Instituto Superior
de Engenharia
Politécnico de Coimbra

DEPARTMENT OF SYSTEMS AND COMPUTER ENGINEERING

Risk Assessment for Progression of Diabetic Nephropathy Based on Patient History Analysis

Dissertation to fulfill the Master's degree in Informatics Engineering

Specialization in Intelligent Data Analysis

Author

Francisco Gabriel Fonseca Mesquita

Supervisor

Simão Pedro Mendes Cruz Reis Paredes

Co-Supervisor

Jorge Manuel Oliveira Henriques

INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA
DE COIMBRA

Coimbra, Julho 2023

ABSTRACT

Diabetic nephropathy (DN) is one of the most common complications in patients with diabetes. It is a chronic disease that progressively affects kidney function and can potentially lead to renal impairment. Digitalization has allowed hospitals to store patients' information in electronic health records (EHRs). The application of ML algorithms to this data can allow the prediction of the risk in the evolution of these patients leading to better management and treatment of the disease. The main objective of this work is to create a predictive model taking advantage of the patient's history present in the EHR data. To achieve this goal, the largest Portuguese dataset from patients with DN followed for 22 years by the Associação Protectora dos Diabéticos de Portugal (APDP) was applied in this work. A longitudinal approach was developed in the data preprocessing phase, enabling the data to be used as input for sixteen distinct ML algorithms. After evaluating and analyzing the respective outcomes, the Light Gradient Boosting Machine was identified as the optimal model, exhibiting good predictive capabilities. This conclusion was supported not only by assessing several classification metrics on train, test, and unseen data, but also by evaluating its performance for each stage of the disease. Moreover, the models were analyzed using feature ranking plots and a comprehensive statistical analysis. Furthermore, the interpretability of the results through the SHAP method and the deployment of the model using Gradio and Hugging Face servers are also presented. Through the integration of ML techniques, an interpretation method and a web application that provides access to the ML model, this research offers an effective approach that may anticipate DN evolution, empowering healthcare professionals to make informed decisions for personalized patient care and disease management.

Keywords: Diabetic Nephropathy, Electronic Health Records, Machine Learning, Longitudinal Analysis, Risk Prediction

RESUMO

A nefropatia diabética (ND) é uma das complicações mais comuns em doentes com diabetes. Trata-se de uma doença crónica que afeta progressivamente os rins, podendo resultar numa insuficiência renal. A digitalização permitiu aos hospitais armazenar as informações dos doentes em registos de saúde eletrónicos (RSE). A aplicação de algoritmos de Machine Learning (ML) a estes dados pode permitir a previsão do risco na evolução destes doentes, conduzindo a uma melhor gestão da doença. O principal objetivo deste trabalho é criar um modelo preditivo que tire partido do historial do doente presente nos RSE. Foi aplicado neste trabalho o maior conjunto de dados de doentes portugueses com DN, seguidos durante 22 anos pela Associação Protetora dos Diabéticos de Portugal (APDP). Foi desenvolvida uma abordagem longitudinal na fase de pré-processamento de dados, permitindo que estes fossem servidos como entrada para dezasseis algoritmos de ML distintos. Após a avaliação e análise dos respetivos resultados, o Light Gradient Boosting Machine foi identificado como o melhor modelo, apresentando boas capacidades de previsão. Esta conclusão foi apoiada não só pela avaliação de várias métricas de classificação em dados de treino, teste e validação, mas também pela avaliação do seu desempenho por cada estádio da doença. Para além disso, os modelos foram analisados utilizando gráficos de feature ranking e através de análise estatística. Como complemento, são ainda apresentados a interpretabilidade dos resultados através do método SHAP, assim como a distribuição do modelo utilizando o Gradio e os servidores da Hugging Face. Através da integração de técnicas ML, de um método de interpretação e de uma aplicação Web que fornece acesso ao modelo, este estudo oferece uma abordagem potencialmente eficaz para antecipar a evolução da ND, permitindo que os profissionais de saúde tomem decisões informadas para a prestação de cuidados personalizados e gestão da doença.

Palavras-chave: Nefropatia Diabética, Registos de Saúde Eletrónicos, Machine Learning, Análise Longitudinal, Previsão de Risco

EPIGRAPH

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are.
If it doesn't agree with experiment, it's wrong.

Richard Feynman

AGRADECIMENTOS

A realização desta dissertação de mestrado foi um grande desafio que apenas foi possível ultrapassar com o apoio e contribuição de diversas pessoas. A todas eu deixo aqui os meus sinceros agradecimentos.

Ao meu orientador, professor Simão Paredes, e ao meu coorientador, professor Jorge Henriques, agradeço o constante acompanhamento, disponibilidade e ajuda ao longo da realização deste trabalho. Acreditaram em mim desde o primeiro momento e foram sem dúvida muito importantes para a finalização deste trabalho com sucesso.

À Associação Protectora dos Diabéticos de Portugal - APDP agradeço a disponibilização dos dados utilizados neste estudo. Em especial, quero mencionar o Doutor Rogério Ribeiro que não só acompanhou todo o projeto, como acrescentou ao mesmo todo o seu conhecimento e experiência clínica, algo que enriqueceu bastante o trabalho.

Aos meus pais e irmão deixo um agradecimento especial pelo papel essencial que tiveram ao longo de todo este percurso. Isto não seria possível sem o seu incondicional apoio. Eles mais do que ninguém sabiam o quanto isto era importante para mim e proporcionaram-me tudo para que o finalizasse da melhor forma.

Ao meu colega e amigo José Maurício agradeço o seu apoio constante, ajudando consideravelmente não só através de troca de ideias como também em termos motivacionais. Mencionar ainda o meu amigo Luís Chaves que me apoiou e ajudou ao longo de toda esta jornada académica.

Ao professor Gonçalo Marques estou muito grato por proporcionar o primeiro contacto com a APDP e por acreditar sempre que eu seria capaz de fazer um bom trabalho. Todo o seu apoio, motivação e amizade foram essenciais para superar este desafio.

E, a todas as pessoas que aqui não foram mencionadas, mas que de alguma forma contribuíram para o sucesso deste trabalho, os meus sinceros agradecimentos.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence
APDP	Associação Protectora dos Diabéticos de Portugal
AUC	Area Under Curve
BMI	Body Mass Index
CKD	Chronic Kidney Disease
DL	Deep Learning
DN	Diabetic Nephropathy
DT	Decision Tree
EHR	Electronic Health Record
ESRD	End Stage Renal Disease
ET	Extras Trees
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GBM	Gradient Boosting Machine
GFR	Glomerular Filtration Rate
HDL	High-Density Lipoprotein
HF	Hugging Face
KNN	K Nearest Neighbor
LDA	Linear Discriminant Analysis
LDL	Low-Density Lipoprotein
LightGBM	Light Gradient Boosting Machine
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naïve Bayes
NLP	Natural Language Processing
QDA	Quadratic Discriminant Analysis
RF	Random Forest
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting Machine

1 INTRODUCTION

Diabetes is a prevalent public health challenge that has an impact on quality of life and mortality rates. The number of people with diabetes increased from 108 million to 422 million between 1980 and 2014 [1]. In Europe, 6.2% of adults had diabetes in 2019, with Cyprus, Portugal, and Germany having the highest rates [2]. Most patients with both type 1 and type 2 diabetes struggle to achieve proper metabolic control, leading to complications such as retinopathy, neuropathy, and nephropathy [3].

Diabetic Nephropathy (DN) is a chronic disease in which the function of the kidneys deteriorates, reducing their ability to remove wastes and toxins from the bloodstream while also affecting the water balance in the body. DN is considered a progressive disease that usually worsens over time until the kidneys cannot function on their own, which is known as end-stage renal disease (ESRD) [4]. In developed countries, half of all cases of ESRD are due to DN, and the cost of treating ESRD patients is very high [5].

Digitalization has allowed hospitals to store the complete history of patient appointments in a database, resulting in the availability of EHRs. These data are longitudinal because they are collected over time and include multiple patient records at different points in time. The application of Machine Learning (ML) techniques to analyze EHR data can provide valuable insights and enable the development of ML models that can predict the risk of DN evolution, aiding physicians in the diagnosis and ultimately improving the quality of healthcare [6], [7].

Founded in 1926, the Portuguese Diabetes Association (Associação Protectora dos Diabéticos de Portugal), commonly known by its acronym APDP, is the world's oldest diabetes association and dean of the International Diabetes Federation (IDF) member associations. APDP is involved in studies and projects with national and international entities focused on diabetes and associated complications such as DN [8]. The EHR data used in this study were provided by the APDP and were collected from patients with diabetes followed for 22 years in their facilities. Furthermore, all the steps presented throughout this study have been validated by a physician associated with APDP to ensure clinical validity and significance of all the developed approaches.

1.1 Objectives

The objective of this work is to develop a predictive model using ML techniques applied to EHR data, with the aim of accurately identifying the progression risk of DN.

It seeks to use a longitudinal approach that can take advantage of the temporality associated with the EHR data. The incorporation of the patient's history provides information on the patient's evolution over time, which is essential for an accurate risk assessment and management of DN.

Furthermore, this work presents a simple and intuitive way of interpreting each prediction made so that there is transparency and understanding of the model.

1.2 Main contributions

This dissertation's main contributions are:

- In-depth exploration of both the clinical aspects concerning DN disease and the technical aspects related to ML. The valuable collaboration and ongoing support from APDP enabled the inclusion of clinical validation and guidance. Through this close collaboration, we established a well-defined approach to address the research problem.
- Exploration and application of several preprocessing steps considering the longitudinal nature of EHR data. The methodology was designed with the purpose of taking advantage of the patient's history and appropriately preparing the data to feed the ML models.
- The experimental setup was based on recommendations and guidelines defined in the literature as those that ensure better validation of the performance results of the ML models. In addition, a detailed analysis of performance by disease stage is presented, which we believe is essential and rarely presented in the literature.
- A simple and intuitive interpretation method is proposed using the SHapley Additive exPlanations (SHAP) method. It allows the user to understand the logic behind the result given by the ML algorithm. In addition, the model is deployed, and a web application is developed, enabling readers to access and use the model, obtaining both prediction outcomes and their corresponding interpretations.

1.3 Task management

This study was carried out over approximately one year (September 2022 to July 2023). Five main tasks can be identified from its realization:

- **Task 1:** Understanding the problem, surveying the state of the art, and writing the literature review.
- **Task 2:** Exploration and analysis of data made available by APDP.

- **Task 3:** Development of an approach using the data. Implementation, combination, and comparison of different techniques.
- **Task 4:** Consistent validation of the implemented models.
- **Task 5:** Writing the master's thesis.

Table 1.1 shows the time distribution of each of the tasks over the one-year period of this work:

Table 1.1: Timeline of tasks undertaken.

	Sep		Oct		Nov		Dec		Jan		Feb		Mar		Apr		May		Jun		Jul	
T1																						
T2																						
T3																						
T4																						
T5																						

1.4 Document structure

This document is structured in 7 different chapters:

Chapter 2 presents the theoretical background knowledge on disease (DN) and Artificial Intelligence (AI).

Chapter 3 presents the challenges inherent in EHR data, an overview of ML applied to EHR in a clinical setting, and a review of the literature made specifically on the application of ML to predict the evolution of DN.

Chapter 4 describes the entire methodology defined, with all steps presented in detail.

Chapter 5 provides a comprehensive presentation of the results using classification metrics to assess model performance both overall and by disease stage.

Chapter 6 presents a comprehensive critical analysis of the results, accompanied by a detailed exploration of the findings and inherent limitations of this research.

Chapter 7 concludes this work with a summary of the study and its findings, as well as different research directions that can be explored in the future.

2 BACKGROUND KNOWLEDGE

The application of Artificial Intelligence (AI) algorithms in the healthcare field opens many doors to better and more efficient medical services, increasing the quality of the service provided, consequently raising people's quality of life.

The first sub-section provides an overview of DN, including its epidemiology, pathophysiology, risk factors, clinical manifestations, and other several aspects. The second sub-section provides a brief background on AI and ML, covering fundamental concepts such as the origins of these algorithms, the steps involved in their construction, and important aspects like interpretation and potential applications. Understanding the intersection of ML and DN is crucial to developing accurate predictive models and helping clinicians make informed decisions. By combining these two areas of knowledge, it is possible to uncover insights that can contribute to early intervention and better management of this condition.

2.1 Diabetic Nephropathy

Diabetes is a chronic disease characterized by high levels of glucose in the blood, which can cause serious damage to the heart, veins, eyes, nerves, and kidneys. Type 1 diabetes is characterized by an autoimmune response that impairs the ability of the pancreas to produce sufficient insulin. On the other hand, type 2 diabetes occurs when the body becomes resistant to the effects of insulin, resulting in reduced glucose uptake by cells. Eventually, the pancreas may produce less insulin, worsening the condition [9]. Gestational diabetes is another type of diabetes that occurs during pregnancy, leading to high blood sugar levels in affected women [10].

Diabetes is one of the most prevalent diseases in the world, with the World Health Organization (WHO) estimating that 422 million suffered from diabetes in 2014. This corresponds to an increase of 314 million affected people compared to 1980 [1]. Numerous complications can result from this disease, and these can be classified into macrovascular or microvascular. Cardiovascular disease, heart stroke, and peripheral vascular disease are examples of macrovascular complications. Microvascular complications are mainly nerve damage (neuropathy), eye damage (retinopathy), and kidney damage (nephropathy) [11]. Although the various complications are serious and reduce the patient's quality of life, DN will be discussed more specifically in this study.

The kidneys are two bean-shaped organs (10 to 15 centimeters long). Normally, people live with two, but it is possible to live a quality life with only one functioning kidney. On very rare occasions, it is even possible to be born with 3 kidneys and remain equally healthy. They play an essential role in filtering out undesirable products and excess fluids. Not only that, but they also act in maintaining the body's acid-base balance and regulating the levels of water, salts, and important minerals

such as sodium, calcium, phosphorus, and potassium in the bloodstream. Each kidney is made up of a large number of filtration units called nephrons. Each nephron filters a tiny portion of blood. Each nephron is made up of a filter called a glomerulus. In the glomerulus, unnecessary products and extra fluids that the body does not need are filtered out. It is these components that will make up the urine, while the cleaned and filtered blood is returned to the body [12].

Prolonged exposure of the glomerulus to blood with great sugar levels can cause high damage to these filters leading to DN or diabetic kidney disease (DKD). In DN, the damaged blood vessels in the kidneys become leaky, allowing proteins like albumin to pass into the urine, a condition known as proteinuria. Initially, the loss of small amounts of albumin, called microalbuminuria or incipient nephropathy (about 30 to 300mg per day), is not easily detectable. However, as the disease progresses, more albumin is excreted (>300mg per day), leading to macroalbuminuria or overt nephropathy, which can indicate a decline in kidney function and the severity of DN.

2.1.1 Epidemiology

When looking at the epidemiology of DN, there are several important pieces of information to highlight. The probability of patients with diabetes developing DN is approximately 1.75% (95% CI: 1.62-1.89) [13]. DN is the most common cause leading to end-stage renal disease (ESRD), that is, the final stage of kidney disease in which the kidneys are no longer capable of functioning and the patient needs dialysis or a kidney transplant. It is also associated with increased morbidity and mortality among patients with DN [14]. Portugal has one of the highest incidences of dialysis in Europe and ranks among the highest in the world [15]. Although the reasons are unknown, there is also other alarming data. According to a 2008 study, about 17% of the population with diabetes in Portugal also suffered from DN [16]. Nevertheless, the prevalence rate is still below that reported worldwide, where about 20 to 40% of patients with diabetes also have DN [17]. Deng et al. present a comprehensive report on the evolution of DN worldwide from 1999 to 2019 [18]. As of 2019, DN is seventh in terms of prevalence, fourth as the leading cause of mortality, and sixth as the cause of disability on a global scale. There are differences between patients with DN associated with Type 1 Diabetes and patients with DN associated with Type 2 Diabetes. Although statistically differentiated in the work of Deng et al., both increased significantly in terms of number of patients, deaths, and disability adjusted life years (DALYs).

2.1.2 Pathophysiology

The pathophysiological mechanisms in the development of DN are multifactorial. Hyperglycemia or high blood glucose is the factor responsible for structural and functional changes in the kidneys [19]–[21]. Persistent high blood sugar levels trigger inflammation and oxidative stress within the kidneys. Inflammation is the body's

response to injury or infection, but in the case of DN, it becomes chronic. This ongoing inflammation contributes to further damage to the kidneys. Oxidative stress occurs when there is an imbalance between harmful molecules called reactive oxygen species (ROS) and the body's antioxidant defenses. Excess ROS production in DN leads to damage to kidney cells. Furthermore, the kidneys' hormonal system called the renin-angiotensin-aldosterone system (RAAS) becomes activated. This system helps regulate blood pressure and fluid balance in the body. However, in DN, activation of RAAS can cause constriction of blood vessels within the kidneys, leading to reduced blood flow. This, in turn, worsens kidney damage.

Over time, continued injury, inflammation, and oxidative stress cause scarring and fibrosis in the kidneys. This fibrosis disrupts the normal structure and function of the kidneys, impairing their ability to effectively filter waste products from the blood. As a result, kidney function gradually declines, and, if left untreated, it can progress to ESRD.

Management of diabetes by controlling blood sugar and blood pressure is essential to slow the progression of DN. In addition, interventions targeting inflammation, oxidative stress and the RAAS system can also be used to mitigate kidney damage and maintain kidney function.

2.1.3 Clinical presentation

The clinical presentation of DN can vary depending on the stage of the disease. In earlier stages, symptoms may go unnoticed. As the condition worsens, there are several clinical conditions that may become evident. DN is typically defined by the presence of proteinuria or a decline in renal function, indicated by a reduced glomerular filtration rate (GFR). It is important to note that patients with modest or no albuminuria may progress to ESRD. DN progresses through different states, each representing a different level of damage and decline in kidney function. These different stages of DN are shown in Figure 2.1.

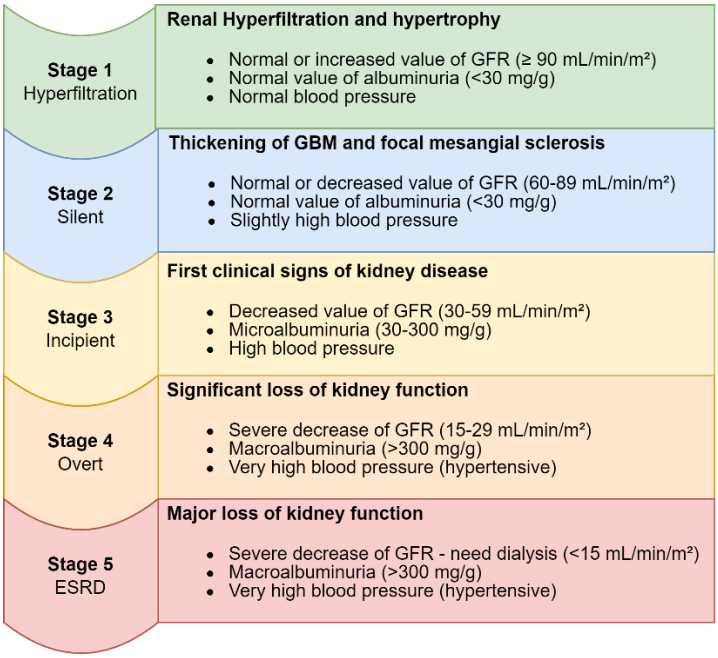

Figure 2.1: Different stages of DN, based on [22].

2.1.4 Risk factors

There are several risk factors that can contribute to the onset and progression of DN. Some factors are modifiable, which means that they can be changed or controlled by changing lifestyle, behavior, or through medical intervention. Others are non-modifiable, meaning that they are inherent characteristics of the individual that cannot be controlled. From reading several papers [23]–[31] resulted Table 2.1 where the main modifiable and non-modifiable DN risk factors are presented. Although many other factors have been identified, these are the ones that appeared almost unanimously in the various papers reviewed.

Table 2.1: Main modifiable and non-modifiable DN risk factors.

Modifiable factors	**Non-Modifiable**
Hyperglycemia	Insulin resistance
High blood pressure / Hypertension	Race
Dyslipidaemia	Age
Obesity	Long duration diabetes
Proteinuria	Genetic factors
High body mass index	
Smoking	

2.1.5 Diagnosis

Diagnosis of DN involves a careful assessment to determine the presence and severity of kidney damage in individuals with diabetes. It is crucial to detect this condition early, as it allows for timely intervention and management.

One common test is to check for the presence of protein in the urine, known as proteinuria. This is done by analyzing a sample of the person's urine. Proteinuria is a key indicator of kidney damage and can be an early sign of DN. Additionally, blood tests are conducted to measure creatinine and estimate the GFR, which reflects how well the kidneys are functioning.

Another alternative test that can be performed is the analysis of the albumin/creatinine ratio (ACR). Healthy kidneys filter creatinine from the blood, which is a chemical waste product. The ratio between the two measures indicates the amount of albumin excreted in relation to the amount of creatinine. A higher ACR value suggests increased albuminuria and is indicative of kidney damage or dysfunction.

Imaging tests, such as ultrasound, may be recommended in some cases to further assess the structure and identify any abnormalities. These tests help healthcare professionals understand the extent of kidney damage and determine the appropriate course of treatment. In some specific cases, a kidney biopsy may be necessary to take a sample of its tissue for further examination under a microscope [32].

2.1.6 Treatment

To prevent or slow the progression of DN, there are usually different treatments that patients can and are medically advised to follow:

- **Glycemic control:** The control of blood sugar levels so that they remain stable within values considered healthy. Although even with good glycemic control there may be progression in DN [33], several studies point out that strict control of blood sugar levels may be able to delay the onset or development of DN [34]–[36]. This control, when done too intensively, can lead to adverse reactions, which is why there are now international guidelines that recommend a unique and careful analysis of each individual and adapt the treatment to the patient's characteristics. Medications such as metformin, sulfonylureas, and insulin are commonly prescribed to achieve optimal glycemic control [37].
- **Blood pressure control:** Due to the nature of the disease, intraglomerular hypertension is of great importance for DN. Similarly to glycemic control, more individualized treatment according to the characteristics of the patient to avoid hypotensive episodes [38]. Medications called angiotensin converting enzyme inhibitor (ACEI) and angiotensin receptor blocker (ARB) are commonly used to lower blood pressure and protect the kidneys [39].
- **Reduce high cholesterol:** Cholesterol-lowering drugs, called statins, are used to treat high cholesterol, and reduce protein in the urine. Shen et al. [40] concluded that statins decrease albuminuria and urinary albumin excretion rates, but its effectiveness is time-dependent (duration of treatment) and also

depends on the type of diabetes the patients have, being most effective in type 2 diabetes.

For the purpose of increasing the effectiveness of these treatments, some changes in habits and lifestyle are usually suggested to complement medical treatment. Such changes may include exercising more to maintain a balanced weight combined with physical exercise, eating healthier, quitting smoking, and reducing salt intake, among other suggestions [41].

If the patient is in the latter state, already in renal dysfunction, the doctor is likely to suggest other, more severe treatment options:

- **Kidney dialysis:** Dialysis is a medical procedure that removes waste products and excess fluids from the bloodstream when the kidneys cannot perform that function. There are two main types: hemodialysis and peritoneal dialysis. Hemodialysis involves externally filtering the blood using a machine. In contrast, peritoneal dialysis utilizes the body's own abdominal lining (peritoneum) to filter the blood internally.
- **Transplant:** A surgical procedure that consists of transferring a healthy kidney from a living or deceased donor to a person whose kidneys no longer function.

2.1.7 Complications

DN is a complication resulting from diabetes, but DN itself can lead to several other complications in the human body. When not detected early or properly treated, it can be the source of several other associated complications. Just as DN develops over months or years, so do the various associated complications:

- **Cardiovascular disease:** DN patients are more likely to have cardiovascular problems and the risk is greater as the patient's nephropathy progresses [42].
- **Anemia:** DN usually causes a decrease in red blood cells, leading to anemia. It is recognized as a frequent DN complication that can increase the risk of cardiovascular and microvascular complications [43].
- **Bone mineral metabolism disorders:** Minerals imbalances such as calcium and phosphorus can occur in DN patients. In Chen et al. work, patients with Type 2 diabetes mellitus show an imbalance in bone mineral metabolism and that DN makes this worse [44].
- **Pregnancy complications:** Even with good blood glucose control, women with DN have a higher risk of maternal complications, such as preeclampsia and the need for cesarean birth [45].
- **Increased infections:** With the immune system weakened and kidney function affected, DN can cause patients to be affected by various infections, especially urinary tract infection (UTI) [46].

2.1.8 Prognosis: clinical risk score models

Strategies for assessing the patient's DN stage discussed earlier, such as urine testing, estimating GFR, or measuring ACR are useful, but clinically they do not provide the information needed to determine the risk of onset or progression of kidney disease. Identifying patients at increased risk for the onset or progression of DN can help to prevent disease progression more effectively and greatly alleviate the burden created in healthcare systems.

Clinical risk score models are the solution and bring great value as risk assessment and stratification tools. These models use a combination of clinical variables and biomarkers to predict the likelihood of developing DN or its progression to more severe stages. These models are created from studies and validation processes that lead to a standard way of predicting individual risk based on clinical evidence.

The main example of this type of model is the kidney disease: Improving Global Outcomes (KDIGO) guidelines. By estimating the GFR (eGFR) and using albuminuria value, this model can assess the risk of developing DN. The risk is stratified into low risk,, moderate risk, high risk and very high risk [47]. The model can be seen in Figure 2.2.

Prognosis of CKD by GFR and albuminuria categories: KDIGO 2012

GFR categories (ml/min per 1.73 m²) Description and range			Persistent albuminuria categories Description and range		
			A1	A2	A3
			Normal to mildly increased	Moderately increased	Severely increased
			<30 mg/g <3 mg/mmol	30–300 mg/g 3–30 mg/mmol	>300 mg/g >30 mg/mmol
G1	Normal or high	≥90			
G2	Mildly decreased	60–89			
G3a	Mildly to moderately decreased	45–59			
G3b	Moderately to severely decreased	30–44			
G4	Severely decreased	15–29			
G5	Kidney failure	<15			

Green, low risk (if no other markers of kidney disease, no CKD); yellow, moderately increased risk; orange, high risk; red, very high risk.

Figure 2.2: KDIGO model, based on [47].

Although widely applied and the standard test for predicting future kidney function decline, the KDIGO model fails with some regularity, especially when there is a need to make risk predictions over an extended time period. This is shown by the PromarkerD blood test. It is a recent test that can predict the risk of a patient with type 2 diabetes developing DN over the next 4 years, and also estimates the risk of patients having DN at the time of the test. The test uses three biomarkers: ApoA4, CD5L, IGFBP3, and three clinical factors: Age, HDL cholesterol, and eGFR. The results presented by this test show improved performance and the ability to predict

the risk of DN over the next four years [48]. The result of the test is shown in the Figure 2.3.

Despite these results, the fact that this is still a recent test and that it is a proprietary algorithm must be considered, and no detail has been found on how the data are analyzed to arrive at the risk score presented.

Figure 2.3: PromarkerD test results, based on [48].

Another proprietary algorithm is the one used by the KidneyIntelX model, used in Renalytix laboratories across EUA. It is an innovative diagnostic platform that uses ML algorithms to assess the risk and progression of kidney disease. It is a test for adult patients with type 2 diabetes and kidney disease at stage 1 to 3 who are at low, intermediate, or high risk for rapid progressive decline in kidney function. This risk is calculated using ML, three biological factors: TNFR1, TNFR2, KIM-1, and eight clinical factors: eGFR, ACR, also called UACR, serum calcium, HbA1c, systolic blood pressure, platelets, and AST enzyme. Along with the risk guidelines/recommendations are given on what the patient should do to improve their clinical situation. Several studies have been carried out in clinical validation of the KidneyIntelX test [49]–[51]. This test result is illustrated on Figure 2.4.

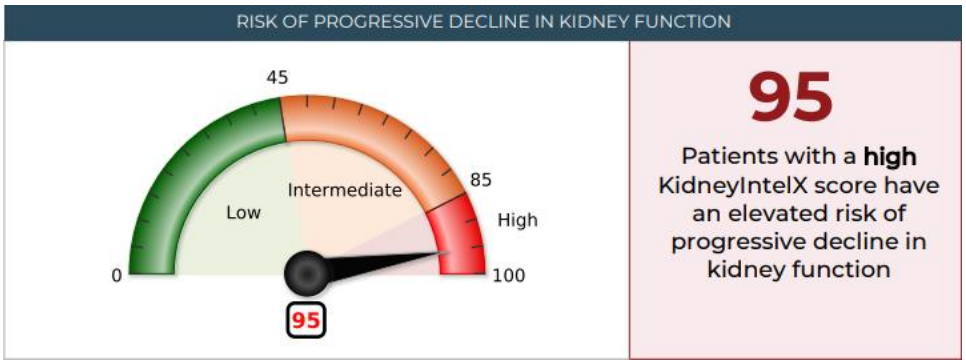

Figure 2.4: KidneyIntelX risk of progressive decline in kidney function, based on [52].

A simpler alternative is to use the kidney failure risk equation (KFRE) originally proposed by Tangri et al. in 2011 [53]. KFRE is capable of predicting kidney failure at 2 and 5 years in patients with stage 3 to 5 kidney disease. It uses 4 clinical variables: ACR, gender, age, and GFR, but additional information can give a better and more trustworthy result. The predicted risk may be non-existent, low, or medium-high. It is not a proprietary algorithm, and anyone can access the predictions through the online calculator provided on their website [54]. Although the KFRE has been

validated in more than 30 countries around the world, there is no available information on its official application in any healthcare center or clinic.

All the information presented throughout this chapter is essential to give context and background knowledge for the practical case that will be presented throughout this document. After presenting the basic knowledge of DN, it will be used from here on in a more direct way applied to the problem we want to solve in particular. The next subchapter presents the background knowledge about AI and ML.

2.2 Artificial Intelligence and Machine Learning

Throughout history, there has been a great deal of debate about the definition of AI [55]. It can have different meanings, and there is no widely accepted definition. In simple terms, we can say that AI is a field within computer science and technology that aims to build machines with intelligence, enabling them to perform tasks that typically rely on various aspects of human intelligence. It involves developing algorithms, models, and systems that can analyze and interpret data, learn from experiences, make informed decisions, and solve complex problems.

The emergence of the term AI as non-fictional takes us back to the early 1950s when different scientists and researchers considered the possibility of having machines capable of human-like intellectual abilities [56]. One of these scientists was Alan Turing, a British mathematician, who proposed the question "Can machines think?" in the famous paper "Computing machinery and intelligence" [57]. Alan Turing proposed the Turing test as a means to assess whether a machine can demonstrate intelligent behavior that is indistinguishable from that of a human. The test involves a scenario where an individual engages in conversations with both machines and humans, with the twist that they are unaware of the identity of each participant. The goal is for that person to guess whether they are talking to a machine or a human. It concludes that a machine is capable of thinking if it succeeds in the imitation game, that is, if it can pass itself off as a human by giving believable, well-constructed answers that usually require human-like intelligence. Later, in 1956, John McCarthy officially introduced the term "Artificial intelligence (AI)" at a conference at Darthmouth College. This was a major milestone that gave birth to the scientific field of AI [58].

Since then, the advancement of AI has been extensive, resulting in a heightened level of curiosity and the emergence of AI researchers and practitioners throughout the globe. Within the wide-ranging spectrum of AI, one of the most prominent areas is ML. Often used almost synonymously, AI and ML, although related, have different meanings. ML is a subset of AI that studies algorithms and techniques that allow a machine to learn from data and make predictions without being explicitly programmed to do so. ML algorithms can analyze and identify patterns, relationships, and insights from large datasets, allowing computers to recognize complex patterns, make predictions, and solve problems [59]. Within ML algorithms

there is also a sub-domain called Deep Learning (DL). These algorithms will be presented below, but it is possible to see the relation between AI, ML and DL in the Figure 2.5.

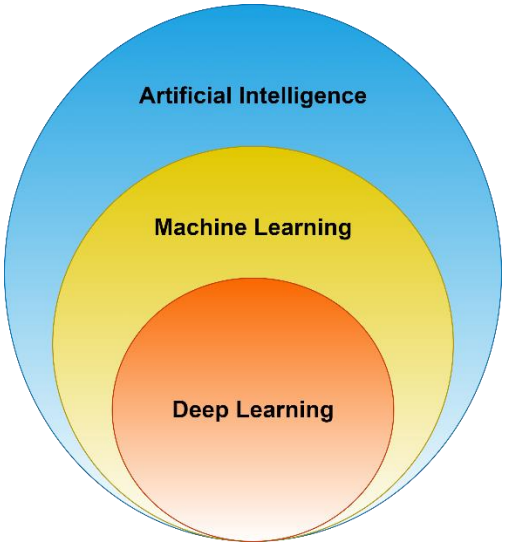

Figure 2.5: AI, ML and DL.

There are essentially 3 types of learning [60]:

- **Supervised Learning:** The objective is to solve a problem using a set of labeled examples. These examples consist of input data (x) and the corresponding output labels (y). The objective is to train an ML model (f(x) = y) by learning from these input-output pairs. The model adjusts its parameters to approximate the underlying function that connects the inputs to the desired outputs, allowing the model to take predictions on new instances based on the learned patterns.
- **Unsupervised Learning:** Unsupervised ML algorithm is used to analyze data and reveal concealed patterns without depending on pre-established labels or specifications. The training data consists only of variables x, and the algorithm aims to extract significant information and effectively cluster similar data points together.
- **Reinforcement Learning:** Here, we do not have either predefined inputs or outputs. Instead, we just describe the current situation, set a goal, and provide a list of possible actions along with their limitations. Through trial and error, the model learns which actions lead to a greater reward and will be penalized for wrong actions.

Figure 2.6 shows a representation of the various types of learning.

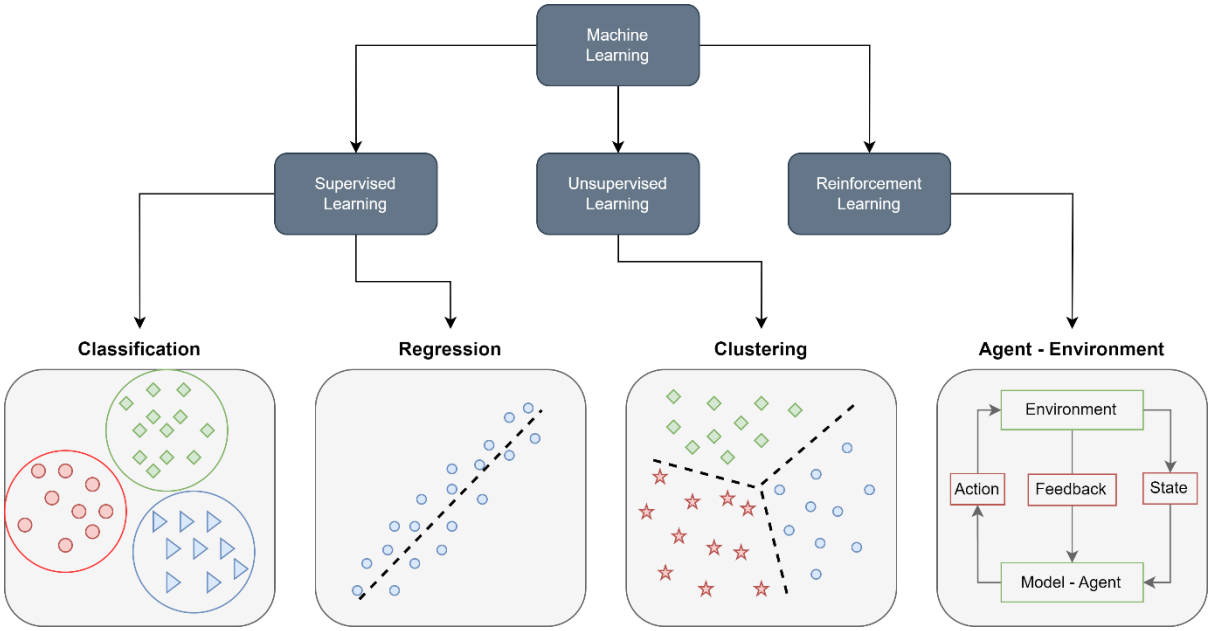

Figure 2.6: Types of learning in ML.

This work is exclusively focused on supervised learning, that can be divided into two main branches: classification and regression. Simply put, regression algorithms seek to predict a numerical value, while classification algorithms seek to predict predefined classes or categories.

A classification or regression problem consists of several main steps such as data preprocessing, training, evaluating, and selecting the best model and, if necessary, implementing interpretation methods to analyze and understand the prediction made. In the next chapter, the various preprocessing techniques will be introduced.

2.2.1 Data preprocessing

The construction of a predictive ML model is strongly influenced by a number of decisions made in data preparation, transformation, and cleaning. This is an essential step that directly impacts the quality of the model, especially in areas where data is more subject to poor quality, as is the case in the clinical area [61].

The required algorithms to circumvent the most frequent problems vary depending on the nature of the data we are working with, but there are essentially four main categories of preprocessing, as described by García et al. [62]:

- **Data cleaning:** Handle noisy and inconsistent data.
- **Data integration:** Different data sources should be combined.
- **Data transformation:** Conversion of raw data into a more appropriate format to be supplied to a given ML model.
- **Data reduction:** Select a subset with the most important features, retaining only the essential information for ML modeling.

For each category there are many different techniques and algorithms. Which is best depends on the nature of the data and the problem. Figure 2.7 summarizes the different categories and algorithms used in each.

