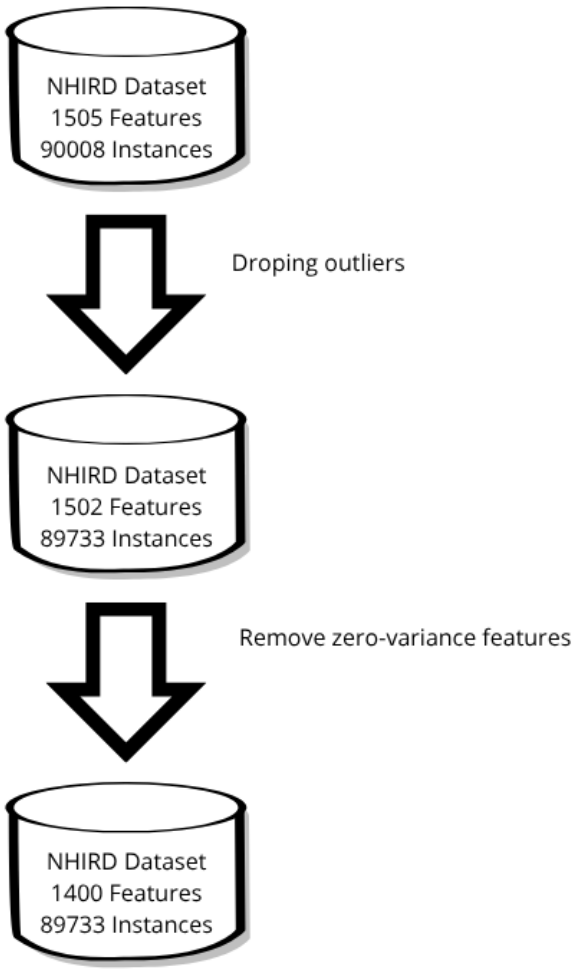


Figure 2.4: NHIRD dataset preprocessing pipeline

2.4.4 Feature selection

For the UCI dataset, the feature space is relatively limited, with only 24 original attributes. Therefore, applying feature selection to reduce dimensionality is not necessary. Additionally, since the model trained on this dataset will serve as a base model for transfer learning, it is preferable to retain as many informative features as possible. This ensures greater representation capacity and improves the generalization potential of the learned model.

In contrast, the NHIRD dataset has a much larger feature space—initially composed of 1,505 features. To identify the most relevant attributes and reduce this dimensionality to 48 features (to match the UCI feature extractor input size), we applied several feature selection methods.
Before performing advanced feature selection, we first removed zero-variance features using the VarianceThreshold technique. This preprocessing step eliminates features with the same value across all samples, reducing the number of features from 1,505 to approximately 1,400.

Subsequently, we applied different feature selection methods, including Mutual Information (MI), ANOVA (F-test), and LightGBM-based importance ranking. All three approaches produced comparable sets of top features, with slight variations in importance scores.
After evaluating the stability and interpretability of the results, the Mutual Information-based selection was chosen to generate the final 48-feature set used in the transfer learning process.

Mutual information

This method involves using the mutual information classifier Algorithme to select features with the highest mutual information (MI) with the target variable.

Diabetes mellitus, Essential hypertension, Gout, Disorders of lipoid metabolism, and Chronic glomerulonephritis are the most prominent comorbidities. The most prominent medications are Sulfonamides, Sulfonylureas, Angiotensin II receptor blockers (ARBs), Biguanides and Dihydropyridine derivatives.

The following table show the most prominent features sorted by MI score.

Table 2.3: Top Features by Importance

Feature	**Details**	**Importance Score**
250	Diabetes mellitus	0.022190
C03CA	Sulfonamides	0.017595
A10BB	Sulfonylureas	0.017444
C09CA	Angiotensin II receptor blockers	0.015826
A10BA	Biguanides	0.015363
C08CA	Dihydropyridine derivatives	0.015337
401	Essential hypertension	0.013846
274	Gout	0.013058
272	Disorders of lipoid metabolism	0.012767
M04AA	Preparations inhibiting uric acid production	0.011156

This feature selection method was adopted to create the training and testing subsets used to train the NHIRD-based model, as illustrated in figure 2.5 The method's ability to capture linear relationships was a key reason for its selection. Its effectiveness is further supported by the resulting feature set, which aligns well with established medical knowledge and was validated by a hospital nephrologist.

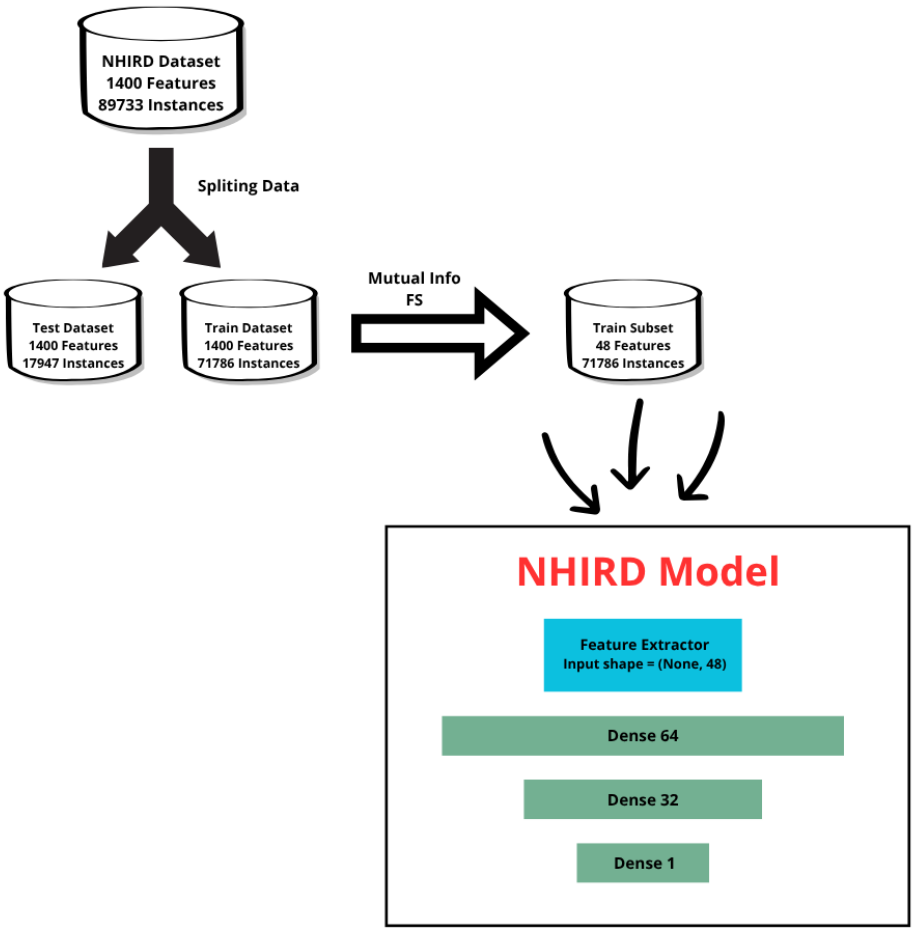


Figure 2.5: Feature Selection process to creat train subset

2.4.5 Model training on UCI CKD

Our base model, trained on the laboratory-oriented UCI dataset, is a simple Deep Neural Network (DNN) designed to predict the presence or absence of Chronic Kidney Disease (CKD). Before selecting this architecture, several machine learning models were implemented and evaluated. The DNN demonstrated the best performance while maintaining simplicity and generalizability, making it well-suited for transfer learning.
The model architecture consists of three dense (fully connected) layers interleaved with two dropout layers to prevent overfitting.
The following table presents a summary of the model's architecture and parameter details.

Table 2.4: The architecture of the UCI DNN model

Layer (type)	**Output Shape**	**Parameters**
dense27 (Dense)	(None, 16)	784
dropout7 (Dropout)	(None, 16)	0
dense28 (Dense)	(None, 8)	136
dropout8 (Dropout)	(None, 8)	0
dense29 (Dense)	(None, 1)	9
Total params: 931		
Trainable params: 929		
Non-trainable params: 0		
Optimizer params: 2		

The base model trained on the UCI dataset demonstrated strong performance, particularly in detecting early stages of CKD. This effectiveness motivated its use as a feature extractor in our transfer learning approach for the NHIRD model.
The following table presents the performance metrics of the UCI-based DNN model.

Model	**Accuracy**	**Precision**	**Recall**
UCI-DNN	0.9812	0.9901	0.9800

Table 2.5: Results for the UCI data

2.4.6 Transfer learning strategy

Transfer learning [18] [21]is the improvement of learning on new task throught the transfer of knowledge from a related task, that has already been learned. Taking a real-life example, learning to ride a bicycle is very difficult and requires learning from scratch how to maintain balance, how to steer the wheel. Once learning how to ride the bicycle, learning how to ride a motorcycle will not be difficult, and it will not be necessary to learn from scratch how to maintain balance and other skills, where riding bicycle skills are transferred, and learning how to ride a motorcycle now is easier.

In the machine learning context, transfer learning is a technique that enable algorithms to learn a new task by using pre-trained models. The following figure shows the transfer learning steps.

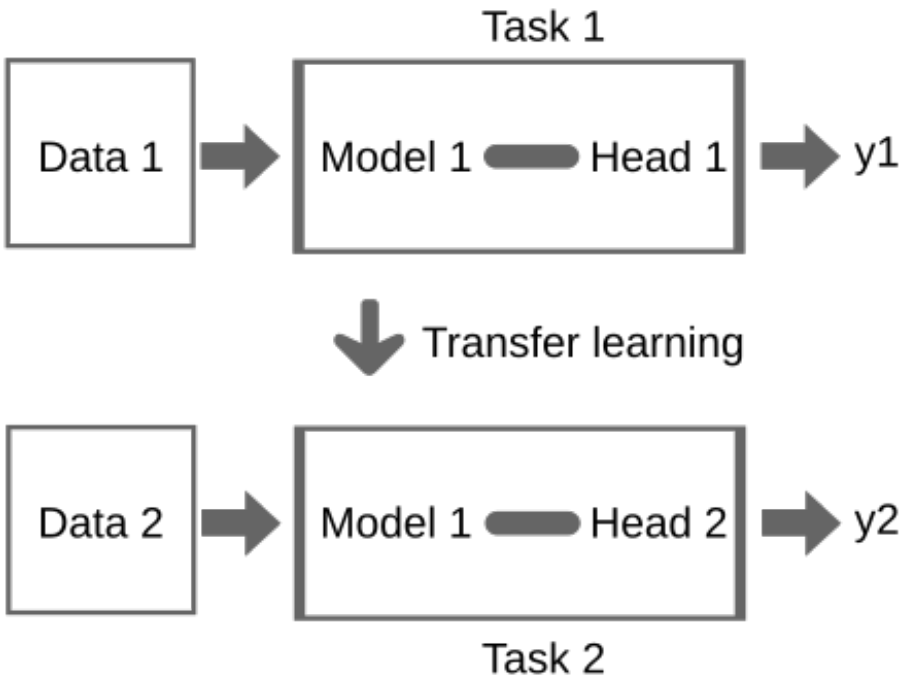


Figure 2.6: Transfer Learning

To transition from early-stage detection of CKD to predicting its onset six months in advance, we applied a transfer learning strategy using our pre-trained DNN model built on the UCI dataset. The original model, designed for binary classification of CKD based on laboratory data, demonstrated strong capability in early detection. However, our goal is now to specialize this model to forecast CKD progression ahead of time using clinical data from the NHIRD dataset.

In transfer learning, the approach depends on how similar the source and target tasks are. When the tasks are closely related—as in our case—we typically remove only the output layer of the pre-trained model and retain the rest of the architecture. We then append new layers, including a new output layer, to adapt the model to the target task. Since our new objective (predicting CKD six months before occurrence) is highly similar to the original task (early detection), we opted to keep all layers except the final one, which we replaced with a new classifier adapted to the prediction task.

The rationale behind this is rooted in how deep learning models work: the deeper the layer, the more abstract the features it captures. Therefore, the intermediate layers of the UCI model are likely to encode useful high-level representations of CKD progression. However, the final output layer of the original model was specifically trained to detect current CKD status—not future risk. By removing it and fine-tuning the rest of the model, we allow the new layers to learn how to leverage these extracted features for forward-looking prediction.

We chose transfer learning over training a model from scratch on NHIRD data because NHIRD lacks laboratory biomarkers, which are essential for building a rich feature space. Transfer learning helps bridge this gap by reusing the knowledge learned from lab-based data and adapting it to clinical-only data.

In our implementation, we first imported the UCI-trained model as a base model. We then extracted all layers except the output layer to form what we refer to as a feature extractor. This sub-model was integrated into the architecture of the NHIRD model. Finally, we appended new layers and fine-tuned the full model, allowing the transferred knowledge to adapt to the clinical prediction task.

A code example of how the feature extractor was defined is shown in the following figure.

```
feature_extractor = Model(inputs=baseModel.get_layer('dense_27').input,
                          outputs=baseModel.get_layer('dropout_8').output)
```

Figure 2.7: Feature Extractor creating code

2.4.7 Evaluation protocol

We split the NHIRD dataset into 80% for training and 20% for testing. From the training set, we generated 72 batches, each containing 400 samples, in order to train 72 distinct models. Each of these models was then evaluated on the entire test set to assess generalization performance.

After training, we analyzed the classification reports of all models. Any model that misclassified samples from the NOCKD class was excluded from further consideration. As a result, only five models were retained for final evaluation.

To select the best-performing model, we used the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) as our primary metric. This metric illustrates the trade-off between Recall (True Positive Rate or Sensitivity) and Precision (or Specificity) across various classification thresholds. Given the imbalance in the dataset, we prioritized Precision over the False Positive Rate in our evaluations.

Figure 2.8 presents the AUC-PR plots for the top five models trained on the 6-month prediction task. Among them, three models achieved an equal AUC-PR score of 0.43. From these, we selected the model with the highest Recall score (0.92, batch 15) as the best-performing model for the final prediction task.

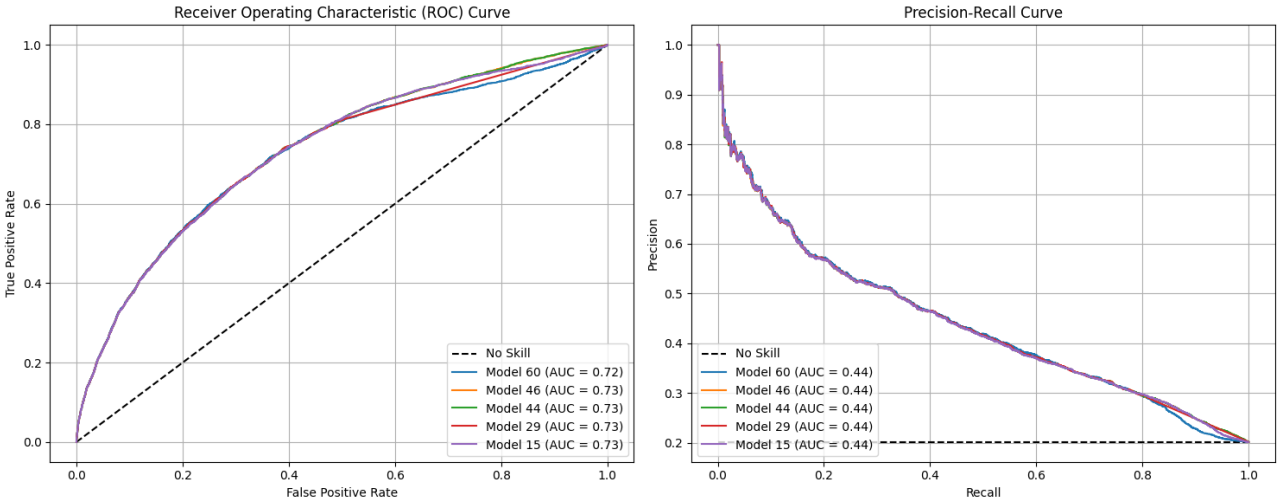


Figure 2.8: ROC and PR curves of best 6 months models.