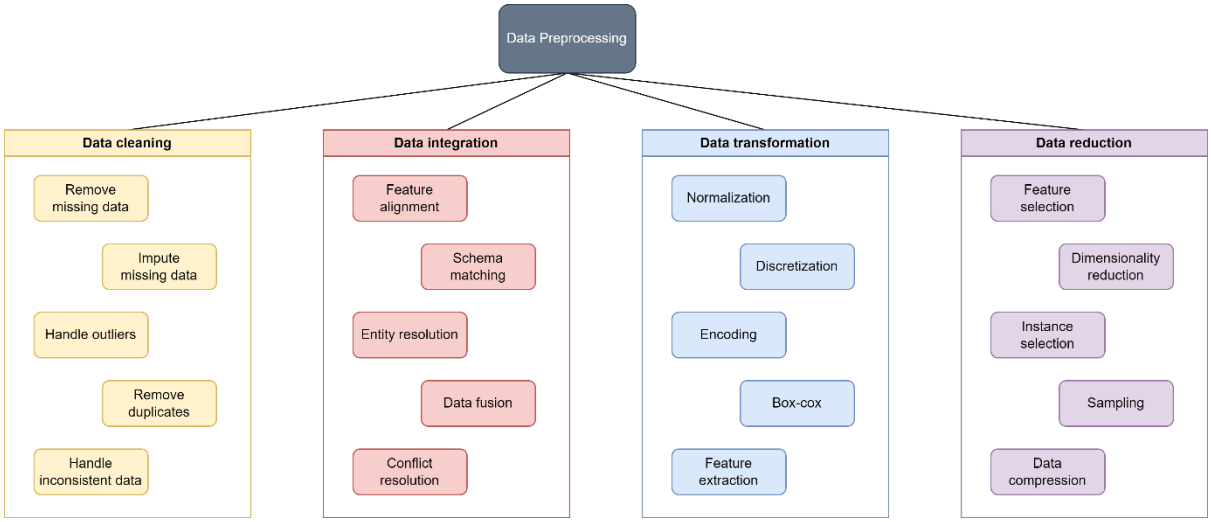

Figure 2.7: Data preprocessing main categories and respective techniques.

2.2.2 Model training, evaluation, and selection

After preparing the data, we move on to the ML modulation phase, where we develop, evaluate, and select the best algorithm to construct our ML model and address our problem effectively. In the field of ML, there are many different types of algorithms. One specific category is called DL, which is focused on using neural networks with multiple layers. These neural networks are designed to work in a similar way to how our brains process information. These networks can autonomously learn complex patterns in the data, where the layers closest to the input are used to learn the most basic patterns, and the layers closest to the output learn the more complex patterns, often abstracted to the human eye. A complete overview of the ML and DL methods can be found in the work of Sarker [63] and Shinde et al. [64].

There are a tremendous number of ML algorithms. Some of the most commonly used classification ML algorithms found in the literature are: Artificial Neural Networks (ANNs), K-nearest neighbors (K-NN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), Gradient boosting (GBM), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Light gradient boosting (LightGBM), K Nearest Neighbor (KNN), and Linear Discriminant Analysis (LDA) [65], [66].

When training an algorithm, it tries to learn the patterns associated with the data and their associated class or target. This will allow the model to predict the class after being trained and applied to new instances. The performance of these algorithms is measured through several metrics resulting from the comparison between the predicted values and the real values of the class present in the data. This comparison

is usually presented in the form of a confusion matrix. Considering a binary problem where class can be positive (1) or negative (0), the respective confusion matrix can be represented by the one shown in Figure 2.8.

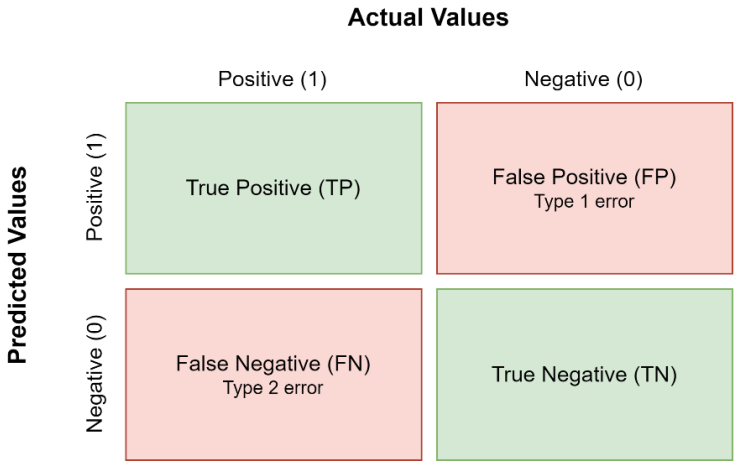

Figure 2.8: Confusion matrix example.

Through the matrix, different metrics with different meanings can be derived. These are essential to understand the effectiveness of the model in solving a given problem. The most commonly used metrics in classification problems are presented in Table 2.2.

Table 2.2: Most commonly used classification metrics.

Metric	**Explanation**	**Formula**
Accuracy	Proportion of correctly classified instances out of the total number of instances given.	$\frac{(TP + TN)}{(TP + FP + TN + FN)}$
Recall / Sensitivity	Proportion of positive instances correctly identified.	$\frac{TP}{(TP + FN)}$
Precision	Proportion of correctly identified positive instances out of all positive predicted instances.	$\frac{TP}{(TP + FP)}$
F1-Score	It measures the effectiveness of a model in terms of simultaneously correctly predicting positive instances (precision) and capturing all positive instances (recall).	$\frac{2 * precision * recall}{precision + recall}$
AUC	Area Under Curve – it measures how well the model separates classes using True Positive Rate values (TPR) and False Positive Rate values (FPR).	N/A
MCC	Matthews Correlation Coefficient – it measures the quality of binary classification predictions by assessing the balance between true and false positives and negatives.	$\frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$

Regardless of the nature of the problem and the type of prediction, there is always one main goal in building a predictive model: predicting something based on the inputs provided. To achieve this, the model should be able to generalize its predictions beyond the data on which it was trained. There are several techniques to

make that possible, which are presented in detail in Raschka's work [67]. In general, the recommended techniques evaluate the model in three different sets. Training, validation, and testing:

- **Training set:** Used to train the model, containing examples with input features and respective label/target. Moreover, it is possible to measure the model's performance during training.
- **Testing set:** It is a portion of data used to evaluate model performance after the hyperparameter tuning process. It prevents the model from overfitting and overestimating results after adjusting its parameters.
- **Validation set:** Unseen data in the training and validation phase. Its purpose is crucial, as it enables the evaluation of the model's ability to generalize and perform well on real-world data.

2.2.3 Model Interpretation

The use of ML models has been growing dramatically and is now in almost all fields of our society, from the technology used in people's daily lives to health, finance, transportation, industry, and many other fundamental sectors. As the ability of models to perform complex tasks increases, it is essential that there is also an increase in their transparency and interpretability. From this extreme necessity comes the field of Explainable Artificial Intelligence (XAI).

If the medical field was seen as an example, AI decision support systems offer a powerful opportunity to improve clinical care around the world. But all this decision-making power must have a logical basis that is perceivable to the person who will ultimately make the decision, the physician. This is something that XAI researchers have been introducing and working on for the past few years [68]–[70].

The purpose of XAI is to show the logic and reason behind a prediction made by the ML model to a person in a way that the person can clearly understand the output [71]. Normally, explainability is divided into 3 different degrees: Pre-modeling explainability, interpretable model, and post-modeling explainability [72]. Figure 2.9 presents a representation of these three different types of interpretabilities.

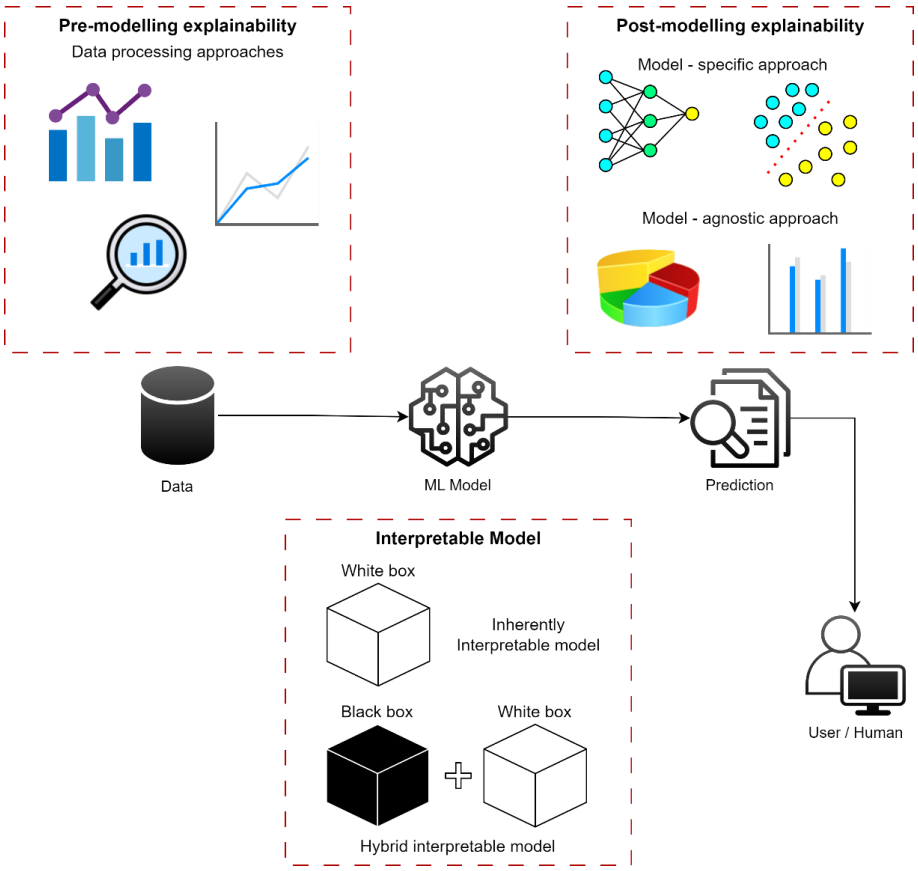

Figure 2.9: Representation of the three main types of explainability.

The use of each of these approaches often depends on several factors, such as the nature of the problem, the type of data, the models chosen, and the practical application the prediction will have. These types of explainability can be defined as follows:

- **Pre-modeling explainability:** Refers to all the knowledge and understanding that precedes the construction of the ML model. It involves a series of processing steps aimed at gaining knowledge of the domain, the data and all the steps needed to properly build a training set. Exploratory data analysis, transformation, and summarization techniques are part of this process.
- **Interpretable model:** This is a type of interpretability associated with the model, where it is possible to understand a certain result just by looking at the summary and parameters of the algorithm. On the one hand, this explainability can be inherent to the model, as, for example, in linear models or decision trees. On the other hand, we can have a hybrid model, which combines an inherently unexplainable model (black-box) with a model possessing inherent explainability (white-box). For example, a hybrid model could combine a black box model like neural networks or SVM with interpretable rule-based models such as decision trees and logistic regression [73].

- **Post-modeling explainability:** This type of explainability technique aims to address the black box nature of complex ML models, which are often difficult to interpret due to their high dimensionality and complex internal processes, often amounting to trillions of parameters [74]. These are divided into model-agnostic and model-specific approaches. As the name implies, the first type of approach is model independent and can be applied to any ML algorithm, such as the Local Interpretable Model-agnostic Explanations (LIME) [68] and the SHAP techniques [75]. Model-specific approaches are model-specific techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) for Convolutional Neural Networks (CNNs) [76] or the attention mechanism in Recurrent Neural Networks (RNNs).

Although efficient interpretability techniques already exist, there is still a lot of difficulty in building a fully transparent and interpretable AI system with great predictive performance. This is due to several challenges and limitations that still persist. One of them is the general inability of the current algorithms to actually provide a concrete reason behind the prediction, which justifies the low acceptance of AI systems in healthcare settings [77]. The well-known trade-off between accuracy and interpretability is another reason for why this is very difficult to achieve. Generally, more complex models such as deep neural networks get better results but are much less explainable than simple models of an interpretable nature such as a decision tree. Effectively communicating complex AI models and their explanations to non-expert users is a challenge. The explanations should be presented in a format that is understandable, meaningful, and induces trust. Bridging the gap between technical concepts and user comprehension is essential for successful adoption and acceptance of AI systems.

Explainability depends on all steps and decisions made throughout the ML pipeline. Despite the existing limitations, if we make optimal decisions for a given problem, it is possible to create a predictive system that performs well and is understandable, providing good result explanations. Explainable Artificial Intelligence (XAI) is a field that has experienced significant advancements in recent years. Given the expanding presence and impact of AI in society, it is certain that XAI will continue to be a highly researched topic in the upcoming years.

2.2.4 Applications

With the birth of the digital age, there has been an exponential increase in the amount of data generated and stored. This was the main driver for technical developments in IA, and today there are systems supported with ML models in almost all areas of society. Using different algorithms, ML enables data-driven decision-making, automation, and optimization. Several benefits may arise from the increased investment and growing performance associated with these systems, making them able to solve some of the most difficult challenges in education, finance, manufacturing, healthcare, military, cybersecurity, and many others.

One of the main branches where AI has the greatest potential and applicability is in education. It can adapt learning to each student through their background (adaptive learning) and personalize the learning experience by identifying the most efficient methodology for each student (personalized learning). Holmes et al. [78] describe many other ways in which AI can have a great impact on education.

Financial institutions use ML algorithms to analyze market trends, predict stock prices, and detect fraudulent activities. By applying advanced models to vast amounts of financial data, AI systems have enhanced risk assessment, improved trading strategies, and enabled real-time fraud detection. The work of Lin shows the brutal impact that AI had not only on finance, but also on the law system [79].

In cybersecurity, ML is used to detect and prevent cyber threats. By analyzing network traffic, user behavior, and system logs, ML algorithms can identify anomalies, detect intrusions, and improve threat intelligence, strengthening the security posture of organizations and individuals.

The impact that ML has had on manufacturing is also very considerable. These algorithms power the development of branches such as predictive maintenance, quality control, failure detection, process optimization, supply chain management, and even product management, design, and innovation.

Among the various fields harnessing AI's capabilities, the medical field stands out as an area with immense disruptive potential. The application of AI in medicine has two main branches: the virtual branch and the physical branch. The virtual branch consists of the application of ML and DL to create systems (software) for physician decision support, as well as patient management and treatment. The physical branch focuses more on real and tangible objects, medical devices, with strong ties to the area of robotics [80].

Current applications of ML in medicine can be found in the areas of cardiology, pulmonary medicine, endocrinology, nephrology, gastroenterology, neurology, cancer, and general image-based diagnostics [81]. There are multiple clinical aspects in which the application of ML has grown greatly and may have a significant impact in the near future.

ML algorithms are capable of analyzing X-ray, magnetic resonance (MRI), and computerized tomography (CT) images to assist in diagnosis. This allows a fast and automated detection of anomalies, tumors or other types of problems that could go unnoticed by the human eye. ML algorithms focused on medical imaging have been extensively researched, with several significant advances in recent years [82].

The rapid prognosis and diagnosis of a disease is crucial to start treating and mitigating its effects as early as possible. By analyzing large amounts of patient data, an ML model may be able to learn to identify certain patterns that may indicate the presence of a disease even before it manifests or progresses to more advanced stages. A clear case of this is the progress that has happened when ML is applied to electronic health records (EHRs). This allows professionals to create decision

support systems to predict adverse events, evolution of a disease, possible unexpected reactions to medications, and even suggest a possible treatment path to follow [83], [84]. This opens the door to AI-supported medicine guided by patient data.

Another medical area where ML algorithms can make a huge impact is in drug discovery and development. The application of ML can accelerate this process by analyzing chemical, biological, and other types of data. Not only that, but potentially predict the efficiency of a drug still being studied, identify possible new therapies, and also optimize drug creation both in terms of components and efficacy [85].

With the increase in portable devices that people carry around on a daily basis and the amount of data they continuously generate, there is a great opportunity to use ML for remote patient monitoring. Through vital sign data, activity level, and other health indicators, ML algorithms can alert in real time to different health risks, bad behaviors, or even positive feedback for each activity that contributes to patient treatment [86].

The closer the primary health care systems are to the people, the better chance they have of accessing this essential service. Health chatbots and virtual assistants have the potential to dramatically reduce this distance to the virtual branch and, therefore, in a sense, almost instantaneous. These apps can provide basic health advice, schedule, and manage appointments, check symptoms, and do something similar to a triage service, among other potential functions. This can not only help patients but also reduce the burden on healthcare professionals [87].

Throughout this section, it was presented the background knowledge needed to understand the necessary context about AI, ML and DL. The next chapter presents a literature review where the concepts of AI and ML are combined with DN in order to create predictive models capable of tracing the risk of disease evolution in various patients.

3 LITERATURE REVIEW

This chapter focuses on the integration of ML techniques with EHR data to develop a predictive model to assess the risk of DN onset or progression. Chapter 3.1 sets the stage by providing a comprehensive overview of the general medical context in which ML techniques are applied to EHR data. Chapter 3.2 provides a review of the literature that explores previous studies proposing different approaches to create predictive ML models for the risk of onset or evolution of DN using EHR data.

3.1 Context

Digitalization has allowed hospitals to store the complete history of patient appointments in a database, resulting in the availability of EHRs. These data are longitudinal because they are collected over time and include multiple patient records at different points in time. Due to the progressive nature of many diseases, a longitudinal approach is usually required to fully assess their development and impact [88]. Given the chronic and long-term nature of diseases such as DN, it is crucial to consider the temporal dimension of patient data and not overlook its importance [89]. The timely implementation of a DN risk assessment may delay or even prevent its progression, which would certainly reduce the number of people with ESRD [90].

EHRs can be defined as longitudinal electronic records that capture a wide range of patient health information from multiple care delivery settings. This covers a range of data, including patient demographics, progress notes, medical issues, prescribed medications, vital signs, medical history, immunization records, laboratory results, radiology reports, and potentially other types of data [91].

While the potential in the use of EHR data is very high for advances in the predictive ability of ML models, the various challenges associated with this type of data must be taken into account. Some major challenges found in the literature [92], [93] include the following:

- **Data irregularity:** Irregular time intervals between recordings present a significant challenge to ML models, as they lack a consistent structure both across different patients and within individual patients. This temporal irregularity, although potentially containing valuable information, is not easily handled by most ML architectures, which are designed for data with fixed and regular time intervals. As a result, effectively incorporating and leveraging this irregular temporal aspect becomes a major obstacle in building accurate ML models for EHR data analysis.
- **Data sparsity:** A record in the context of EHR data represents a medical event or data input, such as a medical appointment or examination. However, it is important to note that these records often lack information on some

variables. Additionally, it is observed that patients who are in better overall health tend to have fewer medical visits compared to those who are in poorer health. This discrepancy in healthcare utilization leads to incomplete or non-existent patient information within the records. Even when information is present, it often varies significantly between different patients, making it challenging to establish consistent and comprehensive patient profiles across the EHR dataset.

- **Data heterogeneity:** EHR data encompasses a wide range of patient records and can represent diverse conditions and varying outcomes. Within this rich and heterogeneous dataset, the identification of patient sub-cohorts, characterized by more closely related groups of individuals, holds the potential to enhance downstream analyses such as cohort analysis and personalized medicine.
- **High-Dimensionality:** Usually, this type of data contains several types of information, which leads to high dimensionality. This is a challenge when creating predictive models due to the higher complexity and also the "curse of dimensionality", which states that as the number of variables increases, it becomes more difficult to make meaning of the data and draw reliable conclusions [94].

Despite these limitations, different ML algorithms have been applied to EHR data to create systems capable of being of practical use in medical assistance. These algorithms can have an impact on disease diagnosis, risk stratification, decision support systems, and even allocation of clinical resources.

ML models are used to assist in the diagnosis of various diseases using EHR data. They learn patterns and associations in the data to help identify potential diagnoses or assist in differential diagnosis. Garcelon et al. [95] has reviewed the use of EHR data and ML algorithms to create predictive models that can identify rare diseases in patients. Although general models capable of identifying several diseases are very difficult considering both the inherent characteristics of the data and the nature of the problem, several approaches capable of identifying a specific disease are presented for both tabular data and clinical images. Guo et al. present an approach capable of applying ML models to predict heart failure using variables derived from EHR data, which include demographic records, medical notes, lab tests, and images [96]. Another example can be seen in the work of Sun et al., who presents a comparison between five ML algorithms applied to EHR data to diagnose diabetic retinopathy [97]. There are also many other works that have succeeded in training ML algorithms on EHR data to create predictive models applied to different diseases, e.g. diabetes disease [98], chronic myelogenous leukemia [99], pulmonary hypertension [100], and neurodegenerative diseases [101].

ML models use EHR data to stratify patients into different risk groups based on factors such as disease prevalence, comorbidities, and genetic markers. This helps to target preventive measures and personalized interventions. There is some work that

applies ML algorithms to various EHR data sets in order to create models that can predict to which risk group a patient belongs. Zeiberg et al. trained a risk stratification ML model capable of predicting the likelihood of acute respiratory distress syndrome (ARDS) at different points during a patient's stay in the hospital [102]. In another approach, Yang et al. propose an ML-based stratification system to identify pregnant women at risk of hyperglycemia, which means developing gestational diabetes [103]. Some other work could be mentioned, such as an ML model capable of stratifying hospitalized patients by the risk of developing acute kidney injury (AKI) [104], risk stratification of patients with chest pain using ML dimensionality reduction techniques [105] or the use of time series and ML models to stratify individuals into nonalcoholic fatty liver disease (NAFLD) [106].

ML models help healthcare professionals make evidence-based decisions by analyzing EHR data. They can provide recommendations for treatment plans, medication selection, and drug development. There are several studies done in this area. Christopoulou et al. presented a study showing the ability of DL methods applied to EHR data to identify drugs, associated medication entities, and interactions among them. This is essential to prevent adverse drug events [107]. In another study made by Issa et al. [108] it is discussed the use of computational strategies, particularly ML and DL methods to model biological processes, identify new disease-relevant targets, and discover associations between drugs and their effects. The idea presented consists of using these methods for drug repurposing in oncology. Other works can be highlighted, such as prediction of treatment effect [109], and predict the outcome of antidepressant treatment in patients with depression [110].

Another function of ML models applied to EHR data can be to optimize resource allocation, such as predicting patient demand, optimizing hospital bed utilization, and improving procedure scheduling. This is shown in the work of Avati et al. where ML is applied to EHR data to improve access to palliative care and facilitate timely interventions for patients in need [111]. In another study, Levin et al. showed that using ML based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length of stay [112]. Works like the use of ML techniques to predict duration of hospitalization in COVID-19 patients [113] or to predict the demand for inpatient beds [114] could also be covered here.

Everything presented in the previous paragraphs shows that there is a multitude of possibilities in which ML applied to EHR data can improve healthcare, both for patients and for physicians, nurses, and medical staff. However, this work focuses on the use of ML models combined with EHR data and applied to DN. In the next chapter, we will present the literature review applied to our specific case study.

3.2 ML Models to Predict Diabetic Nephropathy

Within the scope of this master's thesis, a literature review was conducted in order to identify different longitudinal approaches used to create an ML model capable of predicting the development or onset of DN over time. This article entitled *"Machine Learning techniques to predict the risk of developing diabetic nephropathy: a literature review"* has been submitted to the Journal of Diabetes & Metabolic Disorders [115], but the publication is still pending on the acceptance of the reviewer. This paper can be consulted in Appendix A.

The application of ML techniques to analyze EHR data can provide valuable insights and enable the development of ML models that can predict the risk of developing DN or progressing to higher stages, aiding physicians in the diagnosis and ultimately improving the quality of healthcare [116], [117]. There are many studies done on the use of ML to identify cases of DN. However, the focus of this research is to identify and study the approaches used in clinical EHR data collected over a period of time and the corresponding risk prediction of DN progression.

This literature review aims to answer the following research question:

- **RQ:** What are the most effective machine learning techniques used to construct a model that uses the temporal information in diabetic patients' EHR data to predict the development of DN or progression to higher stages?

3.2.1 Materials and methods

Three databases were used for this literature review: Scopus, Web of Science, and PubMed. These are three of the most popular and reliable sources of scientific information [118]. Only articles written in English and published between January 2015 and December 2022 were included. The search query used was:

- "((diabetes) AND ((machine learning) OR (deep learning)) AND ((time) OR (temporal) OR (time series)) AND (predict) AND ((kidney disease) OR (nephropathy)))".

Figure 3.1 describes the methodology used throughout the process. The first step (Identification) resulted in a total of 164 papers. Based on the references of some of these papers, a further 11 were identified as potentially important, resulting in 175 papers for further analysis. These 11 additional articles were referenced by papers identified in the first stage. During the screening phase, 48 duplicates were removed. In addition, 85 papers were excluded by title and 14 by abstract. These were removed because they did not relate to the intended topic; this phase reduced the original 175 to 28 papers. Of these, only 11 were eligible according to the various criteria defined.

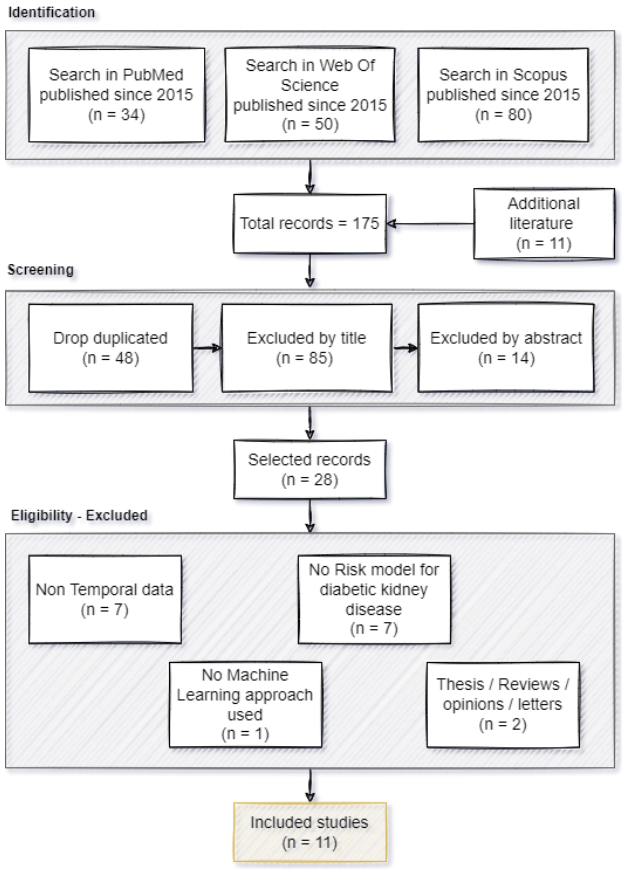

Figure 3.1: Methodology.

Table 3.1 shows a summary of the excluded articles, the criteria, and a brief explanation of the exclusion criteria.

Table 3.1: Papers excluded according to defined criteria.

Papers	**Criteria**	**Brief explanation**
[119]–[125]	Non-temporal data	Excluded papers did not include temporal data, i.e. data from patients followed up during a specific time window with information collected during that time.
[126]–[132]	No risk model for DN	We select articles that predict the risk of progressing or developing DN. Articles that only classify whether patients have the disease or not were excluded.
[133], [134]	Thesis / Review / Opinions / Letters	As these papers are reviews of the literature, this type of paper is not included.
[135]	No ML approach	This paper has used a scoring system that defines the factors that contribute most to the development of DN. Although it is a risk model, it is not an ML approach.

It should be noted that although the keyword "deep learning" was included in the search query, none of the 11 selected papers used DL techniques to solve the

problem. With this in mind, we will focus only on approaches that use ML algorithms.

Following the procedure outlined in Figure 3.1, 11 articles were included in this review. AI applied to temporal clinical data has the potential to improve the way a patient with diabetes is managed according to their risk of developing DN. The different approaches are presented according to different questions: i) which features are most important, ii) what kind of ML models have been created, iii) which ones perform better, and iv) other relevant aspects. The papers selected for this review, together with a summary of their main aspects, are listed in Table 3.2. It is possible to note that most of the articles were published in the last 2-3 years, showing a rapid growth in the application of ML to the management of diabetes-related conditions, taking advantage of the available large amount of clinical data. For the selected articles, information is provided on the source of the data, the importance of the variables for prediction, the approaches used to create the risk models, their interpretation methods, and their performance.

3.2.2 Data Sources

As mentioned earlier, EHR data can contain several types of data. The selected papers, in addition to clinical variables, use, in some cases, Omics-based biomarkers. These can be defined as a molecular signature that is identified using omics data and used to predict the presence or risk of a particular disease or condition, or to monitor the response to a particular treatment. Omics can be divided into different research areas such as proteomics (proteins), transcriptomics (RNA), genomics (genes), metabolomics (metabolites), lipidomics (lipids), and epigenomics (methylated DNA) [136].

The integration of omics data with clinical data can significantly improve the ability to analyze and predict complex diseases using ML. Such integrated analysis can help create models that can clearly explain diseases, enabling real knowledge that leads to improved treatment and a better quality of life for patients [137]. The work of Al-Sari et al. [138] is a very good example of the benefits of combining Omics data with clinical data. The performance of some of the models, which had previously been built using only clinical data, increased significantly when Omics data were included. In this case, metabolites, ketones, and sugar derivatives were used. In general, the integration of molecular data will lead to better prognostic models, as demonstrated in several works [139]–[142]. Despite the many benefits of integrating this type of data, there are some challenges. Sometimes, even when these data are available, they are very difficult to handle, process, analyze, and finally integrate. This requires specialized knowledge in the branches of mathematics, statistics, biology, and computer science [143].

Table 3.2: Summary of studies included in this review.

Paper	Dataset	Pre-processing	ML Model	Performance
Singh et al. (2015) [144]	EHR data of patients in the Mount Sinai Hospital and Mount Sinai Faculty Practice Associates in New York City. From 6435 patients, 12 337 examples were extracted.	Feature selection and generation. Numerical predictors discretization into four bins based on the quartiles of the corresponding predictor and then map them into binary variables.	Multitask Logistic Regression (MLTR)	≈ 68.3% AUC for Threshold of 10% ≈ 71.2% AUC for Threshold of 20%
Dagliati et al. (2018) [145]	943 T2DM patients in charge of the ICSM hospital and followed for more than 10 years.	Data imputation with the MissForest technique and some variables were not considered because imputation errors were too high.	LR	3 years: 70.1% AUC 5 years: 73.4% AUC 7 years: 72.1% AUC
Makino et al. (2019) [146]	Dataset with 64 059 T2DM patients. From that, authors extracted structural, text, and longitudinal data.	Under sampling minority class, several data transformation steps are used to summarize the last 180 days EMR records and create longitudinal data variables.	LR	AUC: 74.3%
Romero et al. (2019) [147]	Data were provided by the NHLBI, sponsor of the ACCORD trial. There were 10 251 T2DM patients from 77 clinical centers in the United States and Canada.	SMOTE technique used to balance target, feature selection using the information gain metric.	RF	88.7 % Accuracy
Sarkosh et al. (2020) [148]	Clinic of Imam Khomeini Hospital Complex (IKHC) dataset with 10 636 T2DM patients followed from 10 years (2012-2021).	Feature selection using Recursive Feature, elimination (RFECV) and RF method, imputation or drop missing values	LR	75.5% AUC
Aminian et al. (2020) [149]	287 438 T2DM patients from Cleveland Clinic's EHRs followed between 1998 and 2017. Two different groups were created: 2287 patients undergoing metabolic surgery and 11 435 matched non-surgical patients.	Missing data imputed using multivariate imputation by chained equations (MICE), variables with more than 25% missing values or no predictive value were removed.	RF	Surgical patients: 73% AUC Nonsurgical patients: 76% AUC
Song et al. (2020) [150]	University of Kansas Medical Center's HERON clinical data repository with 35 779 T2DM patients.	Features with less than 1% representation were removed and missing values imputed.	GBM	AUC: 83%, 78% and 82% in predict DN in 2, 3 and 4 years, respectively.
Chan et al. (2021) [151]	BioMe Biobank at the Icahn School of Medicine at Mount Sinai and the Penn Medicine Biobank data sources. Population of 1146 T2DM with both EHR data and biomarkers.	Data harmonization, only variables in more than 70% participants were included, feature selection based on SHAP values, and missing data imputation.	RF	AUC: 77%
Allen et al. (2022) [152]	111 046 EHRs of T2DM patients that represents more than 700 healthcare sites from USA between 2007 and 2020.	Standardization, impute missing values.	RF	74.8% AUC for any DN stage, 82.3% for stage 3-5, 82.1% for stage 4-5
Dong et al. (2022) [153]	Data from PLA General Hospital with 2809 T2DM patients that were followed from 2008 to 2019.	Drop features with missing data > 25%, missing values imputation with RF, feature selection using RFE.	LightGBM	AUC: 81.15%
Al-Sari et al. (2022) [138]	T1D cohort in Steno Diabetes Center Copenhagen (SDCC) with 537 patients with follow-up data. Later, blood molecular data with 965 features was also included.	Remove high correlated features, outliers, and clinical variables with no predictive power on metabolic data. Feature selection based on SHAP values.	RF	DN model with only clinical data: 92% of AUC, DN model with clinical and omics: 99% AUC

3.2.3 Feature Importance

Most of the selected studies used different methods to understand which variables have the greatest influence on the outcome when predicting risk. Some of these techniques were used to perform feature selection to remove redundant and irrelevant variables, which can potentially lead to better performance [154].

The work of Chan et al. [151] and Al-Sari et al. [138] used SHAP to understand how each feature contributes to the model's predictions by estimating the amount that each variable contributes to the predicted value of an output. This allows them to ensure that they are selecting the most optimal set of variables for the task.

Recursive Feature Elimination (RFE) is an iterative method that can recursively remove the least important features from a dataset and build a model on the remaining attributes. It iterates until the desired number of features is obtained. As presented in Sarkosh et al. [148] and Dong et al. [153], this technique is very useful for selecting a subset of features that aggregates the most important features from a larger dimensional space. In both cases, a variant of this method, Recursive Feature Elimination with Cross-Validation (RFECV), is applied. It uses cross-validation to evaluate the performance of the model at each iteration.

A very similar approach was adopted by Makino et al. [146] and Dagliatti et al. [145] with their LR stepwise feature selection method based on the Akaike information criterion (MLC). Stepwise feature selection is a method of selecting a subset of features by iteratively adding or removing variables. The MLC is a trade-off between model goodness and complexity, and measures the relative quality of a statistical model [155]. It can be used in stepwise feature selection to evaluate the performance of the model at each step and decide which feature to add or to remove. Although it appears similar to the RFE method, this technique trains on the selected subset of features at each step and can use either forward selection or backward elimination, whereas RFE trains on all features and removes the least important feature at each step.

Aminian et al. [149] computed the relative importance of each feature in the final model using MLC for the regression models and the Concordance index (C-Index) for the RF models. The C-Index is a metric that considers the temporal dependence associated with the model result and can be used to rank features by importance or even to analyze the global performance of the model.

A simpler and faster approach based on Univariate feature selection to select the most relevant variables was proposed by Singh et al. [144]. These features are selected based on univariate statistical tests between the feature and the target variable and do not consider dependencies and relations between features.

Song et al. [150] adopted a slightly different approach, using the GBM classifier because it uses an embedded method of feature selection during model training. This allows the most important features to be selected and the model retrained using only these variables.

Table 3.3 shows the clinical variables that were mentioned in more than three papers as one of the most prominent variables able to give high predictive power to the model for analyzing the emergence or development of DN, and their respective importance.

Table 3.3: Most important clinical variables identified.

Papers	Feature	Meaning
[138], [147], [151], [153]	eGFR or GFR	Glomerular filtration rate (GFR) measures how well the kidneys work. eGFR is an estimate, usually calculated using the Modification of Diet in Renal Disease (MDRD) equation and the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation.
[138], [147], [151], [153]	UAlb or Alb	Albumin levels in the blood. Low levels of this protein are called hypoalbuminemia, and high levels are known as hyperalbuminemia
[138], [145], [148], [153]	HbA1c	Glycated hemoglobin (HbA1c) measures glucose levels over the past 2 to 3 months.
[147]–[149], [151]	UACR or ACR	Laboratory tests are used to detect proteinuria, the presence of protein (usually albumin) in the urine.
[148]–[150], [153]	Age	In some articles, it is the age of the patient, in others it is the age at which the patient started to be followed.
[145], [148], [149], [153]	BMI	Body Mass Index uses a person's height and weight to calculate an estimate of body fat.

3.2.4 Cross-sectional studies

Cross-sectional studies are a type of observational research design that aims to collect information about a population or a specific group at a certain moment in time. In this context, this chapter describes various approaches for constructing an ML model capable of predicting the risk of developing DN. However, it is important to note that the presented approaches focus primarily on cross-sectional or static methods, which do not fully exploit the temporal aspect of the EHR data. Only an overview of these methods will be given, as there is more interest in presenting details for longitudinal methods in Chapter 3.2.5. These methods attempt to deal with the time factor associated with the EHR data and are therefore of greater importance to this study.

Dong et al. [153] used data from non-DN patients at baseline who were followed for three years. It used baseline features, and the binary classification predicts the presence or absence of DN within 3 years. Romero et al. [147] followed a similar strategy, but defined eight different time windows for the 7 years of patient follow-up data. Each window corresponds to one year of data, except for the first two windows, which correspond to only 6 months each. Dagliatti et al. [145] also used a binary outcome variable but for three different time thresholds of 3, 5, and 7 years to predict the risk of DN.

Aminian et al. [149] used data from both surgical and non-surgical patients with T2DM. Multivariate time-to-event regression and RF ML models were created to predict the 10-year risk of developing DN. The 10-year risk of morbidity and mortality was estimated for patients with and without metabolic surgery. Chan et al. [151] uses clinical data and biomarkers to generate risk probabilities. The authors named the whole system IntelKidneyX, presented before in Chapter 2.1.8.

Al-Sari et al. [138] and Makino et al. [146] did almost the same as the previously cited papers. The former designated the outcome as progressor or non-progressor, while the latter classified it as worsening or stable.

Unlike the works presented above, Allen et al. [152] are able to predict 3 different outcomes, DN progression to any stage, DN progression to stages 3-5, and DN progression to stages 4-5.

Figure 3.2 provides a general overview of the different approaches described above.