The following table presents the Accuracy, Recall, and AUC scores of the best model for the 6-month prediction task. In this experiment, we compared four models:

1. Transfer Learning – DNN-based: A Deep Neural Network (DNN) was trained on the UCI dataset, and its learned feature extractor was reused for prediction on the NHIRD dataset using a DNN architecture.

2. Transfer Learning – CNN-based: A Convolutional Neural Network (CNN) was trained on the UCI dataset, and its feature extractor was transferred for prediction on the NHIRD dataset using a CNN architecture.

3. From Scratch – DNN-based: A DNN model trained directly on the NHIRD dataset without any transfer learning.

4. From Scratch – CNN-based: A CNN model trained directly on the NHIRD dataset without transfer learning.

This comparison aims to evaluate the effectiveness of transfer learning by comparing models that benefit from prior knowledge (learned from laboratory data) against those trained solely on clinical data.

Type	**Model**	**Accuracy**	**Recall**	**AUC**
With TL	DNN	0.39	0.92	0.73
	CNN	0.57	0.80	0.74
Without TL	DNN	0.71	0.59	0.74
	CNN	0.78	0.44	0.73

Table 2.6: Results for 6 months prediction data

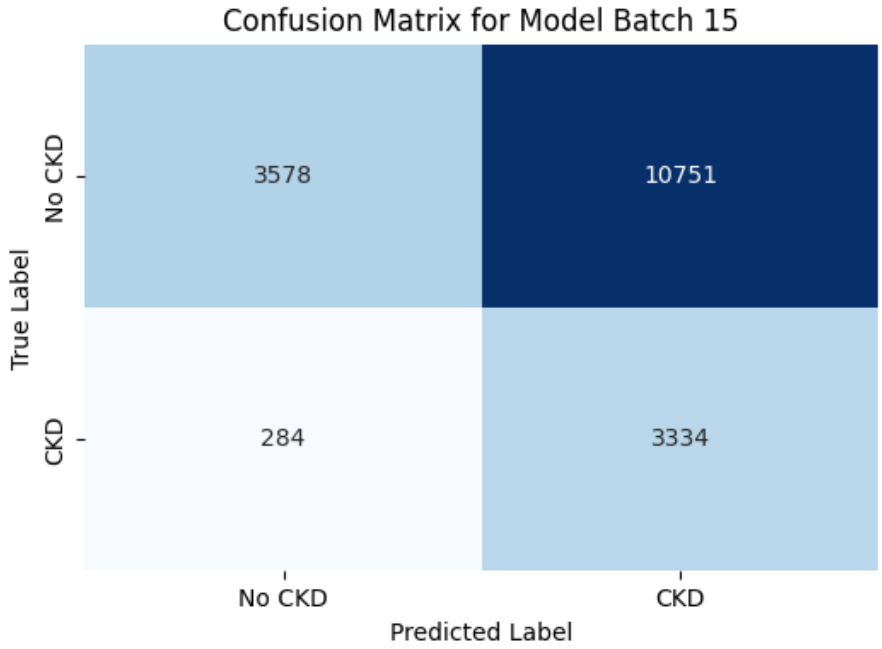


Figure 2.9: Confusion Matrix of the 6 months model.

Study	Model	Accuracy	Recall	AUC
Our model	DNN	0.39	0.92	0.73
	CNN	0.57	0.80	0.74
[29]	LR	0.73	0.66	0.76
	RF	0.72	0.65	0.76
	DT	0.73	0.62	0.74

Table 2.7: Results comparison with the work of [29].

2.5 Discussion

Our proposed model demonstrated strong potential in detecting Chronic Kidney Disease (CKD) cases, particularly through its high Recall value of 0.92, indicating that 92% of actual CKD cases were correctly identified. This high True Positive Rate (TPR) is a significant strength, especially in the context of class-imbalanced datasets, where false negatives (i.e., undetected CKD cases) are far more costly than false positives. In clinical settings, missing a CKD case could delay diagnosis and treatment—hence, prioritizing Recall is both justified and essential.

However, as expected in recall-optimized models, this came at the cost of lower Precision, meaning that a higher proportion of the predicted CKD cases were false positives. This is a common trade-off in imbalanced classification tasks and reflects the reality that over-predicting CKD is less harmful than missing it. The use of balanced batches and class weighting (with a doubled penalty for misclassifying CKD cases) contributed to this recall-focused behavior.

Although the AUC-ROC score reached 0.73, suggesting moderate overall discrimination, it may overestimate performance in imbalanced settings. This is because the ROC curve relies on the False Positive Rate (FPR), which is impacted by the large number of True Negatives (TN), making FPR appear lower than it truly is. In contrast, the AUC-PR score provided a more realistic view of performance under class imbalance, revealing challenges in maintaining high precision while preserving recall. This insight confirms our decision to prioritize AUC-PR and Recall over AUC-ROC and Accuracy.

Notably, the transfer learning (TL) strategy enhanced the model's ability to detect CKD cases, compared to models trained from scratch. This was observed in both Recall and Accuracy improvements. Despite the absence of laboratory data in the NHIRD dataset, the transfer of knowledge from the biologically rich UCI dataset allowed our model to make more clinically informed predictions, validating the added value of transfer learning in bridging the data-type gap.

Another strength lies in the model simplicity: our architectures (DNN) were not overly deep or computationally expensive, which makes them deployable in real-world clinical environments. While we did not exploit temporal features in the NHIRD dataset, we compensated through smart feature selection and cross-domain transfer learning.

When compared to prior work, such as the study in [29], which relies on clinical features using 6-month aggregated data, our method achieved superior Recall scores, reinforcing the idea that transfer learning can compensate for the lack of lab data.

This work shows that prioritizing high recall, even at the cost of precision, is an effective strategy for CKD screening tools, especially when deployed for patients already flagged as clinically suspicious. This minimizes missed diagnoses, aligning well with the practical needs of nephrology departments.

2.6 AI web-Application for CKD Prediction

RenalGuardian is the name of the web application that we have developed to use multu-source data based on transfer learning for CKD prediction.
Firstly, we used Html, Css and JavaScript in the front-end side to provide users a friendly interface. The back-end side was based on the Python programming language, which provide us the ability to use the Flask library to link the interfaces, and to load the NHIRD model to use it to generate predictions, as it is trained and saved using the same programming language.

The figure 2.10 present the main page, which provide details about our system, and several buttons to explore the whole website.

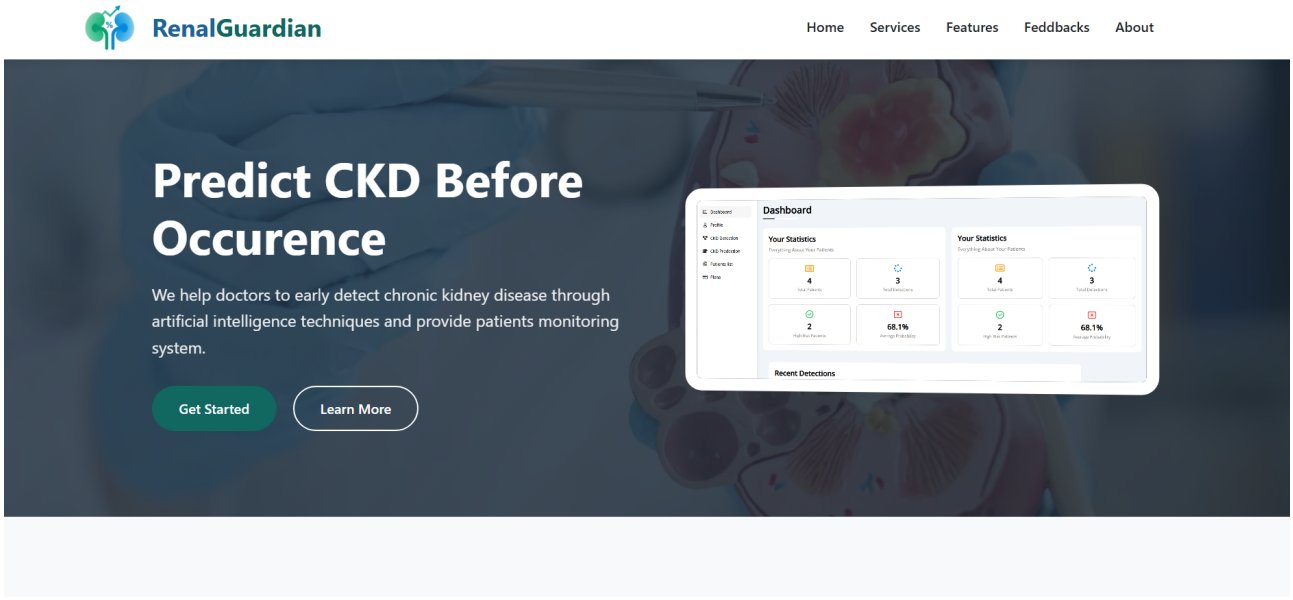


Figure 2.10: The main page.

As soon as the user clicks the Get Started button, he will be redirected to the login/signup page to finish login or registration process as shown in the figure 2.11.

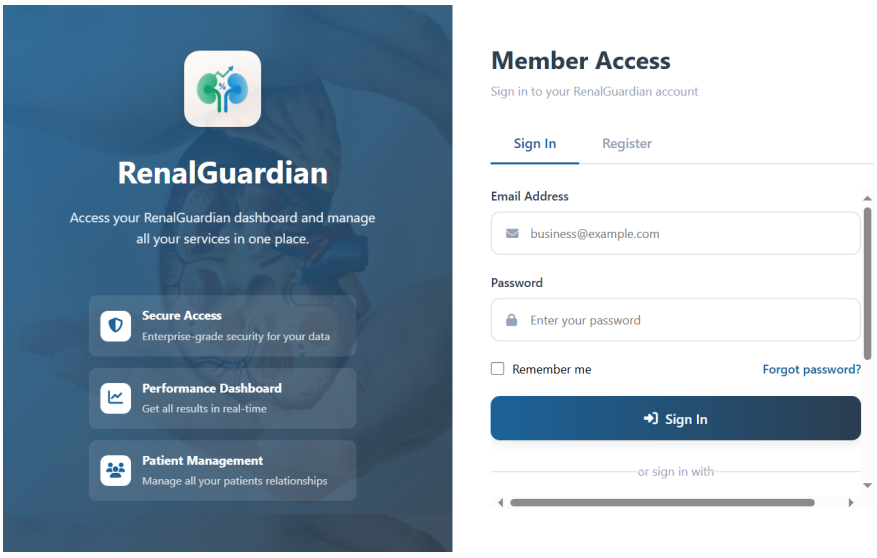


Figure 2.11: The login/signup page.

Once connected, we provide a dashboard contains many important information (figure 2.12).

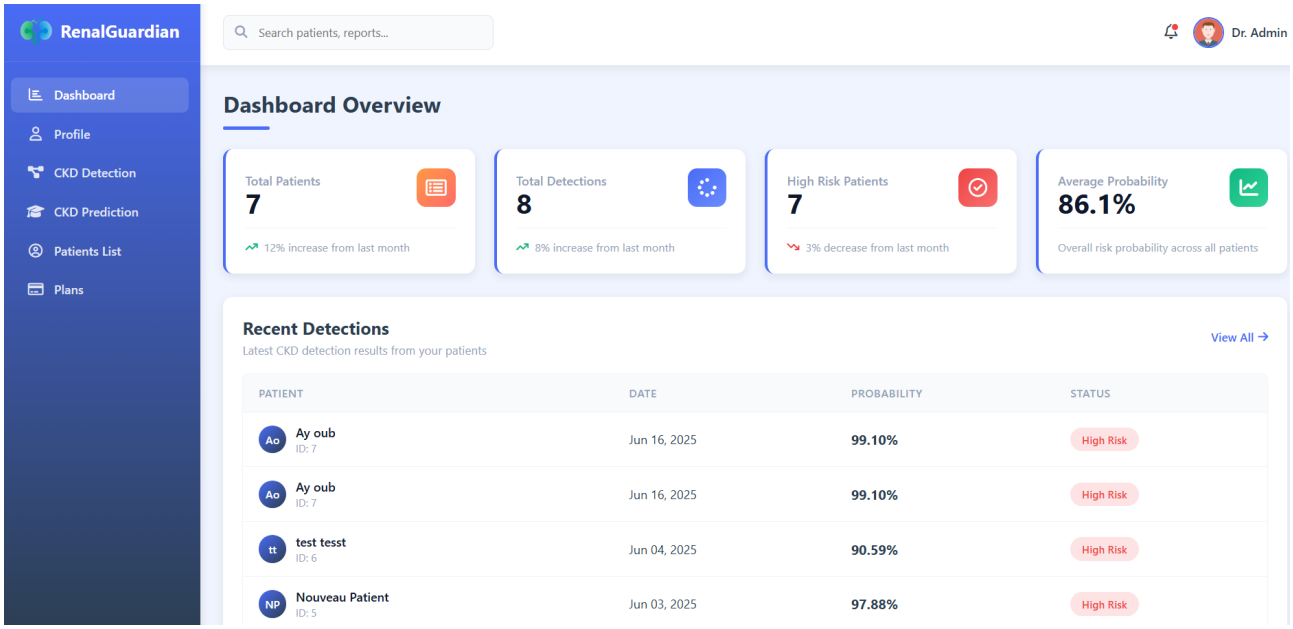


Figure 2.12: The dashboard interface.

2.6.1 Prediction scenario

The core of our system involves using transfer learning from biological data to demographic and clinical data to generate a prediction probability. Firstly, The user must provide the demographic data, include the name, age, sex and other information. Then, adding clinical data is an easy and clear process, the user can add as many as possible comorbidities with the corresponding number of diagnoses, the same thing for the medications.

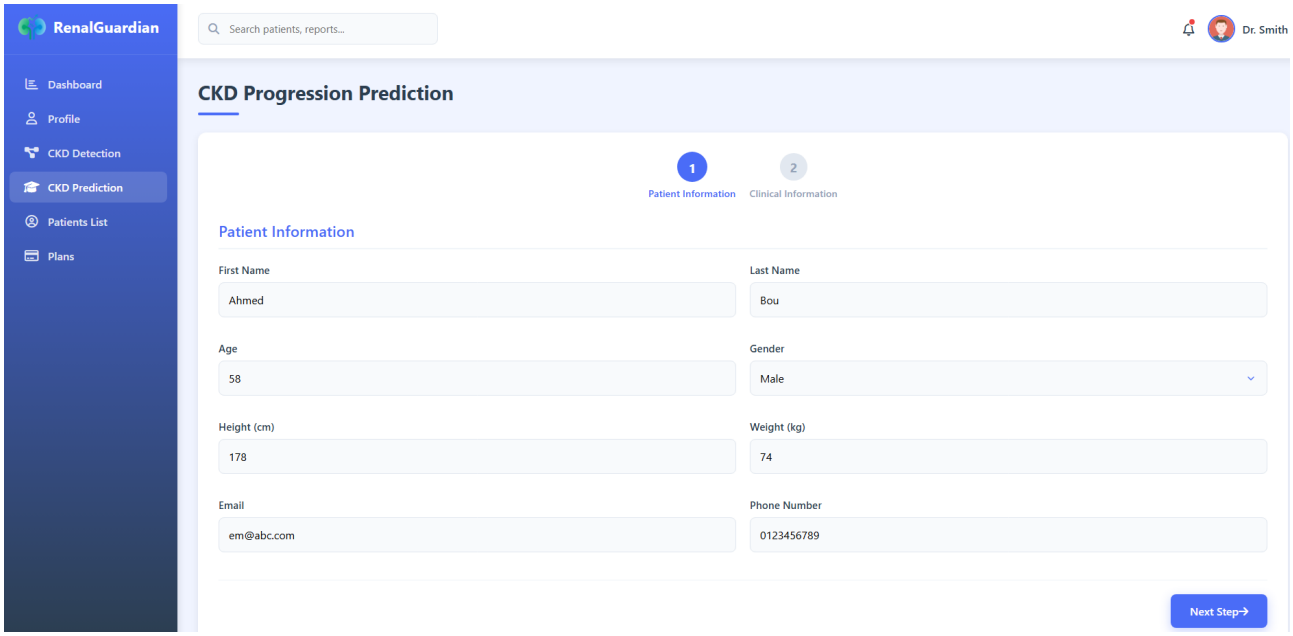


Figure 2.13: Adding demographic information.