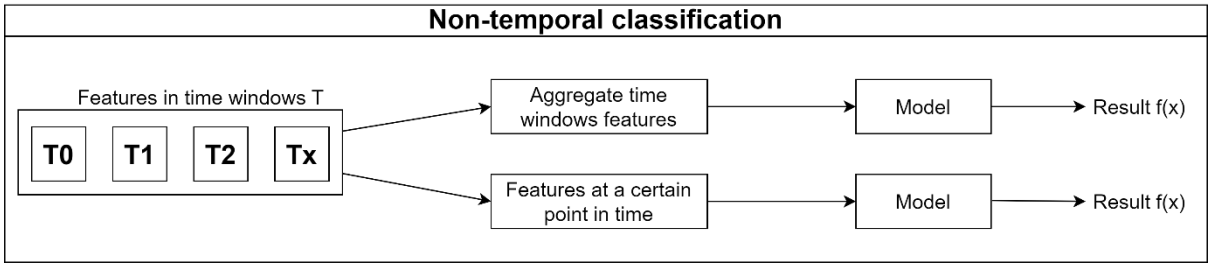

Figure 3.2: Non-temporal/Static approaches.

3.2.5 Longitudinal studies

Longitudinal studies, in contrast to cross-sectional designs, track participants or entities over an extended period, collecting data at multiple time points. These studies enable researchers to observe changes, trends, and patterns over time, making them valuable for understanding the dynamic nature of DN variables. Different temporal approaches have been proposed to deal with EHR and provide risk prediction for DN. Within the remaining selected articles, the following approaches were used: stacked temporal, multitask temporal, discrete survival, and landmark boosting.

The stacked temporal technique was used in both Singh et al. [144] and Song et al. [150] work. It aggregates the data within each time window and uses the data from all time windows to make a final, unique prediction. The T time windows, with F features in each, result in only one time window with T multiplied by F features. One of the disadvantages of this technique is that the larger the temporal space considered, the higher the dimensionality of the data, which can lead to a large overfitting. In Figure 3.3, the physician appointments within each time window are aggregated to form a one-dimensional space, which is then fed into the model and a prediction is obtained.

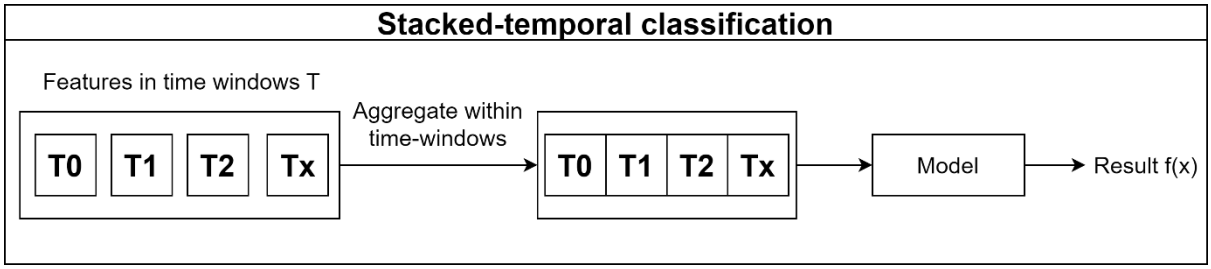

Figure 3.3: Stacked temporal approach.

The multitask temporal method was also proposed in the paper of Sing et al. The authors decided to predict the outcome for each time window separately. Each time window must have at least five physician appointments within that time. When predicting the risk of DN for a new patient, each time window with five or more appointments is used, and the result is the average of the different results obtained in each time window. This stratification of the problem is shown in Figure 3.4, where it can be seen that the ML model operates independently in each time window, and the result is the average of the different results obtained.

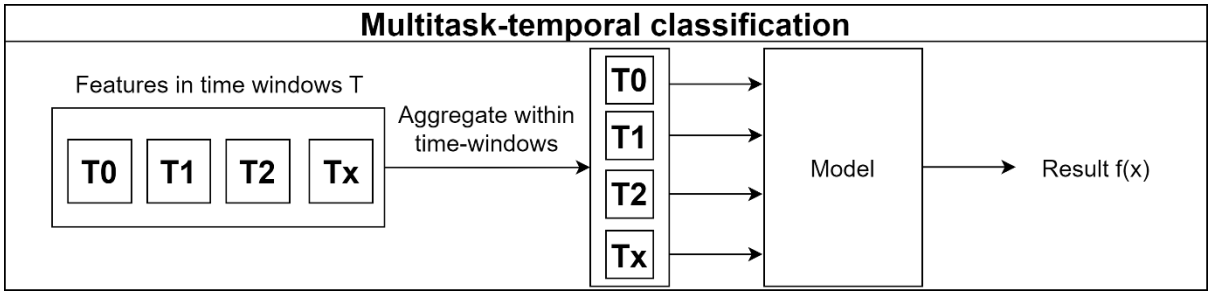

Figure 3.4: Multitask temporal approach.

Discrete survival and landmark boosting are two techniques mentioned in the paper by Song et al. The first makes an individual prediction in each time window, with no overlap between windows. A disadvantage of this technique is that it assumes that there is no relationship between examples in different time windows, even if they come from the same patient. This can be seen in Figure 3.5.

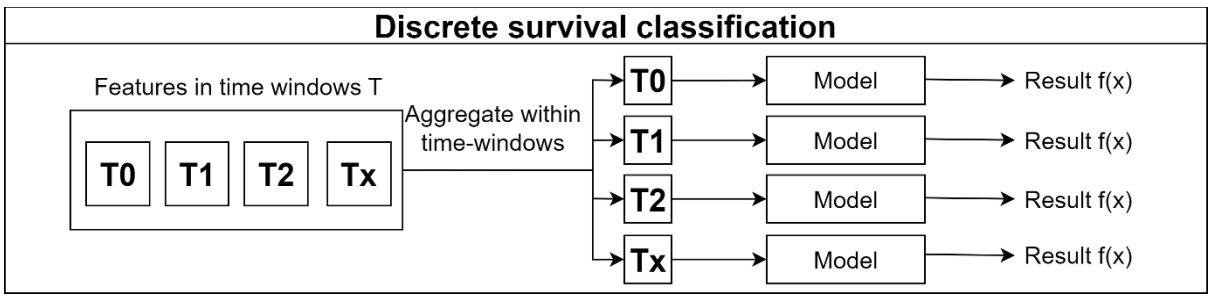

Figure 3.5: Discrete survival classification.

On the other hand, landmark boosting is very similar to discrete survival, but in each time window T, the prediction made in the previous time window T – 1 is also considered. In effect, there is a transfer of knowledge between the time windows, making each prediction more accurate. This can be seen in the representation of the

approach shown in Figure 3.6, where each model receives not only the features corresponding to a time window, but also the prediction made in the previous time window.

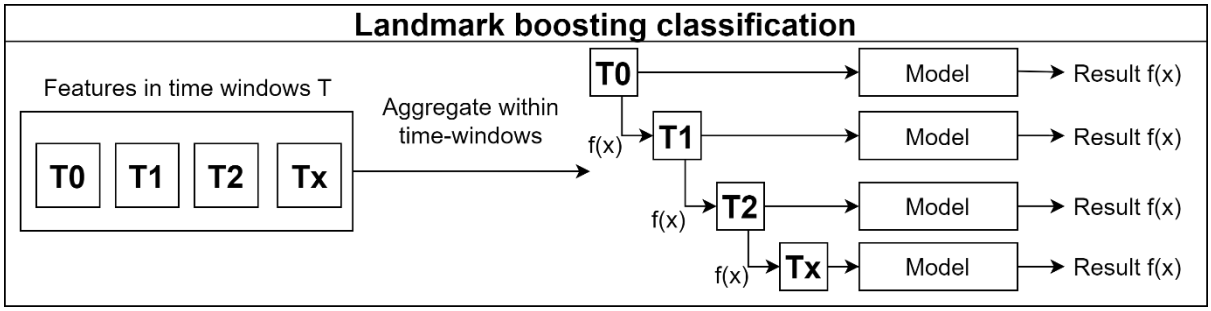

Figure 3.6: Landmark boosting classification.

3.2.6 Performance and Interpretation

This chapter discusses the types of models most commonly used to predict the onset or development of DN. It also presents the main interpretation techniques used and a performance comparison.

Considering the selected papers, five different classifiers were proposed: RF, LR, LightGBM, GBM, and Multi-Task Logistic Regression (MTLR). In Figure 3.7, we can see that the method most selected was RF, followed by LR, and finally LightGBM, GBM, and MTLR, which were selected only once.

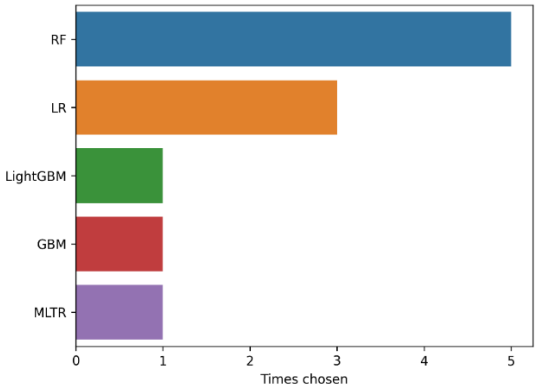

Figure 3.7: Most used ML classifiers in proposed methods.

It is possible to identify three main techniques to interpret the results generated by the predictive model: i) SHAP values, ii) monograms, and iii) decision tree visualization.

SHAP values were proposed by Lundberg et al. in 2017 to analyze the model predictions [156]. It calculates the importance of each feature for a given prediction, where each feature can have a positive or negative impact on that specific prediction. The contribution of features can be local (each observation) or global (set of

observations). In this particular case, local explanations aim to show the reasons that lead to a certain result generated by the model for a specific patient. On the other hand, global explanations aim to show which variables were most important for the overall predictions of the model. These are calculated by aggregating the different local explanations.

Nomograms are graphical representations of LR models. They work like scoring systems, where each feature is assigned a certain number of points according to its value, and the result varies according to the number of points accumulated in the sum of the different features [157].

Finally, some of the articles used only tree-based models because they can be interpreted directly by visual inspection of the associated decision tree. RF is an ensemble of many independent trees, and the output is based on the outputs of multiple decision trees. By looking at the different decision trees, it is possible to see which features are used to make predictions, the importance of each feature, and the overall patterns of predictions [158].

Some papers predict the onset of DN, some predict the worsening, and some authors predict the worsening for specific stages of the disease. In addition, there are papers where the result corresponds to only one specific time window, while others implement a different prediction for each time window, considering a certain number of years. This heterogeneity makes it difficult to compare their performance directly. Table 3.4 provides detailed information on each of the proposed methods.

Table 3.4: Details and performance of proposed methods.

Method	Time range	Outcome variable	Performance metrics
RF [152]	5 Years	Multiclass (DN advance to any stage, DN advance to stage 3-5, and DN advance to stage 4-5)	Any stage - AUC: 0.748, Sensitivity: 0.7, Specificity: 0.662. DN stage 3-5 - AUC: 0.823, Sensitivity: 0.750, Specificity: 0.739. DN stage 4-5 - AUC: 0.821, Sensitivity: 0.751, Specificity: 0.712
RF [149]	10 Years	Binary target (morbidity and mortality risks)	AUC: 0.76
RF [151]	5 Years	Binary (ESRD or no ESRD)	AUC: 0.77
LR [145]	3, 5 and 7 years	Binary (development or absence of DN)	3, 5 and 7 years. Best result (3 years): Accuracy: 0.647, Sensitivity: 0.820, Specificity: 0.730 and AUC: 0.808.
LightGBM [153]	3 years	Binary (DN presence or absence)	Accuracy: 0.768, Sensitivity: 0.741, Specificity: 0.797 and AUC: 0.815.
LR [146]	6 months	Binary (DN stable or aggravation)	Accuracy: 0.701, AUC: 0.743
RF [138]	Non-defined	Binary (DN progression or no progression)	Accuracy: 0.96, AUC: 0.96
RF [147]	8 time-windows at a max of 7 years	Binary on each time window (development or absence of DN)	Accuracy: 0.887
LR [148]	5 years	Binary (DN presence or absence)	AUC: 0.758
MLTR [144]	5 years – time windows of 6 months	Binary on each time window (development or absence of DN)	≈ 68.3% AUC for Threshold of 10% ≈ 71.2% AUC for Threshold of 20%
GBM [150]	2, 3 and 4 years	Binary on each time window (development or absence of DN)	Best result (2 years): AUC: 0.830

The main findings that have emerged from this review of the literature are the following:

- There is very little work that takes full advantage of the time factor inherent in the EHR data. The works of Sing et al. [144] and Song et al. [150] are an exception. In fact, the landmark boosting method proposed in the Song et al. paper was the approach that took more advantage of the time factor. It not only predicts the risk in each time window, but also takes into account the result produced in the previous time window. Although this approach attempts to exploit the full temporal potential of EHR data, it could still be improved, as it considers all records as independent, but, in fact, they are not because the patient has multiple records (appointments).
- Combining omics data with clinical data can help better predict the risk of DN over time, as confirmed in the work of Al-sari [138]. Soon, this type of data will be linked to disease risk models because the information they contain is really valuable to increase the predictive power of the different risk models.
- Another important concern with clinical risk models is interpretability. Almost all the proposed models were selected not only because of their good performance, but also because they allow interpretation of the respective results.
- The vast majority of selected articles have been published recently (in the last 3 years), demonstrating the importance of studying existing clinical data (EHR) through longitudinal analyses, and the potential that these approaches can have in supporting patient follow-up and medical decision making.

Despite the great capabilities and improvements that these proposed models can potentially bring to medical care, the several papers reviewed have limitations that are clearly stated by the authors. Some of the most cited limitations are as follows:

- The patient sample was clinic-based rather than population-based, which means that the model was only tested on a particular dataset, extracted from the population of a particular hospital/clinic. Furthermore, in most studies, there is no external validation dataset, leading to great uncertainty about generalization to a wider population. Cabitza et al. [159] show how external validation is essential to build robust predictive models in medicine.
- Small data samples, too much missing data and missing important features. Models trained on a small amount of data can result in poor generalizability and lead to incorrect conclusions being drawn. Too much missing data can affect the consistency of the data across different visits by a given patient.
- Most selected papers assume independence among examples, which is inaccurate since multiple records from a single patient are present. Considering inter-record dependency is crucial to unlock the potential of temporal EHR data, leading to more powerful and accurate predictive models. Song et al. [150] simulated some inter-record dependency by passing the prediction between time windows.

The reviewed literature suggests that despite the potential of using ML techniques to fully exploit the temporal dimension of EHR data to predict the risk of developing or progressing to DN, this has not yet been fully achieved. Many of the techniques studied have limited use of the temporal dimension and richness of patient records available in EHR data. Many of these works have limited temporal use and fail to take into account the richness of patient records in EHR data. Approaches that rely solely on baseline values or aggregate different clinical visits into a single record neglect the temporal aspect. Some longitudinal approaches are in some way incomplete, with predictions separated by time windows and lacking inter-window correlation. However, the Landmark Boosting approach by Song et al. stands out by establishing correlations between time windows, predicting the disease state in the current window based on the previous window.

In summary, all the papers included in this review were generally able to arrive at a workable risk model for the onset or development of DN using a variety of techniques. There are a small number of longitudinal studies in the area of DN that translate into the creation of a predictive ML model capable of performing well, being interpretable, and properly validated for clinical application. There is also a need for these approaches to consider the patient's history, adding the temporal factor, which can be a key element to achieve not only better results, but also more reliable ones.

This review of the literature is of great importance to this work and serves as a fundamental pillar upon which the research is built. It plays a crucial role in establishing the context and justifying the significance of the chosen research topic. By conducting a comprehensive review of existing scholarly works, it was possible to gain a deeper understanding of the current state of knowledge in ML applied to DN disease. This examination helps identify gaps or limitations within the existing literature, setting the stage for this study.

4 MATERIALS AND METHODS

This chapter provides a comprehensive description of the data used in the study, along with the exploratory data analysis (EDA) performed. Furthermore, the preprocessing steps used to prepare the data for the ML algorithms are presented. The subsequent stages, which involve the creation and interpretation of the ML models, are also elaborated upon.

4.1 Methodology

The developed methodology comprises 5 main phases (Figure 4.1):

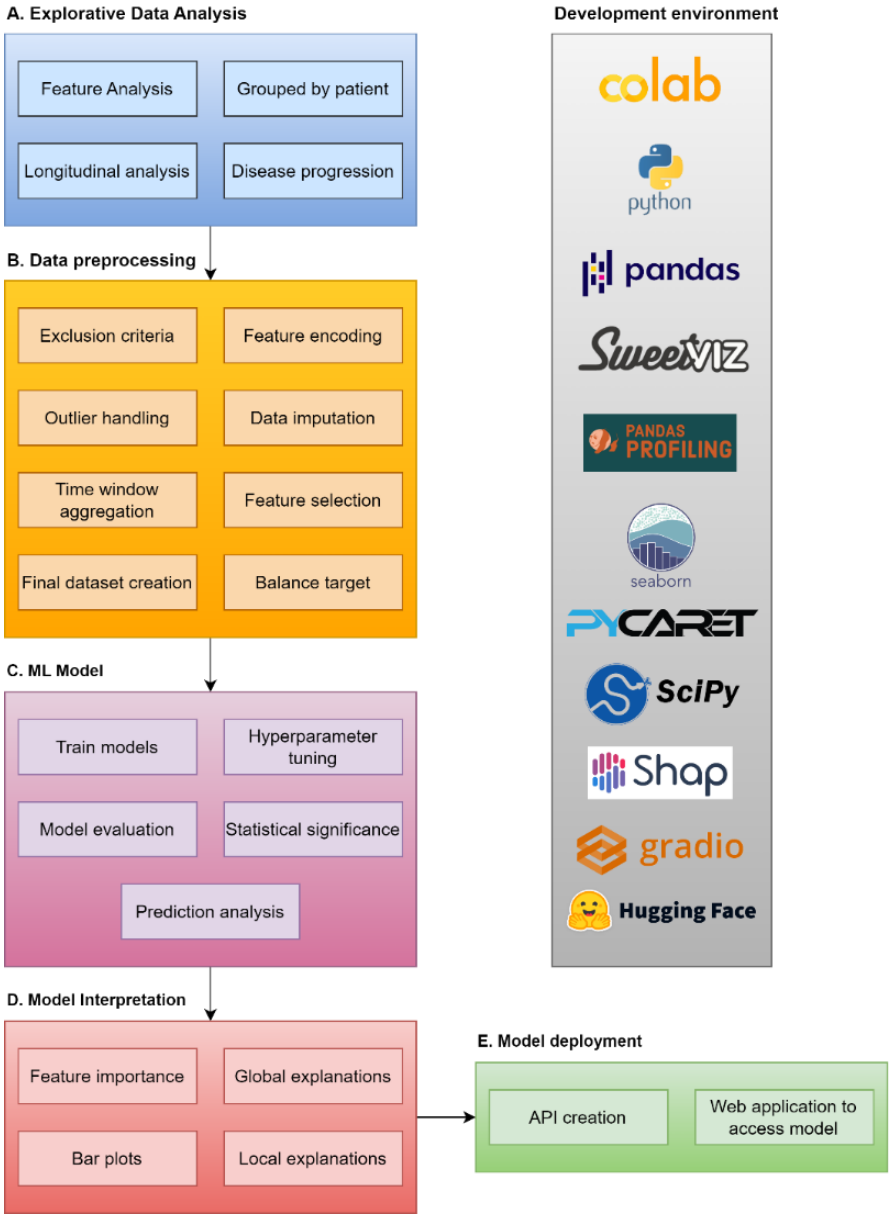

Figure 4.1: Methodology.

- **Explorative data analysis:** Several analyses are needed to understand the data. The process of examining and visualizing the data will serve to start designing the solution to the problem.
- **Data preprocessing:** Clean, transform and organize the raw data in order to be able to shape the solution and provide the data to the ML models.
- **ML Model:** Train diverse ML models on labeled data, tune hyperparameters and evaluate performance on different steps.
- **Model interpretation:** Understand and explain the predictions made by the ML model, presenting relevant information in an intuitive way that clearly shows the logic behind the process.
- **Model Deployment:** Make the trained ML model available to be accessed by anyone and operational to make predictions in real time.

In the upcoming chapters, the methodology will be presented in detail, specifically applied to the data used in this study. Nevertheless, it is crucial to emphasize that the methodology can be extended and applied to other EHR datasets as well.

4.2 Explorative data analysis

The dataset applied in this study was provided by APDP - Associação Protectora dos Diabéticos de Portugal. This organization maintains an electronic database that stores clinical and patient information. Annually, APDP treats an average of 18 000 patients, conducting over 200 000 medical appointments.

The APDP dataset offers a rich and diverse representation of DN patients. It contains demographic information and medical examination results from patients treated at APDP facilities. No external information from other EHR banks was included. The complete dataset provided corresponds to 413 097 clinical records of 21 284 patients, followed over 22 years (1998 to 2020). The data was delivered in an excel sheet, where each row represents a doctor's appointment for a particular patient. The extensive nature of this dataset provides a solid foundation for conducting in-depth analyses and drawing meaningful insights related to DN.

4.2.1 Feature analysis

The data consists of 39 columns. Table 4.1 shows an analysis of all the characteristics present in the data (features). It is important to note that MDRD (Modification of Diet in Renal Disease indicator) and CKD-EPI (Chronic Kidney Disease Epidemiology collaboration) are both essential variables that estimate the glomerular filtration ratio (eGFR). However, the calculation behind each one is different, depending on several factors such as the age and gender of the patient [160]. This study assigns more emphasis to the CKD-EPI since it is more recent than MDRD

as well as a more recommended way to calculate the eGFR [25]. Therefore, it is an essential variable to determine whether the patient has DN and in which stage.

Table 4.1: Dataset feature description.

Feature	Data type	Description	Unique values	Missing data (%)
ID	Number	Patient identification	21 284	0%
Race	Text	Patient Race	6	0%
Age	Number	Patient age	71 518	0%
Sex	Text	Patient gender	2	0%
Diabetes duration	Number	Years since diabetes onset	68 629	0%
Date of registration	Date	Appointment date	6023	0%
Weight	Number	Patient weight	1679	38%
Abdominal circumference	Number	Patient abdominal perimeter	264	40%
BMI	Number	Patient body mass index	210	50%
Systolic BP	Number	Systolic blood pressure	157	97%
Diastolic BP	Number	Diastolic blood pressure	97	97%
Pulse pressure	Number	Pulse pressure	237	27%
Potassium	Number	Potassium	435	57%
Total cholesterol	Number	Total cholesterol	3052	70%
LDL	Number	low-density lipoprotein cholesterol	3892	68%
HDL	Number	high-density lipoprotein cholesterol	1056	71%
Non-HDL	Number	Non-HDL Cholesterol	3581	71%
Triglycerides	Number	Triglycerides	6391	70%
Hba1c	Number	Glycated Hemoglobin	1041	44%
Albuminuria	Number	Urine albumin	29 584	76%
Proteinuria	Number	Urine protein	2389	28%
Creatinuria	Number	Urine creatinine	17 239	88%
Creatinine	Number	Serological creatinine	891	48%
MDRD	Number	Glomerular filtration rate estimate	170	51%
MDRD Stage	Text	MDRD Staging	5	51%
Delta MDRD	Number	MDRD variation since last visit	168	56%
Delta MDRD /t	Number	MDRD variation over time	193	56%
CKD-EPI	Number	Glomerular filtration rate estimate	83 090	48%
Medicacion (ATC)	Text	ATC-coded pharma drugs	122 276	29%
Medicacion (CFT)	Text	CFT-coded pharma drugs	153 241	29%
Medicacion (active principle)	Text	Set of active principle of drug	145 272	30%
Ophthalmic complications	Boolean	Ophthalmological complications	2	83%
Cardiovascular complications	Boolean	Cardiovascular complications	2	83%
Podiatric complications	Boolean	Podiatric complications	2	83%
Neurological complications	Boolean	Neurological complications	2	83%
Peripheral vascular disease	Boolean	Peripheral vascular disease	2	83%
Nephrological complications	Boolean	Nephrological complications	2	83%
Other complications	Text	Indicates other complications	5263	97%

Table 4.2 shows the analysis of the variable that gives us information about the stage of the disease in each record (target).

Table 4.2: Stages of nephropathy.

CKD	CKD-EPI (eGFR)	Description
Stage 1	≥ 90	Normal condition
Stage 2	60-89	Light
Stage 3	45-59	Light to moderate
Stage 3.5	30-44	Moderate
Stage 4	15-29	Serious
Stage 5	<15	Terminal (ESRD)

Considering all the 413 097 records encompassing 21 284 patients in the dataset, approximately 48% of them, equivalent to 198 156 records, had missing values for the target variable (CKD). Figure 4.2 shows the distribution of this variable. These missing values were then imputed using the last available disease stage for each patient. This process is presented in Chapter 4.3.4.

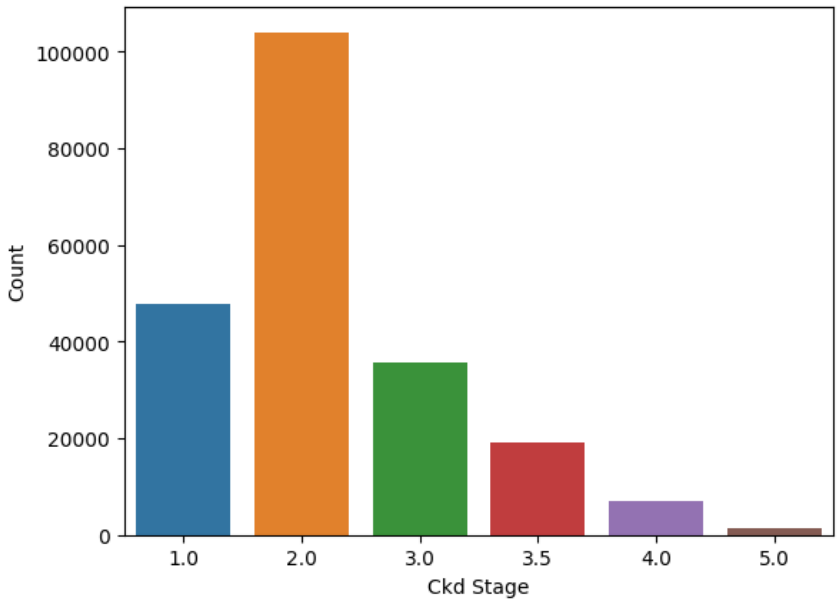

Figure 4.2: Original target distribution.

In order to understand the relationships between the different variables present in the data, the correlation matrix was used. This matrix presents in terms of a statistical metric (correlation) how two variables are linearly related. A direct correlation indicates that as one variable increases, the other tends to increase as well, and this correlation is greater the closer the value is to 1. When one variable exhibits a tendency to decrease alongside the other, it signifies an inverse correlation, and this correlation is smaller the closer the value is to -1. The closer the value is to zero, the less relationship there is between the two variables. The correlation matrix can be seen in Figure 4.3.

An analysis of the correlation matrix reveals interesting patterns and confirms the clinical relationships between the variables. It is possible to see that clinically related measures or measures that belong to the same domain are strongly correlated, such as weight, abdominal circumference, and BMI. Another example can be given by the features HDL, Non-HDL, and Triglycerides. Through correlation, it is possible to see the impact of the numerical variables on the target (CKD). The features Age, Abdominal circumference, Systolic BP, Potassium, Albuminuria, and Creatinine are variables with considerable direct correlation to disease stage. On the other side, Diastolic BP, Pulse pressure, MDRD and CKD-EPI have a considerable inverse correlation with the target as well.

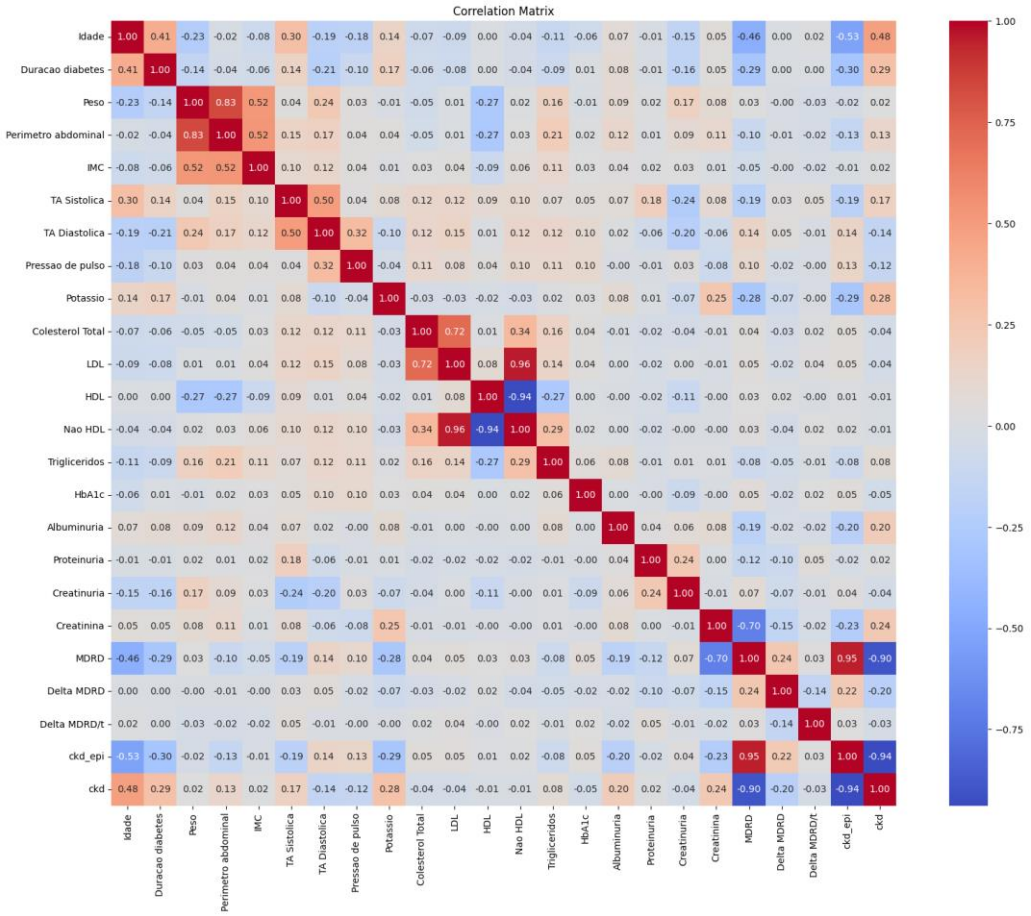

Figure 4.3: Correlation matrix.

4.2.2 Patient analysis

In order to conduct various data analyses, a new variable is generated to indicate the time window in which the doctor's appointment is scheduled. The time window is defined in years and represents the time that has passed from the day the patient started being followed until the date of the appointment. The patients were followed for 22 years, and therefore there are 22 different time windows. Figure 4.4 shows an example of defining the time windows for a patient with a total of six visits.

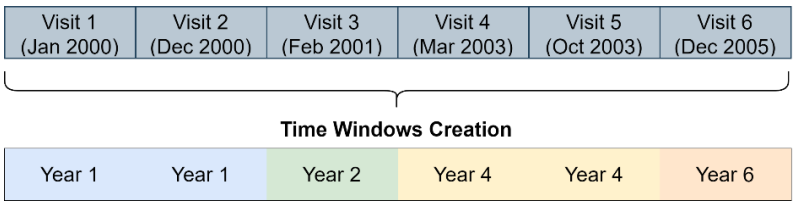

Figure 4.4: Example of how time windows are defined.

The distribution of patient medical visits over the years can be seen in Figure 4.5. It is possible to see that during the first year of follow-up there is a higher volume of consultations, and then this number tends to decrease over time due to various reasons, such as the patient's lack of interest, a stable situation that does not need constant follow-up, or in extreme cases the patient's death.

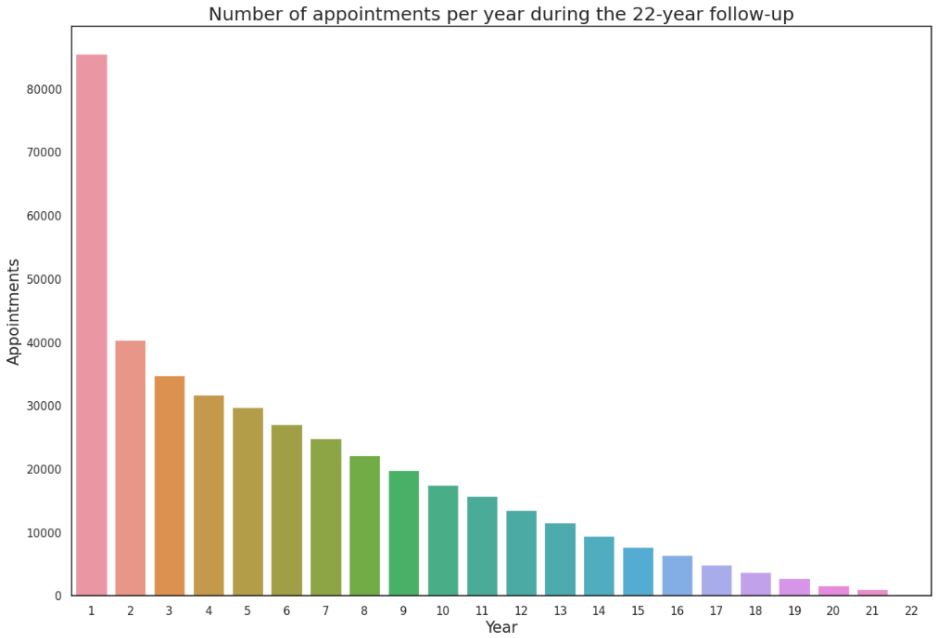

Figure 4.5: Medical appointments over the 22 years of follow-up.

Figure 4.6 shows that the same tendency holds when we look at the number of patients over the various time windows. In the initial time window, the data contains the entire patient population with approximately 21 284 individuals. However, in the final time window, only 144 patients remained, indicating a substantial 99.3% decrease in the number of patients followed over the course of 22 years. No new participants were introduced to the study; the original cohort of 22 284 patients was continuously monitored from the beginning to the end.

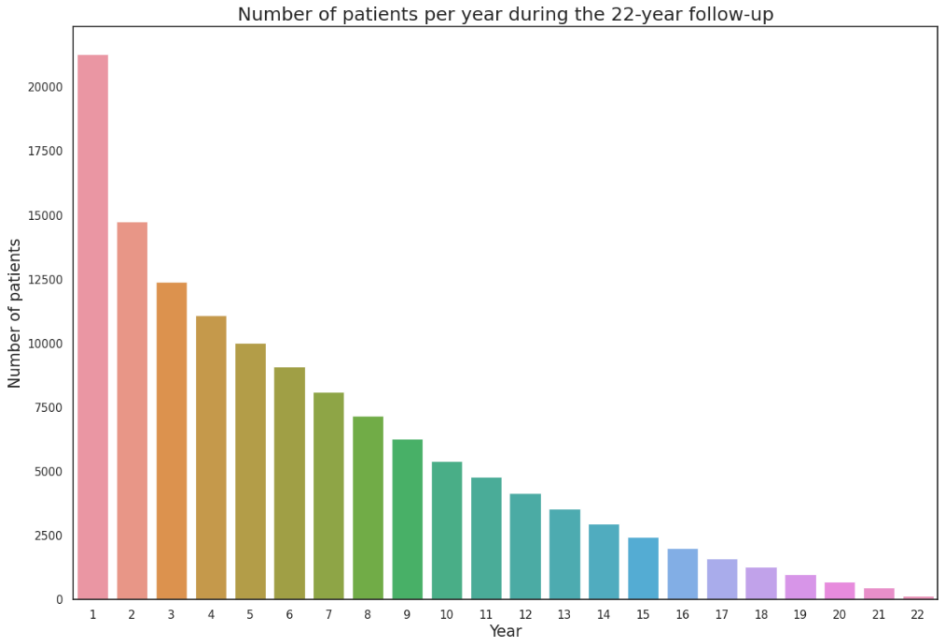

Figure 4.6: Number of patients per year during the 22-year follow-up.

On average, each patient has about 19 records. The maximum number of appointments for a patient is 247, while the minimum is 2 appointments. The ten patients with the highest number of consultations are shown in Figure 4.7.

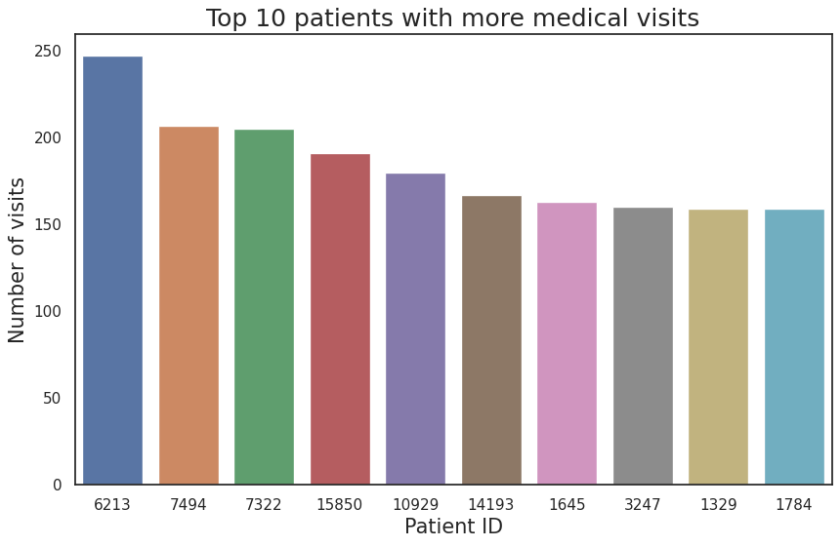

Figure 4.7: Patients with the highest number of appointments.

The data was explored in depth to understand the population and have as much knowledge as possible to shape future decisions to be made when designing the solution. In the next chapter, the diverse analysis performed within the scope of disease progression are presented.

4.2.3 Disease progression

Understanding patterns and crucial insights regarding the evolution of the dependent variable is a fundamental aspect of data analysis. The target variable in this study denotes the stage of DN in the patients, and several analyses were conducted to enhance comprehension of the problem at hand.