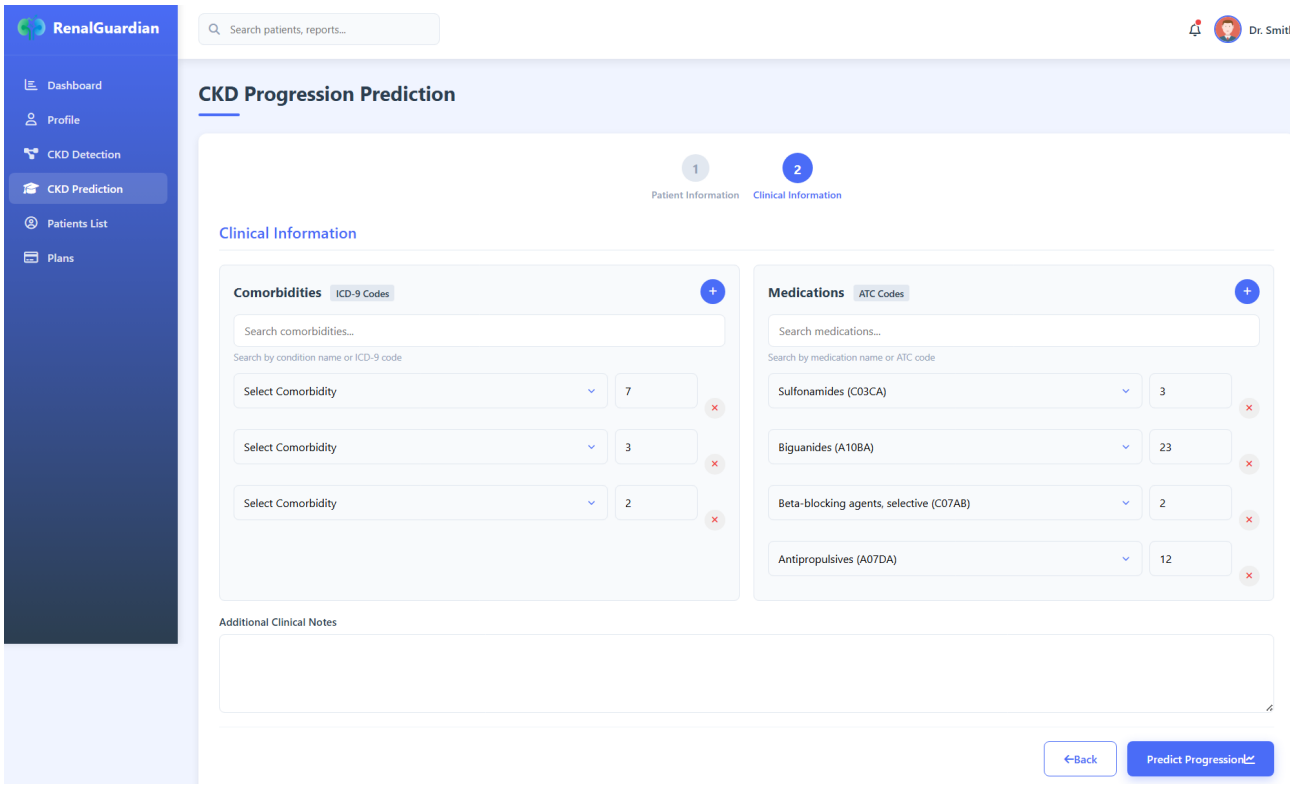


Figure 2.14: Adding clinical information.

After filling the two forms sections, and when clicking on Predict button, our TL-NHIRD model receive the patient data from the front-end form, generate a prediction probability,

then send the probabilty back to the front-end to display prediction results. Based on this probability, we classify the case as Low Risk, Normal Risk, or High Risk.

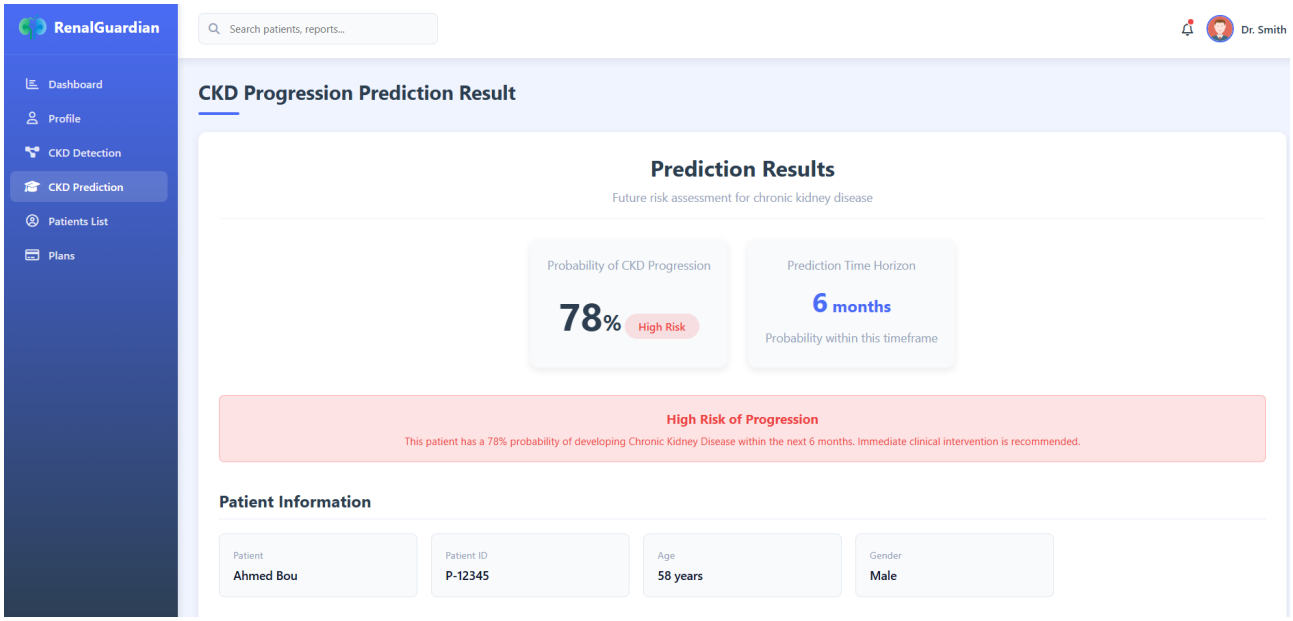


Figure 2.15: Prediction results.

2.6.2 Detection scenario

A similar process will be used by the user to get CKD current stage based on Age, clinical, and laboratory data. Two forms section must be filled, to finally get the probability of being an early stage CKD patient or NOT.

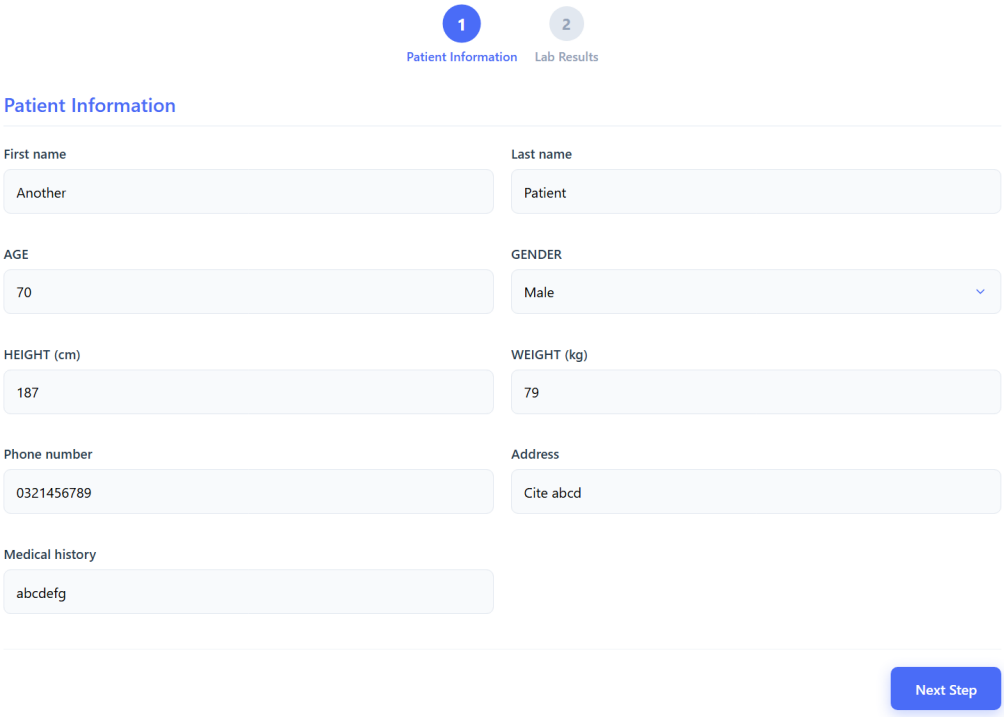


Figure 2.16: Adding demographic information.

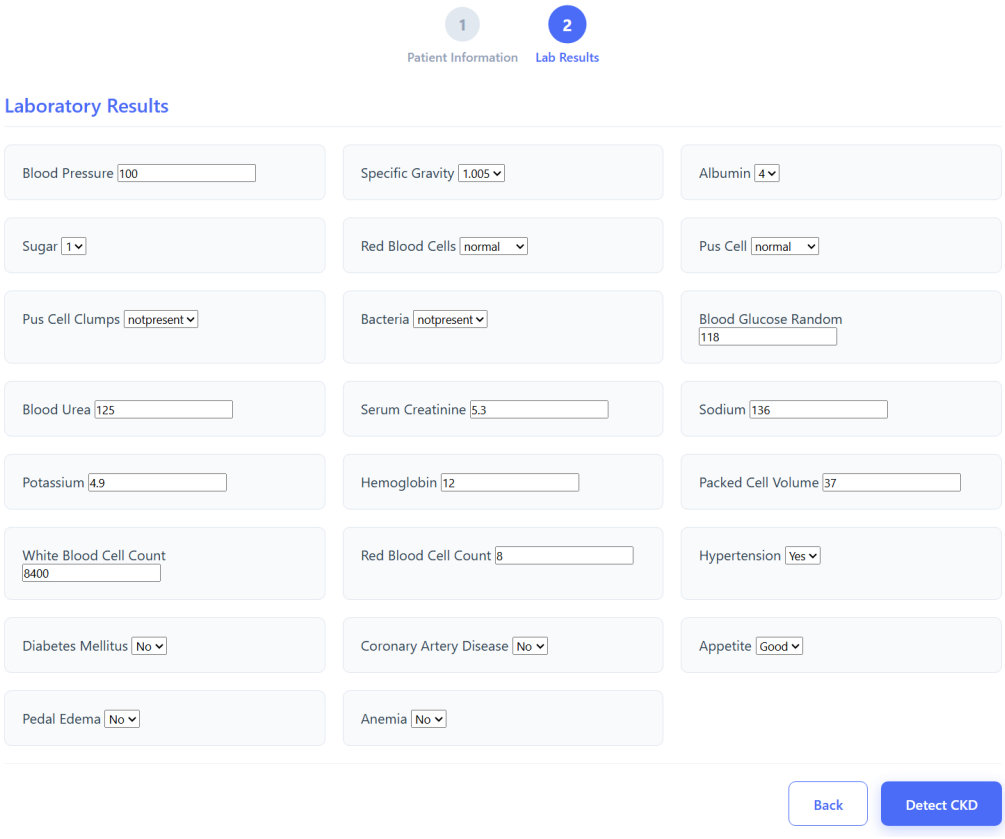


Figure 2.17: Adding clinical and biological information.

Finaly, results page will contain a probability of being CKD patient, and we calculate the eGFR value using the MDRD equation, and we detect the stage basing on the eGFR value. Additionally, the results page contains the details provided, some clinical recommendations, and some useful buttons.

Figure 2.18: Early Detection results.

2.7 Conclusion

This chapter presented the methodology, data sources, and technical implementation underpinning our approach to early CKD prediction. We began by discussing the motivation behind leveraging both biological and clinical data and outlined our system architecture designed to bridge the gap between laboratory-rich datasets and administrative claims data.

We introduced the two main datasets used in our study—the UCI laboratory dataset and the NHIRD clinical dataset—detailing their characteristics, preprocessing pipelines, and challenges. Feature selection techniques were applied to reduce dimensionality and improve model

efficiency, particularly for the NHIRD dataset.

Our approach included training a base model using a Deep Neural Network (DNN) on the UCI dataset, which served as a foundation for transfer learning. The knowledge learned from laboratory-based prediction was transferred to a DNN-based model trained on the NHIRD dataset to predict CKD occurrence six months in advance.

We also developed and presented an AI-based application integrating these models, aimed at assisting healthcare professionals in detecting and predicting CKD. This tool is designed to support clinical decision-making through user-friendly interfaces and actionable prediction outputs.

Through careful model design, batch processing, and evaluation, we implemented and fine-tuned multiple models. The use of transfer learning demonstrated its potential in enhancing CKD prediction performance, especially in data-limited clinical environments.

General Conclusion

Chronic Kidney Disease (CKD) represents a growing global health concern, characterized by its silent progression and serious complications in late stages, including end-stage renal disease (ESRD). The socio-economic and medical impact of CKD is profound—ranging from the financial burden of dialysis and transplantation to the deterioration of patient quality of life and increased mortality. Consequently, early prediction and risk stratification have become critical to reducing morbidity, improving care delivery, and optimizing healthcare resources.

Despite advancements in machine learning (ML) for medical prediction, the field of CKD prediction still faces important challenges—primarily due to fragmentation in data sources. On one hand, large-scale clinical and claims datasets (e.g., Taiwan's NHIRD) offer wide population coverage but lack essential laboratory biomarkers required for accurate clinical decision-making. On the other hand, laboratory-oriented datasets (e.g., the UCI CKD dataset) provide high-resolution biological data but are limited in size and scope, restricting their generalizability.

To address this gap, our research proposes a hybrid machine learning framework based on Transfer Learning (TL). The key idea was to leverage the diagnostic power of laboratory data—by training a base Deep Neural Network (DNN) on the UCI dataset—and transfer the learned knowledge to enhance prediction on the NHIRD clinical dataset, which is more scalable but lacks lab values. This integration allows us to simulate the benefits of biological markers in settings where such data is unavailable.

Our experiments confirmed the effectiveness of this cross-domain knowledge transfer, particularly in improving the recall score, which reached 92% in the 6-months-ahead prediction task. High recall is essential in medical contexts, where failing to identify a patient at risk (false negatives) can have life-threatening consequences. Although the model faces challenges in precision due to class imbalance, the high sensitivity ensures that potential CKD patients are identified early for further clinical evaluation.

Moreover, this work contributes methodologically by:

- Demonstrating a practical pipeline that combines data preprocessing, feature selection, and TL.

- Proposing an efficient model architecture with reduced computational complexity.

- Validating the model with domain experts through scenario-based evaluation.

From an applied perspective, our approach offers a scalable solution for healthcare providers

to monitor at-risk populations, especially in low-resource environments where lab testing is limited. It also paves the way for future systems that dynamically integrate both clinical and biological streams of data.