DN is a progressive disease that tends to worsen over time. This is evidenced by the percentage-based evolution of patients in different stages of the disease presented in Figure 4.8. There is an observable increase in the number of patients in advanced stages (3 or above), accompanied by a corresponding decrease in the earlier stages (1 and 2).

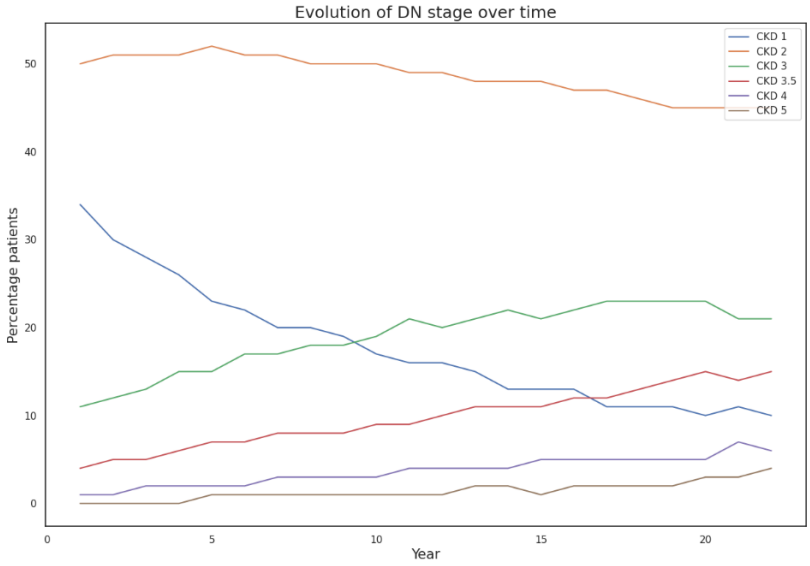

Figure 4.8: Evolution of DN stage over time.

Although often portrayed as such, DN is not always sequential in its evolution, and the patient may skip stages, and despite being uncommon, it is possible for a patient in stage 1 to worsen to more advanced stages without passing through intermediate stages such as stages 2 and 3. Figure 4.9 shows the most common disease developments within one year. From the analysis of this graph, it is possible to see that normally a patient in a time window of one year remains on same stage, but there may be cases in which the patient advances or moves back several stages. This shows some of the non-linear nature of the problem.

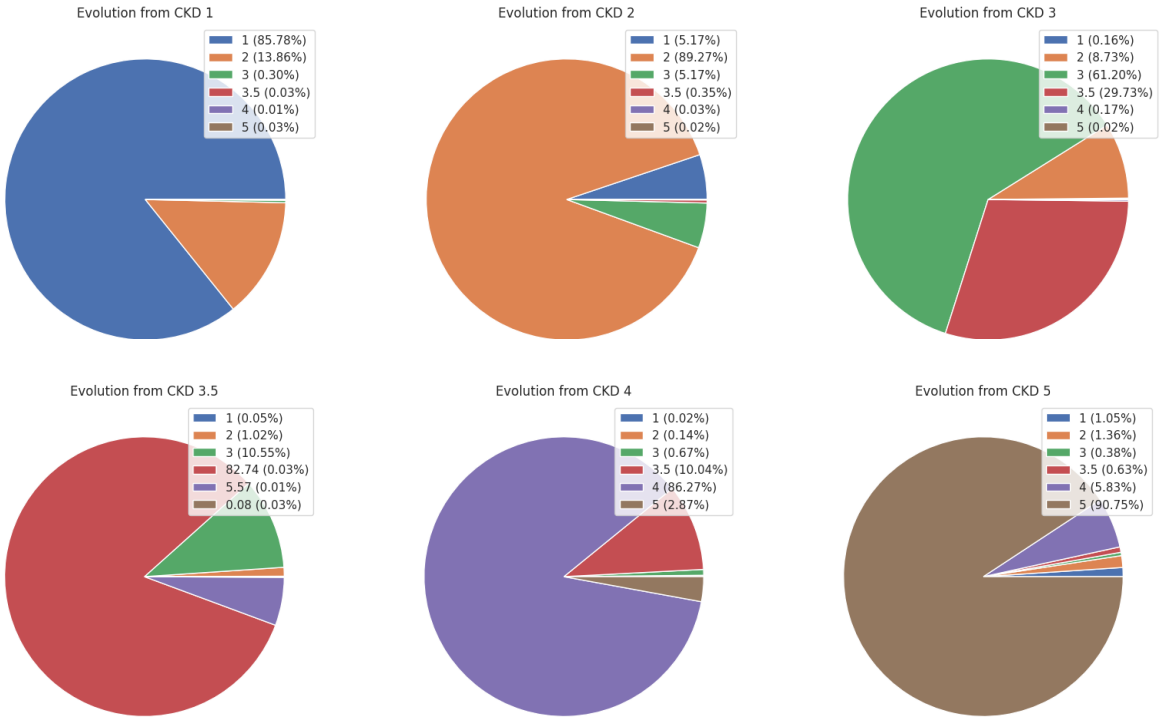

Figure 4.9: Most common disease developments within 1 year.

The analyses conducted in this chapter have provided invaluable initial insights into the data, its significance, the characteristics of the population under study, and the longitudinal aspects associated with it. The findings have greatly enhanced the understanding of the underlying problem, playing a crucial role in shaping, and laying the foundation for the next chapters.

4.3 Preprocessing

Data preprocessing is a crucial step in ensuring the quality and reliability of the data used in any research study. Throughout this chapter, the various methods used to clean, integrate, transform, and reduce data will be presented. All these steps were taken to create a carefully prepared dataset that shapes the proposed solution and allows it to be used with ML algorithms.

4.3.1 Exclusion criteria

Through the literature review presented in Chapter 3, it was possible to see that all approaches included the creation of time windows to solve the problem of irregularity associated with the EHR data. Most of the works create annual time windows, but it is possible to observe that in some cases 6-month time windows were created. Taking this into account, all patients with a follow-up time of less than 6 months were removed from this study. Therefore, 1374 patients were discarded. Additionally, patients with fewer than three consultations have hardly any associated temporality, and therefore they were discarded. Thus, a total of 366 patients were removed.

In addition to the establishment of the exclusion criteria for patients, the number of variables present in the data was also minimized. The characteristics Medication (ATC), Medication (CFT), Medication (active principle) and Other complications are made up of text, and it is necessary to apply natural language processing techniques (NLP) to derive value from them. Although the information from these features is potentially interesting for the problem, the high cardinality associated with them makes it very difficult to apply and extract information from them using NLP methods [161]. For this reason, they were discarded for this study.

All features related to MDRD have been removed because CKD-EPI equation is also present in the data, being recognized as more accurate [25]. Therefore, MDRD, MDRD Stage, Delta MDRD and Delta MDRD/t variables were not considered. Features whose information has no value for the creation of the predictive model have also been removed, these being the ID and the Registration date.

After applying these criteria, 19 544 from the initial 21 284 patients and 29 from the initial 39 features were considered in the development of this work.

4.3.2 Feature encoding

Categorical features represent qualitative characteristics or groupings, but many ML algorithms require numerical input. By encoding these features, we ensure that the algorithms can effectively understand and analyze the data. This is extensively explored and demonstrated in [162].

Encoding was applied to both the Sex and Race characteristics. The first was transformed into a binary variable. If the patient is male, then the value is zero. If the patient is female, then the value of the variable is equal to 1. On the other hand, the race feature was transformed into a numeric feature. The six categories were transformed into integer values (0, 1, 2, 3, 4, 5). This transformation was done linearly due to the low number of categories (low cardinality). If the feature had large cardinality, other encoding techniques could be considered, such as one hot encoding.

4.3.3 Outlier handling

Outliers are data points that deviate significantly from the majority of observations in a dataset. Detecting and handling these values is one of the most important processes in an ML problem. If not handled properly, outliers can affect the performance and reliability of the data analysis and model performance [163]. By properly handling outliers, it is possible to improve the integrity and quality of the dataset, ensuring that subsequent analyses and ML algorithms are not influenced by these extreme values.

In the clinical context, and, more precisely in the context of this study, it is necessary to distinguish between an outlier in the study population and a clinical outlier. An outlier within the study population refers to a patient with non-normal values in a specific variable, which may still hold medical validity. On the other hand, a clinical outlier represents an extreme value that deviates significantly from the expected range within the broader clinical setting. These clinical outliers present values that are very abnormal or even impossible to occur from a clinical perspective. This distinction is necessary because the outliers in the study population must be retained to ensure diversity in the data. Clinical outliers should be removed because they may consist of insertion errors, measurement failures, or other types of problems.

To detect the outliers, a combination of different methods was used.

- **Distribution analysis:** The statistical distribution of the variables is analyzed with the objective of identifying extreme values that should be considered outliers. For this purpose, both histograms and box plots were used.
- **Z-Score:** It is a test used to detect outliers on univariate data. The z score tells how many standard deviations from the mean your score is. The threshold value is then set for which a certain value is considered an outlier. For this study, a threshold value of 3 was set, which means that all values whose z-

score is greater than 3 or less than -3 will be considered outliers. The optimal value for the threshold depends on the data and the problem to be worked on, but in general a value of 3 is used [164].

$$Zscore\ (i) = \frac{x_i - \bar{x}}{SD} \tag{4.1}$$

- **Interquartile Range (IQR):** It identifies outliers based on the spread of data within the middle 50% of a distribution, specifically the range between the first quartile (Q1) and the third quartile (Q3). This range is called the interquartile range (IQR). An outlier is detected when it lies more than λ times (usually 1.5 times) in the interquartile range from the median [165]:

$$|x_i| > (\tilde{x} + \lambda iqr) \tag{4.2}$$

These methods were used to visualize and analyze each feature. Figure 4.10 shows an example of the application of these methods to the BMI variable.

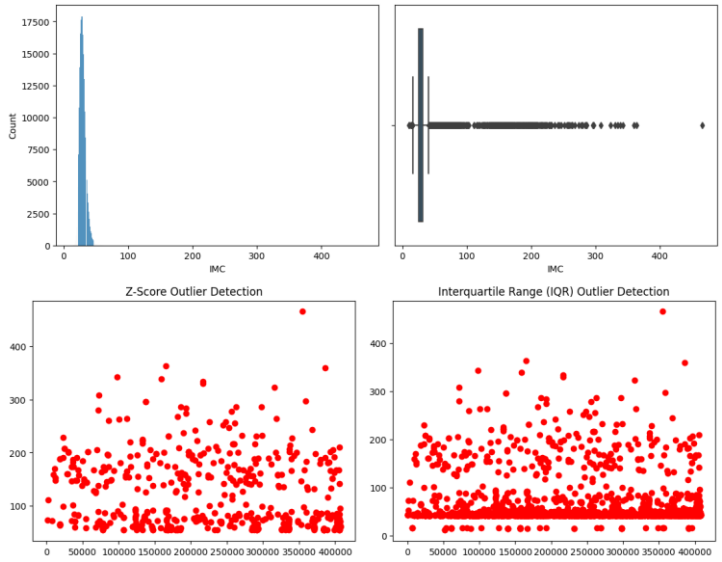

Figure 4.10: Outlier detection techniques.

The decision to consider a value as an outlier was made based on this analysis and on the clinical guidelines consulted. To ensure that no plausible values were treated as an outlier, a longitudinal analysis was performed to see if those values made sense given the temporal evolution of the feature. Figure 4.11 shows a representation of this longitudinal analysis.

Evolution of feature X over time									
35.2	41.6	37.4	39.6	48.1	150	38.4	47.3	50.2	54.2

Outlier

Evolution of feature Y over time									
75.3	81.3	110.8	125.1	129.4	150	142.7	110.4	85.1	90.1

Normal value

Figure 4.11: Representation of the longitudinal analysis performed in the treatment of outliers.

A total of 620 outliers were identified. Table 4.3 illustrates the distribution of these outliers across various features and the corresponding percentage they represent. When calculating the outlier percentages for each feature, missing values are excluded from consideration.

Table 4.3: Outliers detected in each feature and their proportion.

Feature	Outliers detected	Corresponding percentage
BMI	262	0.13%
Pulse pressure	102	0.03%
Total cholesterol	31	0.03%
LDL	60	0.05%
HDL	16	0.01%
Non-HDL	29	0.02%
Triglycerides	28	0.02%
Hba1c	30	0.01%
Albuminuria	1	<0.01%
Proteinuria	61	0.71%

All detected outliers will be changed through imputation strategies defined and presented in the next chapter. This allows all the information to be kept and no data needs to be removed.

4.3.4 Data imputation

The high number of missing data in both the independent variables (features) and the dependent variable (target) presented in the chapter 4.2.1 leads to the need for data imputation. There are essentially three types of missing values [166]:

- **Missing completely at random (MCAR):** Missing values occur completely randomly throughout the dataset. It is independent of observed and unobserved data, that is, there is no concrete difference between patients with missing values and patients without any missing values. For example, the doctor may forget to record a certain value during a consultation.
- **Missing at random (MAR):** Missing values can be explained or predicted by other variables in the dataset. In other words, the probability of missingness depends on the observed data but not on the missing data. For example, elderly people usually tend to measure their blood pressure regularly. In this case, missing blood pressure values can be age-related.
- **Not missing at random (MNAR):** Missing values depend on both the missing values and the observed values. For example, people with low cholesterol levels tend to measure their cholesterol less often.

These three types of missing values are represented in Figure 4.12.

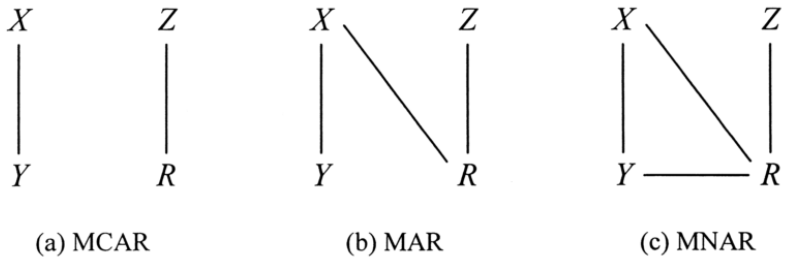

Figure 4.12: three types of missing data mechanisms: MCAR, MAR, and MNAR. The data includes variables X (observed values) and Y (missing values). Z represents the cause of missing values and R is an indicator variable that distinguishes missing and observed values in Y, in other words, missingness. Based on [167].

Alternatively, the Figure 4.12 can be represented mathematically:

$$MCAR = P(P|A) \tag{4.3}$$
$$MAR = P(R|X,Z) \tag{4.4}$$
$$MNAR = P(R|X,Y,Z) \tag{4.5}$$

The data presented in this study have missing values of the MNAR type. Missingness is associated with unobserved factors and missing data itself. Certain clinical measurements are made in only a few consultations, depending on multiple unobservable factors such as the effectiveness of the patient's treatment, the patient's willingness to undergo a certain exam, the doctor's opinion, the necessity of making a certain measurement considering the patient's condition, and other factors. In addition, it may also depend on the missing values themselves, for example, a patient with normal triglyceride values tends to measure this value less often than a patient with high or low values.

Missing values or null values in EHR data are a direct result of the temporality associated with it. It would be very complicated to measure all variables in each patient visit, resulting in a large number of measurements that are left undone when looking at the data from an overall perspective. Taking this into account, the proposed solution, medically validated on the APDP side, involves assuming that when a value is missing, the most recent available value should be considered as unchanged. In other words, any missing value is assumed to remain the same as the value measured on the previous visit, or, if not available, the most recent visit where the value was measured. In cases where no previous value is available, the logic is reversed and the immediately following value is used.

This logic is described by the forward and backward fill techniques. These techniques were applied to both the features and the target. It is important to emphasize that these techniques were implemented for each patient, and that there is no relationship between data from different patients. Figure 4.13 shows an example of the application of these algorithms.

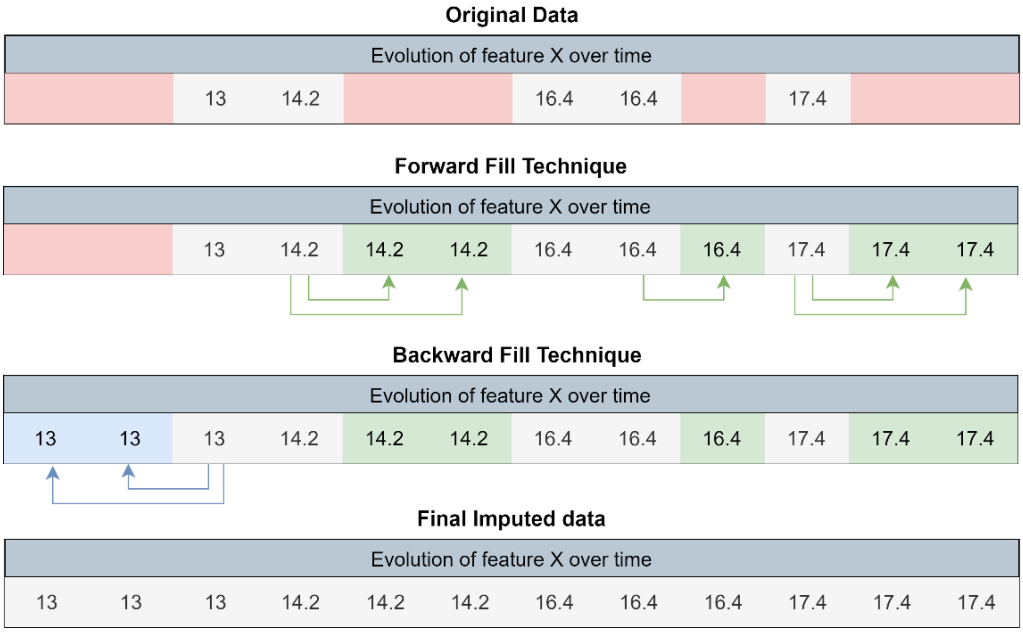

Figure 4.13: Forward and Backward fill technique to impute data.

Using this strategy, all the missing values related to the target problem, the DN stage, were successfully imputed. This imputation process did not change the distribution of our target variable, as evidenced by the unaltered distribution displayed in Figure 4.14.

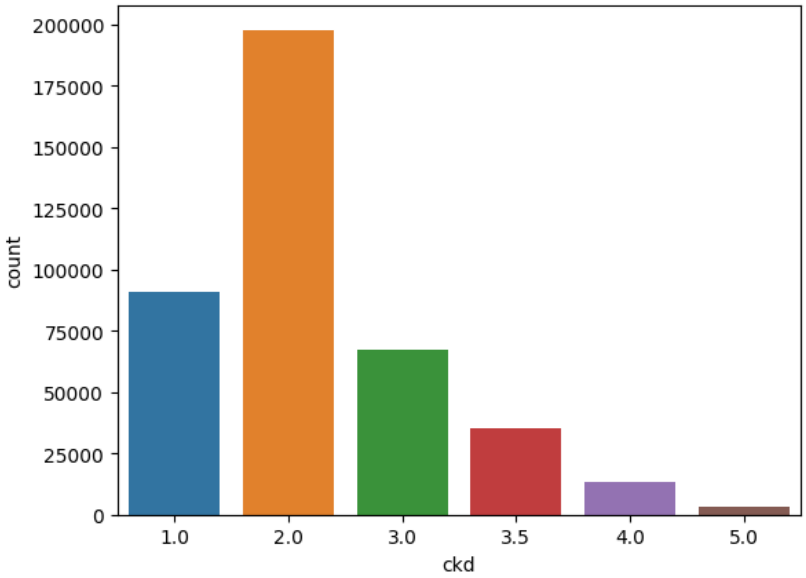

Figure 4.14: Target distribution after imputation of missing values.

Although this strategy has filled most of the missing values, there are, however, patients in whom a particular variable did not have a single value available, making the application of forward and backward propagation unfeasible. In these cases, two different strategies were applied according to the type of variable. For the binary variables, the negative value was imputed in all records, assuming that if there is no value it is because the patient does not suffer from any of the conditions indicated

by the binary variable, such as neurological, podiatric, ophthalmological, among other complications.

For numerical features, an imputation strategy was applied through a stratified mean by disease stage, also medically validated by APDP. Initially, the average is calculated for each feature grouped by disease stage, that is, for each feature six different average values are obtained (one for each disease stage). Then the imputation of the values to the patient data is done by looking at the stage to which each record corresponds. In Figure 4.15, it is possible to visually perceive how this method works.

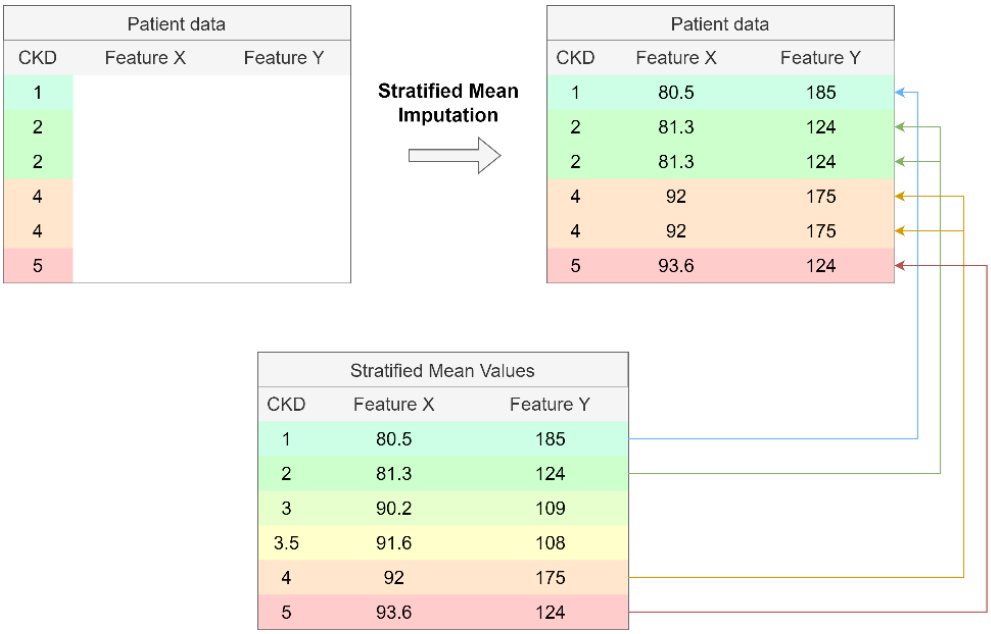

Figure 4.15: Stratified mean imputation technique.

Categorical features did not need imputation as they did not have any missing values. This can be seen in Table 4.1 on chapter 4.2.1. A total of 250 patients were excluded from the analysis due to missing or negative values in the variable representing diabetes duration. Imputing or inferring these values using the same stratified mean technique employed earlier was deemed inappropriate, considering the nature of the variable. As a continuous measure indicating the duration of diabetes for each patient, it is not meaningful or advisable to impute it based on average values across different disease stages.

After this imputation process, there are now 19 294 patients with 29 fully filled features. However, the data are still high in dimensionality and too poorly shaped to be fed into an ML model. The next chapters present the steps taken to solve this.

4.3.5 Time window aggregation

Time windows are implemented in the EHR data to introduce some structure and organization. Typically, data of each patient is segmented into time windows, ranging from 2 to 22 years. If a patient has 22 consecutive time windows, it indicates that

they have had at least one annual visit for 22 years. However, it is not mandatory for patients to have consultations every year, resulting in scenarios where a patient may have, for example, three non-consecutive time windows. These time windows might correspond to the first year, the fifth year, and the tenth year of follow-up, highlighting the irregularity in their visit pattern. Not only that, but the number of appointments per time window or year can vary, from one to dozens or even hundreds. This chapter presents the strategy created to aggregate visits within the time window, resulting in only one record per year for each patient.

Looking at the literature review Chapter 3.2, it can be observed a tendency in the works regarding the creation of time windows and the aggregation of visits within each respective window. Aggregation within time windows reduces the complexity of the data by condensing multiple records into summarized information for each window. This simplification enables data analysis and the identification of trends, patterns, and changes over time.

For this purpose, aggregations using statistical measures were used. Different aggregations were used:

- **Numerical aggregation:** To aggregate numerical variables, the median was used. The median was used instead of the mean because of its greater robustness to extreme values [168].
- **Binary aggregation:** Binary variables are variables that mark the appearance of other types of complications in patients. While the patient does not suffer from the complication, the value of the variable is always negative. When the patient presents symptoms of a certain complication, the variable associated with it becomes positive. Therefore, in this case, it is essential to use the last available record available to avoid losing information about the appearance of new complications at the end of the time window.
- **Target aggregation:** For the target, the statistical mode was used. This measure allows extracting the most common patient condition during the time window. When two values have the same frequency, the most recent is chosen. This statistical measure is used due to the clinical character of the disease, where in a time span of less than a year it is very unlikely to have major evolutions in staging, and therefore the use of statistical mode will not lead to great loss of important information.

Although the choice of statistical measures was made to minimize information loss, it is inevitable that important information is discarded in this step. This is the "price to pay" to be able to shape the data and give it structure and constant temporality. Figure 4.16 shows a representation of the different aggregation techniques.

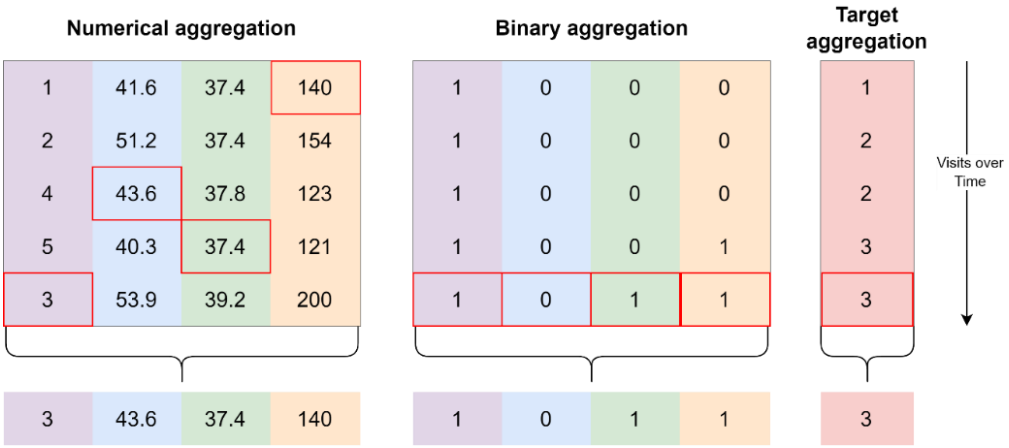

Figure 4.16: Types of values aggregation per time window based on statistical measures.

All patients now have the same number of records per time window, but between them there are still different numbers of time windows. An ML algorithm is trained with several examples, all in the same format, with the same number of variables, and the same prediction objective. The next chapter will show the steps taken to shape data.

4.3.6 Shaping Data

Before transforming and shaping the data to be served to ML models, it is necessary to define the temporality to be used in the proposed solution. The data to be considered must be sequential, specifically consisting of consecutive temporal windows. Consequently, an analysis was performed to determine the number of patients with a range of 2 to 22 consecutive time windows. It is important to remember that there is only a defined temporal window if the patient has had at least one visit in that time frame. In addition, it is important to denote that the analysis was done by looking at any sequence throughout the patient's records, two consecutive years of data does not necessarily indicate that they represent the initial two years; they can be situated within any temporal segment of the patient's records. This analysis is presented in Figure 4.17.

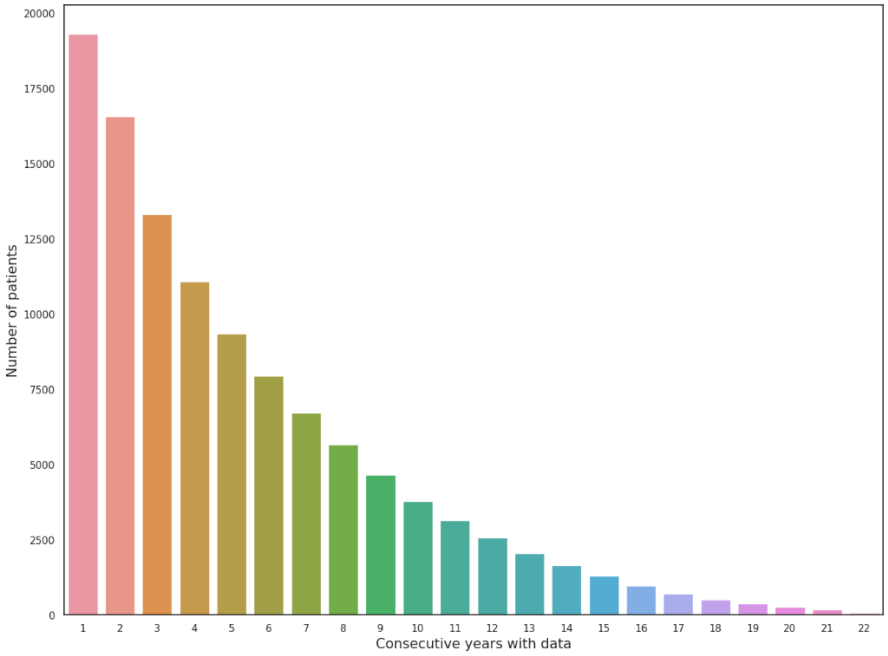

Figure 4.17: Analysis of patients with consecutive years of data ranging from 2 to 22 years of follow-up.

The longer the patient's history with consecutive years with data is considered, the fewer patients we have available and the more complicated it is to transform the data to fit a possible solution to the problem at hand. On the basis of this analysis and in agreement with the APDP, it was then defined that the solution would be to predict the risk of the disease's evolution in the next year, i.e., considering a history of x years, predict the stage of the disease in the following year. This is beneficial because the greater the temporal range in the prediction, the smaller the amount of data. At the clinical level, it is also more important to predict risk in a short period of time to be more reliable, and the annual range is adequate for the patient to take action to minimize potential predictions of worsening of the DN.

Considering the temporality associated with the data is something defined as a goal in this study. However, after the analysis presented in Figure 4.17, it is noticeable that considering a very high patient history will greatly reduce the amount of data and consequently the predictive ability of the model. Additionally, if a long time period is considered, it will be difficult for the model to have great clinical applicability because it will require that the patient be followed for a long period of time and consecutively, which is rather unusual. Taking all this into account, it was decided to use two years of the patient's history to predict the patient's DN stage in the following year.

To shape data, 3 years of records are used, two years of patient history where all variables are considered, and the third year that gives the stage of the disease in the following year, that is, the target of the problem. Out of the total number of patients, only 13 316 individuals had records in three consecutive years, as illustrated in Figure 4.17, and these patients were subsequently taken into consideration. Any sequence

of 3 consecutive years of each patient was considered independently, a concept that will be referred to hereafter as the patient journey. This is represented in Figure 4.18.

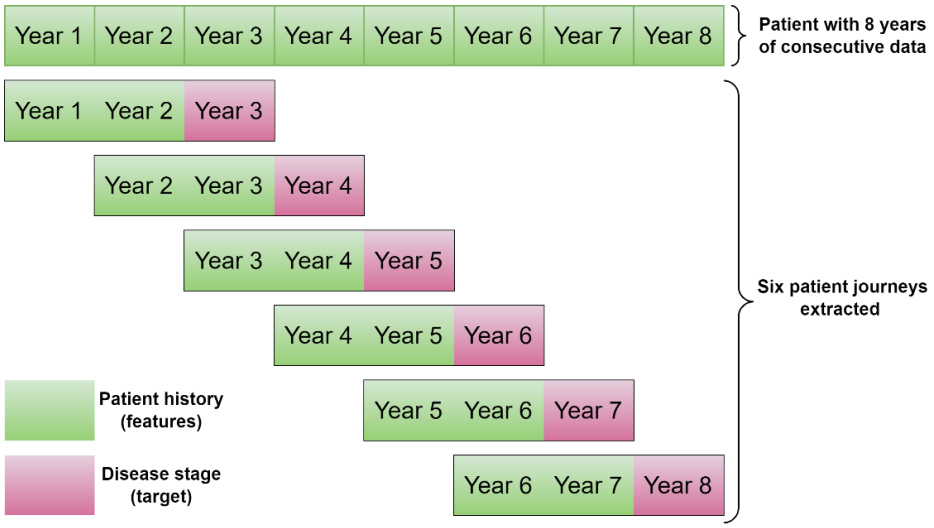

Figure 4.18: Shaping data - extracting data instances or patient journeys from one patient.

Each extracted patient journey or instance will correspond to an example for the model to train. From the 13 316 patients, a total of 79 822 different instances were extracted. The data has been transformed into a format suitable for inputting into an ML model. However, each data instance now consists of 58 variables, where 29 features are from the first year and another 29 corresponding to the second year of history. The target corresponds to the disease stage in the third year, the following year. Not only is there this high dimensionality, but there is also the need to balance the target, and this will be explored in the next chapter.

4.3.7 Target imbalance

Target balancing is an important piece in ML problems, having significant implications for the development of accurate and reliable models. It is a common occurrence, especially in real-world data, that leads to disproportionality in the class distribution. This unbalance of classes leads to biased models with overestimated and not properly validated performances.

Solving this class unbalancing is essential for two main reasons [169]:

- **Fair treatment of all classes:** Helps the model not to favor one class over another. By having the same number of examples for each class, the model learns equally about each output, usually benefiting its performance.
- **Robustness and generalizability:** The low representation of some classes over others leads to the model becoming unable to learn about minority ones and then unable to generalize that knowledge to unseen data.

This kind of unbalance in the class can easily be evidenced by looking at the distribution of the target in the data resulting from the previous processes. More

advanced stages such as 3.5, 4 and 5 show extremely low representation when compared to the earlier stages. This is represented in Figure 4.19.

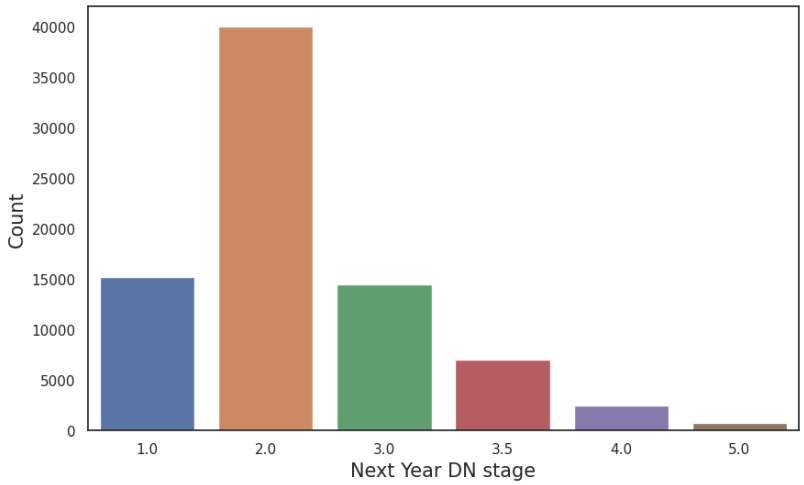

Figure 4.19: Target distribution after shaping data.

Different balancing approaches have been considered to try to minimize the impact of unbalancing, such as:

- **Undersampling original target:** The classes remain the same six, but the majority classes are randomly reduced to the same number as the minority class. In the end, all classes are left with about 766 instances. Making a total of 4596 patient records.
- **Undersampling three-class target:** The classes are transformed into a set of 3 other classes: increase, maintain, or decrease. After that, the majority classes are randomly reduced to the same number as the minority class. In the end, each class is left with 14 329 instances. This makes 42 987 data points in total.
- **Undersampling binary class target:** Classes are transformed into 2 other classes: stable and aggravation. Stable patients are those who maintain or retreat from a stage of DN, while worsening patients are those who advance to a higher stage of DN. After the majority class (stable) is reduced through random undersampling, each class has 12 348 instances. In total 24 696.

The first approach where the original target is used proved to be extremely inefficient because besides the large amount of discarded data when balancing the classes, there is also a larger number of classes to predict, substantially increasing the complexity of the problem and leading to a huge drop in performance. For this reason, the approach was discarded.

Between the three-class target approach and the binary target approach, the binary approach was selected. In the literature study done and presented in chapter 3, ten of the eleven papers used a binary target to solve the problem, as can be confirmed in that same chapter, in Table 3.4. Additionally, in a meeting with an APDP physician, this choice was medically validated.

This approach forces patients in renal dysfunction or stage 5 to be discarded. Patients at this stage of disease cannot worsen and therefore this type of model is not applicable. There were 603 patients at stage 5, but out of these, only 79 retreated to a lower stage, showing a clear tendency for the patient to remain at stage 5. Clinically, there is no need to apply a predictive model to patients already in renal dysfunction because they are already at a stage where medical care has already identified the problem and the patient is being treated by available and indicated methods.

Although the balancing was done from the reduction of the majority class, this process is not 100% random. The majority class (Stable) has 67 065 instances, while the minority class has 12 348 instances. If the process were totally random, there could be stages of the disease without any representation as shown in Figure 4.20. It is important to note that the stage shown in the figure corresponds to the last stage of the patient. The importance of having diversity in the training examples of patients in the various stages is high, considering not only the generalizability of the model but also the nature of the DN. To counteract this, a partially random approach was then applied when reducing the majority class. This strategy called partially random undersampling is presented in Figure 4.21.