Perspectives

To further enhance the clinical applicability and predictive performance of our CKD prediction system, several strategic directions are envisioned:

First, we plan to refine the transfer learning strategy by incorporating embedding techniques. Rather than directly transferring model weights, we aim to extract and transfer low-dimensional, meaningful representations (embeddings) of both clinical and laboratory features. This approach may capture more abstract and generalizable relationships between data modalities, potentially improving the adaptability and robustness of our model across diverse healthcare datasets.

Second, to overcome the "black-box" nature of deep neural networks and increase clinician confidence, we intend to integrate Explainable AI (XAI) methods. These techniques will help uncover the key contributing factors behind each prediction, offering clinicians clear, interpretable insights into why a specific patient is classified as high-risk. This transparency is critical for fostering trust and supporting clinical decision-making.

Finally, recognizing that CKD risk factors and disease progression vary significantly across patient populations, we aim to develop tailored models for specific subgroups. These subgroup-specific models—targeting, for example, diabetic patients, hypertensive individuals, or elderly populations—will allow for more precise predictions, improve early detection within vulnerable cohorts, and support more targeted preventive care and resource planning.

Bibliography

[1] Hebatullah Abdulazeem, Sera Whitelaw, Gunther Schauberger, and Stefanie J Klug. A systematic review of clinical health conditions predicted by machine learning diagnostic and prognostic models trained or validated using real-world primary health care data. *Plos one*, 18(9):e0274276, 2023.

[2] Rehan Ahmad and Basant K Mohanty. Chronic kidney disease stage identification using texture analysis of ultrasound images. *Biomedical Signal Processing and Control*, 69:102695, 2021.

[3] Syed Arslan Ali, Basit Raza, Ahmad Kamran Malik, Ahmad Raza Shahid, Muhammad Faheem, Hani Alquhayz, and Yogan Jaya Kumar. An optimally configured and improved deep belief network (oci-dbn) approach for heart disease prediction based on ruzzo–tompa and stacked genetic algorithm. *IEEE Access*, 8:65947–65958, 2020.

[4] Yanal Ahmad Alkudsi. Chronic kidney disease early prediction using machine learning.

[5] Khaled Alnowaiser. Improving healthcare prediction of diabetic patients using knn imputed features and tri-ensemble model. *IEEE Access*, 12:16783–16793, 2024.

[6] Deema Mohammed Alsekait, Hager Saleh, Lubna Abdelkareim Gabralla, Khaled Alnowaiser, Shaker El-Sappagh, Radhya Sahal, and Nora El-Rashidy. Toward comprehensive chronic kidney disease prediction based on ensemble deep learning models. *Applied Sciences*, 13(6):3937, 2023.

[7] Merijn Beeksma, Suzan Verberne, Antal van den Bosch, Enny Das, Iris Hendrickx, and Stef Groenewoud. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC medical informatics and decision making*, 19:1–15, 2019.

[8] Katia Berkache, Zohra Bengharez, Bastien Poitier, Djamila Ouabdesslam, Abdelkrim Guerinik, and Mourad Amrane. End-stage kidney disease in sidi bel abbes, algeria: Epidemiological profile of hemodialysis patients from 2015 to 2018. *Clinical Epidemiology and Global Health*, 12:100808, 2021.

[9] Thulasi Bikku. Multi-layered deep learning perceptron approach for health risk prediction. *Journal of Big Data*, 7(1):50, 2020.

[10] Pankaj Chittora, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat, Tulika Chakrabarti, Zbigniew Leonowicz, Michał Jasiński, Łukasz Jasiński, Radomir Gono, Elżbieta Jasińska, et al. Prediction of chronic kidney disease-a machine learning perspective. *IEEE access*, 9:17312–17334, 2021.

[11] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[12] Anne AH de Hond, Artuur M Leeuwenberg, Lotty Hooft, Ilse MJ Kant, Steven WJ Nijman, Hendrikus JA van Os, Jiska J Aardoom, Thomas PA Debray, Ewoud Schuit, Maarten van Smeden, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ digital medicine*, 5(1):2, 2022.

[13] Mohamed Djerioui, Youcef Brik, Mohamed Ladjal, and Bilal Attallah. Heart disease prediction using mlp and lstm models. In *2020 international conference on electrical engineering (ICEE)*, pages 1–5. IEEE, 2020.

[14] Yuhan Dong, Rui Wen, Zhide Li, Kai Zhang, and Lin Zhang. Clu-rnn: A new rnn based approach to diabetic blood glucose prediction. In *2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB)*, pages 50–55. IEEE, 2019.

[15] Erik Dovgan, Anton Gradišek, Mitja Luštrek, Mohy Uddin, Aldilas Achmad Nursetyo, Sashi Kiran Annavarajula, Yu-Chuan Li, and Shabbir Syed-Abdul. Using machine learning models to predict the initiation of renal replacement therapy among chronic kidney disease patients. *Plos one*, 15(6):e0233976, 2020.

[16] Sarah A Ebiaredoh-Mienye, Theo G Swart, Ebenezer Esenogho, and Ibomoiye Domor Mienye. A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease. *Bioengineering*, 9(8):350, 2022.

[17] Samit Kumar Ghosh and Ahsan H Khandoker. Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*, 14(1):3687, 2024.

[18] Asmaul Hosna, Ethel Merry, Jigmey Gyalmo, Zulfikar Alom, Zeyar Aung, and Mohammad Abdul Azim. Transfer learning: a friendly introduction. *Journal of Big Data*, 9(1):102, 2022.

[19] Meng-Hsuen Hsieh, Li-Min Sun, Cheng-Li Lin, Meng-Ju Hsieh, Kyle Sun, Chung-Y Hsu, An-Kuo Chou, and Chia-Hung Kao. Development of a prediction model for colorectal cancer among patients with type 2 diabetes mellitus using a deep neural network. *Journal of clinical medicine*, 7(9):277, 2018.

[20] Yuchen Hua, Thor S Stead, Andrew George, and Latha Ganti. Clinical risk prediction with logistic regression: Best practices, validation techniques, and applications in medical research. *Academic Medicine & Surgery*, 2025.

[21] Mohammadreza Iman, Hamid Reza Arabnia, and Khaled Rasheed. A review of deep transfer learning and recent advancements. *Technologies*, 11(2):40, 2023.

[22] Md Ariful Islam, Md Ziaul Hasan Majumder, and Md Alomgeer Hussein. Chronic kidney disease prediction based on machine learning algorithms. *Journal of pathology informatics*, 14:100189, 2023.

[23] Md Sakib Bin Islam, Md Shaheenur Islam Sumon, Rusab Sarmun, Enamul H Bhuiyan, and Muhammad EH Chowdhury. Classification and segmentation of kidney mri images for chronic kidney disease detection. *Computers and Electrical Engineering*, 119:109613, 2024.

[24] RS Jeena and Sukesh Kumar. Stroke prediction using svm. In *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pages 600–602. IEEE, 2016.

[25] Chandrasekhar Rao Jetti, Rehamatulla Shaik, Sadhik Shaik, and Sowmya Sanagapalli. Disease prediction using naïve bayes—machine learning algorithm. *International Journal of Science & Healthcare Research*, 2021.

[26] Katarzyna Kolasa, Bisrat Admassu, Malwina Hołownia-Voloskova, Katarzyna J Kedzior, Jean-Etienne Poirrier, and Stefano Perni. Systematic reviews of machine learning in healthcare: a literature review. *Expert Review of Pharmacoeconomics & Outcomes Research*, 24(1):63–115, 2024.

[27] Csaba P Kovesdy. Epidemiology of chronic kidney disease: an update 2022. *Kidney international supplements*, 12(1):7–11, 2022.

[28] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*, 2021.

[29] Surya Krishnamurthy, Kapeleshh Ks, Erik Dovgan, Mitja Luštrek, Barbara Gradišek Piletič, Kathiravan Srinivasan, Yu-Chuan Li, Anton Gradišek, and Shabbir Syed-Abdul. Machine learning prediction models for chronic kidney disease using national health insurance claim data in taiwan. In *Healthcare*, volume 9, page 546. MDPI, 2021.

[30] Raman Kumar, Sarvesh Garg, Rupinder Kaur, MGM Johar, Sehijpal Singh, Soumya V Menon, Pulkit Kumar, Ali Mohammed Hadi, Shams Abbass Hasson, and Jasmina Lozanović. A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions. *Frontiers in Artificial Intelligence*, 8:1583459, 2025.

[31] Nuo Lei, Xianlong Zhang, Mengting Wei, Beini Lao, Xueyi Xu, Min Zhang, Huifen Chen, Yanmin Xu, Bingqing Xia, Dingjun Zhang, et al. Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, 22(1):205, 2022.

[32] Shenglong Li and Xiaojing Zhang. Research on orthopedic auxiliary classification and prediction model based on xgboost algorithm. *Neural Computing and Applications*, 32(7):1971–1979, 2020.

[33] Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, 2022.

[34] He-Ren Lou, Xin Wang, Ya Gao, and Qiang Zeng. Comparison of arima model, dnn model and lstm model in predicting disease burden of occupational pneumoconiosis in tianjin, china. *BMC Public Health*, 22(1):2167, 2022.

[35] Srabanti Maji and Srishti Arora. Decision tree algorithms for prediction of heart disease. In *Information and Communication Technology for Competitive Strategies: Proceedings of Third International Conference on ICTCS 2017*, pages 447–454. Springer, 2019.

[36] P Manimaran, R Vignesh, B Vignesh, and G Thilak. Enhanced prediction of lung cancer stages using svm and medical imaging. In *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1334–1338. IEEE, 2025.

[37] MedlinePlus. Vital signs. `https://medlineplus.gov/ency/article/002217.htm`, 2024. Accessed: 2025-05-01.

[38] Lu Men, Noyan Ilk, Xinlin Tang, and Yuan Liu. Multi-disease prediction using lstm recurrent neural networks. *Expert Systems with Applications*, 177:114905, 2021.

[39] K Moon and A Jetawat. Predicting lung cancer with k-nearest neighbors (knn): A computational approach. *Indian J. Sci. Technol*, 17(21):2199–2206, 2024.

[40] Nazmun Nahar and Ferdous Ara. Liver disease prediction by using different decision tree techniques. *International Journal of Data Mining & Knowledge Management Process*, 8(2):01–09, 2018.

[41] K Nirmala, K Saruladha, and Kenenisa Dekeba. Investigations of cnn for medical image analysis for illness prediction. *Computational Intelligence and Neuroscience*, 2022(1):7968200, 2022.

[42] Simon Nusinovici, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, and Ching-Yu Cheng. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, 122:56–69, 2020.

[43] Madhumita Pal and Smita Parija. Prediction of heart diseases using random forest. In *Journal of Physics: Conference Series*, volume 1817, page 012009. IOP Publishing, 2021.

[44] M Pavithra, K Saruladha, and K Sathyabama. Gru based deep learning model for prognosis prediction of disease progression. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 840–844. IEEE, 2019.

[45] P Prabhu and S Selvabharathi. Deep belief neural network model for prediction of diabetes mellitus. In *2019 3rd international conference on imaging, signal processing and communication (ICISPC)*, pages 138–142. IEEE, 2019.

[46] P Ramprakash, R Sarumathi, R Mowriya, and S Nithyavishnupriya. Heart disease prediction using deep neural network. In *2020 international conference on inventive computation technologies (ICICT)*, pages 666–670. IEEE, 2020.

[47] Shishir Rao, Yikuan Li, Rema Ramakrishnan, Abdelaali Hassaine, Dexter Canoy, John Cleland, Thomas Lukasiewicz, Gholamreza Salimi-Khorshidi, and Kazem Rahimi. An explainable transformer-based deep learning model for the prediction of incident heart failure. *IEEE journal of biomedical and health informatics*, 26(7):3362–3372, 2022.

[48] Anjan Nikhil Repaka, Sai Deepak Ravikanti, and Ramya G Franklin. Design and implementing heart disease prediction using naives bayesian. In *2019 3rd International conference on trends in electronics and informatics (ICOEI)*, pages 292–297. IEEE, 2019.

[49] Soundarapandian P. Rubini, L. and P. Eswaran. Chronic Kidney Disease. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5G020.

[50] Derara Duba Rufo, Taye Girma Debelee, Achim Ibenthal, and Worku Gachena Negera. Diagnosis of diabetes mellitus using gradient boosting machine (lightgbm). *Diagnostics*, 11(9):1714, 2021.

[51] NS Safia. Prediction of breast cancer through random forest. *Current Medical Imaging*, 19(10):1144–1155, 2023.

[52] Dina Saif, Amany M Sarhan, and Nada M Elshennawy. Deep-kidney: an effective deep learning framework for chronic kidney disease prediction. *Health Information Science and Systems*, 12(1):3, 2023.

[53] Ahmad Waleed Salehi, Shakir Khan, Gaurav Gupta, Bayan Ibrahimm Alabduallah, Abrar Almjally, Hadeel Alsolai, Tamanna Siddiqui, and Adel Mellit. A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, 15(7):5930, 2023.

[54] Hadrien Salem, Sarah Ben Othman, Marc Broucqsault, and Slim Hammadi. Combining convolution and involution for the early prediction of chronic kidney disease. In *International Conference on Computational Science*, pages 255–269. Springer, 2024.

[55] Francesco Sanmarchi, Claudio Fanconi, Davide Golinelli, Davide Gori, Tina Hernandez-Boussard, and Angelo Capodici. Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review. *Journal of nephrology*, 36(4):1101–1117, 2023.

[56] Suhila Sawesi, Arya Jadhav, and Bushra Rashrash. Machine learning and deep learning techniques for prediction and diagnosis of leptospirosis: Systematic literature review. *JMIR Medical Informatics*, 13:e67859, 2025.

[57] Farah Shahid, Aneela Zameer, and Muhammad Muneeb. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. *Chaos, Solitons & Fractals*, 140:110212, 2020.

[58] VirenViraj Shankar, Varun Kumar, Umesh Devagade, Vinay Karanth, and K Rohitaksha. Heart disease prediction using cnn algorithm. *SN Computer Science*, 1(3):170, 2020.

[59] Xing Song, Lemuel R Waitman, SL Alan, David C Robbins, Yong Hu, Mei Liu, et al. Longitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine: retrospective cohort study. *JMIR medical informatics*, 8(1):e15510, 2020.

[60] Paul E Stevens, Sofia B Ahmed, Juan Jesus Carrero, Bethany Foster, Anna Francis, Rasheeda K Hall, Will G Herrington, Guy Hill, Lesley A Inker, Rümeyza Kazancıoğlu, et al. Kdigo 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney international*, 105(4):S117–S314, 2024.

[61] P Thamilselvan. Lung cancer prediction and classification using adaboost data mining algorithm. *International Journal of Computer Theory and Engineering*, 14(4):149–154, 2022.

Webography

[W1] , https://www.theisn.org/about-isn/mission-vision-values/manifesto/, Last access : 08/03/2025.

[W2] , https://elwatan-dz.com/lutte-contre-linsuffisance-renale-en-algerie-le-depistage- Last access : 08/03/2025.

[W3] , https://ihplus.com/fr/kidney-disease/, Last access 30/05/2025.

[W4] , https://www.statpearls.com/articlelibrary/NursingArticle/28357/, Last access : 16/06/2025.

[W5] , https://www.kidney.org/about/kidney-disease-fact-sheet, Last access : 08/03/2025.

[W6] , https://www.dneph.com/chronic-kidney-disease/stages-of-ckd/, Last access : 08/03/2025.

[W7] , https://rsingla.ca/kidneyUS/, Last access : 16/06/2025.

[W8] , https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease, Last access : 16/06/2025.

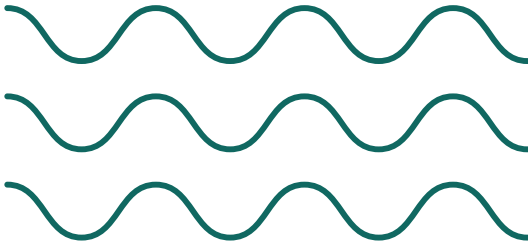
[W9] , https://osf.io/j3gur/e, Last access : 16/06/2025.