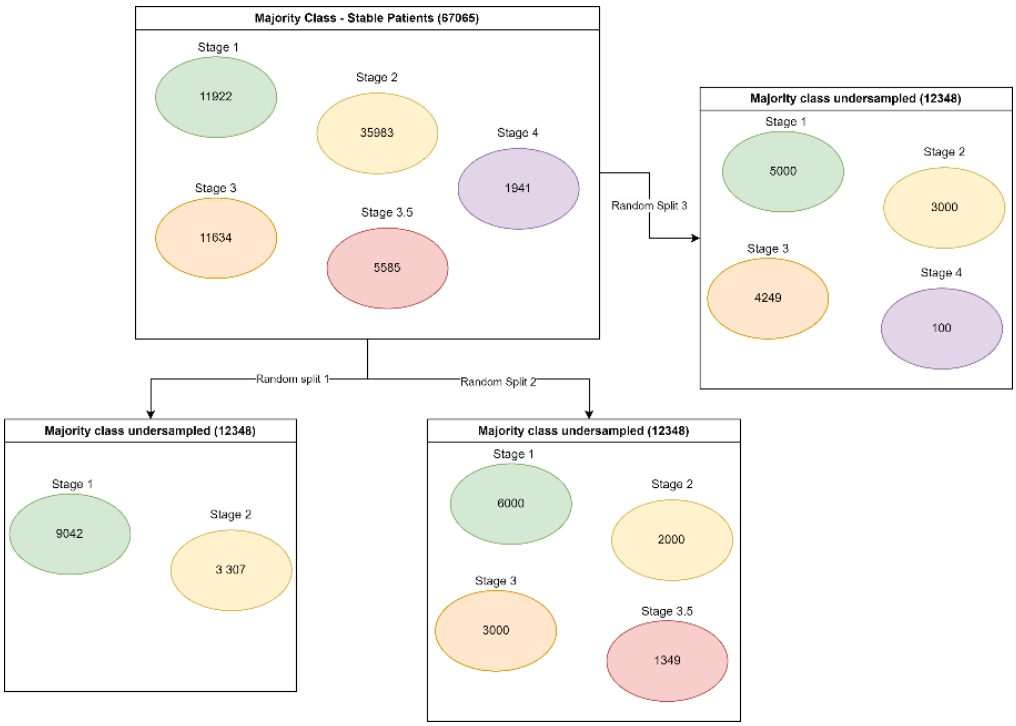

Figure 4.20: Random undersampling of binary target.

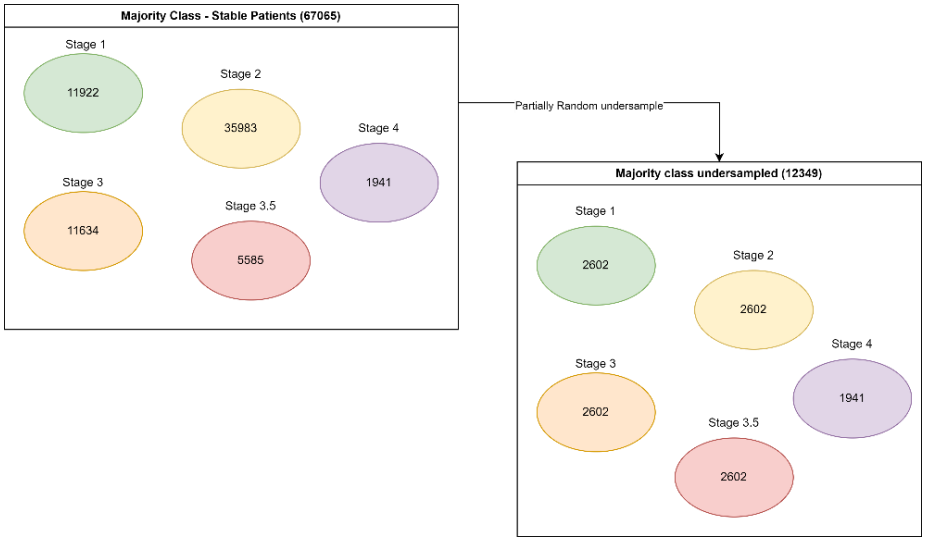

Figure 4.21: Partially random undersampling of binary target.

What has been presented will consist of what will be called approach A from now on. This approach turned out to have some flaws during the validation of its ML model, so a second approach B is proposed. This new approach applies a slightly different balance.

Approach A balances in such a way that patients in the reduced class (stable) are equally distributed over the last stage of the disease, but this approach has a major flaw: the data when grouped by last stage are not balanced in the binary class. This can be seen in Figure 4.22.

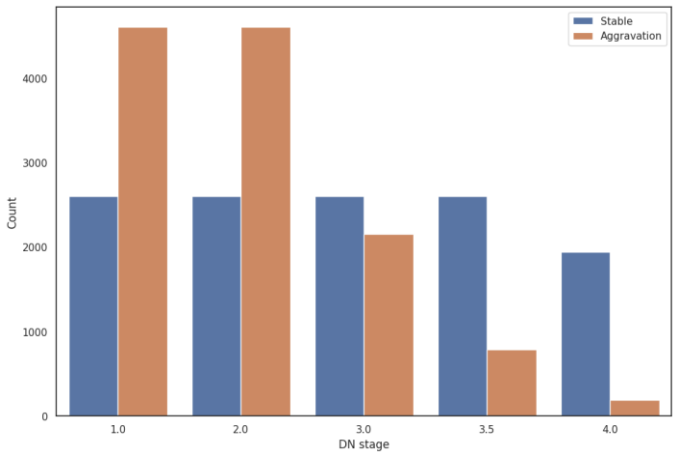

Figure 4.22: Unbalancing of the target by disease stage.

This is a problem because there are classes with virtually no representation per state. Aggravation from stage 4 to stage 5 has practically no representation, and therefore approach A is unable to predict such an event. As a solution, approach B is proposed, which consists of balancing by considering the equality of the class in each stage of the disease. This is done by removing the number of instances for each stage from the majority class in order to match the minority class. In the end, each stage has the same number of instances in which the patient remains stable and worsens.

This reduces the bias of the model and brings greater validity and confidence in the results. Figure 4.23 presents the logic behind this new approach.

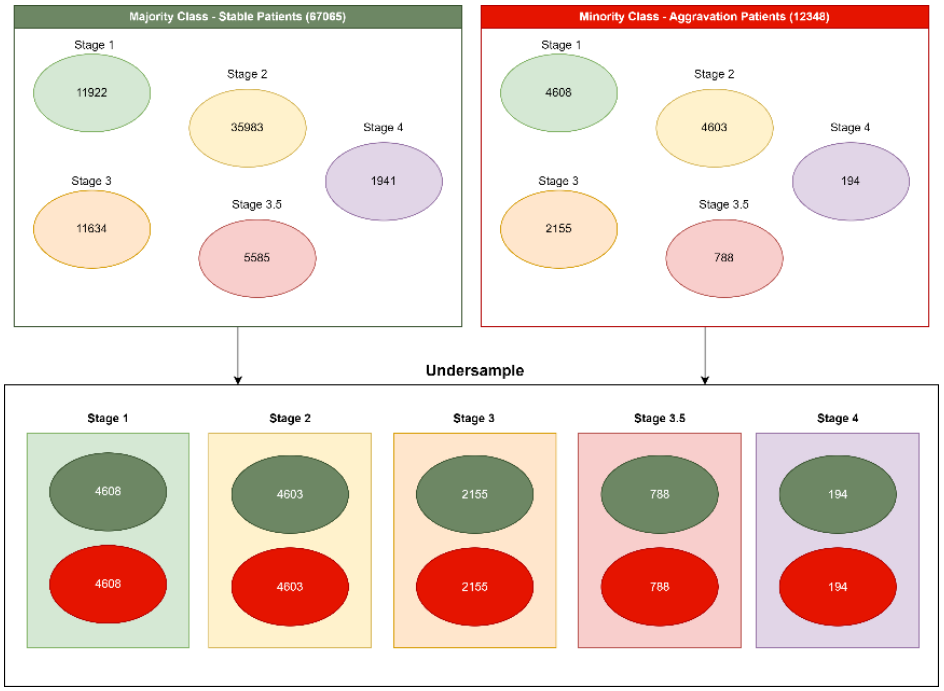

Figure 4.23: Approach B – balance target through class majority undersampling and balancing per DN stage.

This process represents a perfect equality of the target at each DN stage, with exactly the same number of occurrences for each event. The perfectly balanced distribution can be seen in Figure 4.24.

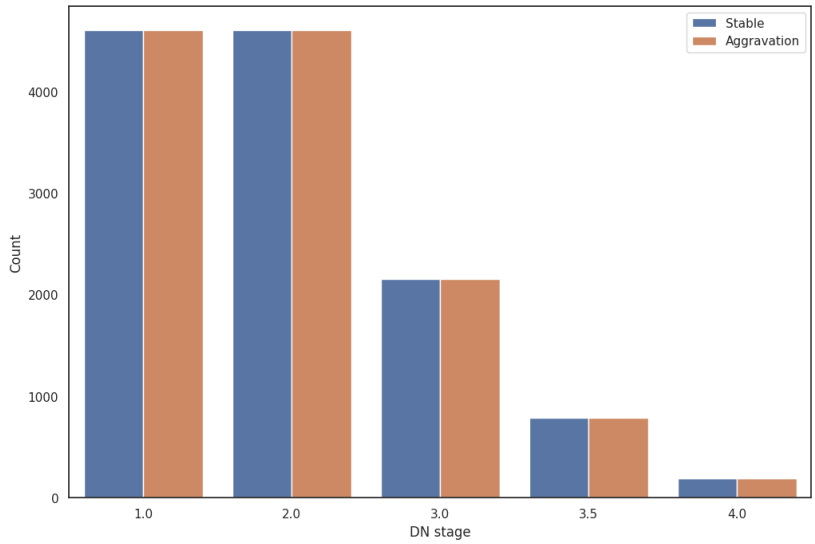

Figure 4.24: Target balanced by disease stage.

This difference in balancing the data by last stage is not only visible in the training and test data, but is also noticeable in the unseen data, as seen in Figure 4.25.

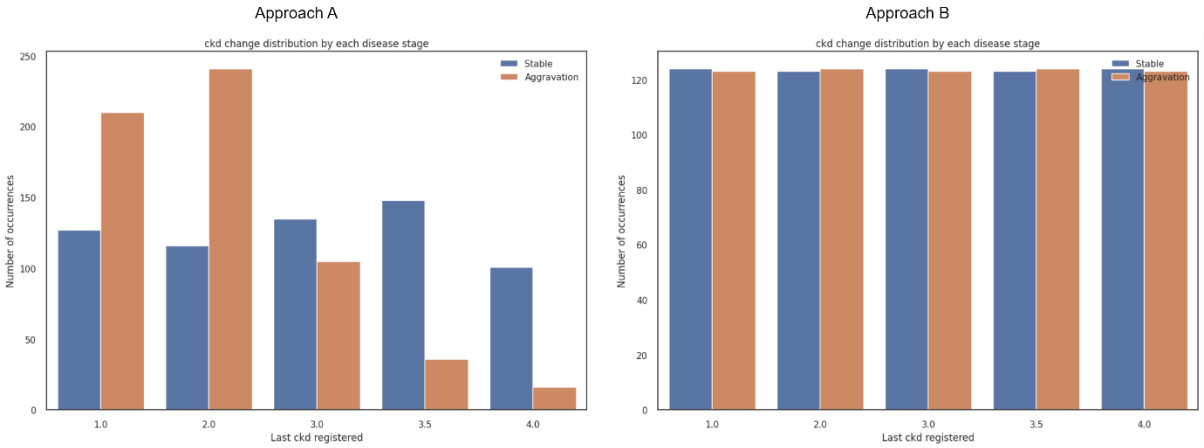

Figure 4.25: Difference between class distribution by disease stage in unseen data

Two different approaches were thus defined. Although the data are almost ready, they still contain a very large number of variables, 29 for each year in the patient's history, making a total of 58 in total. The selection of features with the goal of reducing the high dimensionality of the data is presented in the next chapter.

4.3.8 Feature selection

As mentioned in chapter 3, high dimensionality is one of the biggest challenges of EHR data. The larger the number of variables, the more difficult it becomes to extract knowledge from the data [94]. Not only that, but a model that requires a lot of data to make a prediction makes its application either unfeasible or very poorly applicable because it is unlikely that a patient has all the necessary records available to feed the predictive model.

To make the feature selection, five subsets of variables were selected based on different criteria. These being:

- **Correlation analysis:** Analysis of the correction matrix in order to understand the variables most correlated with the stage of the disease.
- **Literature review:** The literature review conducted as part of this study revealed several variables that are essential to predict the stage of DN. This information is presented in Chapter 3, Table 3.3.
- **Feature ranking:** Several experiments were conducted, and the features with the greatest impact on prediction were evaluated. This impact was calculated using algorithms based on impurity (Gini index or entropy) or information gain for each feature. Additionally, the global feature ranking generated with SHAP values was also used.

In addition to these three criteria, we initially considered dynamic algorithms for feature selection such as recursive feature elimination (RFE), but these algorithms are not adapted to deal with the temporality associated with the problem. The features corresponding to the first year of patient history must also be present in the

second year, and this would not be considered when using more dynamic and automated techniques for selecting the optimal features. This problem can be seen through the illustration in Figure 4.26.

Figure 4.26: Feature selection taking into account the temporality of the variables.

Table 4.4 shows the five subsets created following the various criteria presented above. It is important to note that all subsets have the CKD feature that indicates the stage of the disease. Both this and all variables will have two different values afterward, one for the first year of the patient's history and another for the second year, in order to predict the patient's stage in the following year.

Table 4.4: Subsets of data created based on different defined criteria.

Subset 1	**Subset 2**	**Subset 3**	**Subset 4**	**Subset 5**
Age	Age	Age	Age	Ckd-epi
Race	Race	Race	Duration of diabetes,	Albuminuria
Gender	Gender	Gender	Abdominal	HbA1c
Duration of diabetes	Duration of diabetes	Duration of diabetes	circumference,	Proteinuria
Abdominal	Abdominal	Abdominal	Systolic BP	Age
circumference	circumference	circumference	Pulse pressure	BMI
CKD	Nephrological	Systolic BP	Potassium	CKD
	complications	Total cholesterol	Total cholesterol	
	Cardiovascular	Albuminuria	HDL	
	complications	Proteinuria	Albuminuria	
	CKD	Potassium	CKD-EPI	
		CKD-EPI	Nephrological	
		CKD	complications	
			Cardiovascular	
			complications	
			CKD	

Among the various subsets considered, subset 5 was chosen because it includes all the variables identified in the literature review as the most important to predict the evolution of DN. Subset 5 not only uses fewer variables than most of the other subsets, but also allows ML models to perform better. This selection of features was further validated by APDP.

The data is now in a format capable of solving many of the original problems present in EHR databases. Time windows have been used to address the irregularity of the data. With the imputation of data and aggregation of annual records, part of the

problem of data sparsity has been solved. With the feature selection presented, the data also has low dimensionality. Figure 4.27 illustrates the final dataset that will be used to create the predictive model. This process will be detailed in the Chapter 4.4.

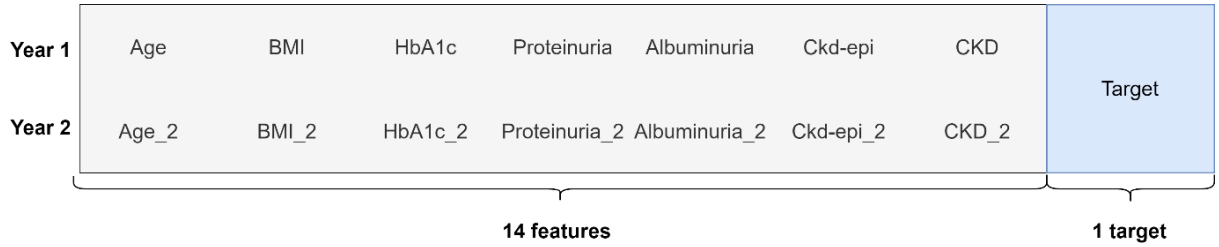

Figure 4.27: Format of the final dataset to be used to modulate the solution and create the predictive ML model.

4.3.9 Data normalization

An important step in the data preprocessing pipeline is data normalization. Its purpose is to scale numerical data features to a uniform range. It ensures that all features have similar scales, making it easier for ML algorithms to process the data effectively and produce accurate results. Data normalization takes each feature and transforms it so that they all fall within a common range, such as 0 to 1 or -1 to 1. In this way, the ML algorithm treats all features equally, preventing any single feature from dominating the learning process simply because of its larger scale [170].

Although there are several algorithms to normalize the data, Z-Score was chosen [171]. This technique transforms all numerical values into a range between 0 and 1. Figure 4.28 illustrates the Z-Score technique used for data normalization.

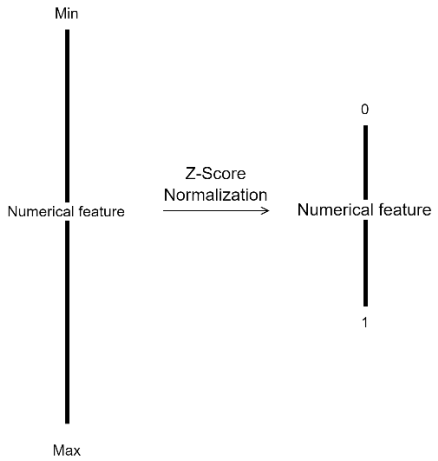

Figure 4.28: Representation of z-score normalization technique

Now, all the features are on the same scale and the data are ready to be supplied to the ML model. All the details about the creation of the model are presented in the next chapter.

4.4 ML Model

This chapter introduces the experimental setup used to construct the ML models. The process of training, hyperparameter tuning, and evaluation of the algorithms will be described in detail. To complement this, the reasoning behind the choice of the best model will be presented, including how the analysis of the classification metrics, the results produced by the model and the statistical significance between them was performed.

4.4.1 Experimental setup

The entire experimental setup was performed using the PyCaret framework in version 3.0.1 and Python in version 3.7. PyCaret uses several embedded ML libraries such as scickit learn, Catboost, LightGBM, Optuna, and others. It is considered one of the best low code ML frameworks [172].

At the beginning, the dataset was divided into two separate sets: 5% was allocated for validation/unseen data, while the remaining 95% was used for training and testing purposes. The training and test data were then divided into 70% for training and 30% for test. This ratio of 3:7 is the most recommended to use at this step [173]. This resulted in 1235 instances that will not be part of the model training and testing process, being unseen data. 16422 instances for the model to train and 7039 for testing.

Sixteen ML algorithms were tested: DT, Catboost, RF, Extra Trees (ET), Ridge, XGBoost, LR, LightGBM, GBM, LDA, Multilayer Perceptron (MLP), NB, KNN, AdaBoost, SVM and Quadratic Discriminant Analysis (QDA). Initially, all these classifiers were trained with the default hyperparameters.

From these sixteen algorithms tested, the top five were chosen based on their performance in the training data. Cross-validation (CV) with K=10 was used to train the models. This corresponds to 10 iterations where in each, 90% of the data is used for training and 10% for testing. This allows for greater robustness and reliability of the results obtained [67]. Even using the CV method to train models, there was a need to have a test set with 30% of the data so that there is no data in the training that is also used in the testing (data leakage), leading to a false generalizability of the model [174]. Therefore, it is essential to ensure that training, test, and unseen data are used independently in each of the model training and evaluation phases.

After training the models, the five classifiers with the best performance were chosen to go through the hyperparameter tuning phase. From these models, those that show a large performance loss in the test set were discarded, this being a sign of overfitting the model [175].

After training, optimizing the parameters, and testing the different algorithms, they were all retrained with the full data (train and test) in order to increase the number of instances on which the model trains (totaling 23461 data points). The selection of

the best classifier was done based on the performance evaluation in all three steps, train, test, and unseen data. Not only that, but several analyses were taken into account, such as analysis of the result grouped by stage, statistical significance, and distribution of importance of the variables for prediction. Figure 4.29 shows all the steps made in the experimental setup of this study. In the next chapters, more details will be given about some steps such as hyper parameter tuning, model evaluation, and statistical analysis.

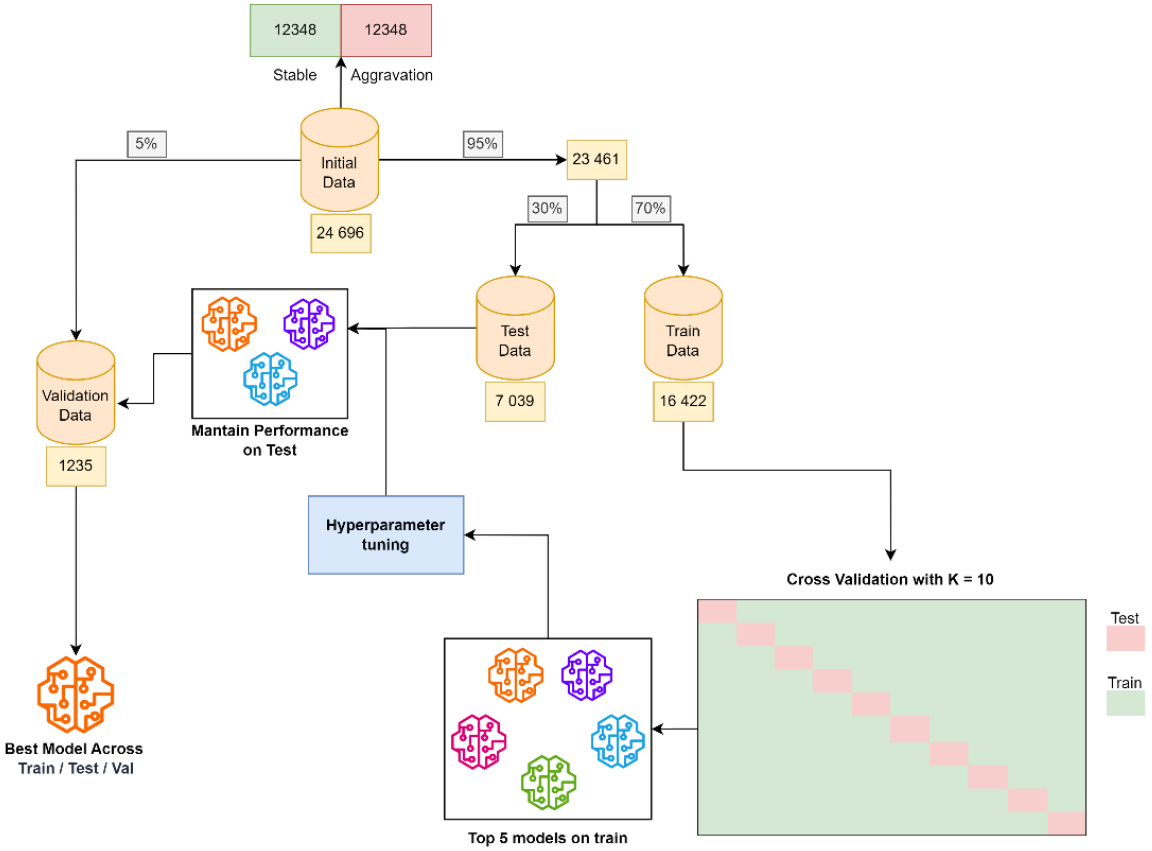

Figure 4.29: Experimental setup.

4.4.2 Hyperparameter tuning

This crucial aspect, known as hyperparameter tuning or model tuning, has been highlighted by Probst et al., showing its significant impact on performance [176]. For this purpose, the RandomGridSearch algorithm was used. This algorithm randomly selects different combinations of hyperparameters within predefined ranges and then tests the performance of the algorithm for each combination [177].

When choosing the tuning method, different characteristics such as efficiency, flexibility, performance, and popularity were considered. RandomizedSearchCV was the algorithm that met these criteria and was therefore chosen to be used in this study.

To determine the optimal parameters, cross-validation using K = 10 folds was used. The use of cross-validation during the hyper parameter adjustment phase is a widely

adopted practice [178]. The tuning process was applied to both approach A and approach B. The default hyper parameters and the new tuned hyperparameters of each of the models for each of the approaches is presented in Appendix B.

4.4.3 Model evaluation

The way in which the models are trained has already been described, and this chapter describes how the models were evaluated at each stage of the experimental setup.

The following metrics were used to evaluate the performance of the models: Accuracy, Recall, Precision, F1 Score and AUC. Each of these metrics gives us valuable information about the performance of the models, but in a clinical context there is a need to give more importance to the Recall metric, also referred to as sensitivity. In predicting the evolution of a disease, it is essential to predict all patients who worsen. If worsening patients are defined as positive and those who remain stable as negative, the detection of a negative patient as positive is not as penalizing as the opposite. This is demonstrated in several medical studies in the literature [179]–[181].

In both training and testing, performance was evaluated exclusively based on the classification metrics. In the end, the models are subjected to validation on unseen data in the training and testing processes. Here, performance is evaluated not only in terms of general performance metrics, but also through an evaluation by patient's last disease stage, because a model can have very good results but be unable to predict the evolution of a disease that is in a certain stage. It is this analysis that leads to the creation of two different approaches (A and B).

4.4.4 Statistical significance

When comparing the performance of different classifiers, it is important to consider statistical significance. This helps to determine whether any observed differences in performance are due to random chance or whether they are truly meaningful.

McNemar's test [182] is used as the statistical test. This test is used to see the statistical difference in performance between two classifiers. It involves constructing a contingency table in the form of a 2x2 matrix. This matrix represents the correct and incorrect predictions made by both models. Figure 4.30 shows a representation of this matrix. The null hypothesis can then be translated as $P(B) = P(C)$, which means that the two models have equivalent performance. If the test rejects the null hypothesis (p value < 0.05) then the hypothesis that the performance of the models is equal is rejected.

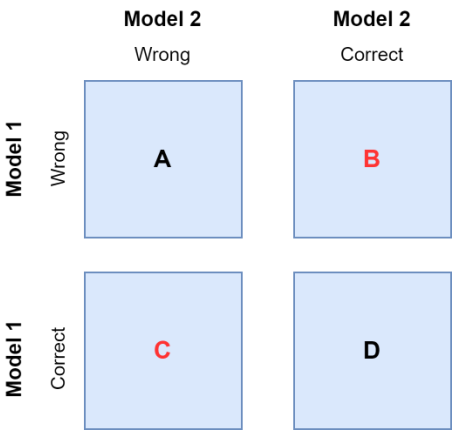

Figure 4.30: McNemar's contingency table, based on [182].

In this study, McNemar's test version with the continuous correction proposed by Edwards et al. is used [183]. To calculate McNemar's test statistic, commonly referred to as 'chi-squared', we can use the following formula:

$$x^2 = \frac{(|B - C| - 1)^2}{(B + C)} \tag{4.6}$$

If the calculated test statistic value exceeds the critical value of 3.84 in a 95% confidence interval, it can be concluded that the two methods exhibit significant differences in their performance [184]. After assigning a significance threshold, usually a value of 0.05, it is also possible to obtain the exact p-value [185]:

$$p = 2\sum_{i=b}^{n} \binom{n}{i} 0.5^i (1 - 0.5)^{n-i} \tag{4.7}$$

This test helps to analyze the results of the various models. If two models reject the null hypothesis, then there are differences in their performance, which means that there is a difference between choosing one or the other. Otherwise, the choice of the best model should be based on factors other than just performance because then they are equivalent.

4.5 Model Interpretation

In addition to a good performance, it is essential to be able to show the logic behind the prediction made. An interpretable model is essential in sensitive applications such as medicine. SHAP method was selected to interpret model. It is a method proposed by Lunderg et al. [68] being able to explain the importance of each variable for each observation (local interpretation) or for a set of observations (global interpretation). Although it has specific interpretation algorithms for tree-based, linear, and neural network models, it also has a model agnostic interpretation method, which means that it can be applied to any ML algorithm. It is one of the most widely applied interpretation methods and has been implemented several times in predictive models based on clinical data [186]–[188].

Global interpretations were used to understand the internal associations of the model, which features were more important overall to the target. This type of interpretation is very useful to understand any possible bias or problem inherent to the model. Local interpretation, on the other hand, is essential to provide interpretability and justification for the prediction given for each patient. This type of interpretation should be focused on the physician or patient as a way to be able to understand the logic behind each output, and it must be presented in an understandable format. This difference can be evidenced in Figure 4.31.

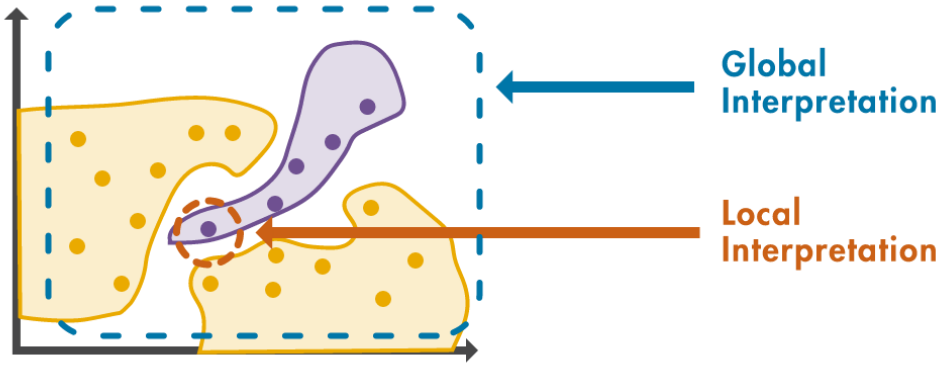

Figure 4.31: Global vs local interpretation, based on [189].

The SHAP method decomposes the output predicted by the model as the sum of the impact of the different features. A feature can have a positive or negative contribution to the result. Each contribution has an associated weight, and the greater the weight, the more important the associated feature is in the prediction. This is illustrated in Figure 4.32.

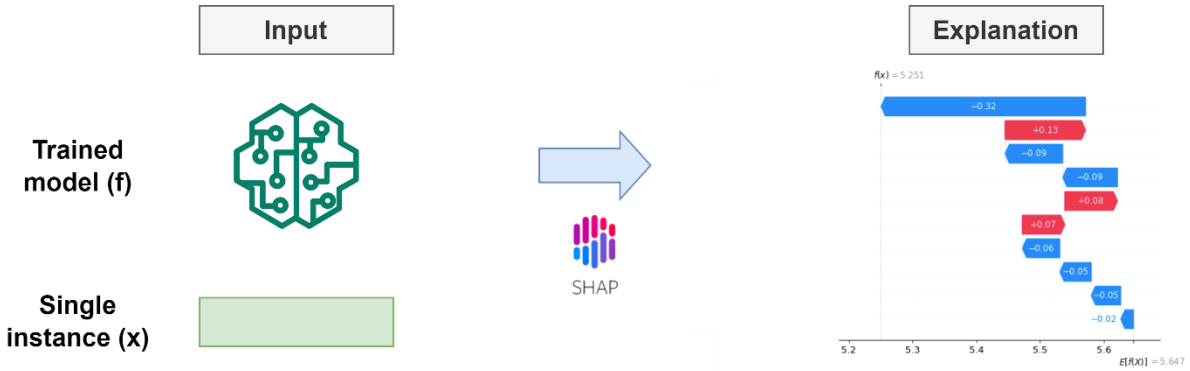

Figure 4.32: Representation of SHAP values use on interpret model or single instance.

4.6 Model deployment

To provide access to the ML model through an interactive interface, a powerful Python library called Gradio is used. Through a simplified form, Gradio allows you to create interactive interfaces where users provide their inputs (data) and receive the model's prediction in real time. Gradio version 3.36.1 was applied [190].

The purpose of deploying the model in this study is not only to give the user the possibility to make predictions with the data provided itself but also to provide the actual explanation behind this result. Gradio has several ways to publicly share the application with anyone, and there are two main ways to do this. The first is to use a proxy to the local server and create a public link that is accessible by anyone. This is done automatically by Gradio, but the link expires in 3 days. This method is represented in Figure 4.33. The second method uses the Hugging Face (HF) spaces for the application host, where all the necessary infrastructure is on the HF side. This allows for a permanent link to the application. Figure 4.34 illustrates this process.

The second method was used by hosting the Gradio application on the HF spaces in order to allow access to anyone who is interested in trying the ML predictive model proposed in this study[1].

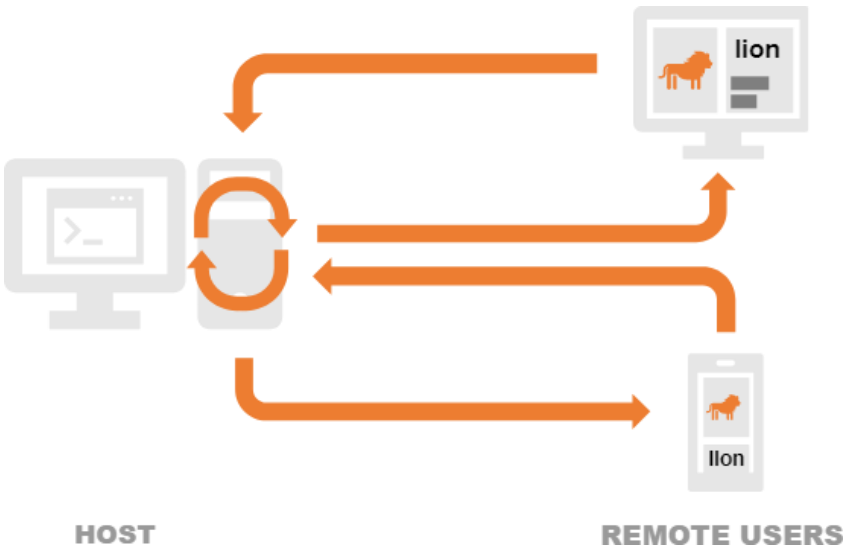

Figure 4.33: Gradio ML model hosted on local server, based on [190].

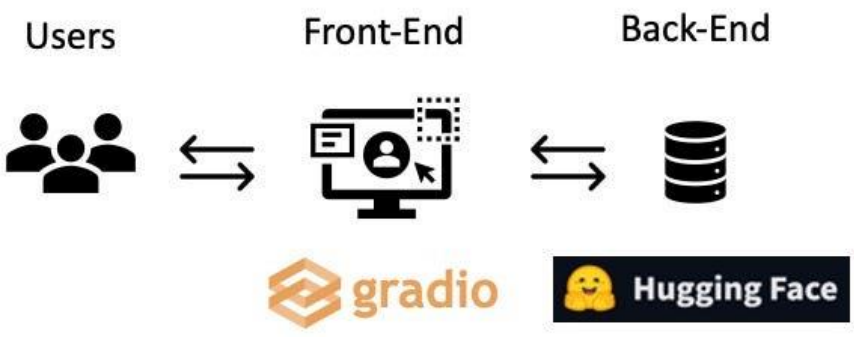

Figure 4.34: Gradio ML model hosted on HF servers, based on [191].

[1] Application is available at: https://huggingface.co/spaces/Fmesquita17/DN_Evolution

5 RESULTS

This chapter presents the results obtained in all the steps of the experimental setup. Initially, the results of approach A are presented, clearly showing the associated and detected issues. Then the results of the second proposed approach, named approach B, are presented. In approach A, the problem target is balanced by having an equal number of records for patients who worsen and those who remain stable. On the other hand, in approach B, it improves upon approach A by balancing the target with respect to the current disease stage of the patient. This ensures an equal number of examples for patients who worsen and remain stable at each stage of the disease. Once the different results are presented, the interpretation of the proposed model and its deployment is shown.

5.1 Approach A: Balanced target distribution

Sixteen ML classifiers were initially trained. Their performance on the training data is presented in Table 5.1.

Table 5.1: Performance of ML algorithms on train set – approach A.

Model	Accuracy	Recall	Precision	F1 Score	AUC	MCC
GBM	**0.7583**	0.7836	**0.7452**	0.7639	**0.8346**	**0.5175**
Catboost	0.7575	0.7816	0.7450	0.7628	0.8344	0.5158
LightGBM	0.7570	**0.7895**	0.7407	**0.7642**	0.8317	0.5153
MLP	0.7561	0.7819	0.7431	0.7619	0.8306	0.5131
Adaboost	0.7541	0.7724	0.7454	0.7576	0.8242	0.5087
RF	0.7536	0.7691	0.7454	0.7570	0.8228	0.5076
ET	0.7490	0.7603	0.7427	0.7514	0.8224	0.4981
XGBoost	0.7457	0.7704	0.7325	0.7519	0.8188	0.4922
LR	0.7362	0.7137	0.7464	0.7297	0.8116	0.4728
LDA	0.7336	0.7073	0.7457	0.7260	0.8115	0.4678
Ridge	0.7324	0.7060	0.7446	0.7248	0.0000	0.4655
QDA	0.7320	0.7648	0.7174	0.7397	0.8033	0.4655
SVM	0.7228	0.6833	0.7415	0.7109	0.0000	0.4472
KNN	0.6839	0.6678	0.6890	0.6781	0.7374	0.3681
DT	0.6697	0.6607	0.6718	0.6661	0.6697	0.3395
NB	0.6309	0.7296	0.6085	0.6634	0.6601	0.2677

Even with all the default hyperparameters, there are already classifiers with good results, especially in terms of Recall and AUC. The five best models are chosen for parameter tuning. For this choice, the various metrics are considered, with a major focus on Recall. Taking this into account, the models chosen for parameter tuning were: GBM, Catboost, LightGBM, MLP and Adaboost. The performance of these models after parameter tuning on the test set is shown in Table 5.2

Table 5.2: Performance of ML algorithms after hyperparameters tuning on test set – approach A.

Model	Accuracy	Recall	Precision	F1 Score	AUC	MCC
GBM	0.7552	**0.7753**	0.7446	**0.7597**	0.8400	**0.5109**
Catboost	0.7544	0.7679	0.7469	0.7573	0.8394	0.5090
LightGBM	0.7532	0.7714	0.7436	0.7572	0.8383	0.5069
MLP	**0.7554**	0.7397	**0.7628**	0.7511	**0.8403**	**0.5109**
Adaboost	0.7539	0.7622	0.7490	0.7556	0.8329	0.5080

By comparing the results obtained on the training set with the results obtained on the test set after parameter tuning, it can be observed that the overall performance remained similar. It is possible to denote a considerable drop in the Recall of the MLP model, but the other metrics have persisted very much the same. Given these results, the five models were then retrained with the training and test data to be then tested on the unseen data. The results of these models on unseen data can be seen in the Table 5.3.

Table 5.3: Performance of ML algorithms on unseen data – approach A.

Model	Accuracy	Recall	Precision	F1 Score	AUC	MCC
GBM	0.7377	0.7679	0.7380	0.7527	**0.8255**	0.4740
Catboost	0.7368	**0.7741**	0.7341	0.7536	0.8253	0.4724
LightGBM	**0.7417**	0.7726	0.7414	**0.7567**	0.8238	**0.4821**
MLP	0.7393	0.7632	**0.7424**	0.7527	0.8240	0.4773
Adaboost	0.7377	0.7617	0.7409	0.7512	0.8167	0.4741

In general, a slight drop in performance is noticeable for all models when applied to unseen data. Despite this, the results remain acceptable and do not seem to be indicative of an inability to generalize the knowledge learned by the ML model. An optimal predictive model should be able to distinguish between stable and worsening regardless of the patient's disease stage. To validate the ML model's performance, an analysis was conducted, considering the patient's disease stage. Figure 5.1 - Figure 5.5 present the performance, as well as the confusion matrix of the five classifiers by current stage of the disease.

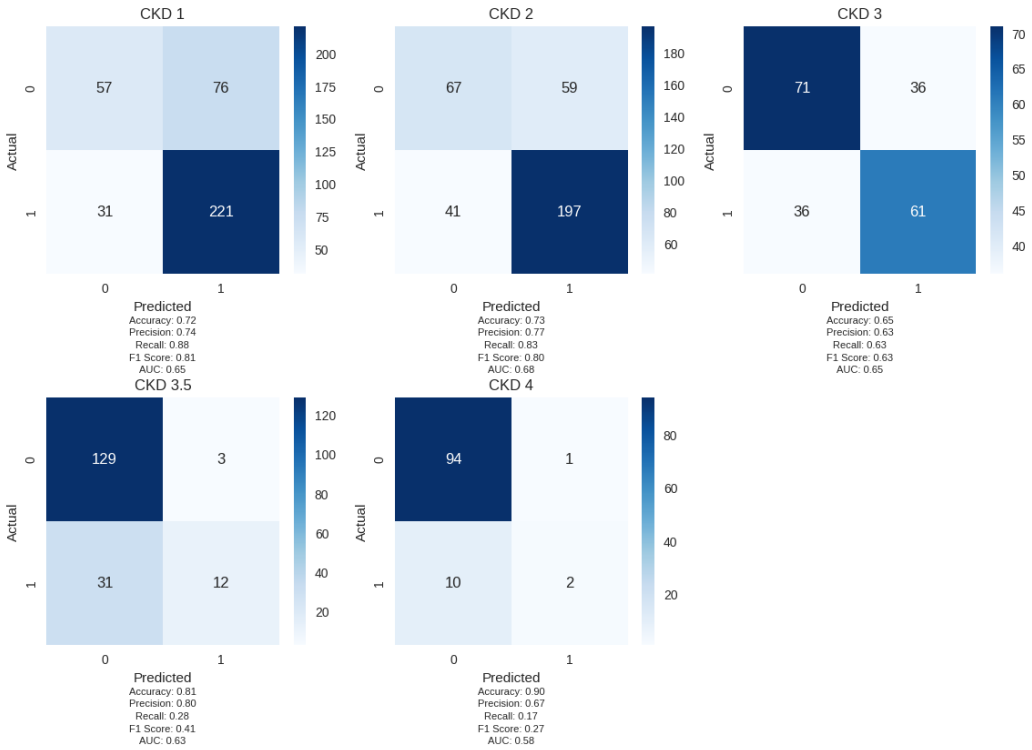

Figure 5.1: GBM classifier performance by current patient stage – approach A.

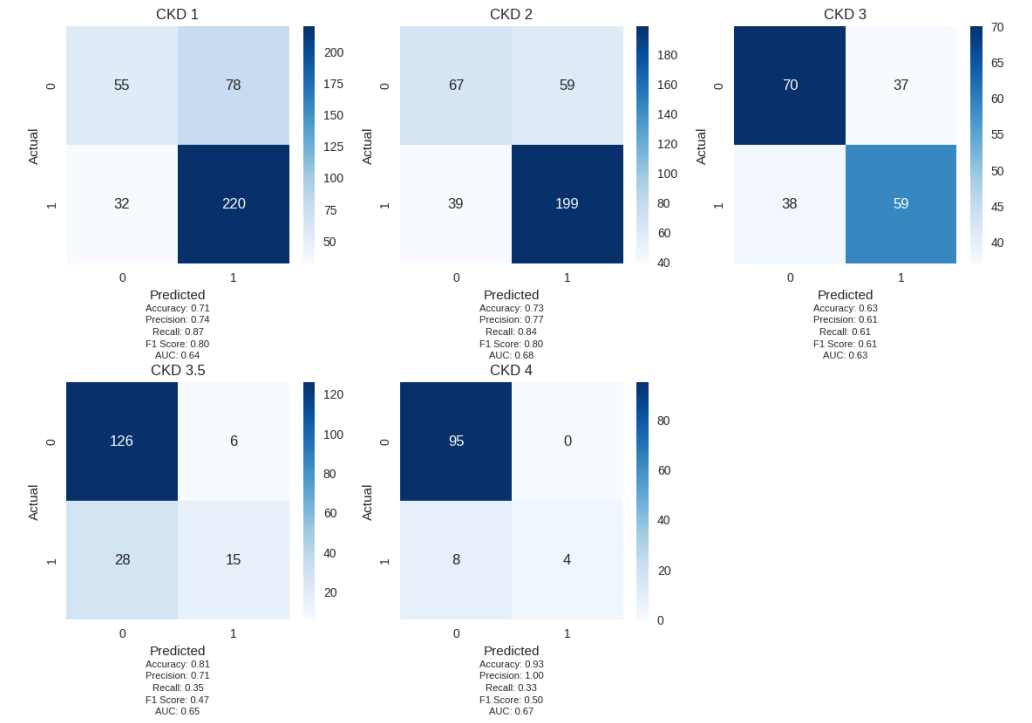

Figure 5.2: Catboost classifier performance by current patient stage – approach A.

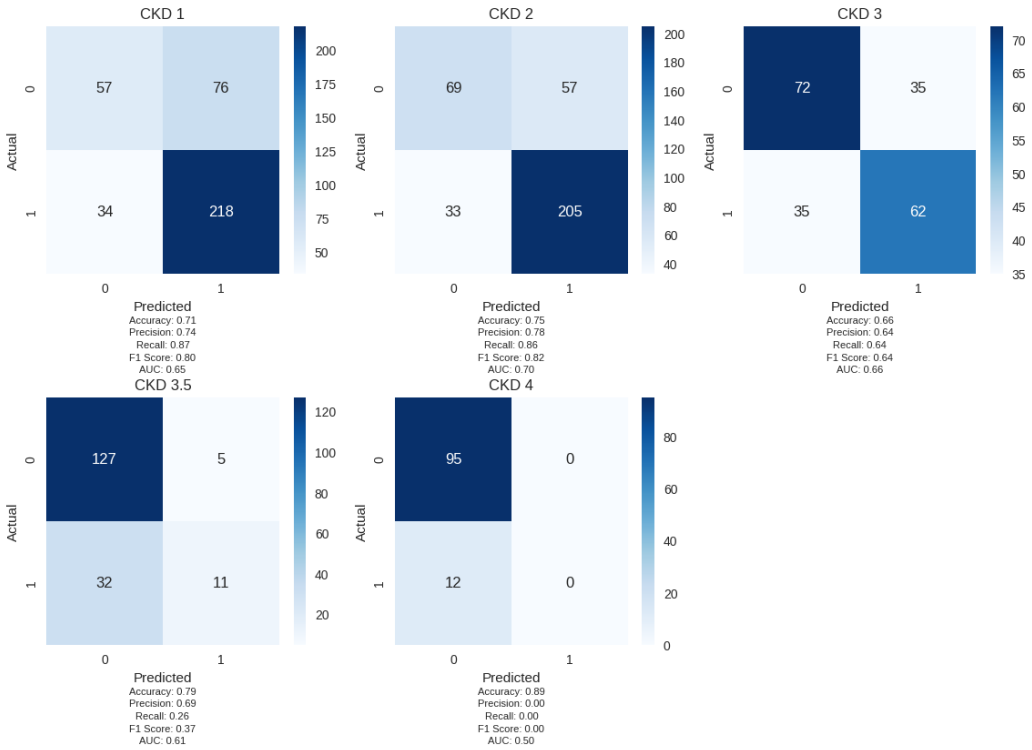

Figure 5.3: LightGBM classifier performance by current patient stage – approach A.

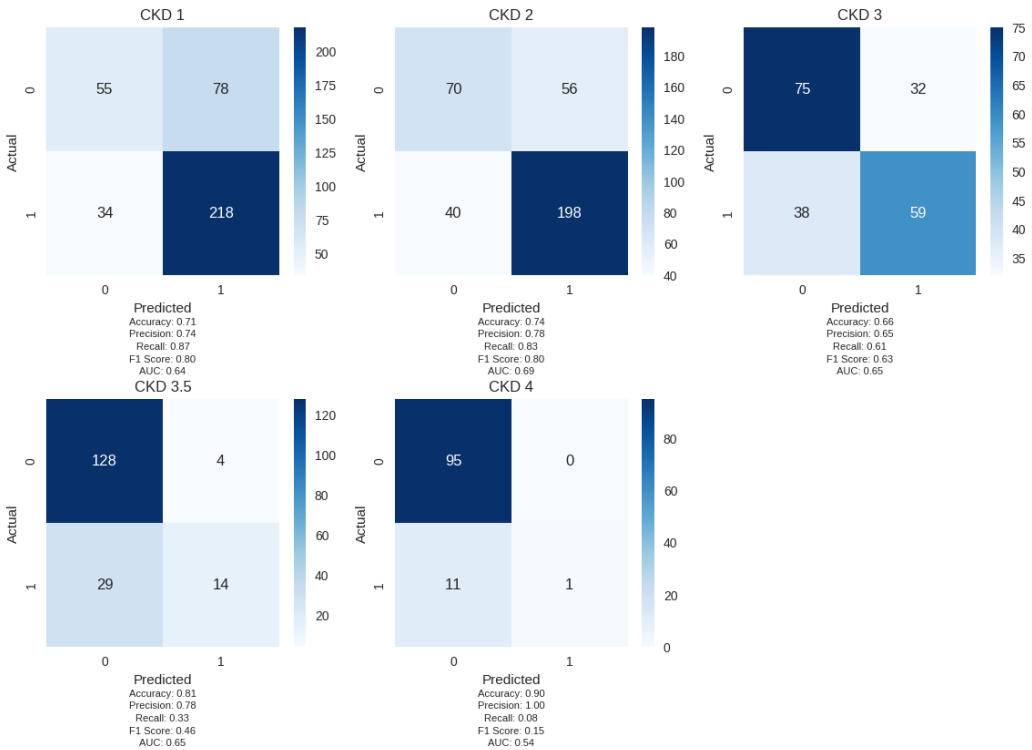

Figure 5.4: MLP classifier performance by current patient stage – approach A.

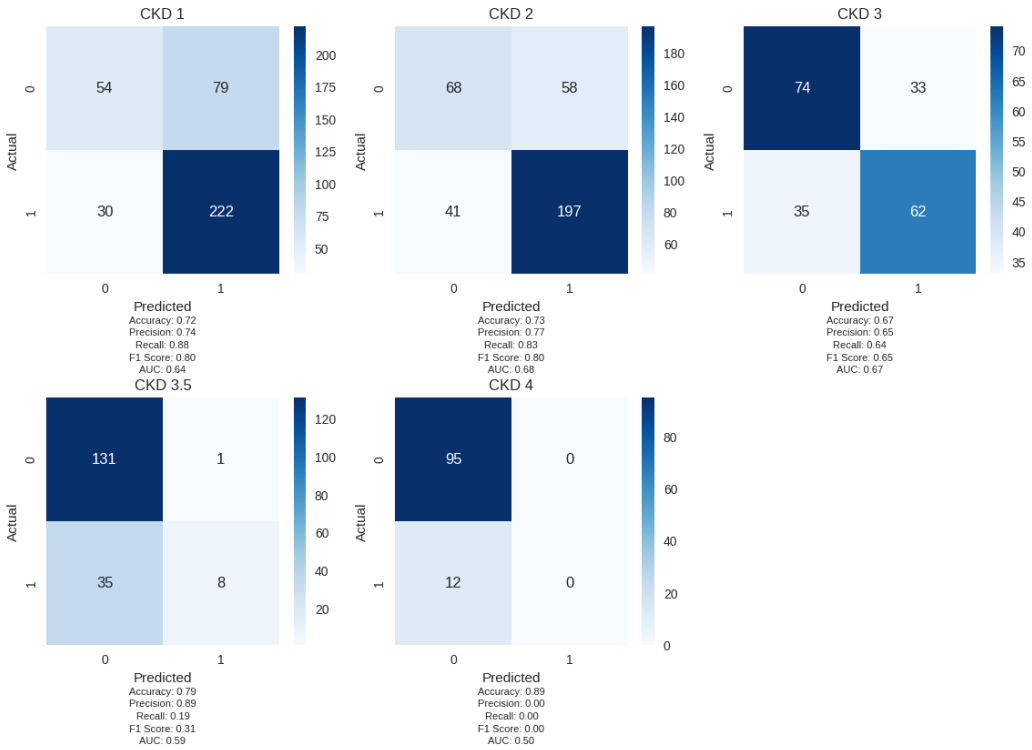

Figure 5.5: Adaboost classifier performance by current patient stage – approach A.

The results of the different models by the last available stage of the patient (current stage) show the inability of the different models to predict the evolution of the disease in the following year when patients are in certain stages. The classifier has a clear tendency to predict that the patient worsens when the patient is in earlier stages (1 and 2) and a clear tendency to predict that the patient remains stable when the patient is in more advanced stages (3,5 and 4). This is a result of the way the data is distributed in the training and test set and in the unseen data, as shown in chapter 4.3.7. To try to bring a predictive capability to any patient at any disease stage, approach B was created, and is presented in the next chapter.

5.2 Approach B: Balanced target distribution by disease stage

This approach, compared to the previous one, brings a total balance in the training and test set and in the unseen data set per disease stage. This allows to train the model with the same number of instances of patients that worsen and that remain stable for each stage of the disease. In the unseen data that serves as validation for the generalization of the model, there are the same number of instances for each stage and the same number of instances that remain stable and worsen. This allows for no bias in the results, and the model can predict any worsening of the disease at any stage.

The same classifiers applied in approach A were trained, and their performance on the training set is shown in Table 5.4.

Table 5.4: Performance of ML algorithms on train set – approach B.

Model	Accuracy	Recall	Precision	F1 Score	AUC	MCC
GBM	**0.7139**	0.7213	**0.7111**	0.7160	0.7902	**0.4281**
Catboost	0.7119	0.7220	0.7081	0.7148	0.7865	0.4241
MLP	0.7109	0.7240	0.7056	0.7145	0.7846	0.4221
LR	0.7101	**0.7315**	0.7018	**0.7162**	0.7798	0.4208
LightGBM	0.7101	0.7153	0.7084	0.7116	**0.7939**	0.4205
LDA	0.7098	0.7197	0.6983	0.7083	0.7796	0.4205
Ridge	0.7095	0.7197	0.6978	0.7081	0.0000	0.4198
Adaboost	0.7085	0.7189	0.7046	0.7115	0.7768	0.4173
RF	0.7026	0.7015	0.7033	0.7022	0.7716	0.4054
SVM	0.7019	0.7077	0.7011	0.7032	0.0000	0.4051
ET	0.7016	0.6988	0.7029	0.7007	0.7696	0.4033
XGBoost	0.6990	0.7072	0.6961	0.7014	0.7677	0.3983
KNN	0.6465	0.6307	0.6513	0.6407	0.6936	0.2931
DT	0.6257	0.6226	0.6265	0.6244	0.6257	0.2516
QDA	0.5943	0.2611	0.7850	0.3889	0.7596	0.2533
NB	0.5522	0.2639	0.6251	0.3696	0.6216	0.1285

The five models chosen for hyperparameter tuning were: GBM, Catboost, MLP, LR and LightGBM. Other models could have been chosen because their performance is quite similar, such as LDA, Ridge and even Adaboost, but algorithms such as LR and LightGBM are more frequently proposed in the literature. The results of the selected models in the test set after parameter tuning are presented in Table 5.5.

Table 5.5: Performance of ML algorithms after hyperparameters tuning on test set – approach B.

Model	Accuracy	Recall	Precision	F1 Score	AUC	MCC
GBM	0.7092	0.7233	**0.7035**	0.7133	**0.7791**	0.4185
Catboost	0.7064	**0.7267**	0.6983	0.7122	0.7768	0.4130
MLP	0.7073	0.7261	0.6999	0.7128	0.7789	0.4150
LR	0.6978	0.7222	0.6887	0.7050	0.7678	0.3961
LightGBM	**0.7093**	0.7250	0.7030	**0.7138**	0.7764	**0.4189**

The test results after parameter tuning generally show a slight drop in performance metrics, but nothing considerable, with some occasions even showing an increase. All models were thus retrained with all available data (train and test set), and their performance evaluated on unseen data. The results are presented in Table 5.6.

Table 5.6: Performance of ML algorithms on unseen data – approach B.

Model	Accuracy	Recall	Precision	F1 Score	AUC	MCC
GBM	0.7142	0.7488	0.7000	0.7236	0.7836	0.4294
Catboost	0.7036	0.7245	0.6952	0.7095	0.7732	0.4077
MLP	0.7101	**0.7647**	0.6858	**0.7275**	**0.7887**	0.4239
LR	0.6858	0.7374	0.6681	0.7011	0.7602	0.3737
LightGBM	**0.7166**	0.7342	**0.7089**	0.7213	0.7806	**0.4335**

The results on unseen data generally remained similar, and even improved for some algorithms such as LightGBM and MLP.

As with approach A, this does not necessarily mean that the model has predictive power for all patients, regardless of their disease stage. The performance of the

different models by disease stage is shown in Figure 5.6 - Figure 5.10. The difference in the validation and results obtained when compared to approach A is easily noticeable. These results increase the guarantee of generalization of the model regardless of the stage of the disease in which the patient is.

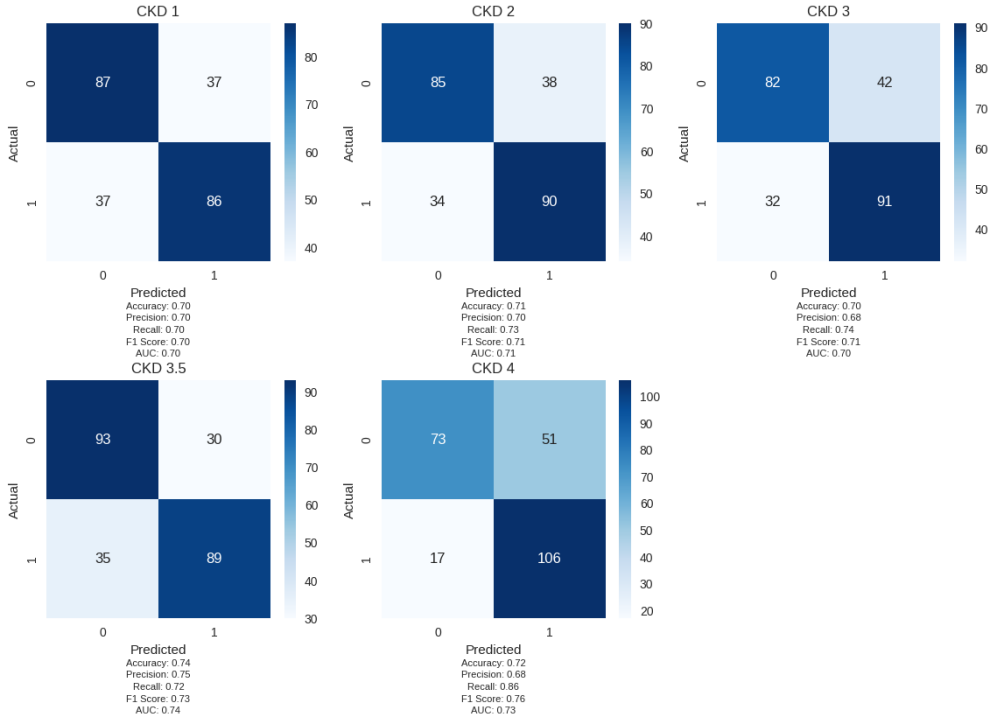

Figure 5.6: GBM classifier performance by current patient stage – approach B.

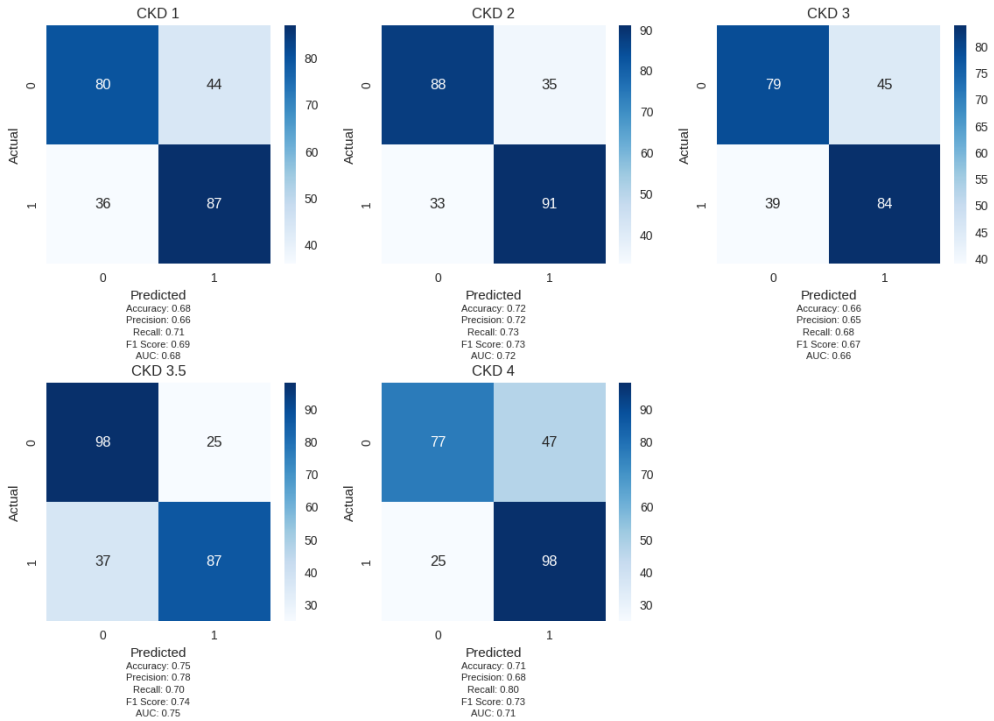

Figure 5.7: Catboost classifier performance by current patient stage – approach B.

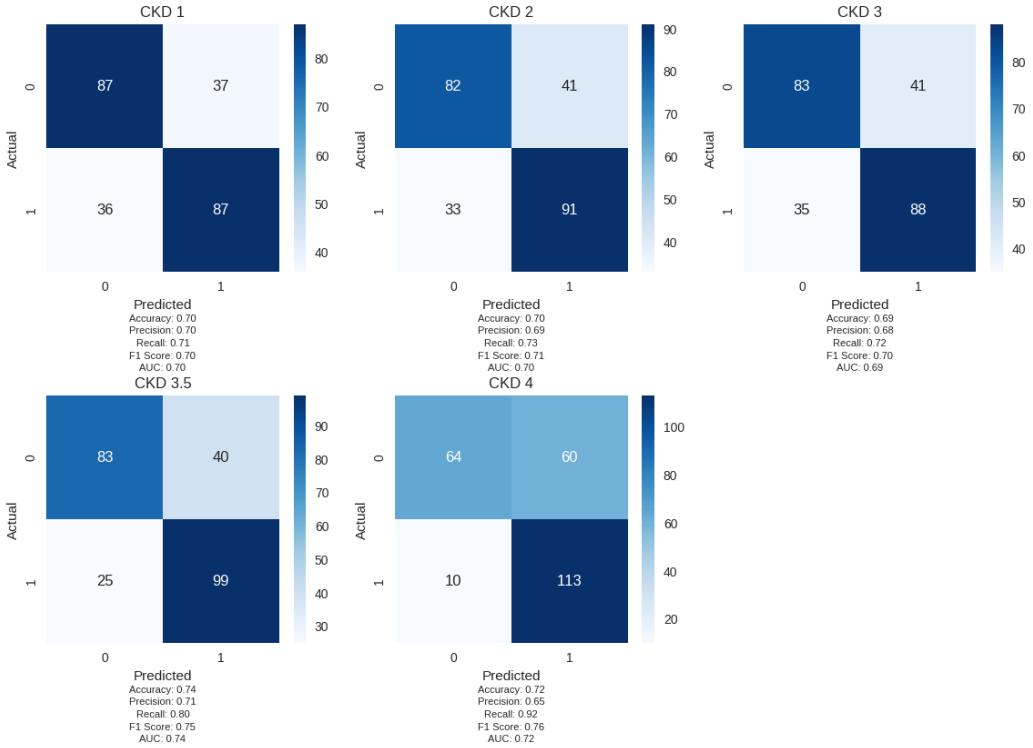

Figure 5.8: MLP classifier performance by current patient stage – approach B.

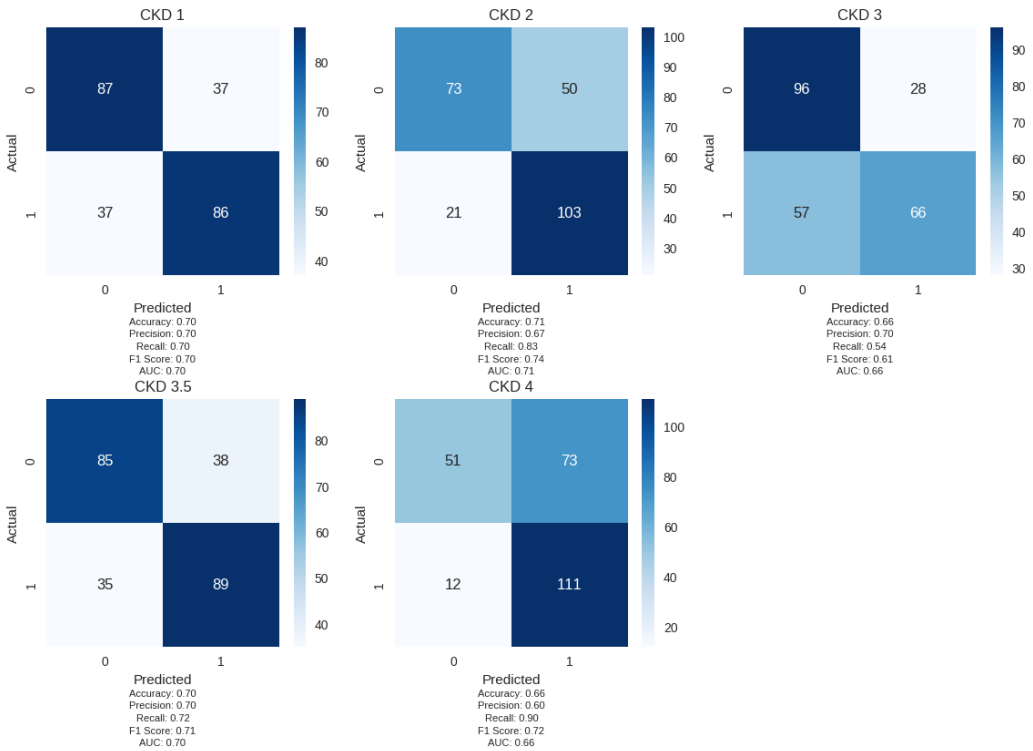

Figure 5.9: LR classifier performance by current patient stage – approach B.

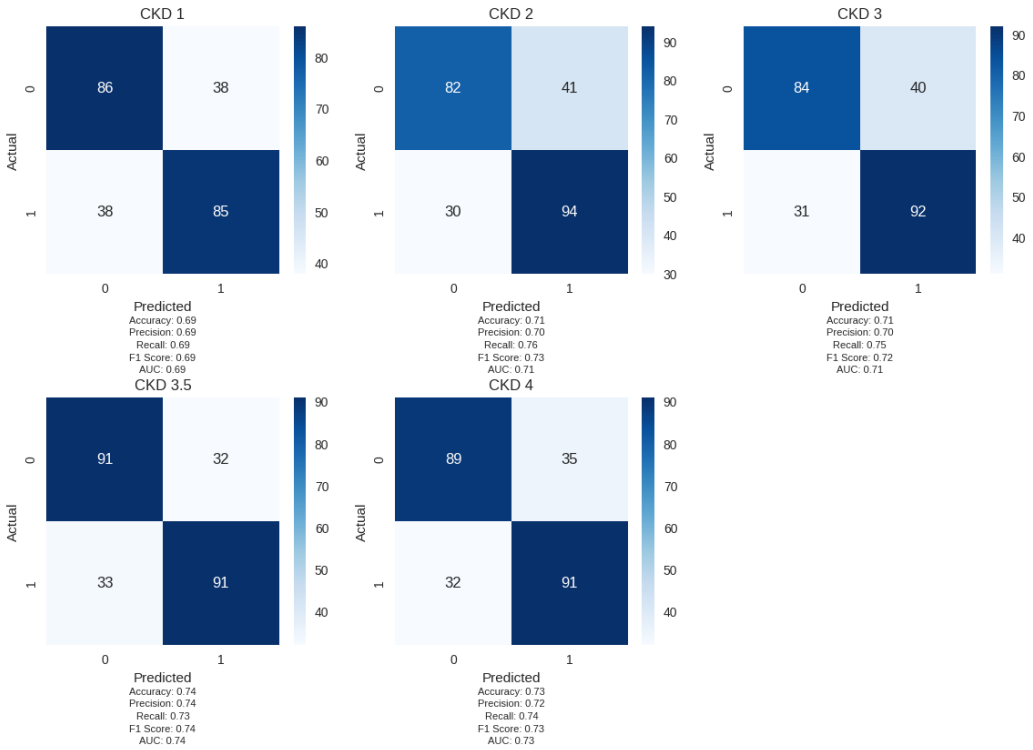

Figure 5.10: LightGBM classifier performance by current patient stage – approach B.

5.3 Proposed model

After analyzing all the results shown above on approach B, the selection of the best model was based on its performance across the training set, test set, and unseen data, with more emphasis on the results of the performance metrics obtained on the unseen data. Furthermore, beyond the overall performance, the examination of performance based on different disease stages was considered. To facilitate this analysis, the statistical significance analysis between the performance of the different models, as presented in Appendix C.1, and the ranking of features provided in Appendix C.2, were also taken into careful consideration.

That said, the model considered to have the best performance is the LightGBM. It was able to maintain its performance throughout the various stages of the experimental setup. It is also the best model to predict the outcome of the disease in more advanced stages, also maintaining good performance in earlier and intermediate stages.

5.3.1 Interpretation

The interpretation can be divided into global and local aspects. To assess the importance of features for the final output, the beeswarm SHAP plot is commonly employed, which represents a prevalent form of global interpretation facilitated by the SHAP library [192].

As can be seen in Figure 5.11, the beeswarm plot presents the features with the most significant impact and identifies how their values influence the model's output. A negative impact on the output raises the likelihood of predicting a negative (stable) class, while a positive contribution enhances the probability of predicting a positive (aggravation) class. High values in features such as age, albuminuria_2 and HbA1c lead to an increased risk of worsening DN. On the other hand, it is possible to observe that the higher the patient's current stage (ckd_2), the lower the probability of worsening.

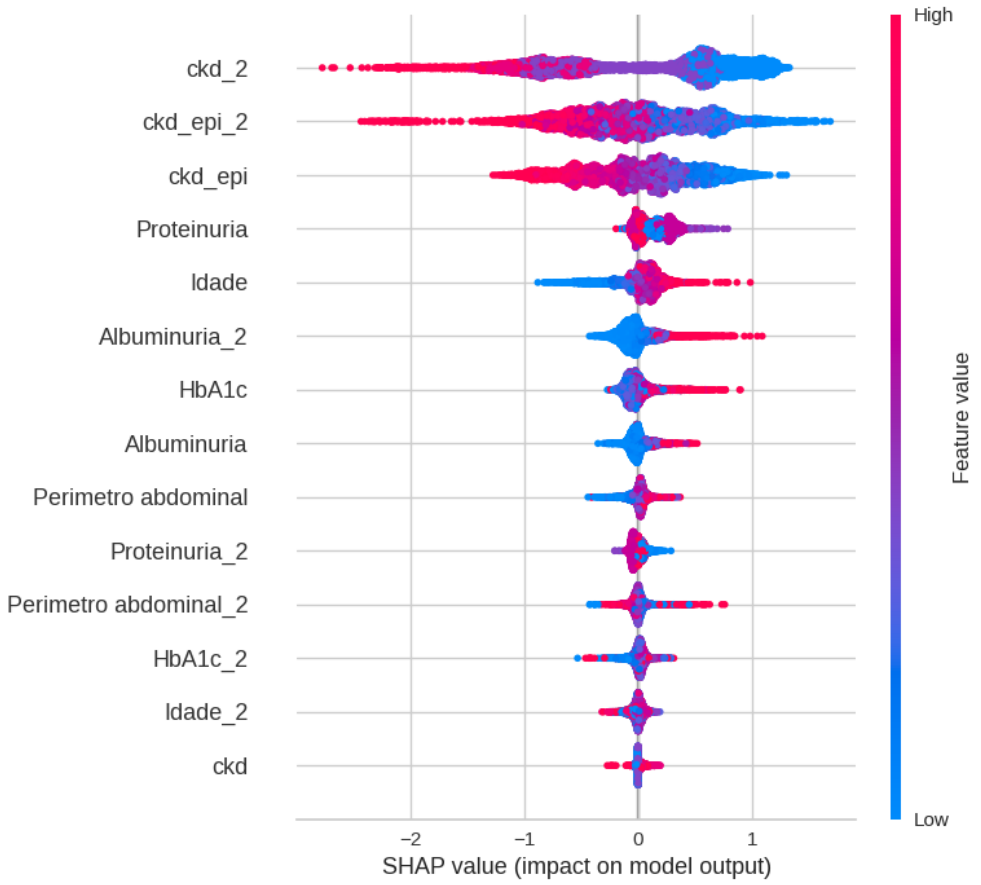

Figure 5.11: Global interpretation using Beeswarm SHAP plot.

As a local interpretation, that is, for an individual prediction, two different plots were used: waterfall and force plot. Both present the same information but with different layouts. The waterfall plot can be seen in Figure 5.12 and the force plot in Figure 5.13. Like the beesworm SHAP plot, both the waterfall and the force plot present the impact of a feature as positive (in red) or negative (in blue).

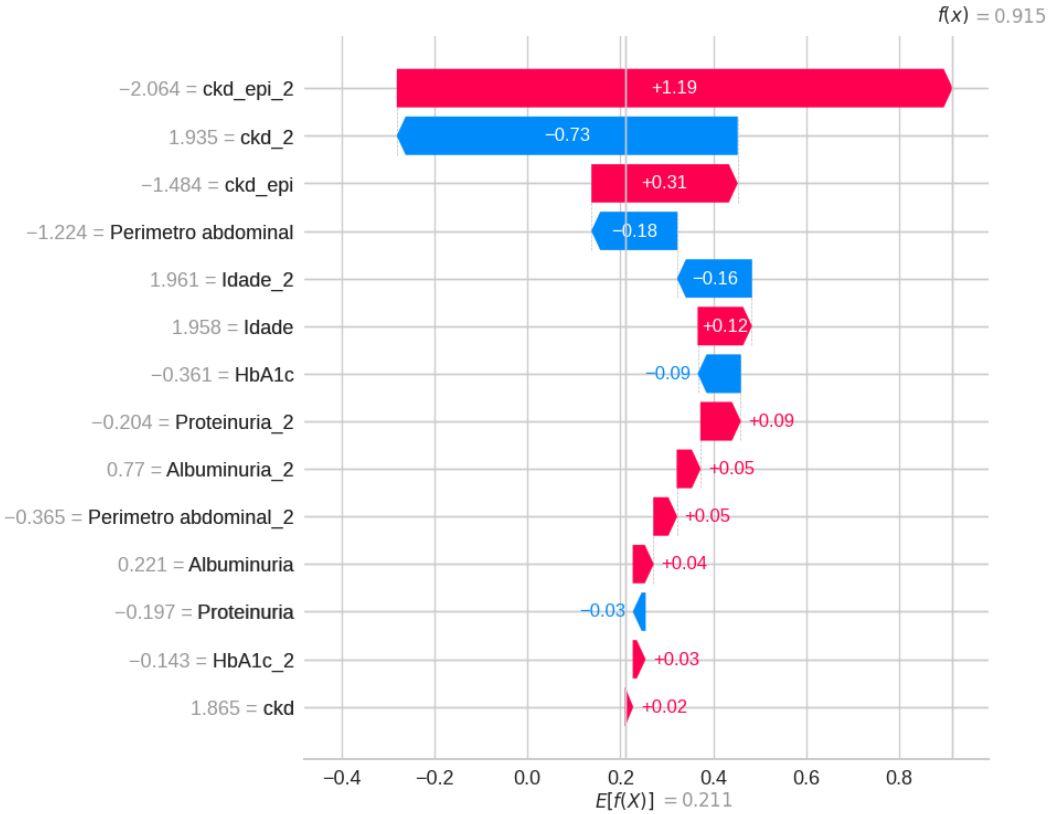

Figure 5.12: Local interpretation using SHAP waterfall plot.

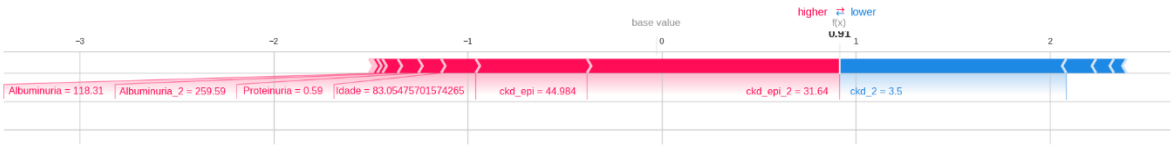

Figure 5.13: Local interpretation using SHAP force plot.

5.3.2 Web application deployment

To allow the model to be tested by those who wish to do so, a web application was created using Gradio and the HF servers. The user provides the variables needed for the predictive model, that is, 14 features in total, 7 for each year. After submitting the necessary data, the user has access to the forecast (stable or aggravation), the confidence that the model has associated with this forecast (in percentage) and the two local interpretation graphs presented in the previous chapter.