[W10] , https://figshare.com/articles/dataset/6711155?file=12242270, Last access : 16/06/2025.

UNIVERSITY
8 MAY 1945
GUELMA

RenalGuardian

STARTUP ANNEX



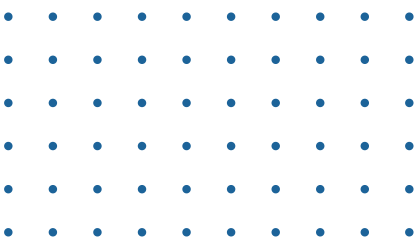
Project details and budget projections for RenalGuardian developement

Supervised by

Dr. Hiba ABDELMOUMENE

Presented by

Ilyas GUETTAF



INFORMATION CARD

SUPERVISION TEAM

Main Supervisor	Faculty	Speciality	Skills
Dr. Hiba ABDELMOUMENE	Mathematics, Computer Science and Material Sciences	Computer Science	IT, Project Management, AI-based decision support systems

PROJECT TEAM

Student	Faculty	Speciality	Skills
Ilyas GUETTAF	Mathematics, Computer Science and Material Sciences	Computer science	Software Developement, ML and DL techniques, UI/UX design for medical applications

TABLE OF CONTENTS

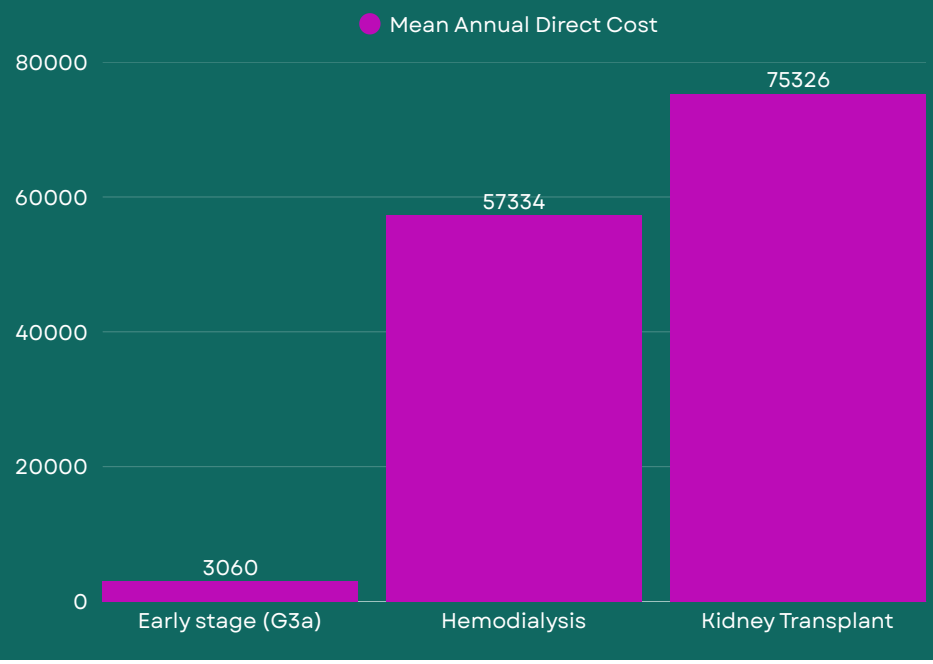
PROJECT PRESENTATION

PROJECT IDEA

Problematic

CKD is a state of progressive loss of kidney function, ultimately resulting in the need for renal replacement therapy, such as dialysis or transplantation. It is an unsuspected disease because it most often develops silently. A late diagnosis of this asymptomatic disease will lead to serious and costly complications. Most of the CKD population resides in the earlier asymptomatic stages (1-3).

Annual direct costs per patient skyrocket as CKD advances, placing immense strain on health economies. The cost of end-stage treatments like dialysis is orders of magnitude higher than managing early-stage disease.

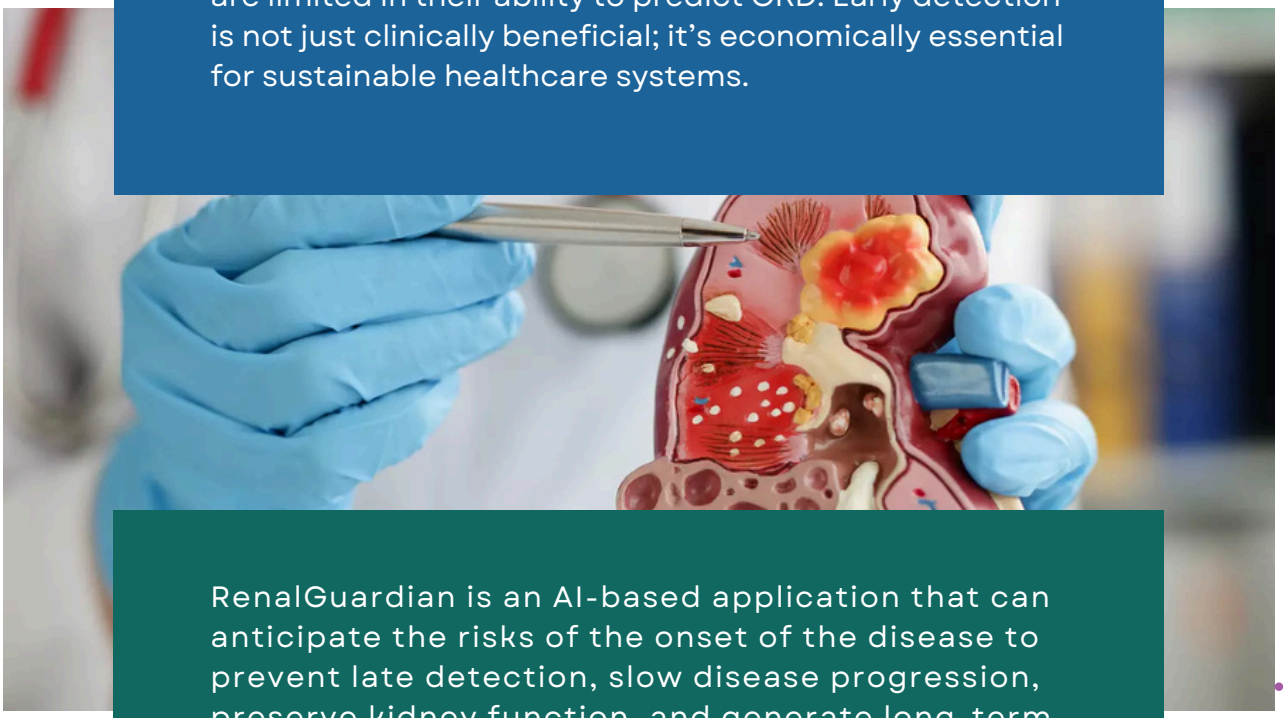


PROJECT PRESENTATION

PROJECT IDEA

Solution

The financial burden of CKD escalates as the disease progresses. Early detection is important and may allow many advantages. Experts need intelligent tools to predict and assess risks for early intervention and preventive care, where traditional diagnostic methods are limited in their ability to predict CKD. Early detection is not just clinically beneficial; it's economically essential for sustainable healthcare systems.



RenalGuardian is an AI-based application that can anticipate the risks of the onset of the disease to prevent late detection, slow disease progression, preserve kidney function, and generate long-term cost savings for both patients and healthcare systems. The end user will be the domain expert (nephrologist), where we will provide an easy-to-use interface that can help him manage the values and observe the results.

PROJECT PRESENTATION

SUGGESTED VALUE

Enhanced Clinical Decision Support

It makes the work of experts and the support easier

Less hospitalization

Less damage may not require hospitalization

Financial profit

An economic benefit for the country's health system by reducing health costs

Preserving Patient Health

improves patients' quality of life and extends healthy longevity

Early & Precise Risk Detection

AI's ability to identify individuals at risk for CKD significantly earlier than traditional methods

PROJECT PRESENTATION

WORKING TEAM

GUETTAF Ilyas

A computer science student, he has medical experience through a diabetes diagnosis and treatment project, aiming to expand his medical projects by implementing machine learning techniques.

OBJECTIVES

Short and Mid-Term

- Train and test a machine learning model that will be the core of RenalGuardian.
- Launsh RenalGuardian after developping.
- Making collaborations.

Long Term

- Collection of more data to create a more generalizable model.
- Clinical Validation.
- Going beyond the boundaries of an application; an AI powred clinic.

PROJECT PRESENTATION

IMPLEMENTATION SCHEDULE

Firstly, we start with searching and analyzing studies, collect required data to develop our machine learning model, after desining a prototype, we present our prototype to the experts, specifically nephrologists, and refine it using experts feedbacks and reviews.

Then, we develop our system, and present it, aiming to obtain more feedbacks that will be the base of our system optimization.

Finaly, after refine our system and optimize it, and as it is desgned for a sensitive domain, scenario-based evaluations and expert-centered validation phase are key steps to ensure the system meets key clinical criteria without requiring access to real-time patient data or undergoing clinical trials at this stage.

PHASE 1
Research, Market Analysis and Data Collection

PHASE 2
Developement

PHASE 3
Initial Test and Feedbacks

PHASE 4
Optimizations Feedbacks-based

PHASE 5
Final Testing And Scenario-based validation

PHASE 6
Launching RenalGuardian

PROJECT PRESENTATION

IMPLEMENTATION SCHEDULE

Phases	1	2	3	4	5	6	7	8	->
1									
2									
3									
4									
5									
6									

INNOVATIVE ASPECTS

NATURE OF INNOVATIONS

The innovation involves combining different data sources (health insurance data and laboratory records) to extract the best disease prediction features. We compare two different machine learning models: one that relies only on non-laboratory data, and the other that relies on laboratory data, which helps understand the extent to which expensive tests can be replaced by models based on easily accessible data. RenalGuardian offers a system that helps and assists healthcare and health insurance platforms, allowing Clinical and identifying the groups most vulnerable to infection. Dual-layer model : detection + prediction.

FILEDS OF INNOVATION

E-Health : Real-Time CKD Prediction

Introduces real-time ckd detection and prediction capabilities, allowing experts to detect and respond to disease at his early stages. This approach protects individuals more effectively and prevent disease to progress.

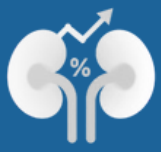
IA for predictive health systems : ML and DL

Identify complexe and cashed risk factors that traditional methods can not easily detect, using Deep learning capacities.

UI/UX Designe

Embodying ideas and continuous refinement of the system interface to be more agreeable and easy to use.

STRATEGIC MARKET ANALYSIS



RenalGuardian Buisness Model Canvas 0

Key Partnerships

Clinics and hospitals
Doctors
Investors

Key Activities

Plateform development
Sales and marketing
Instant responses

Key Resources

Website
Doctors / knowledge
Brand

Value Propositions

Doctors
- Decision help

Patients
- Stats following
- Alerts

Hospitals & Clinics
- Orienting patients
- Avoiding Dialyze

Customer Relationships

Trust
Maintenance
Friendly use

Channels

Website
Social Media
Conferences
Word of mouth

Customer Segments

Doctors
Patients
Hospitals
Clinics

Cost Structure

Data collection
Development
Legal/insurance

Revenue Streams

Premium subscription
Membership program

STRATEGIC MARKET ANALYSIS

MARKET SEGMENT

Government agencies

Including Hospitals and Insurance Companies, RenalGuardian provides healthcare professionals with clear, actionable, and data-driven insights at the point of care, empowering them to make more informed and personalized management decisions for at-risk individuals. Our targeting of state institutions is driven by economic reasons, as the cost of dialysis and transplantations is high.

Private agencies

Another parralel market is private clinics and diagnostic centers. Traditional methods are still used, RenalGuardian will provide a decision support system, as well as patient monitoring and management system. Collaborations with such agencies allow us to benefit from their experiences to make a more robust system.

STRATEGIC MARKET ANALYSIS

MEASURING COMPETITION INTENSITY

National Level

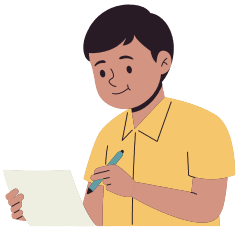
In the national level, we can only find traditional systems that rely on medical examinations alone, but they are expensive and may not be available to all patients .

International Level

Many innovative companies are already leveraging AI for early disease detection. AinnovaTech is an AI platform to forecast kidney function deterioration and recommend intervention timing. HEALWELL AI are deploying an AI "clinical co-pilot" technologies for early detection of chronic diseases, including CKD. Also, The licensing agreement between Aptar Digital Health and AstraZeneca aims to enhance patient outcomes through the early detection of hard-to-detect diseases like chronic kidney disease.

STRATEGIC MARKET ANALYSIS

MARKETING STRATEGIES



Pilot hospitals

Offering our product to hospitals or clinics for first use, this will make RenalGuardian recommanded.

Free trial

A free trial periods to attract entreprises and convince them to buy our product.

Social Media Power

Social media is one of the most strong market places. Also, offering official accounts make it easy for us to be more professional.

Partnership and events

Participing in events allow us to present our product in front of investors, and make collaborations.

Word of mouth

One of the strongest strategies, based on our employers capacities to present our product.

PRODUCTION AND ORGANIZATION PLAN

PRODUCTION PROCESS

1

IDENTIFY REQUIREMENTS AND PLANING

Identifying needs and tools used in medical sector, by collaborating with experts from hospitals and clinics.

2

SYSTEM PROTOTYPING AND ARCHITECTURE

Designing a User Interface to explain prediction scenario, and developing a detailed system architecture.

3

RENALGUARDIAN DEVELOOPEMENT

create a responsive and user-friendly interface, linked to our machine learning model, both providing several services.

4

TESTING

Unit Testing, integration testing, and user acceptance testing.

5

DEPLOYMENT AND APPLICATION LAUNCHING

Deploy the application on secure production environment.

6

MONITORING AND MAINTENANCE

Launching RenalGuardian

PRODUCTION AND ORGANIZATION PLAN

THE MAIN SUPPLIERS

Data & Clinical Partnerships

Medical Data Providers:
Establishing strong relationships with hospitals, clinics, and research institutions to access patient health records to validate our ML model's performance within the Algerian population.

Clinical Validation & Medical Expertise:
Expert's deep knowledge is needed for feature engineering, RenalGuardian clinical validation and integration into Algerian hospitals and clinics.

Technology & AI Development Tools

Cloud Computing
Our AI-based application require relying on cloud service providers for scalability and data security.

Specialized Software and Development Tools
A robust Database Management System is crucial to manage data, as equal as Development and Collaboration Tools, which are essential for the engineering team

Regulatory Compliance and Legal Expertise

Data Privacy and Healthcare Law Specialists
Accordinge to Law No. 18-07, legal counsel specializing in healthcare IT and data privacy will ensure RenalGuardian handles sensitive information legally and ethically

PRODUCTION AND ORGANIZATION PLAN

JOB POSITIONS

Lead AI/ML & Data Engineer
This crucial role designs, develops, and optimizes the core AI models for CKD prediction. They are also responsible for establishing robust data pipelines, ensuring data quality, and managing the machine learning operations (MLOps) lifecycle from data ingestion to model deployment and monitoring. This role combines the core AI expertise with the foundational data infrastructure work.

Full-Stack Software Developer
Responsible for building both the user-facing application (frontend) and the server-side logic (backend) that connects the AI models to the end-users. This includes developing secure APIs, managing databases, and creating intuitive interfaces for clinicians and potentially patients. They ensure the seamless functionality and user interaction of the entire platform.

Clinical AI & Data Validation Specialist
A bridge between medical expertise and technical development. This specialist provides critical clinical insights, aids in interpreting and validating medical datasets, ensures the AI model's outputs are clinically meaningful and safe, and helps integrate the application into real-world healthcare workflows. Their input is vital for the product's medical accuracy and adoption.

Product & UX Lead
Defines the product vision, strategy, and roadmap. This role is responsible for understanding user needs (clinicians, patients) through research, translating those needs into clear product requirements, and overseeing the user experience (UX) design to ensure the application is intuitive and effective. They drive what gets built and why.