A random patient present in the unseen dataset was used as an example. The data input for this example patient can be seen in Figure 5.14. In the application ten examples are already provided, randomly taken from the unseen data. The example presented here is among those ten and can be viewed in detail in the app.

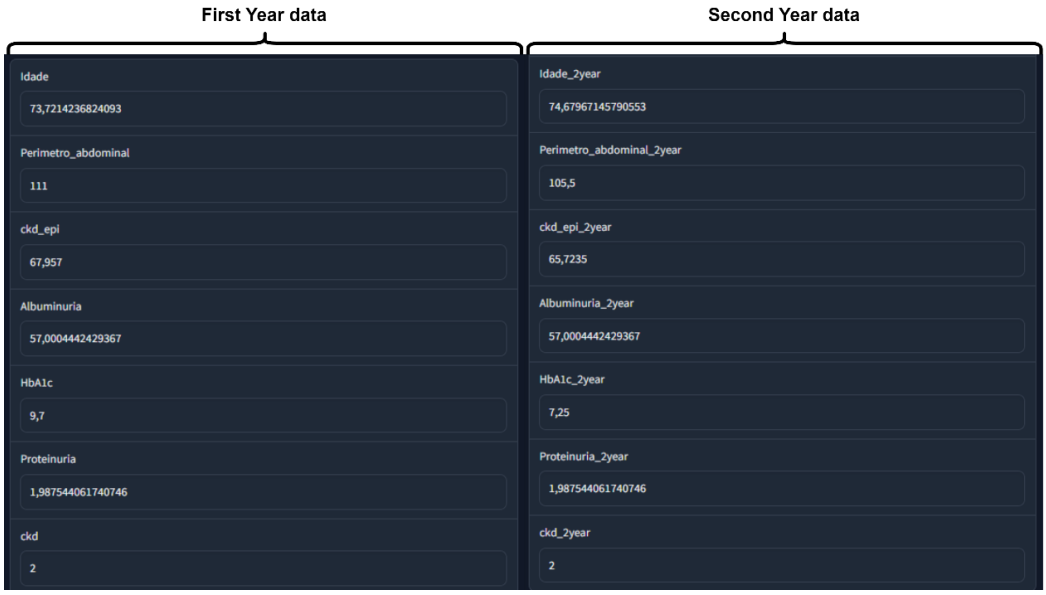

Figure 5.14: Example of data input in the created application.

After submitting the data, it will be processed and passed through the ML pipeline where the values are normalized and provided to the ML model. The user receives on his side, in a few seconds, the prediction, the certainty associated with this result and the two SHAP local interpretation plots. The output is shown in Figure 5.15.

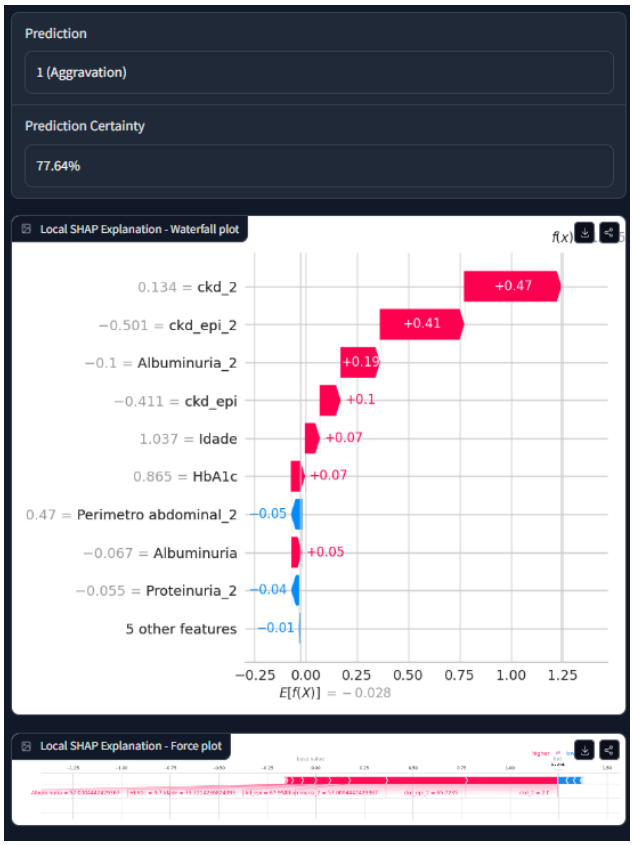

Figure 5.15: Example of output generated by created application.

6 DISCUSSION

This study presents an approach capable of predicting the risk of evolution of DN within one year, considering two years of patient history. During this work, the literature on DN was examined in detail to gain insights into various approaches employed for similar issues. This knowledge helped and grounded several steps, such as data analysis and preprocessing, as well as the creation and comparison of several ML models.

Approach A presents models with interesting results on classification metrics when performing a general analysis of training, test, and unseen data. Initially, these models demonstrate the ability to predict the evolution of DN in one year. However, a more detailed analysis of their performance, considering the stage of the disease, reveals a limitation in predicting aggravation events in more advanced stages. This limitation is due to the data distribution in the training, test, and unseen data sets.

To address this limitation, Approach B modifies the data distribution on train, test, and unseen data. This results in a more robust model, capable of maintaining performance regardless of patient characteristics. Although the overall results of Approach B may fall slightly behind those of Approach A, a significant difference is observed when examining the results by disease stage. Approach B consistently maintains its predictive ability across different stages, showing a slightly higher predictive ability for the more advanced stages such as stages 3.5 and 4.

By analyzing the SHAP plots and feature rankings, it becomes evident that certain variables hold significant importance in the predictive model. The current stage of the disease in the second year of the patient's history (ckd_2) and the calculated eGFR values (ckd_epi and ckd_epi2) have high relevance on prediction. Also, it is noticeable that the model considers the difference in eGFR from one year to the next. A drop in the eGFR value (variable ckd_epi) indicates in most cases a strong likelihood of disease aggravation.

The main findings of this work were:

- **Temporal importance in DN risk prediction:** When assessing a patient's risk for DN evolution, doctors consider the patient's history and the temporal changes in certain clinical values. This study highlights the significance of temporality and the impact it may have on creating better predictive models. This importance of temporality is evident in Figure 5.12 and Figure 5.13, where the prediction indicates the patient's aggravation, primarily based on the decrease in ckd_epi over one year.
- **Better and more in-depth analysis of ML models:** This study highlights the necessity of further validating the performance of predictive models. It is crucial to remember that good results alone do not guarantee a high predictive capacity [193]. Therefore, careful validation of the results is essential to identify and mitigate any biases associated with the model, ensuring its clinical

applicability. It is also necessary to emphasize the need to validate the results presented by using one or several different data sets. Through this external validation it is possible to ensure that the proposed model is able to generalize what it has learned [194].

- **End-to-end ML data pipeline:** From data preprocessing to model training, evaluation and deployment, each step was carefully designed to take into account the temporal patterns and dependencies present in the data. This end-to-end pipeline enables us to effectively capture the temporal dynamics of DN progression and utilize it for accurate predictions using ML techniques.

6.1 Strengths and Limitations of the Study

This study has several strengths and limitations. As strengths, it is possible to identify the following:

- **Clinical context:** This study as well as all the review of the literature were carried out together with the constant support of APDP. This close collaboration allowed us to understand the various aspects of DN, which enabled the formation of a step-by-step approach that aligned with the specific problem's characteristics and underlying nature.
- **Selection of the proposed model:** The entire experimental setup was designed to follow the best practice recommendations for training ML models [67]. Also, an analysis of performance by patient stage was also made, resulting in two different approaches. In addition to performance, statistical significance and ranking of features were also considered. This brings greater confidence in the results, leading to a proposed model that is more robust and better prepared to make predictions on new data.
- **ML model interpretation and deploy:** Making the model accessible to anyone through a simple and intuitive interface is essential, but, in some ways, a complex task as described in [195]. Not only has this been done in this study, but the user is also provided with interpretability plots that show the logic behind a particular prediction. Access to the model and the interpretation of the results produced by it leads to greater transparency of the study [196].

Despite the various strengths presented, this study also has some associated limitations:

- **Inability to take full advantage of temporality:** The architecture of the implemented ML models leads to the need to adapt the data to be less sparse, more uniform and with well-defined structures. This has led to a great difficulty in taking advantage of all the temporality associated with the patient's history. With a greater exploitation of the temporality and evolution associated with each patient, better and more robust ML classifiers can be created [197].

- **Loss of information in data:** Despite considering a time frame of just 2 years, it was necessary to condense multiple consultations within a year into a single record. This approach aimed to mitigate information loss, although this loss is unavoidable, particularly for patients with numerous records within a specific year. It is likely that greater use of all of a patient's historical information will lead to better predictive performance and therefore loss of information is indicated here as a limitation to the work.
- **Performance of ML models:** Despite undergoing various optimization phases, including adjustments to data distribution and hyperparameters tuning, the proposed model ultimately falls short of achieving optimal performance. The accuracy results indicate a success rate of approximately 72%, implying that it fails to make accurate predictions roughly one out of every three attempts. While achieving high performance with real EHR raw data can be challenging, it is important to acknowledge that the modest performance of the model remains a limitation in this study.
- **Dataset shift:** The model in this study is trained using data exclusively from patients treated at the APDP clinic. It is crucial to understand that this model may not perform as effectively in individuals from different populations. This limitation is known as "dataset shift," where the model's performance can be affected by differences between the training data and the target population [198].

Despite the limitations presented, it is possible to state that the proposed model seems to be able to predict the evolution of DN in one year with acceptable effectiveness considering 2 years of patient history.

7 CONCLUSION

Patients with type 1 or type 2 diabetes suffer from various complications, with DN being one of the most severe. It is the major cause of ESRD worldwide leading to a lower quality of life and a huge burden on both individuals and health systems. The possibility of risk profiling of patients may lead to better control of DN and consequently to a significant decrease of patients in advanced stages of the disease. The objective of this study was to develop a predictive ML model capable of predicting the risk of DN progression using patients' EHR data. For this purpose, a dataset from APDP was used, encompassing 21,284 patients who were followed for over 22 years, resulting in approximately 413,097 recorded visits.

This study proposes a predictive model able to predict the evolution of DN in one year using information related to the last two years of the patient. ML model architectures are not designed to deal with temporal and sequential data, and therefore it was necessary to shape and structure the data in order to create models with good predictive ability. This was possible through a longitudinal approach where several data preprocessing steps were done considering the nature of EHR data and the temporality associated with patient history. In addition, the performance of the proposed model was validated in multiple steps, ensuring that it has good predictive capacity and remains constant regardless of the patient's stage.

Just as important as a model with predictive ability is the ability to show the reason behind a prediction. For this, the SHAP method was used, which shows in a visual and intuitive way which variables contributed either positively or negatively to the predicted outcome. To finalize the work, the model was deployed so that any user could access it, make a prediction, and have access not only to the outcome predicted, but also to its interpretation.

Overall, this study presents a significant contribution in the field of DN risk prediction by developing a predictive ML model using EHR data. The longitudinal approach, robust model performance, interpretability, and user-friendly deployment create the basis for improved disease management and enhanced patient outcomes.

7.1 Future research directions

As future work, several different directions can be suggested:

- **Use of DL architectures:** As mentioned earlier, classical ML architectures were not designed to handle temporal and sequential data. DL has demonstrated success in health risk prediction, especially for patients with chronic and progressing conditions like DN [199], [200]. Stage-Aware Neural Networks (StageNet) [200] and Time-aware LSTM networks [201] are two of the architectures identified as having the greatest potential to better

incorporate the time factor and mitigate the challenges of EHR data. This can consequently lead to better performing models.

- **Incremental Learning:** With the increase in data and especially its diversity, ML models can be expected to perform better. The ability of a model to learn new knowledge without forgetting what it has previously learned is referred to as continual, lifelong, or incremental learning. This type of approach would allow the model to continuously learn by adapting to new data as it becomes available [202].
- **Make use of information from text features:** In this work, textual features were discarded due to the need to apply NLP techniques and the high cardinality of these variables. However, it is expected that the full utilization of these features will enrich the data and bring relevant information capable of improving the predictive capacity of the classifiers.
- **Overcome the "curse of dimensionality":** Different feature dimension reduction techniques can be explored in order to understand if it is possible to keep most of the information available while maintaining and even improving the ability of ML algorithms to learn and detect patterns in the data. Some works have tried to overcome the "curse of dimensionality" by using these techniques [203]–[205].

REFERENCES

[1] "Diabetes." https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed Jun. 16, 2023).

[2] OECD, *Health at a Glance: Europe 2020: State of Health in the EU Cycle*. Paris: Organisation for Economic Co-operation and Development, 2020. Accessed: Oct. 29, 2022. [Online]. Available: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2020_82129230-en

[3] Z. T. Bloomgarden, "Diabetes Complications," *Diabetes Care*, vol. 27, no. 6, pp. 1506–1514, Jun. 2004, doi: 10.2337/diacare.27.6.1506.

[4] P. Fioretto, I. Barzon, and M. Mauer, "Is diabetic nephropathy reversible?," *Diabetes Res. Clin. Pract.*, vol. 104, no. 3, pp. 323–328, Jun. 2014, doi: 10.1016/j.diabres.2014.01.017.

[5] "Diabetic Kidney Disease: A Report From an ADA Consensus Conference | Diabetes Care | American Diabetes Association." https://diabetesjournals.org/care/article/37/10/2864/30796/Diabetic-Kidney-Disease-A-Report-From-an-ADA (accessed Oct. 29, 2022).

[6] J. Wong, M. Murray Horwitz, L. Zhou, and S. Toh, "Using Machine Learning to Identify Health Outcomes from Electronic Health Record Data," *Curr. Epidemiol. Rep.*, vol. 5, no. 4, pp. 331–342, Dec. 2018, doi: 10.1007/s40471-018-0165-9.

[7] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, Jan. 2017, doi: 10.1016/j.csbj.2016.12.005.

[8] A. P. dos D. de P.- Apdp, "Associação Protectora dos Diabéticos de Portugal - APDP," 2019, Accessed: Jul. 19, 2023. [Online]. Available: https://apdp.pt/

[9] T. Tuomi, "Type 1 and Type 2 Diabetes: What Do They Have in Common?," *Diabetes*, vol. 54, no. suppl_2, pp. S40–S45, Dec. 2005, doi: 10.2337/diabetes.54.suppl_2.S40.

[10] T. A. Buchanan and A. H. Xiang, "Gestational diabetes mellitus," *J. Clin. Invest.*, vol. 115, no. 3, pp. 485–491, Mar. 2005, doi: 10.1172/JCI24531.

[11] A. D. Deshpande, M. Harris-Hayes, and M. Schootman, "Epidemiology of Diabetes and Diabetes-Related Complications," *Phys. Ther.*, vol. 88, no. 11, pp. 1254–1264, Nov. 2008, doi: 10.2522/ptj.20080020.

[12] A. S. Reddi and K. Kuppasani, "Kidney Function in Health and Disease," in *Nutrition in Kidney Disease*, L. D. Byham-Gray, G. M. Chertow, and J. D. Burrowes, Eds., in Nutrition and Health. Totowa, NJ: Humana Press, 2008, pp. 3–15. doi: 10.1007/978-1-59745-032-4_1.

[13] S. Hussain, M. Chand Jamali, A. Habib, M. S. Hussain, M. Akhtar, and A. K. Najmi, "Diabetic kidney disease: An overview of prevalence, risk factors, and biomarkers," *Clin. Epidemiol. Glob. Health*, vol. 9, pp. 2–6, Jan. 2021, doi: 10.1016/j.cegh.2020.05.016.

[14] Y. Chen, K. Lee, Z. Ni, and J. C. He, "Diabetic Kidney Disease: Challenges, Advances, and Opportunities," *Kidney Dis.*, vol. 6, no. 4, pp. 215–225, Mar. 2020, doi: 10.1159/000506634.
[15] M. Prata, "Incidence in dialysis and chronic kidney disease prevalence in the Portuguese population," *Port. J. Nephrol. Hypertens.*, vol. 35, pp. 63–63, Apr. 2021, doi: 10.32932/pjnh.2021.04.122.
[16] J. Vinhas *et al.*, "Prevalence of Chronic Kidney Disease and Associated Risk Factors, and Risk of End-Stage Renal Disease: Data from the PREVADIAB Study," *Nephron Clin. Pract.*, vol. 119, no. 1, pp. c35–c40, Jun. 2011, doi: 10.1159/000324218.
[17] O. Gheith, N. Farouk, N. Nampoory, M. A. Halim, and T. Al-Otaibi, "Diabetic kidney disease: world wide difference of prevalence and risk factors," *J. Nephropharmacology*, vol. 5, no. 1, pp. 49–56, Oct. 2015.
[18] Y. Deng *et al.*, "Global, Regional, and National Burden of Diabetes-Related Chronic Kidney Disease From 1990 to 2019," *Front. Endocrinol.*, vol. 12, 2021, Accessed: Jun. 16, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fendo.2021.672350
[19] P. B. Vinod, "Pathophysiology of diabetic nephropathy," *Clin. Queries Nephrol.*, vol. 1, no. 2, pp. 121–126, Apr. 2012, doi: 10.1016/S2211-9477(12)70005-5.
[20] C. Mora-Fernández, V. Domínguez-Pimentel, M. M. de Fuentes, J. L. Górriz, A. Martínez-Castelao, and J. F. Navarro-González, "Diabetic kidney disease: from physiology to therapeutics," *J. Physiol.*, vol. 592, no. 18, pp. 3997–4012, 2014, doi: 10.1113/jphysiol.2014.272328.
[21] C.-Y. Jung and T.-H. Yoo, "Pathophysiologic Mechanisms and Potential Biomarkers in Diabetic Kidney Disease," *Diabetes Metab. J.*, vol. 46, no. 2, pp. 181–197, Mar. 2022, doi: 10.4093/dmj.2021.0329.
[22] G. Wang *et al.*, "The analysis of risk factors for diabetic nephropathy progression and the construction of a prognostic database for chronic kidney diseases," *J. Transl. Med.*, vol. 17, no. 1, p. 264, Aug. 2019, doi: 10.1186/s12967-019-2016-y.
[23] N. Samsu, "Diabetic Nephropathy: Challenges in Pathogenesis, Diagnosis, and Treatment," *BioMed Res. Int.*, vol. 2021, p. e1497449, Jul. 2021, doi: 10.1155/2021/1497449.
[24] M. Oshima *et al.*, "Trajectories of kidney function in diabetes: a clinicopathological update," *Nat. Rev. Nephrol.*, vol. 17, no. 11, Art. no. 11, Nov. 2021, doi: 10.1038/s41581-021-00462-y.
[25] K. Tziomalos and V. G. Athyros, "Diabetic Nephropathy: New Risk Factors and Improvements in Diagnosis," *Rev. Diabet. Stud.*, vol. 12, no. 1–2, 2015, doi: 10.1900/RDS.2015.12.110.
[26] M. C. Thomas *et al.*, "Diabetic kidney disease," *Nat. Rev. Dis. Primer*, vol. 1, no. 1, Art. no. 1, Jul. 2015, doi: 10.1038/nrdp.2015.18.
[27] R. Z. Alicic, M. T. Rooney, and K. R. Tuttle, "Diabetic Kidney Disease: Challenges, Progress, and Possibilities," *Clin. J. Am. Soc. Nephrol.*, vol. 12, no. 12, p. 2032, Dec. 2017, doi: 10.2215/CJN.11491116.

[28] D. Lizicarova, B. Krahulec, E. Hirnerova, L. Gaspar, and Z. Celecova, "Risk factors in diabetic nephropathy progression at present," *Bratisl. Med. J.*, vol. 115, no. 08, pp. 517–521, 2014, doi: 10.4149/BLL_2014_101.
[29] P. Rossing and M. Frimodt-Møller, "Clinical Features and Natural Course of Diabetic Nephropathy," in *Diabetic Nephropathy: Pathophysiology and Clinical Aspects*, J. J. Roelofs and L. Vogt, Eds., Cham: Springer International Publishing, 2019, pp. 21–32. doi: 10.1007/978-3-319-93521-8_2.
[30] B. Satirapoj, "Nephropathy in Diabetes," in *Diabetes: An Old Disease, a New Insight*, S. I. Ahmad, Ed., in Advances in Experimental Medicine and Biology. New York, NY: Springer, 2013, pp. 107–122. doi: 10.1007/978-1-4614-5441-0_11.
[31] "Nefropatia Diabética," *APIR - Associação Portuguesa de Insuficientes Renais*. https://www.apir.org.pt/nefropatia-diabetica/ (accessed Jun. 17, 2023).
[32] M. Deem, J. Rice, K. Valentine, J. E. Zavertnik, and M. Lakra, "Screening for diabetic kidney disease in primary care: A quality improvement initiative," *Nurse Pract.*, vol. 45, no. 4, p. 34, Apr. 2020, doi: 10.1097/01.NPR.0000657316.97157.e4.
[33] G. L. Bakris, "Recognition, pathogenesis, and treatment of different stages of nephropathy in patients with type 2 diabetes mellitus," *Mayo Clin. Proc.*, vol. 86, no. 5, pp. 444–457, May 2011.
[34] V. Perkovic *et al.*, "Intensive glucose control improves kidney outcomes in patients with type 2 diabetes," *Kidney Int.*, vol. 83, no. 3, pp. 517–523, Mar. 2013, doi: 10.1038/ki.2012.401.
[35] "Intensive Blood Glucose Control and Vascular Outcomes in Patients with Type 2 Diabetes | NEJM." https://www.nejm.org/doi/10.1056/NEJMoa0802987?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200www.ncbi.nlm.nih.gov (accessed Jun. 18, 2023).
[36] "Effect of intensive diabetes treatment on albuminuria in type 1 diabetes: long-term follow-up of the Diabetes Control and Complications Trial and Epidemiology of Diabetes Interventions and Complications study," *Lancet Diabetes Endocrinol.*, vol. 2, no. 10, pp. 793–800, Oct. 2014, doi: 10.1016/S2213-8587(14)70155-X.
[37] B. Quiroga, D. Arroyo, and G. de Arriba, "Present and Future in the Treatment of Diabetic Kidney Disease," *J. Diabetes Res.*, vol. 2015, p. e801348, Apr. 2015, doi: 10.1155/2015/801348.
[38] "Standards of Medical Care in Diabetes—2015: Summary of Revisions," *Diabetes Care*, vol. 38, no. Supplement_1, p. S4, Dec. 2014, doi: 10.2337/dc15-S003.
[39] C. E. Mogensen, "MICROALBUMINURIA, BLOOD PRESSURE AND DIABETIC RENAL DISEASE: ORIGIN AND DEVELOPMENT OF IDEAS," in *The Kidney and Hypertension in Diabetes Mellitus*, CRC Press, 2004.

[40] X. Shen *et al.*, "Efficacy of statins in patients with diabetic nephropathy: a meta-analysis of randomized controlled trials," *Lipids Health Dis.*, vol. 15, no. 1, p. 179, Oct. 2016, doi: 10.1186/s12944-016-0350-0.
[41] F. Locatelli, B. Canaud, K.-U. Eckardt, P. Stenvinkel, C. Wanner, and C. Zoccali, "The importance of diabetic nephropathy in current nephrological practice," *Nephrol. Dial. Transplant.*, vol. 18, no. 9, pp. 1716–1725, Sep. 2003, doi: 10.1093/ndt/gfg288.
[42] Y. Aso, "Cardiovascular Disease in Patients with Diabetic Nephropathy," *Curr. Mol. Med.*, vol. 8, no. 6, pp. 533–543, Sep. 2008.
[43] E. Ritz, "Anemia and diabetic nephropathy," *Curr. Diab. Rep.*, vol. 6, no. 6, pp. 469–472, Nov. 2006, doi: 10.1007/s11892-006-0081-0.
[44] H. Chen, X. Li, R. Yue, X. Ren, X. Zhang, and A. Ni, "The effects of diabetes mellitus and diabetic nephropathy on bone and mineral metabolism in T2DM patients," *Diabetes Res. Clin. Pract.*, vol. 100, no. 2, pp. 272–276, May 2013, doi: 10.1016/j.diabres.2013.03.007.
[45] K. Bramham, "Diabetic Nephropathy and Pregnancy," *Semin. Nephrol.*, vol. 37, no. 4, pp. 362–369, Jul. 2017, doi: 10.1016/j.semnephrol.2017.05.008.
[46] R. Pang, *Urinary Tract Infection and Nephropathy: Insights into Potential Relationship*. BoD – Books on Demand, 2022.
[47] I. H. de Boer *et al.*, "KDIGO 2020 Clinical Practice Guideline for Diabetes Management in Chronic Kidney Disease," *Kidney Int.*, vol. 98, no. 4, pp. S1–S115, Oct. 2020, doi: 10.1016/j.kint.2020.06.019.
[48] K. E. Peters, S. D. Bringans, W. A. Davis, R. J. Lipscombe, and T. M. E. Davis, "A Comparison of PromarkerD to Standard of Care Tests for Predicting Renal Decline in Type 2 Diabetes".
[49] L. Chan *et al.*, "Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease," *Diabetologia*, vol. 64, no. 7, pp. 1504–1515, Jul. 2021, doi: 10.1007/s00125-021-05444-0.
[50] K. Chauhan *et al.*, "Initial Validation of a Machine Learning-Derived Prognostic Test (KidneyIntelX) Integrating Biomarkers and Electronic Health Record Data To Predict Longitudinal Kidney Outcomes," *Kidney360*, vol. 1, no. 8, p. 731, Aug. 2020, doi: 10.34067/KID.0002252020.
[51] D. Lam *et al.*, "Clinical Utility of KidneyIntelX in Early Stages of Diabetic Kidney Disease in the CANVAS Trial," *Am. J. Nephrol.*, vol. 53, no. 1, pp. 21–31, Jan. 2022, doi: 10.1159/000519920.
[52] E. Wilfinger, "The Kidney Disease Risk Assessment Test," *KidneyIntelX*. https://www.kidneyintelx.com/test/ (accessed Jul. 04, 2023).
[53] N. Tangri *et al.*, "A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure," *JAMA*, vol. 305, no. 15, pp. 1553–1559, Apr. 2011, doi: 10.1001/jama.2011.451.
[54] "The Kidney Failure Risk Equation." http://kidneyfailurerisk.com (accessed Jun. 20, 2023).

[55] P. Wang, "On Defining Artificial Intelligence," *J. Artif. Gen. Intell.*, vol. 10, no. 2, pp. 1–37, Jan. 2019, doi: 10.2478/jagi-2019-0002.
[56] "[PDF] Applicability of Artificial Intelligence in Different Fields of Life | Semantic Scholar." https://www.semanticscholar.org/paper/Applicability-of-Artificial-Intelligence-in-Fields-Shubhendu-Vijay/2480a71ef5e5a2b1f4a9217a0432c0c974c6c28c (accessed May 22, 2023).
[57] A. M. Turing, "Computing Machinery and Intelligence," *Mind New Ser.*, vol. 59, no. 236, pp. 433–460, 1950.
[58] J. Moor, "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years.," *AI Mag.*, vol. 27, pp. 87–91, Jan. 2006.
[59] G. Rebala, A. Ravi, and S. Churiwala, *An Introduction to Machine Learning*. Springer, 2019.
[60] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Mark.*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.
[61] P. Misra and A. Yadav, "Impact of Preprocessing Methods on Healthcare Predictions," *SSRN Electron. J.*, Jan. 2019, doi: 10.2139/ssrn.3349586.
[62] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. in Intelligent Systems Reference Library, vol. 72. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-10247-4.
[63] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 6, p. 420, Aug. 2021, doi: 10.1007/s42979-021-00815-1.
[64] P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Aug. 2018, pp. 1–6. doi: 10.1109/ICCUBEA.2018.8697857.
[65] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, Mar. 2021, doi: 10.1007/s42979-021-00592-x.
[66] O. F. Y. J Akinsola J. E. T. ,. Awodele O. ,. Hinmikaiye J. O. ,. Olakanmi O. ,. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *Seventh Sense Research Group*. https://dev.ijcttjournal.org//archives/ijctt-v48p126 (accessed Jun. 08, 2023).
[67] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning." arXiv, Nov. 10, 2020. Accessed: Jun. 13, 2022. [Online]. Available: http://arxiv.org/abs/1811.12808
[68] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." arXiv, Aug. 09, 2016. doi: 10.48550/arXiv.1602.04938.
[69] W. Samek and K.-R. Müller, "Towards Explainable Artificial Intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., in Lecture

Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 5–22. doi: 10.1007/978-3-030-28954-6_1.
[70] C. C. Yang, "Explainable Artificial Intelligence for Predictive Modeling in Healthcare," *J. Healthc. Inform. Res.*, vol. 6, no. 2, pp. 228–239, Jun. 2022, doi: 10.1007/s41666-022-00114-1.
[71] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, p. eaay7120, Dec. 2019, doi: 10.1126/scirobotics.aay7120.
[72] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3503–3568, Jun. 2022, doi: 10.1007/s10462-021-10088-y.
[73] T. Wang and Q. Lin, "Hybrid predictive models: when an interpretable model collaborates with a black-box model," *J. Mach. Learn. Res.*, vol. 22, no. 1, p. 137:6085-137:6122, Jan. 2021.
[74] P. Villalobos, J. Sevilla, T. Besiroglu, L. Heim, A. Ho, and M. Hobbhahn, "Machine Learning Model Sizes and the Parameter Gap." arXiv, Jul. 05, 2022. doi: 10.48550/arXiv.2207.02852.
[75] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions." arXiv, Nov. 24, 2017. doi: 10.48550/arXiv.1705.07874.
[76] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that?" arXiv, Jan. 25, 2017. doi: 10.48550/arXiv.1611.07450.
[77] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/S2589-7500(21)00208-9.
[78] W. Holmes, M. Bialik, and C. Fadel, "Artificial intelligence in education," in *In: Data ethics : building trust : how digital technologies can serve humanity. (pp. 621-653). Globethics Publications (2023)*, Globethics Publications, 2023, pp. 621–653. Accessed: Jun. 13, 2023. [Online]. Available: https://doi.org/10.58863/20.500.12424/4276068
[79] T. C. W. Lin, "Artificial Intelligence, Finance, and the Law," *Fordham Law Rev.*, vol. 88, p. 531, 2020 2019.
[80] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metab. - Clin. Exp.*, vol. 69, pp. S36–S40, Apr. 2017, doi: 10.1016/j.metabol.2017.01.011.
[81] G. Briganti and O. Le Moine, "Artificial Intelligence in Medicine: Today and Tomorrow," *Front. Med.*, vol. 7, 2020, Accessed: Jun. 13, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2020.00027
[82] G. Currie, K. E. Hawk, E. Rohren, A. Vial, and R. Klein, "Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging," *J. Med. Imaging Radiat. Sci.*, vol. 50, no. 4, pp. 477–487, Dec. 2019, doi: 10.1016/j.jmir.2019.09.005.

[83] Y. Si *et al.*, "Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review," *J. Biomed. Inform.*, vol. 115, p. 103671, Mar. 2021, doi: 10.1016/j.jbi.2020.103671.
[84] "Machine learning approach on healthcare big data: a review." https://www.aimspress.com/article/doi/10.3934/bdia.2020005?viewType=HTML (accessed Jun. 15, 2023).
[85] J. Vamathevan *et al.*, "Applications of machine learning in drug discovery and development," *Nat. Rev. Drug Discov.*, vol. 18, no. 6, pp. 463–477, Jun. 2019, doi: 10.1038/s41573-019-0024-5.
[86] F. Sabry, T. Eltaras, W. Labda, K. Alzoubi, and Q. Malluhi, "Machine Learning for Healthcare Wearable Devices: The Big Picture," *J. Healthc. Eng.*, vol. 2022, p. e4653923, Apr. 2022, doi: 10.1155/2022/4653923.
[87] S. Reddy, J. Fox, and M. P. Purohit, "Artificial intelligence-enabled healthcare delivery," *J. R. Soc. Med.*, vol. 112, no. 1, pp. 22–28, Jan. 2019, doi: 10.1177/0141076818815510.
[88] H. Hund, S. Gerth, D. Lossnitzer, and C. Fegeler, "Longitudinal Data Driven Study Design.," *Stud. Health Technol. Inform.*, vol. 205, pp. 373–377, Aug. 2014.
[89] C. Ponchiardi, M. Mauer, and B. Najafian, "Temporal Profile of Diabetic Nephropathy Pathologic Changes," *Curr. Diab. Rep.*, vol. 13, May 2013, doi: 10.1007/s11892-013-0395-7.
[90] M. C. Thomas *et al.*, "Diabetic kidney disease," *Nat. Rev. Dis. Primer*, vol. 1, no. 1, Art. no. 1, Jul. 2015, doi: 10.1038/nrdp.2015.18.
[91] N. Menachemi and T. H. Collum, "Benefits and drawbacks of electronic health record systems," *Risk Manag. Healthc. Policy*, vol. 4, pp. 47–55, Dec. 2011, doi: 10.2147/RMHP.S12985.
[92] F. Xie *et al.*, "Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies," *J. Biomed. Inform.*, vol. 126, p. 103980, Feb. 2022, doi: 10.1016/j.jbi.2021.103980.
[93] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk Prediction with Electronic Health Records: A Deep Learning Approach," in *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)*, in Proceedings. Society for Industrial and Applied Mathematics, 2016, pp. 432–440. doi: 10.1137/1.9781611974348.49.
[94] V. Berisha *et al.*, "Digital medicine and the curse of dimensionality," *Npj Digit. Med.*, vol. 4, no. 1, Art. no. 1, Oct. 2021, doi: 10.1038/s41746-021-00521-5.
[95] N. Garcelon, A. Burgun, R. Salomon, and A. Neuraz, "Electronic health records for the diagnosis of rare diseases," *Kidney Int.*, vol. 97, no. 4, pp. 676–686, Apr. 2020, doi: 10.1016/j.kint.2019.11.037.
[96] A. Guo, M. Pasque, F. Loh, D. L. Mann, and P. R. O. Payne, "Heart Failure Diagnosis, Readmission, and Mortality Prediction Using Machine Learning and Artificial Intelligence Models," *Curr. Epidemiol. Rep.*, vol. 7, no. 4, pp. 212–219, Dec. 2020, doi: 10.1007/s40471-020-00259-w.

[97] Y. Sun and D. Zhang, "Diagnosis and Analysis of Diabetic Retinopathy Based on Electronic Health Records," *IEEE Access*, vol. 7, pp. 86115–86120, 2019, doi: 10.1109/ACCESS.2019.2918625.

[98] Y. Sun and D. Zhang, "Machine Learning Techniques for Screening and Diagnosis of Diabetes: a Survey," *Teh. Vjesn.*, vol. 26, no. 3, pp. 872–880, Jun. 2019, doi: 10.17559/TV-20190421122826.

[99] R. G. Hauser *et al.*, "A Machine Learning Model to Successfully Predict Future Diagnosis of Chronic Myelogenous Leukemia With Retrospective Electronic Health Records Data," *Am. J. Clin. Pathol.*, vol. 156, no. 6, pp. 1142–1148, Dec. 2021, doi: 10.1093/ajcp/aqab086.

[100] E. Kogan *et al.*, "A machine learning approach to identifying patients with pulmonary hypertension using real-world electronic health records," *Int. J. Cardiol.*, vol. 374, pp. 95–99, Mar. 2023, doi: 10.1016/j.ijcard.2022.12.016.

[101] M. A. Myszczynska *et al.*, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nat. Rev. Neurol.*, vol. 16, no. 8, Art. no. 8, Aug. 2020, doi: 10.1038/s41582-020-0377-8.

[102] D. Zeiberg, T. Prahlad, B. K. Nallamothu, T. J. Iwashyna, J. Wiens, and M. W. Sjoding, "Machine learning for patient risk stratification for acute respiratory distress syndrome," *PLOS ONE*, vol. 14, no. 3, p. e0214465, Mar. 2019, doi: 10.1371/journal.pone.0214465.

[103] J. Yang *et al.*, "Machine Learning-Based Risk Stratification for Gestational Diabetes Management," *Sensors*, vol. 22, no. 13, Art. no. 13, Jan. 2022, doi: 10.3390/s22134805.

[104] Y. Hu *et al.*, "A Simpler Machine Learning Model for Acute Kidney Injury Risk Stratification in Hospitalized Patients," *J. Clin. Med.*, vol. 11, no. 19, Art. no. 19, Jan. 2022, doi: 10.3390/jcm11195688.

[105] N. Liu *et al.*, "Utilizing machine learning dimensionality reduction for risk stratification of chest pain patients in the emergency department," *BMC Med. Res. Methodol.*, vol. 21, no. 1, p. 74, Apr. 2021, doi: 10.1186/s12874-021-01265-2.

[106] O. Ben-Assuli *et al.*, "Stratifying individuals into non-alcoholic fatty liver disease risk levels using time series machine learning models," *J. Biomed. Inform.*, vol. 126, p. 103986, Feb. 2022, doi: 10.1016/j.jbi.2022.103986.

[107] F. Christopoulou, T. T. Tran, S. K. Sahu, M. Miwa, and S. Ananiadou, "Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 1, pp. 39–46, Jan. 2020, doi: 10.1093/jamia/ocz101.

[108] N. T. Issa, V. Stathias, S. Schürer, and S. Dakshanamurthy, "Machine and deep learning approaches for cancer drug repurposing," *Semin. Cancer Biol.*, vol. 68, pp. 132–142, Jan. 2021, doi: 10.1016/j.semcancer.2019.12.011.

[109] J. Chu, W. Dong, J. Wang, K. He, and Z. Huang, "Treatment effect prediction with adversarial deep learning using electronic health records," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 4, p. 139, Dec. 2020, doi: 10.1186/s12911-020-01151-9.

[110] Z. Xu *et al.*, "Using Machine Learning to Predict Antidepressant Treatment Outcome From Electronic Health Records," *Psychiatr. Res. Clin. Pract.*, p. n/a-n/a, Mar. 2023, doi: 10.1176/appi.prcp.20220015.

[111] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah, "Improving palliative care with deep learning," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 4, p. 122, Dec. 2018, doi: 10.1186/s12911-018-0677-8.