Regulatory Affairs Manager
Essential for navigating the complex landscape of medical device regulations. This manager ensures the application complies with all relevant national (e.g., Algerian ANPP) and international (e.g., EU AI Act, FDA) standards. They are responsible for preparing and managing all necessary documentation and submissions for product approval and market access.

PRODUCTION AND ORGANIZATION PLAN

KEY PARTNERS

Government Agencies (ministries of health)

This partner is vital for strategic alignment, market access, and large-scale adoption. He set healthcare policies, control national budgets, and can facilitate regulatory pathways and national-level integration for RenalGuardian within the public health system.

Research Institutions

hese partners bring scientific credibility, deep medical expertise, and access to valuable data for rigorous validation. They are crucial for conducting independent clinical trials, ensuring the AI's efficacy and safety, and fostering a strong evidence base for our technology.

Private agencies

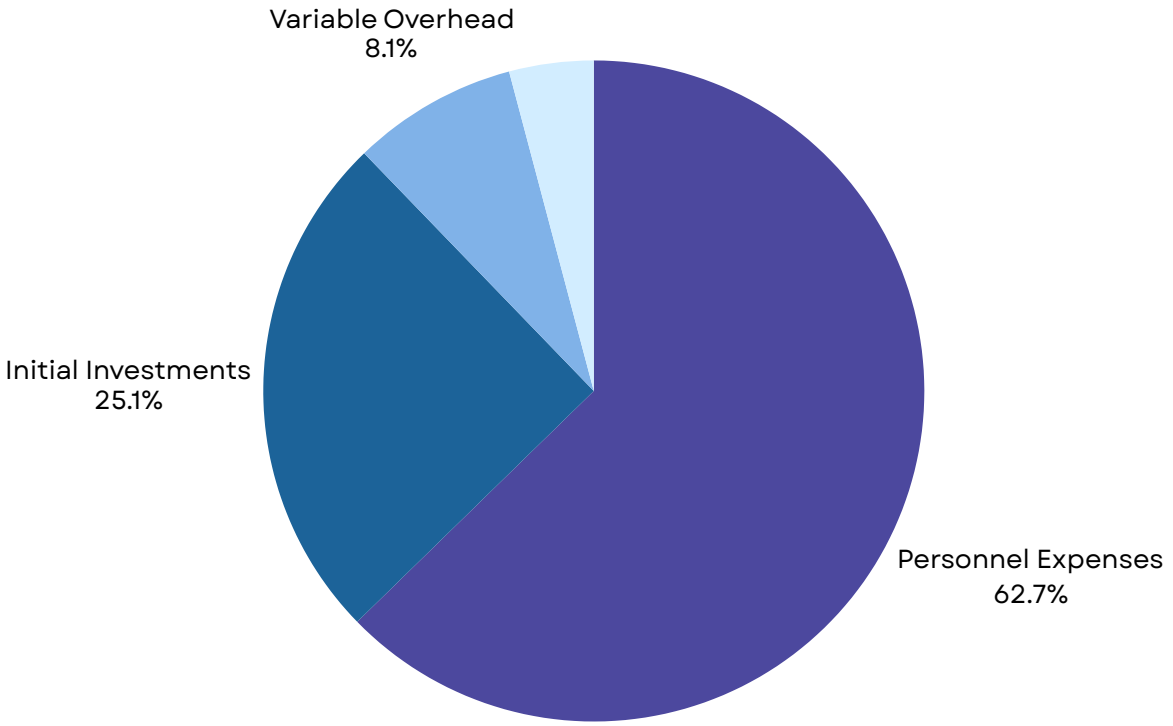
These partners offer agility for early adoption, direct patient interaction, and real-world testing grounds. They are essential for demonstrating the practical value of RenalGuardian in diverse clinical settings, gathering immediate user feedback, and supporting initial market penetration.

FINANCIAL PLAN
COSTS AND CHARGES

Knowing how to distinguish between costs and expenses is important when managing a business or its accounting. Understanding the differences between the two will help us ensure effective financial management at RenalGuardian.

A cost is an amount paid to acquire an asset, this include fixed and variable costs. On the contrary, an expense is not an one-time payment, it is an amount paid regularly towards ongoing business operations.

Our initial year's expenses are dominated by personnel, foundational investments, and operating overheads, totaling approximately 190,080,000 DA



FINANCIAL PLAN PROJECTIONS

Our most significant investment will be in top-tier talent. This table outlines the estimated annual compensation for core roles.

Job Position	Number of Positions	Total Annual (DA)
Lead AI/ML & Data Engineer	1	6,000,000
Full-Stack Software Developer	1	4,800,000
Clinical AI & Data Validation Specialist	1	5,000,000
Product & UX Lead	1	5,500,000
Regulatory Affairs Manager	1	5,200,000
Total Personnel Cost	**5**	**26,500,000**

This table outlines the estimated total annual compensation for our team over five years.

Year 1	Year 2	Year 3	Year 4	Year 5
26,500,000	26,603,000	33,410,000	38,106,000	39,096,000

FINANCIAL PLAN
STATEMENT OF OVERHEADS

- **Fixed Overheads**

These costs remain relatively constant regardless of activity levels, projected with a modest annual increase.

Expense Category	Year 1	Year 2	Year 3
Office Rent & Utilities	3,240,000	3,337,200	3,437,316
Core Software Licenses	1,350,000	1,390,500	1,432,215
Insurance & Legal Retainer	4,725,000	4,866,750	5,012,753
Regulatory Compliance Maintenance	1,350,000	1,390,500	1,432,215
Total Fixed Overheads	10,665,000	10,984,950	11,314,499

FINANCIAL PLAN
STATEMENT OF OVERHEADS

- **Variable Overheads**

These costs fluctuate with business activity, such as user adoption and data processing volume, scaling proportionally with growth.

Expense Category	Year 1	Year 2	Year 3
Cloud Infrastructure (scaling)	6,750,000	8,100,000	9,450,000
Data Acquisition/Licensing (ongoing)	4,050,000	4,860,000	5,670,000
Marketing & Sales Activities	5,400,000	6,480,000	7,560,000
Professional Services (ad-hoc)	2,700,000	3,240,000	3,780,000
Travel & Conferences	2,025,000	2,430,000	2,835,000
Total Variable Overheads	20,925,000	25,110,000	29,295,000

FINANCIAL PLAN

THE INVESTMENTS STATEMENT

These represent one-time or upfront costs necessary to get the product off the ground and achieve initial regulatory milestones. The majority of these are concentrated in Year 1.

Investment Category	Estimated Cost (DA)
Initial R&D Infrastructure Setup (specialized software)	6,750,000
Initial Large-Scale Data Purchase/Licensing	13,500,000
Regulatory Certification & Filing Fees	10,125,000
Legal Fees (Incorporation, IP, early contracts)	5,400,000
Product Design & Prototyping Tools	2,025,000
Working Capital Buffer	27,000,000
Total Initial Investments	**64,800,000**

FINANCIAL PLAN

STATEMENT OF ACTIFS & PASSIFS

This simplified snapshot provides an overview of the company's financial position at the end of the first year, reflecting initial funding, investments, and operational activities.

Investment Category	Estimated Cost (DA)
Liquidity (Cash & Bank Balance)	40,000,000
Equipment and Software (Fixed Assets)	6,750,000
Software Development (Intangible Asset)	50,000,000
Customer receivables (Accounts Receivable - from early revenue)	5,000,000
Total Actifs	**101,750,000**

Investment Category	Estimated Cost (DA)
Accounts Payable	5,000,000
Salaries and Accrued Expenses	2,000,000
Shareholder Equity - reflects initial funding net of losses	94,750,000
Total Passifs & Equity	**101,750,000**

FINANCIAL PLAN

PROJECTED ANNUAL REVENUE

Our revenue model will focus on recurring subscriptions and value-based pricing, aligning our success with improved patient outcomes and healthcare cost savings. We project significant growth as adoption scales.

Our projections show exponential growth driven by increasing market penetration and the compelling value proposition of early CKD prediction.

Year 1	Year 2	Year 3
13,500,000	94,500,000	337,500,000

RENALGUARDIAN PROTOTYPE V-0

- **Getting started**

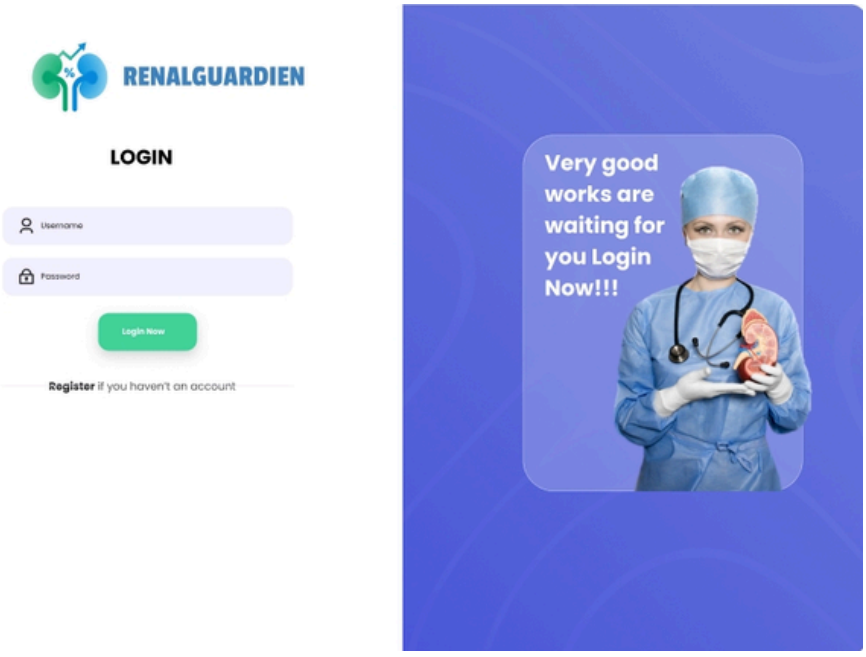


Figure 1 : The login page

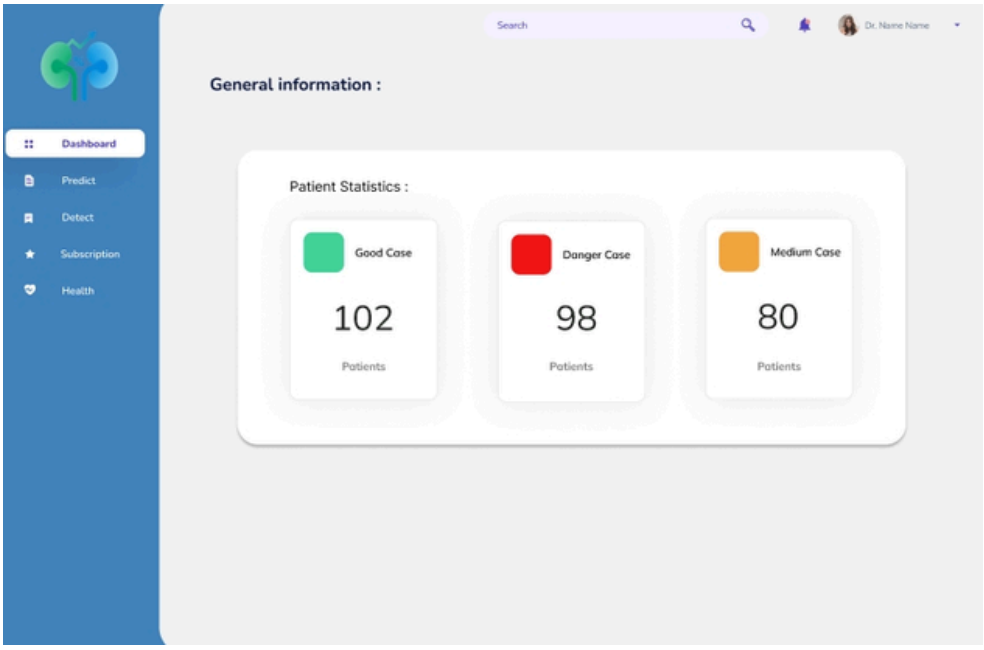


Figure 2 : The dashboard

RENALGUARDIAN PROTOTYPE V-0

- **Detection Scenario**

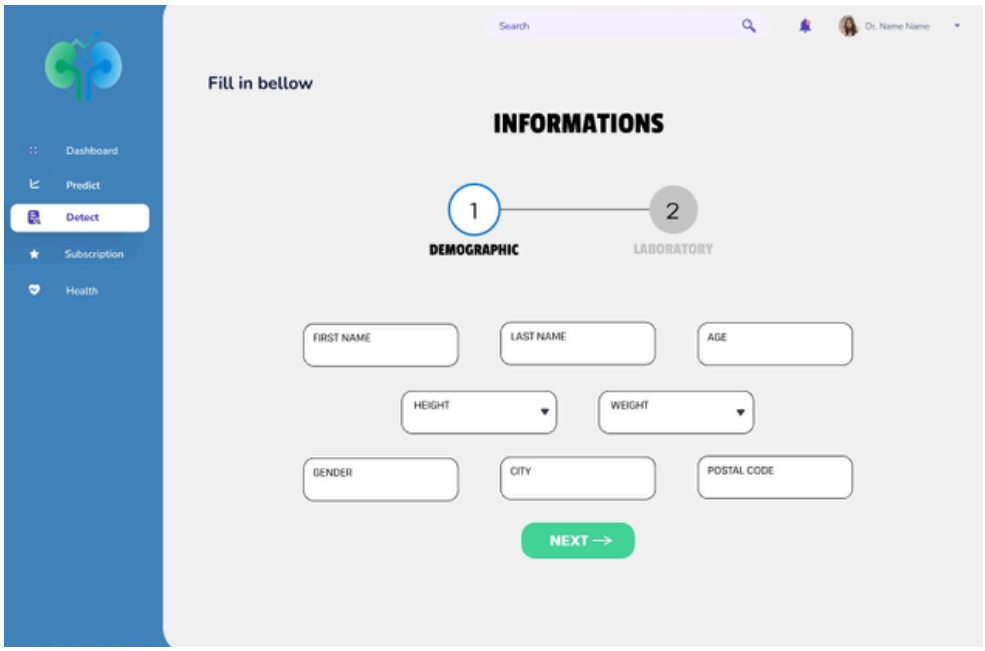


Figure 3 : Demographic Informations

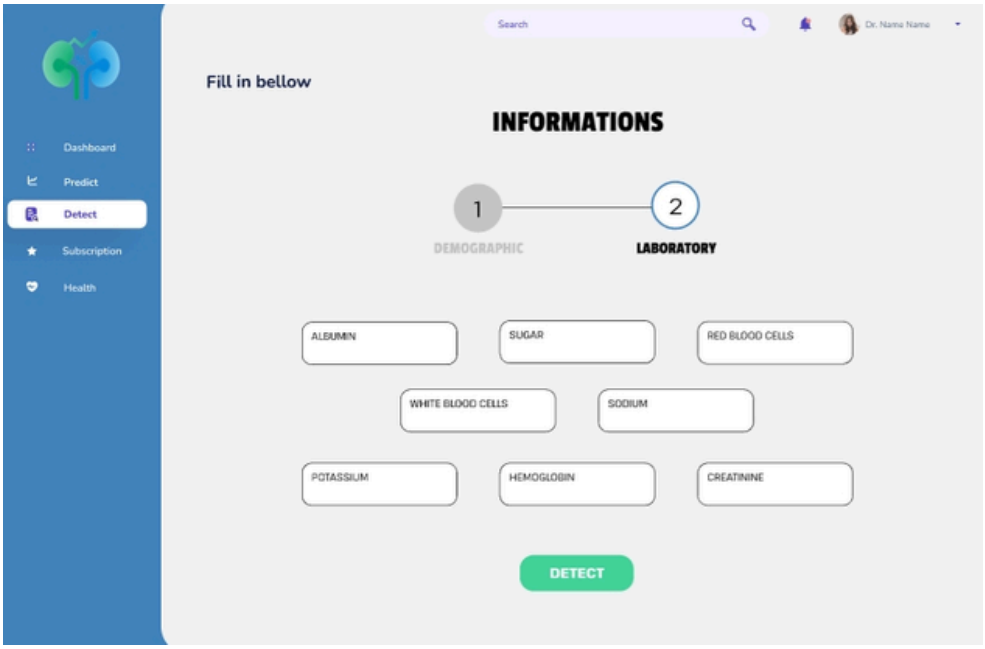


Figure 4 : Laboratory Informations

RENALGUARDIAN PROTOTYPE V-0

- **Detection Scenario**

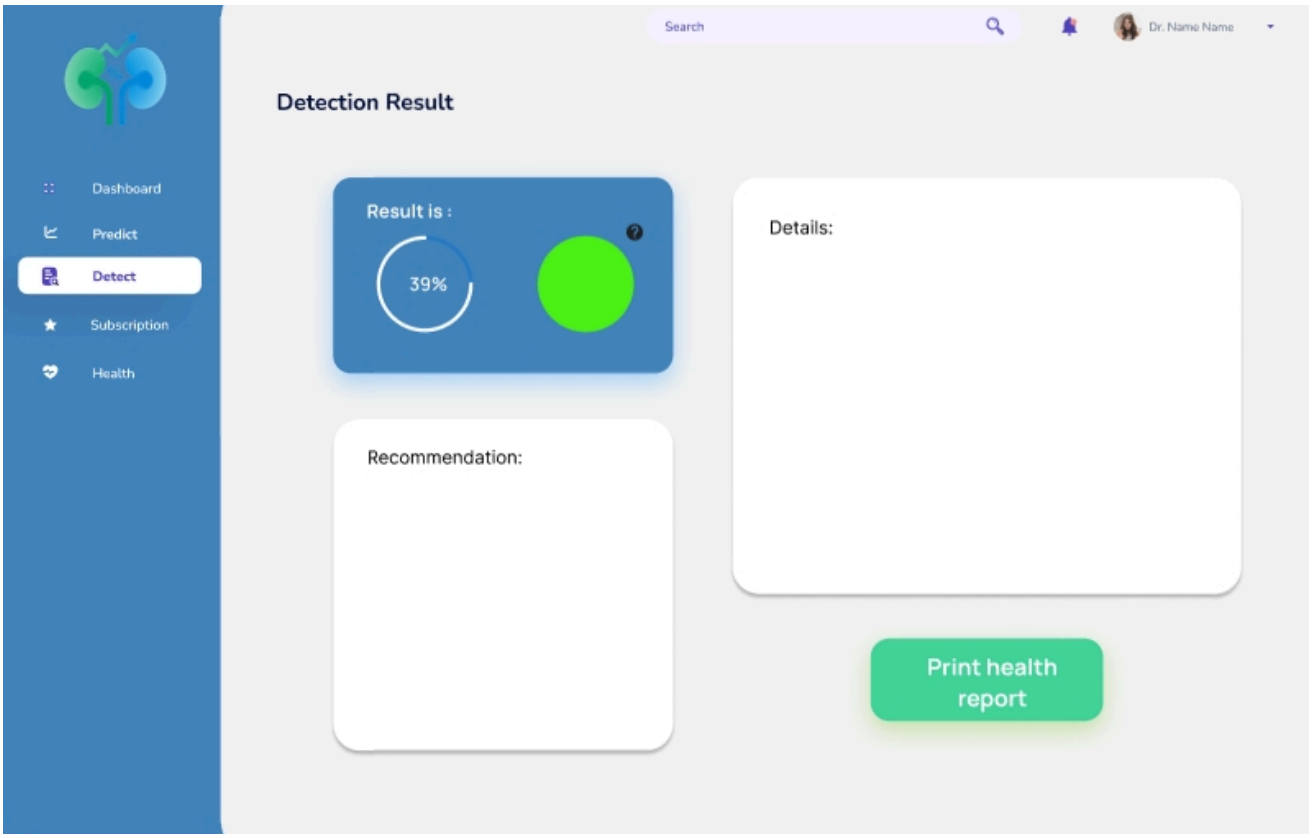


Figure 5 : Detection results

RENALGUARDIAN PROTOTYPE V-0

- **Prediction Scenario**

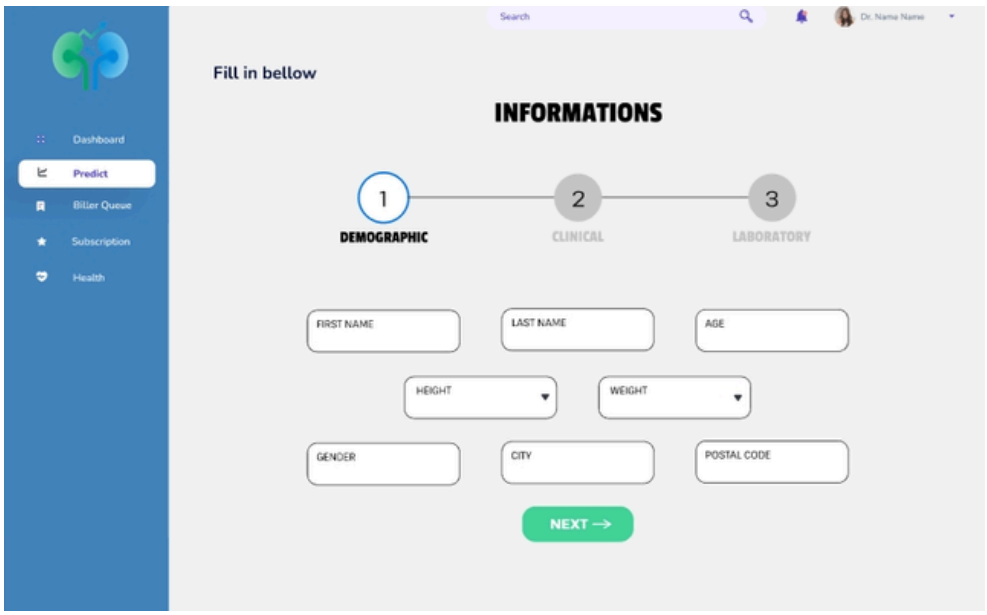


Figure 6 : Demographic Informations

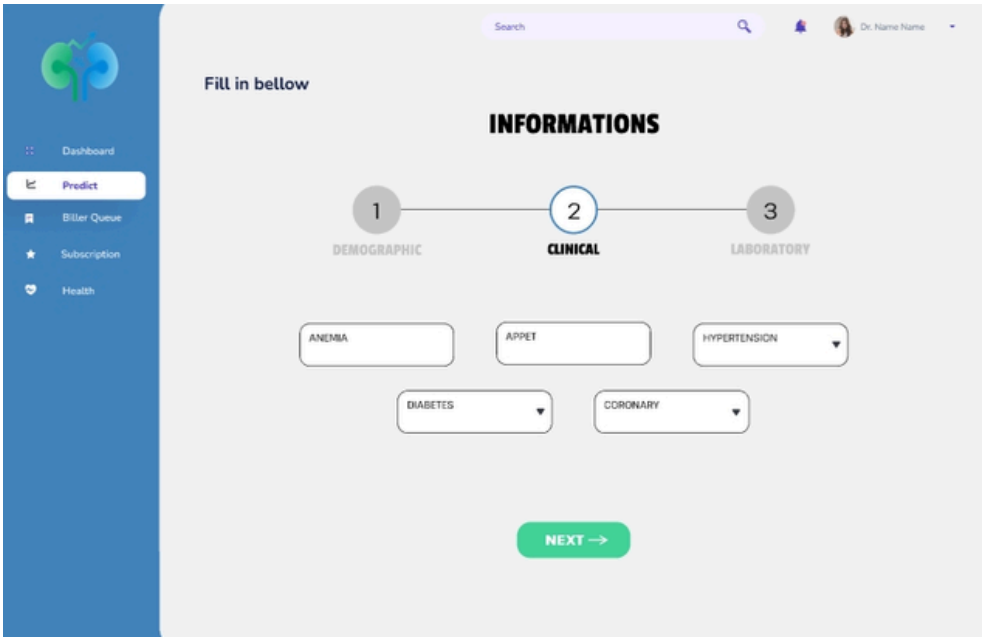


Figure 7 : Laboratory Informations

RENALGUARDIAN PROTOTYPE V-0

- **Prediction Scenario**

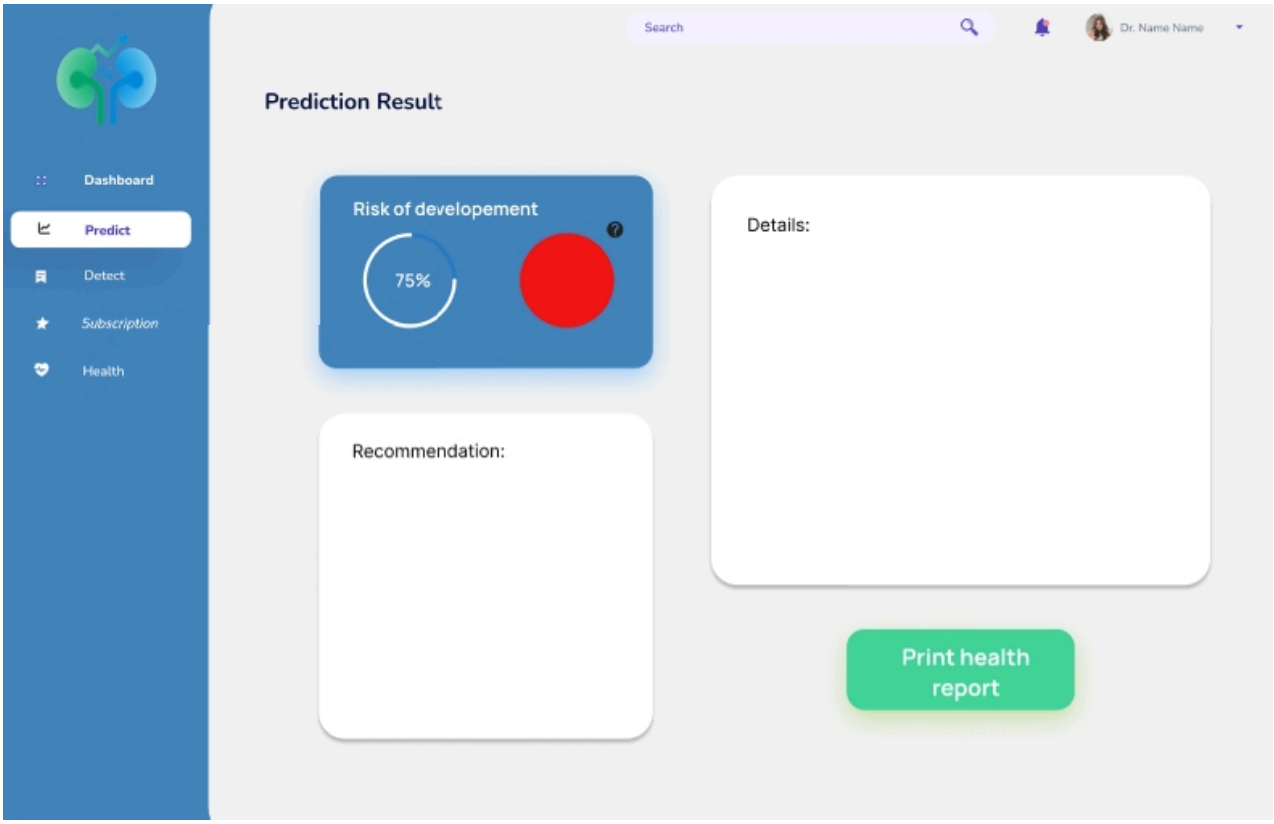


Figure 8 : The prediction results

RENALGUARDIAN
FINAL PROTOTYPE

- **Getting started**

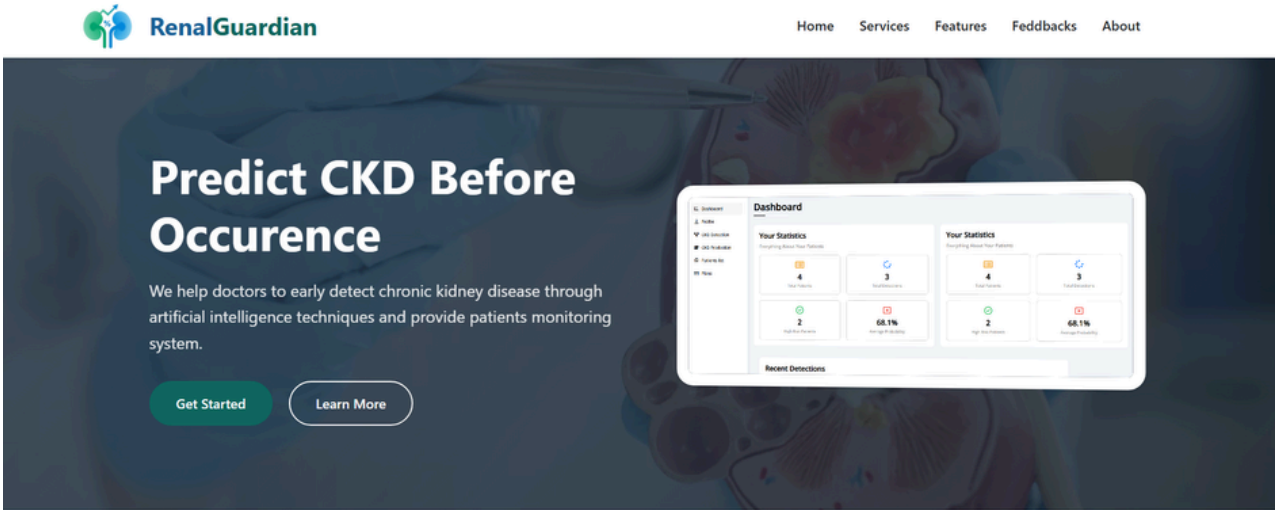


Figure 9 : The main page

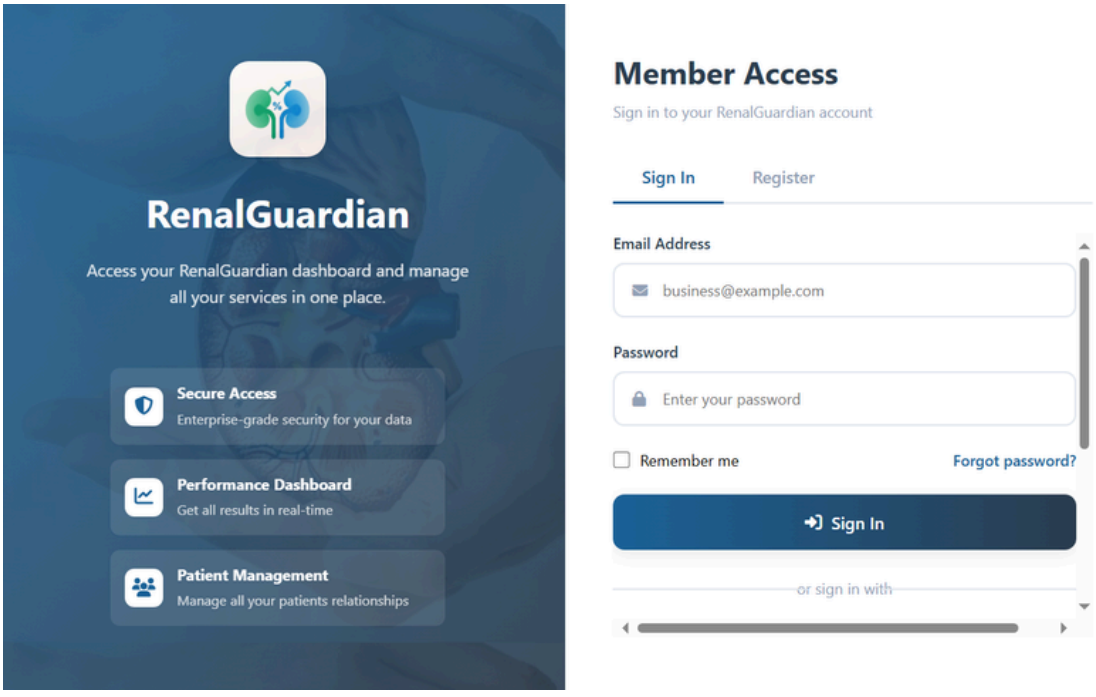


Figure 10 : The login page

RENALGUARDIAN

FINAL PROTOTYPE

- **Detection scenario**

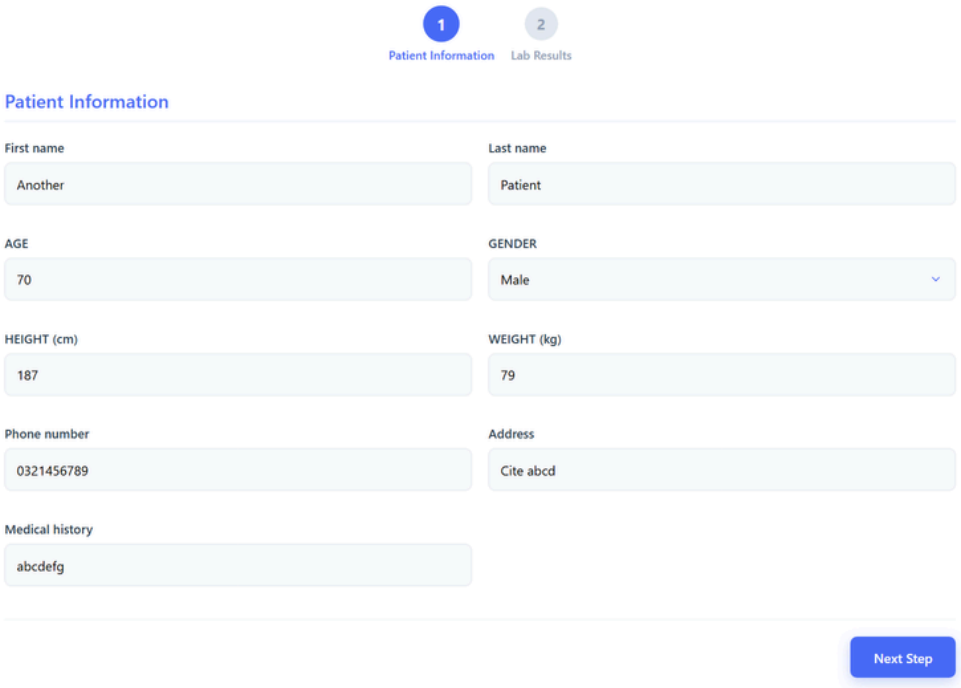


Figure 11 : Adding demographic information

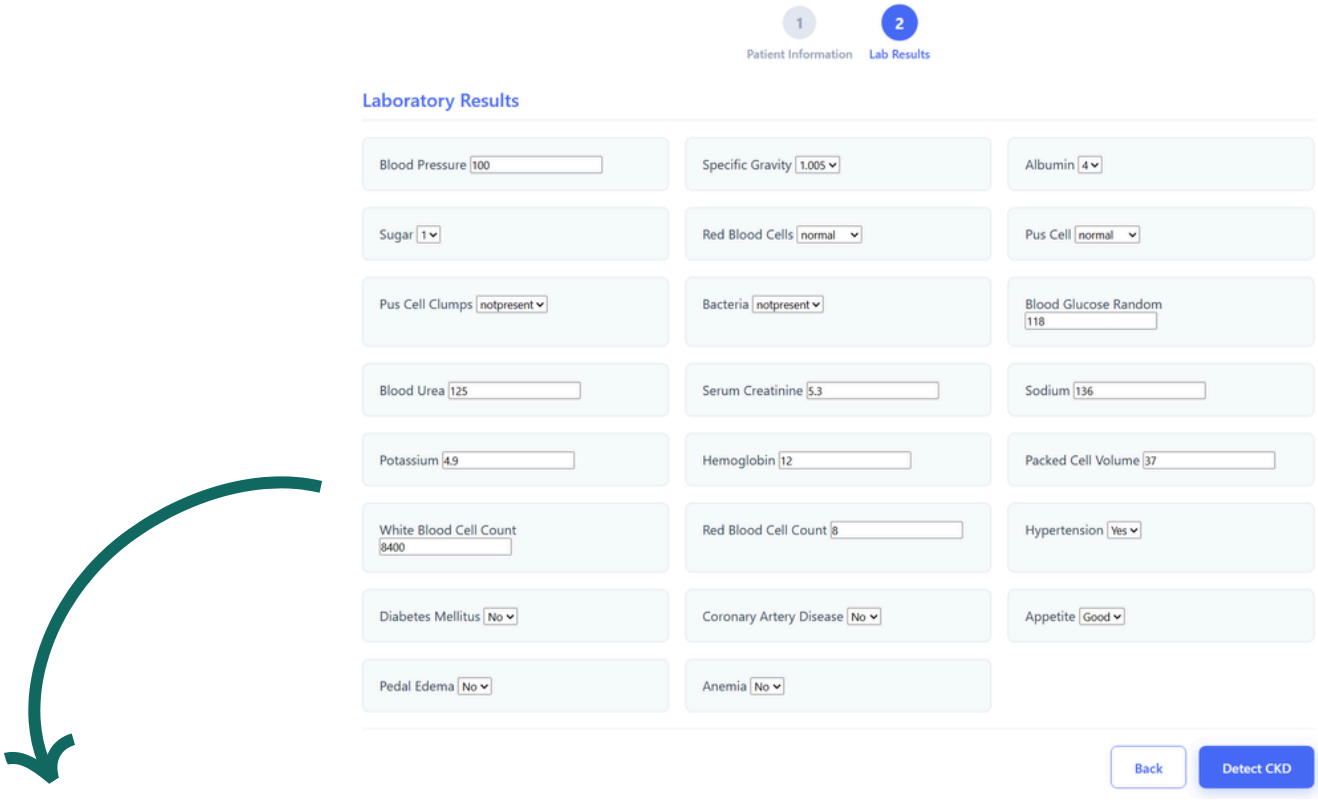


Figure 12 : Adding laboratory information

RENALGUARDIAN
FINAL PROTOTYPE

- **Detection scenario**

Figure 13 : Detection Results

RENALGUARDIAN
FINAL PROTOTYPE

- **Prediction scenario**

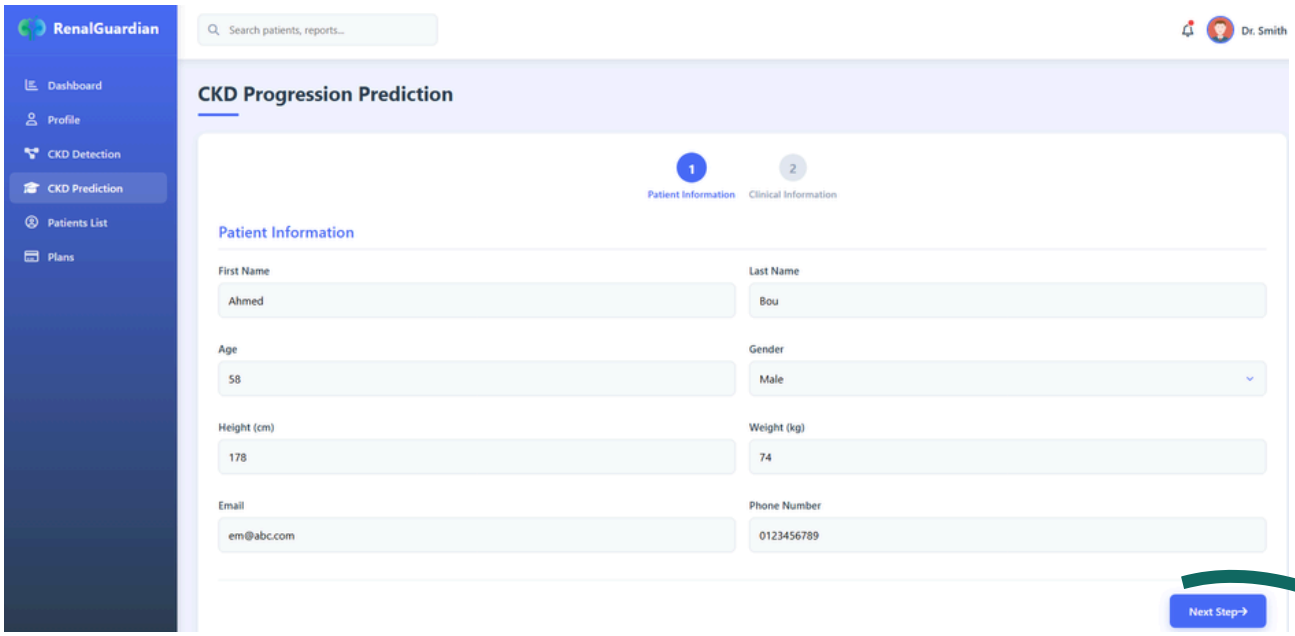


Figure 14 : Adding demographic information

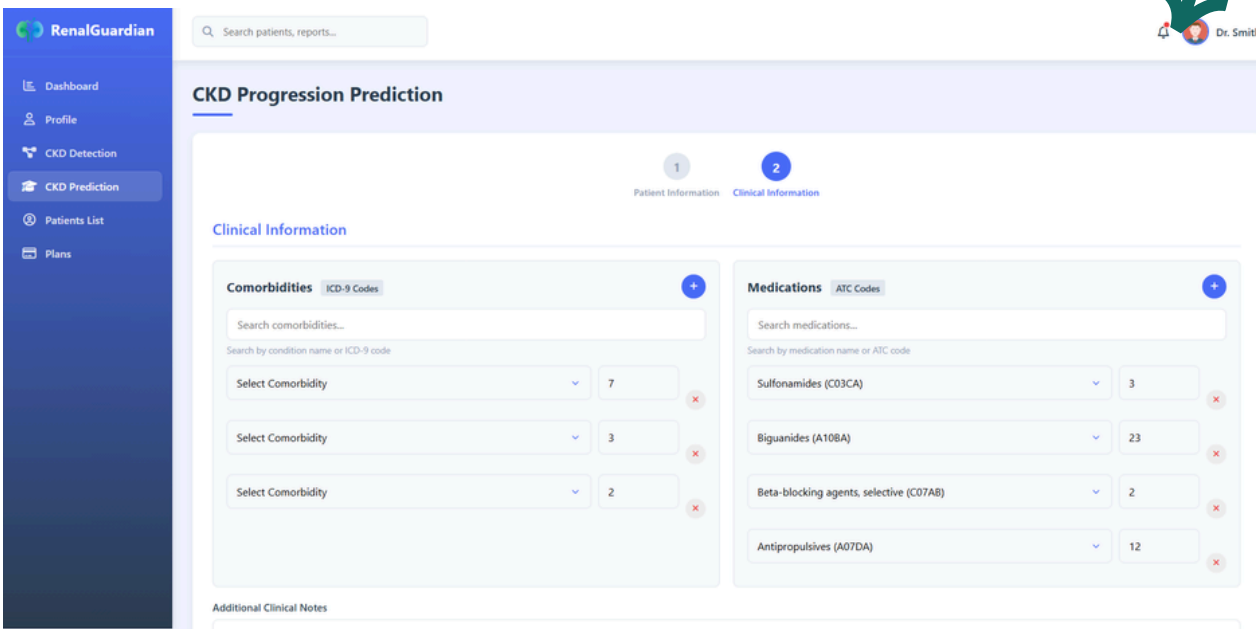


Figure 15 : Adding clinical information

RENALGUARDIAN
FINAL PROTOTYPE

- **Prediction scenario**

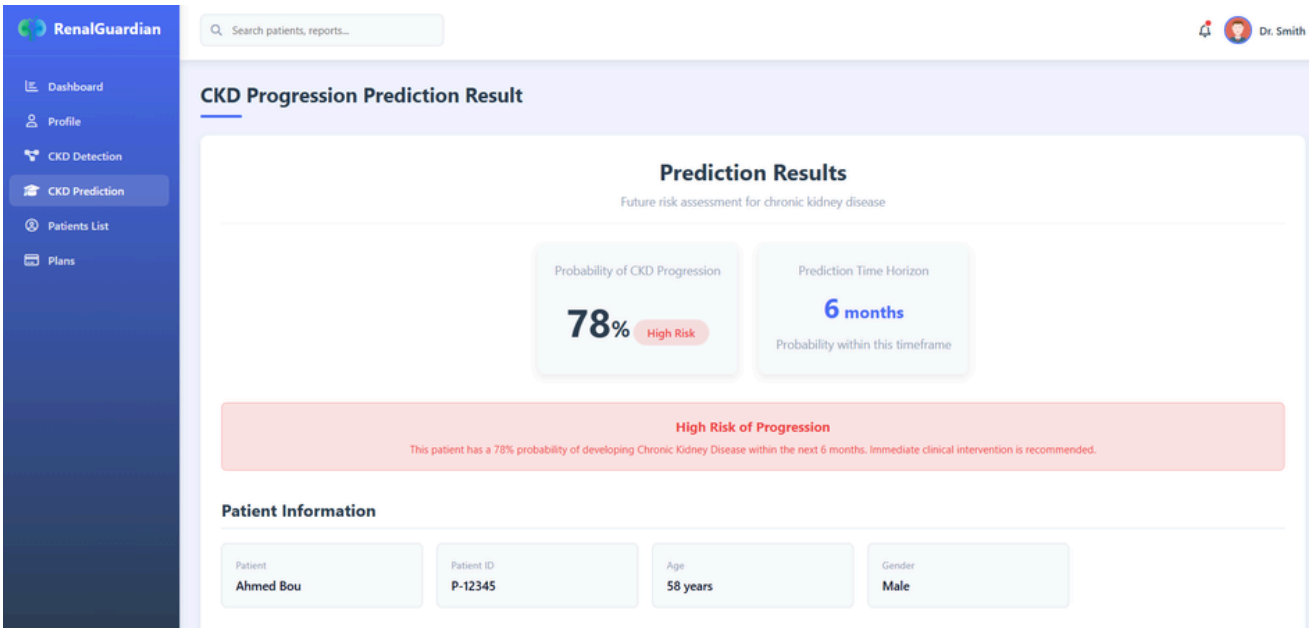
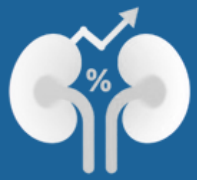


Figure 16 : Predictions Results



RenalGuardian Buisness Model Canvas 1

Key Partnerships

-Medical Institutions for data providing and clinical validation
-Investors
-Cybersecurity Firms

Key Activities

-Plateform development
-Model training & improvement
-Sales and marketing
-Clinical Validation & Regulatory Approval

Key Resources

-Scalable Software Platform
-Intellectual Property (IP)
-Financial Capital for R&D
-Human Capital
-Trained prediction models

Value Propositions

Doctors
Enhanced Real-Time diagnostic precision
Hospitals & Clinics
Enhanced Reputation & Quality of Care
IHC
Making strategies

Customer Relationships

-Trust & Reliability
-Training for healthcare staff
-Maintenance & periodic performance reviews
-Dedicated onboarding support

Channels

-Direct sales to hospitals and clinics
-Website
-Demonstrations at Conferences and Health Tech Events
-Clinical publications

Customer Segments

-Hospitals & Clinics
-Healthcare Insurance Companies
-Nephrologists
-Diagnostic Centers

Cost Structure

R&D Expenses
AI-Software Development
Legal/insurance
Sales & Marketing Expenses
Office & Administrative Overheads

Revenue Streams

Premium subscription
Software licencing
Membership program