[112] S. Levin *et al.*, "Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay," *BMJ Innov.*, vol. 7, no. 2, Apr. 2021, doi: 10.1136/bmjinnov-2020-000420.

[113] J. Ebinger *et al.*, "A Machine Learning Algorithm Predicts Duration of hospitalization in COVID-19 patients," *Intell.-Based Med.*, vol. 5, p. 100035, Jan. 2021, doi: 10.1016/j.ibmed.2021.100035.

[114] M. Tello *et al.*, "Machine learning based forecast for the prediction of inpatient bed demand," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 55, Mar. 2022, doi: 10.1186/s12911-022-01787-9.

[115] "Journal of Diabetes & Metabolic Disorders," *Springer*. https://www.springer.com/journal/40200 (accessed Jun. 24, 2023).

[116] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, Jan. 2017, doi: 10.1016/j.csbj.2016.12.005.

[117] J. Wong, M. M. Horwitz, L. Zhou, and S. Toh, "Using machine learning to identify health outcomes from electronic health record data," *Curr. Epidemiol. Rep.*, vol. 5, no. 4, pp. 331–342, Dec. 2018, doi: 10.1007/s40471-018-0165-9.

[118] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, "Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses," *FASEB J.*, vol. 22, no. 2, pp. 338–342, 2008, doi: 10.1096/fj.07-9492LSF.

[119] X. Song, L. R. Waitman, Y. Hu, A. S. L. Yu, D. Robins, and M. Liu, "Robust clinical marker identification for diabetic kidney disease with ensemble feature selection," *J. Am. Med. Inform. Assoc.*, vol. 26, no. 3, pp. 242–253, Mar. 2019, doi: 10.1093/jamia/ocy165.

[120] P. Connolly *et al.*, "Analytical validation of a multi-biomarker algorithmic test for prediction of progressive kidney function decline in patients with early-stage kidney disease," *Clin. Proteomics*, vol. 18, no. 1, p. 26, Nov. 2021, doi: 10.1186/s12014-021-09332-y.

[121] V. Singh, V. K. Asari, and R. Rajasekaran, "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease," *Diagnostics*, vol. 12, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/diagnostics12010116.

[122] "Using Machine Learning to Predict Diabetes Complications | IEEE Conference Publication | IEEE Xplore." https://ieeexplore.ieee.org/document/9677649 (accessed Dec. 04, 2022).

[123] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "A Machine Learning Approach to Predicting Diabetes Complications," *Healthcare*, vol. 9, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/healthcare9121712.

[124] S. K. David, M. Rafiullah, and K. Siddiqui, "Comparison of Different Machine Learning Techniques to Predict Diabetic Kidney Disease," *J. Healthc. Eng.*, vol. 2022, p. e7378307, Apr. 2022, doi: 10.1155/2022/7378307.

[125] M. Zuo, W. Zhang, Q. Xu, and D. Chen, "Deep Personal Multitask Prediction of Diabetes Complication with Attentive Interactions Predicting Diabetes Complications by Multitask-Learning," *J. Healthc. Eng.*, vol. 2022, p. 5129125, 2022, doi: 10.1155/2022/5129125.

[126] B. P. Swan, M. E. Mayorga, and J. S. Ivy, "The SMART Framework: Selection of Machine Learning Algorithms With ReplicaTions—A Case Study on the Microvascular Complications of Diabetes," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 2, pp. 809–817, Feb. 2022, doi: 10.1109/JBHI.2021.3094777.

[127] P. Novitski, C. M. Cohen, A. Karasik, G. Hodik, and R. Moskovitch, "Temporal patterns selection for All-Cause Mortality prediction in T2D with ANNs," *J. Biomed. Inform.*, vol. 134, p. 104198, Oct. 2022, doi: 10.1016/j.jbi.2022.104198.

[128] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Sci. Rep.*, vol. 6, no. 1, Art. no. 1, May 2016, doi: 10.1038/srep26094.

[129] A. L. Neves *et al.*, "Using electronic health records to develop and validate a machine-learning tool to predict type 2 diabetes outcomes: a study protocol," *BMJ Open*, vol. 11, no. 7, p. e046716, Jul. 2021, doi: 10.1136/bmjopen-2020-046716.

[130] N. H. Chowdhury *et al.*, "Performance Analysis of Conventional Machine Learning Algorithms for Identification of Chronic Kidney Disease in Type 1 Diabetes Mellitus Patients," *Diagn. Basel Switz.*, vol. 11, no. 12, p. 2267, Dec. 2021, doi: 10.3390/diagnostics11122267.

[131] S. Ravizza *et al.*, "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data," *Nat. Med.*, vol. 25, no. 1, pp. 57–59, Jan. 2019, doi: 10.1038/s41591-018-0239-8.

[132] Y. Fan, E. Long, L. Cai, Q. Cao, X. Wu, and R. Tong, "Machine Learning Approaches to Predict Risks of Diabetic Complications and Poor Glycemic Control in Nonadherent Type 2 Diabetes," *Front. Pharmacol.*, vol. 12, p. 665951, Jun. 2021, doi: 10.3389/fphar.2021.665951.

[133] Q. Xu, L. Wang, and S. S. Sansgiry, "A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning," *J. Med. Artif. Intell.*, vol. 3, no. 0, Art. no. 0, Mar. 2020, doi: 10.21037/jmai.2019.10.04.

[134] T. Rahman, S. M. Farzana, and A. Z. Khanom, "Prediction of diabetes induced complications using different machine learning algorithms," Thesis,

BRAC University, 2018. Accessed: Dec. 04, 2022. [Online]. Available: http://dspace.bracu.ac.bd/xmlui/handle/10361/10945

[135] W. Jiang *et al.*, "Establishment and Validation of a Risk Prediction Model for Early Diabetic Kidney Disease Based on a Systematic Review and Meta-Analysis of 20 Cohorts," *Diabetes Care*, vol. 43, no. 4, pp. 925–933, Mar. 2020, doi: 10.2337/dc19-1897.

[136] C. M. Micheel *et al.*, *Omics-Based Clinical Discovery: Science, Technology, and Applications*. National Academies Press (US), 2012. Accessed: Jan. 07, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK202165/

[137] A. Holzinger, B. Haibe-Kains, and I. Jurisica, "Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 46, no. 13, pp. 2722–2730, Dec. 2019, doi: 10.1007/s00259-019-04382-9.

[138] N. Al-Sari *et al.*, "Precision diagnostic approach to predict 5-year risk for microvascular complications in type 1 diabetes," *eBioMedicine*, vol. 80, Jun. 2022, doi: 10.1016/j.ebiom.2022.104032.

[139] K. Jayawardana *et al.*, "Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information," *Int. J. Cancer*, vol. 136, no. 4, pp. 863–874, 2015, doi: 10.1002/ijc.29047.

[140] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer," *Clin. Cancer Res.*, vol. 24, no. 6, pp. 1248–1259, Mar. 2018, doi: 10.1158/1078-0432.CCR-17-0853.

[141] D. Tong *et al.*, "Improving prediction performance of colon cancer prognosis based on the integration of clinical and multi-omics data," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 22, Feb. 2020, doi: 10.1186/s12911-020-1043-1.

[142] R. De Bin, W. Sauerbrei, and A.-L. Boulesteix, "Investigating the prediction ability of survival models based on both clinical and omics data: two case studies," *Stat. Med.*, vol. 33, no. 30, pp. 5310–5329, 2014, doi: 10.1002/sim.6246.

[143] A. Cambiaghi, M. Ferrario, and M. Masseroli, "Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration," *Brief. Bioinform.*, vol. 18, no. 3, pp. 498–510, May 2017, doi: 10.1093/bib/bbw031.

[144] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," *J. Biomed. Inform.*, vol. 53, pp. 220–228, Feb. 2015, doi: 10.1016/j.jbi.2014.11.005.

[145] A. Dagliati *et al.*, "Machine Learning Methods to Predict Diabetes Complications," *J. Diabetes Sci. Technol.*, vol. 12, no. 2, pp. 295–302, Mar. 2018, doi: 10.1177/1932296817706375.

[146] M. Makino *et al.*, "Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning," *Sci. Rep.*, vol. 9, p. 11862, Aug. 2019, doi: 10.1038/s41598-019-48263-5.

[147] V. Rodriguez-Romero, R. F. Bergstrom, B. S. Decker, G. Lahu, M. Vakilynejad, and R. R. Bies, "Prediction of Nephropathy in Type 2 Diabetes: An Analysis of the ACCORD Trial Applying Machine Learning Techniques," *Clin. Transl. Sci.*, vol. 12, no. 5, pp. 519–528, 2019, doi: 10.1111/cts.12647.

[148] S. M. Hosseini Sarkhosh, M. Hemmatabadi, and A. Esteghamati, "Development and validation of a risk score for diabetic kidney disease prediction in type 2 diabetes patients: a machine learning approach," *J. Endocrinol. Invest.*, Sep. 2022, doi: 10.1007/s40618-022-01919-y.

[149] A. Aminian *et al.*, "Predicting 10-Year Risk of End-Organ Complications of Type 2 Diabetes With and Without Metabolic Surgery: A Machine Learning Approach," *Diabetes Care*, vol. 43, no. 4, pp. 852–859, Feb. 2020, doi: 10.2337/dc19-2057.

[150] X. Song, L. R. Waitman, A. S. Yu, D. C. Robbins, Y. Hu, and M. Liu, "Longitudinal Risk Prediction of Chronic Kidney Disease in Diabetic Patients Using a Temporal-Enhanced Gradient Boosting Machine: Retrospective Cohort Study," *JMIR Med. Inform.*, vol. 8, no. 1, p. e15510, Jan. 2020, doi: 10.2196/15510.

[151] L. Chan *et al.*, "Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease," *Diabetologia*, vol. 64, no. 7, pp. 1504–1515, Jul. 2021, doi: 10.1007/s00125-021-05444-0.

[152] A. Allen *et al.*, "Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus," *BMJ Open Diabetes Res. Care*, vol. 10, no. 1, p. e002560, Jan. 2022, doi: 10.1136/bmjdrc-2021-002560.

[153] Z. Dong *et al.*, "Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records," *J. Transl. Med.*, vol. 20, no. 1, p. 143, Mar. 2022, doi: 10.1186/s12967-022-03339-1.

[154] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018, doi: 10.1016/j.neucom.2017.11.077.

[155] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *WIREs Comput. Stat.*, vol. 11, no. 3, p. e1460, 2019, doi: 10.1002/wics.1460.

[156] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Jun. 13, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[157] N. Salkind, *Encyclopedia of Research Design*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2010. doi: 10.4135/9781412961288.

[158] X. Zhao, Y. Wu, D. L. Lee, and W. Cui, "iForest: Interpreting Random Forests via Visual Analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 407–416, Jan. 2019, doi: 10.1109/TVCG.2018.2864475.

[159] F. Cabitza *et al.*, "The importance of being external. methodological insights for the external validation of machine learning models in medicine," *Comput. Methods Programs Biomed.*, vol. 208, p. 106288, Sep. 2021, doi: 10.1016/j.cmpb.2021.106288.

[160] P. Delanaye, E. Cavalier, H. Pottel, and T. Stehlé, "New and old GFR equations: a European perspective," *Clin. Kidney J.*, p. sfad039, Mar. 2023, doi: 10.1093/ckj/sfad039.

[161] P. Cerda and G. Varoquaux, "Encoding High-Cardinality String Categorical Variables," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1164–1176, Mar. 2022, doi: 10.1109/TKDE.2020.2992529.

[162] F. Pargent, "A Benchmark Experiment on How to Encode Categorical Features in Predictive Modeling," Mar. 2019. Accessed: Jul. 05, 2023. [Online]. Available: https://www.semanticscholar.org/paper/A-Benchmark-Experiment-on-How-to-Encode-Categorical-Pargent/8e758a2136f17a0cd05a22c927ff3a0309b259be

[163] K. Singh and S. Upadhyaya, "Outlier Detection: Applications And Techniques," *Int. J. Comput. Sci. Issues*, vol. 9, Jan. 2012.

[164] X. Yang, W. Zhou, N. Shu, and H. Zhang, "A Fast and Efficient Local Outlier Detection in Data Streams," in *Proceedings of the 2019 International Conference on Image, Video and Signal Processing*, in IVSP '19. New York, NY, USA: Association for Computing Machinery, Feb. 2019, pp. 111–116. doi: 10.1145/3317640.3317653.

[165] P. R. Jones, "A note on detecting statistical outliers in psychophysical data," *Atten. Percept. Psychophys.*, vol. 81, no. 5, pp. 1189–1196, Jul. 2019, doi: 10.3758/s13414-019-01726-3.

[166] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, Feb. 2020, doi: 10.1007/s10462-019-09709-4.

[167] K. I. Mokhtar Wan Nor Arifin, Tengku Muhammad Hanis Tengku, *Chapter 13 Missing data | Data Analysis in Medicine and Health using R*. Accessed: Jul. 06, 2023. [Online]. Available: https://bookdown.org/drki_musa/dataanalysis/missing-data.html

[168] T. Pham-Gia and T. L. Hung, "The mean and median absolute deviations," *Math. Comput. Model.*, vol. 34, no. 7, pp. 921–936, Oct. 2001, doi: 10.1016/S0895-7177(01)00109-1.

[169] H. Kaur, H. S. Pannu, and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and

Solutions," *ACM Comput. Surv.*, vol. 52, no. 4, p. 79:1-79:36, Aug. 2019, doi: 10.1145/3343440.

[170] D. K. D. Sree and D. C. S. Bindu, "Data Analytics: Why Data Normalization," *Int. J. Eng. Technol.*, vol. 7, no. 4.6, Art. no. 4.6, Sep. 2018, doi: 10.14419/ijet.v7i4.6.20464.

[171] J. Wang, "Heart Failure Prediction with Machine Learning: A Comparative Study," *J. Phys. Conf. Ser.*, vol. 2031, no. 1, p. 012068, Sep. 2021, doi: 10.1088/1742-6596/2031/1/012068.

[172] U. Gain and V. Hotti, "Low-code AutoML-augmented Data Pipeline – A Review and Experiments," *J. Phys. Conf. Ser.*, vol. 1828, no. 1, p. 012015, Feb. 2021, doi: 10.1088/1742-6596/1828/1/012015.

[173] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation," *Dep. Tech. Rep. CS*, Feb. 2018, [Online]. Available: https://scholarworks.utep.edu/cs_techrep/1209

[174] S. N, S. G, and B. J M, "Data Wrangling and Data Leakage in Machine Learning for Healthcare." Rochester, NY, Aug. 08, 2018. Accessed: Jul. 08, 2023. [Online]. Available: https://papers.ssrn.com/abstract=3708142

[175] J. M. Kernbach and V. E. Staartjes, "Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II-Generalization and Overfitting," *Acta Neurochir. Suppl.*, vol. 134, pp. 15–21, 2022, doi: 10.1007/978-3-030-85292-4_3.

[176] P. Probst, B. Bischl, and A.-L. Boulesteix, "Tunability: Importance of Hyperparameters of Machine Learning Algorithms." arXiv, Oct. 22, 2018. doi: 10.48550/arXiv.1802.09596.

[177] D. Paper, "Scikit-Learn Classifier Tuning from Simple Training Sets," in *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*, D. Paper, Ed., Berkeley, CA: Apress, 2020, pp. 137–163. doi: 10.1007/978-1-4842-5373-1_5.

[178] D. Berrar, "Cross-Validation," 2018. doi: 10.1016/B978-0-12-809633-8.20349-X.

[179] A. Gupta, A. Anand, and Y. Hasija, "Recall-based Machine Learning approach for early detection of Cervical Cancer," *2021 6th Int. Conf. Converg. Technol. I2CT*, pp. 1–5, Apr. 2021, doi: 10.1109/I2CT51068.2021.9418099.

[180] A. Kaplan *et al.*, "Artificial Intelligence/Machine Learning in Respiratory Medicine and Potential Role in Asthma and COPD Diagnosis," *J. Allergy Clin. Immunol. Pract.*, vol. 9, no. 6, pp. 2255–2261, Jun. 2021, doi: 10.1016/j.jaip.2021.02.014.

[181] J. Tohka and M. van Gils, "Evaluation of machine learning algorithms for health and wellness applications: A tutorial," *Comput. Biol. Med.*, vol. 132, p. 104324, May 2021, doi: 10.1016/j.compbiomed.2021.104324.

[182] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/BF02295996.

[183] A. L. Edwards, "Note on the 'correction for continuity' in testing the significance of the difference between correlated proportions," *Psychometrika*, vol. 13, no. 3, pp. 185–187, Sep. 1948, doi: 10.1007/BF02289261.
[184] "McNemar Test - an overview | ScienceDirect Topics." https://www.sciencedirect.com/topics/medicine-and-dentistry/mcnemar-test (accessed Jul. 08, 2023).
[185] "STAT 479 -- Machine Learning (Fall 2018) - Dr. Raschka." https://pages.stat.wisc.edu/~sraschka/teaching/stat479-fs2018/ (accessed Jul. 08, 2023).
[186] C. Chen and H. Seo, "Prediction of rock mass class ahead of TBM excavation face by ML and DL algorithms with Bayesian TPE optimization and SHAP feature analysis," *Acta Geotech.*, vol. 18, no. 7, pp. 3825–3848, Jul. 2023, doi: 10.1007/s11440-022-01779-z.
[187] G. Zhang *et al.*, "A machine learning model based on ultrasound image features to assess the risk of sentinel lymph node metastasis in breast cancer patients: Applications of scikit-learn and SHAP," *Front. Oncol.*, vol. 12, 2022, Accessed: Jul. 08, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2022.944569
[188] "Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP - ScienceDirect." https://www.sciencedirect.com/science/article/pii/S0010482521006077 (accessed Jul. 08, 2023).
[189] Z. Chen, F. Xiao, F. Guo, and J. Yan, "Interpretable machine learning for building energy management: A state-of-the-art review," *Adv. Appl. Energy*, vol. 9, p. 100123, Jan. 2023, doi: 10.1016/j.adapen.2023.100123.
[190] G. Team, "Gradio." https://gradio.app (accessed Jul. 09, 2023).
[191] R. Atienza, "Gradio & Hugging Face for Rapid Deep Learning App Development," *Medium*, Mar. 14, 2022. https://medium.com/@rowel/gradio-hugging-face-for-rapid-deep-learning-app-development-709a78e7ccc0 (accessed Jul. 13, 2023).
[192] "beeswarm plot — SHAP latest documentation." https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html (accessed Jul. 12, 2023).
[193] K. Wagstaff, "Machine Learning that Matters." arXiv, Jun. 18, 2012. doi: 10.48550/arXiv.1206.4656.
[194] F. Cabitza *et al.*, "The importance of being external. methodological insights for the external validation of machine learning models in medicine," *Comput. Methods Programs Biomed.*, vol. 208, p. 106288, Sep. 2021, doi: 10.1016/j.cmpb.2021.106288.
[195] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," *Nat. Biomed. Eng.*, vol. 6, no. 12, Art. no. 12, Dec. 2022, doi: 10.1038/s41551-022-00898-y.

[196] S. Vollmer *et al.*, "Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness," *BMJ*, vol. 368, p. l6927, Mar. 2020, doi: 10.1136/bmj.l6927.

[197] J. Adler-Milstein, J. Everson, and S.-Y. D. Lee, "EHR Adoption and Hospital Performance: Time-Related Effects," *Health Serv. Res.*, vol. 50, no. 6, pp. 1751–1771, 2015, doi: 10.1111/1475-6773.12406.

[198] "Artificial Intelligence in Medicine | NEJM." https://www.nejm.org/doi/full/10.1056/NEJMe2206291 (accessed Jul. 17, 2023).

[199] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018, doi: 10.1109/JBHI.2017.2767063.

[200] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, "StageNet: Stage-Aware Neural Networks for Health Risk Prediction," in *Proceedings of The Web Conference 2020*, Apr. 2020, pp. 530–540. doi: 10.1145/3366423.3380136.

[201] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient Subtyping via Time-Aware LSTM Networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '17. New York, NY, USA: Association for Computing Machinery, Aug. 2017, pp. 65–74. doi: 10.1145/3097983.3097997.

[202] C. S. Lee and A. Y. Lee, "Clinical applications of continual learning machine learning," *Lancet Digit. Health*, vol. 2, no. 6, pp. e279–e281, Jun. 2020, doi: 10.1016/S2589-7500(20)30102-3.

[203] N. Liu *et al.*, "Utilizing machine learning dimensionality reduction for risk stratification of chest pain patients in the emergency department," *BMC Med. Res. Methodol.*, vol. 21, no. 1, p. 74, Apr. 2021, doi: 10.1186/s12874-021-01265-2.

[204] M. F. Kabir, T. Chen, and S. A. Ludwig, "A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction," *Healthc. Anal.*, vol. 3, p. 100125, Nov. 2023, doi: 10.1016/j.health.2022.100125.

[205] À. Hernàndez-Carnerero, M. Sànchez-Marrè, I. Mora-Jiménez, C. Soguero-Ruiz, S. Martínez-Agüero, and J. Álvarez-Rodríguez, "Dimensionality reduction and ensemble of LSTMs for antimicrobial resistance prediction," *Artif. Intell. Med.*, vol. 138, p. 102508, Apr. 2023, doi: 10.1016/j.artmed.2023.102508.

APPENDICES

Appendix A – Literature review paper

This chapter presents the article entitled "*Machine Learning techniques to predict the risk of developing diabetic nephropathy: a literature review*". This article was carried out within the framework of this dissertation with the aim of identifying the work carried out that applies a longitudinal ML approach on EHR data to predict the evolution of DN.

Machine Learning techniques to predict the risk of developing diabetic nephropathy: a literature review

F. Mesquita[1], J. Bernardino[1,2], J. Henriques[2], JF. Raposo[3], RT. Ribeiro[3], S. Paredes[1,2]

ABSTRACT

Purpose: Diabetes is a major public health challenge with widespread prevalence, often leading to complications such as Diabetic Nephropathy (DN) - a chronic condition that progressively impairs kidney function. Machine learning models can exploit the inherent temporal factor in clinical data to better predict the risk of developing DN faster and more accurately than traditional clinical models.
Methods: Three different databases were used for this literature review: Scopus, Web of Science, and PubMed. Only articles written in English and published between January 2015 and December 2022 were included.
Results: We included 11 studies, from which we discuss a number of algorithms capable of extracting knowledge from clinical data, incorporating dynamic aspects in patient assessment, and exploring their evolution over time. We also present a comparison of the different approaches, their performance, advantages, disadvantages, interpretation, and the value that the time factor can bring to a more successful prediction of diabetic nephropathy.
Conclusion: Our analysis showed that some studies ignored the temporal factor, while others partially exploited it. Greater use of the temporal aspect inherent in Electronic Health Records (EHR) data, together with the integration of omics data, could lead to the development of more reliable and powerful predictive models.

Keywords
Diabetic nephropathy, kidney disease, clinical data, risk prediction, machine learning.

I. INTRODUCTION

The widespread prevalence of diabetes is still a major public health challenge, with a significant impact on people's quality of life and an increase in mortality. Between 1980 and 2014, the number of people with diabetes increased almost fourfold, from 108 million to 422 million, according to the World Health Organization [1]. In the European scenario, 6.2% of adults had diabetes in 2019. Cyprus, Portugal, and Germany were the countries with the highest levels, around 9% or more [2]. In addition, the metabolic control needed to delay diabetes complications is not achieved by the majority of patients. As a result, diabetes can cause many complications, including eye problems (retinopathy), nerve damage (neuropathy), and kidney problems (nephropathy) [3].

Diabetic Nephropathy (DN) is a chronic disease in which the function of the kidneys deteriorates, reducing their ability to eliminate wastes and toxins from the bloodstream and affecting the water balance in the body. DN is considered a progressive disease that usually gets worse over time until the kidneys can no longer function on their own, which is known as end-stage renal disease (ESRD) [4]. It is a disease that is usually considered irreversible although it has been observed that with long-term normalization of the diabetic environment, the architecture of the kidney can undergo significant remodelling and the lesions associated with diabetic nephropathy can be reversed [5]. In developed countries, half of all ESRD cases are due to DN, and the cost of treating ESRD patients is very high [6].

Digitalization has allowed hospitals to store the complete history of patient appointments in a database, resulting in the availability of EHRs. These data are longitudinal because they are collected over time and include multiple patient records at different points in time. Due to the progressive nature of many diseases, a longitudinal approach is usually required to fully assess their development and impact [7]. Given the chronic and long-term nature of diseases such as DN, it is crucial to consider the temporal dimension of patient data and not overlook its importance [8]. The timely implementation of a DN risk assessment may delay or even prevent its progression, which would certainly reduce the number of people with ESRD [9].

The dream of machines that can one day be self-learning without explicit programming is an old one [10]. Machine learning (ML) has its roots in the Artificial Intelligence (AI) movement of the 1950s, with a strong emphasis on practical goals and applications, focusing on tasks such as prediction and optimization [11]. In very simple terms, ML uses various algorithms to learn the patterns and relationships present in a dataset and ultimately predict an outcome. We are now experiencing a major and rapid transformation, brought about by significant advances in ML, which is exponentially increasing automation in many areas of society [12].

ML applied to medicine has great potential to support diagnosis by using a significant amount of patient data and

1 Polytechnic Institute of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes - Quinta da Nora, 3030 - 199 Coimbra, Portugal{a2018056868@isec.pt, sparedes@isec.pt, jorge@isec.pt}
2 Center for Informatics and Systems of University of Coimbra, University of Coimbra, Pólo II, 3030-290 Coimbra, Portugal{jorge@isec.pt; jh@dei.uc.pt; sparedes@isec.pt}
3 Education and Research Center, APDP Diabetes Portugal, Rua do Salitre 118-120, 1250-203 Lisboa, Portugal. {rogerio.ribeiro@apdp.pt; filipe.raposo@apdp.pt}

processing it in a fast and intelligent way, helping physicians to make more informed decisions [13]. In fact, ML algorithms can potentially play a crucial role in a faster and more reliable way to diagnose complications associated with diabetes such as DN [14]. The application of ML techniques to analyze EHR data can provide valuable insights and enable the development of ML models that can predict the risk of developing DN or progressing to higher stages, aiding physicians in the diagnosis and ultimately improving the quality of healthcare [15], [16].

There are many studies done on the use of ML to identify cases of diabetic nephropathy. However, the focus of this research is to identify and study the approaches used on clinical EHR data collected over a period of time and the corresponding risk prediction of developing diabetic nephropathy.

This work aims to answer the following research question:

RQ: What are the most effective machine learning techniques used to construct a model that uses the temporal information in diabetic patients' EHR data to predict the development of DN or progression to higher stages?

This literature review was done in a systematic way to ensure that the results are transparent and reproducible, minimizing the bias that would result from the specific choice of studies (cherry-picking) [17].

The main contributions of this work are the following:

- We present and compare different temporal approaches used in clinical data to develop a predictive model that can accurately identify the risk of developing DN or progressing to higher stages in the future. By providing a comprehensive overview of these approaches, we aim to encourage the development of effective predictive models that can help physicians improve patient outcomes.
- We contribute to the understanding of the impact that the temporal factor can have on the prediction of DN by reviewing and comparing static and dynamic approaches.
- We identify the limitations of static and dynamic approaches and highlight the need for further research to improve the accuracy of risk prediction.
- We show that it is already possible to see that the integration of omics data can potentially improve the results and increase the credibility of predicting DN risk.

The remainder of this paper is organized as follows. Section II describes the methodology used to select the articles to be reviewed. Section III presents the results obtained. A discussion of the main findings arising from these results is presented in Section IV. Threats to the validity of this literature review are presented in Section V, while possible future research directions are outlined in Section VI. Finally, Section VII presents the main conclusions.

II. MATERIALS AND METHODS

Three databases were used for this literature review: Scopus, Web of Science, and PubMed. These are three of the most popular and reliable sources of scientific information [18]. Only articles written in English and published between January 2015 and December 2022 were included. The search query used was:

"((diabetes) AND ((machine learning) OR (deep learning)) AND ((time) OR (temporal) OR (time series)) AND (predict) AND ((kidney disease) OR (nephropathy)))".

Figure 1 describes the methodology used throughout the process. The first step (Identification) resulted in a total of 164 papers. Based on the references of some of these papers, a further 11 were identified as potentially important, resulting in 175 papers for further analysis. These 11 additional articles were referenced by papers identified in the first stage. During the screening phase, 48 duplicates were removed. In addition, 85 papers were excluded by title and 14 by abstract. These were removed because they did not relate to the intended topic; this phase reduced the original 175 to 28 papers. Of these, only 11 were eligible according to the various criteria defined.

Table 1 shows a summary of the excluded articles, the criteria, and a brief explanation of the exclusion criteria.

It should be noted that although the keyword "deep learning" was included in the search query, none of the 11 selected papers used Deep Learning (DL) techniques to solve the problem. With this in mind, we will focus only on approaches that use ML algorithms.

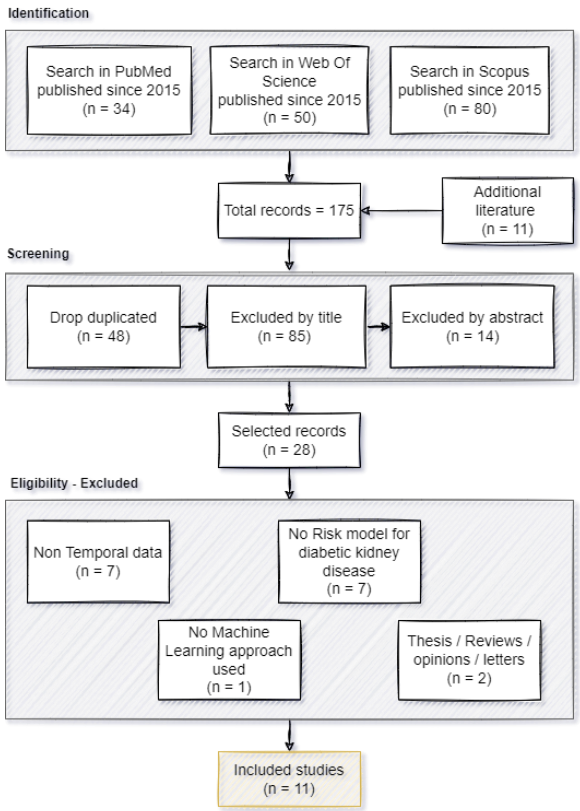

Figure 1: Methodology

TABLE 1 PAPERS EXCLUDED ACCORDING TO DEFINED CRITERIA.

Papers	Criteria	Brief Explanation
[19]–[25]	Non-Temporal Data	Excluded papers did not include temporal data, i.e. data from patients followed up during a specific time window with information collected during that time.
[26]–[32]	No Risk model for DN.	We select articles that predict the risk of progressing or developing DN. Articles that only classify whether patients have the disease or not were excluded.
[33], [34]	Thesis / Reviews / Opinions / letters	As these papers are reviews of the literature, this type of paper is not included.
[35]	No ML approach	This paper has used a scoring system that defines the factors that contribute most to the development of DN. Although it is a risk model, it is not an ML approach.

III. RESULTS

Following the procedure outlined in Figure 1, 11 articles were included in this review. Artificial intelligence applied to temporal clinical data has the potential to improve the way a diabetic patient is managed according to their risk of developing DN. The different approaches are presented according to different questions: i) which features are most important, ii) what kind of ML models have been created, iii) which ones perform better, and iv) other relevant aspects. The papers selected for this review, together with a summary of their main aspects, are listed in Table 2. Looking at Table 2, we can see that most of the articles were published in the last 2-3 years, which shows a rapid growth in the application of ML to the management of diabetes-related conditions, taking advantage of the large amount of clinical data available. For the selected articles, information is provided on the source of the data, the importance of the variables for prediction, the approaches used to create the risk models, their interpretation methods, and their performance.

A. Data Sources

With the emergence and growth of available data, ML models have increased the predictive potential in a wide range of tasks in several application areas. With digitalization, all patient's data is stored in computer databases. In fact, Electronic Health Records (EHRs) contain vital information about the patient, such as their medical history, illnesses, medications, treatment plans, allergies, and other highly relevant information. This type of data helps clinical research enormously by making it easier to access and track patient data [47]. It also allows for temporal and longitudinal analysis of the data, leading to different approaches and more accurate and correct predictive capabilities [48].

In addition to clinical variables, Omics-based biomarkers are often used. These can be defined as a molecular signature that is identified using omics data and used to predict the presence or risk of a particular disease or condition, or to monitor the response to a particular treatment. Omics can be divided into different research areas such as proteomics (proteins), transcriptomics (RNA), genomics (genes), metabolomics (metabolites), lipidomics (lipids) and epigenomics (methylated DNA) [49].

The integration of omics data with clinical data can significantly improve the ability to analyze and predict complex diseases using ML. Such integrated analysis can help create models that can clearly explain diseases, enabling real knowledge that leads to improved treatment and a better quality of life for patients [50]. The work of Al-Sari et al. [46] is a very good example of the benefits of combining Omics data with clinical data. The performance of some of the models, which had previously been built using only clinical data, increased significantly when Omics data were included. In this case, metabolites, ketones, and sugar derivatives were used. In general, the integration of molecular data will lead to better prognostic models, as demonstrated in several works [51]–[54]. Despite the many benefits of integrating this type of data, there are some challenges. Sometimes, even when these data are available, they are very difficult to handle, process, analyze, and finally integrate. This requires specialized knowledge in the branches of mathematics, statistics, biology, and computer science [55].

B. Feature Importance

There are a number of factors that can lead to the onset or development of DN, such as demographic and genetic factors, clinical measurements, laboratory tests, and medical history. Most of the selected studies used different methods to understand which variables had the greatest influence on the final outcome when predicting risk. Some of these techniques were used to perform feature selection to remove redundant and irrelevant variables, which can potentially lead to better performance [56].

The work of Chan et al. [43] and Al-Sari et al. [46] used SHapley Additive exPlanations (SHAP) to understand how each feature contributes to the model's predictions, by estimating the amount that each variable contributes to the predicted value of an output. This allows them to ensure that they are selecting the most optimal set of variables for the task.

Recursive Feature Elimination (RFE) is an iterative method that can recursively remove the least important features from a dataset and build a model on the remaining attributes. It iterates until the desired number of features is obtained. As presented in Sarkosh et al. [40] and Dong et al. [45], this technique is very useful for selecting a subset of features that aggregates the most important features from a larger dimensional space. In both cases, a variant of this method, Recursive Feature Elimination with Cross-Validation (RFECV), is applied. It uses cross-validation to evaluate the performance of the model at each iteration.

TABLE *2*: SUMMARY OF STUDIES INCLUDED IN THIS REVIEW.

Paper	Dataset	Pre-processing	ML Model Proposed	Performance
Singh et al. (2015) [36]	EHR data of patients in the Mount Sinai Hospital and Mount Sinai Faculty Practice Associates in New York City. From 6,435 patients, 12,337 examples were extracted.	Feature selection and generation. Numerical predictors discretization into four bins based on the quartiles of the corresponding predictor and then map them into binary variables.	Multitask Logistic Regression (MLTR)	≈ 68.3% AUROC for Threshold of 10% ≈ 71.2% AUROC for Threshold of 20%
Dagliati et al. (2018) [37]	943 T2DM patients in charge of the ICSM hospital and followed for more than 10 years.	Data imputation with the MissForest technique and some variables were not considered because imputation errors were too high.	Logistic Regression (LR)	3 years: 70.1% AUC 5 years: 73.4% AUC 7 years: 72.1% AUC
Makino et al. (2019) [38]	Dataset with 64,059 T2DM patients. From that, authors extracted structural, text, and longitudinal data.	Under sampling minority class, several data transformation steps are used to summarize the last 180 days EMR records and create longitudinal data variables.	Logistic Regression (LR)	AUC: 74.3%
Romero et al. (2019) [39]	Data were provided by the NHLBI, sponsor of the ACCORD trial. There were 10,251 T2DM patients from 77 clinical centers in the United States and Canada.	SMOTE technique used to balance target, feature selection using the information gain metric.	Random Forest	88.7 % Accuracy
Sarkosh et al. (2020) [40]	Clinic of Imam Khomeini Hospital Complex (IKHC) dataset with 10,636 T2DM patients followed from 10 years (2012-2021).	Feature selection using Recursive Feature, elimination (RFECV) and RF method, imputation or drop missing values	Logistic Regression (LR)	75.5% AUC
Aminian et al. (2020) [41]	287,438 T2DM patients from Cleveland Clinic's EHRs followed between 1998 and 2017. Two different groups were created: 2,287 patients undergoing metabolic surgery and 11,435 matched non-surgical patients.	Missing data imputed using multivariate imputation by chained equations (MICE), variables with more than 25% missing values or no predictive value were removed.	Random Forest	Surgical patients: 73% AUROC Nonsurgical patients: 76% AUROC
Song et al. (2020) [42]	University of Kansas Medical Center's HERON clinical data repository with 35,779 T2DM patients.	Features with less than 1% representation were removed and missing values imputed.	Gradient Boosting Machine (GBM)	AUROC: 83%, 78% and 82% in predict DN in 2, 3 and 4 years, respectively.
Chan et al. (2021) [43]	BioMe Biobank at the Icahn School of Medicine at Mount Sinai and the Penn Medicine Biobank data sources. Population of 1146 T2DM with both EHR data and biomarkers.	Data harmonization, only variables in more than 70% participants were included, feature selection based on SHapley Additive exPlanations (SHAP) values, and missing data imputation.	Random Forest	AUC: 77%
Allen et al. (2022) [44]	111,046 EHRs of T2DM patients that represents more than 700 healthcare sites from USA between 2007 and 2020.	Standardization, impute missing values.	Random Forest.	74.8% AUROC for any DN stage, 82.3% for stage 3-5 82.1% for stage 4-5
Dong et al. (2022) [45]	Data from PLA General Hospital with 2809 T2DM patients that were followed from 2008 to 2019.	Drop features with missing data > 25%, missing values imputation with RF, feature selection using RFE.	LightGBM	AUC: 81.15%
Al-Sari et al. (2022) [46]	T1D cohort in Steno Diabetes Center Copenhagen (SDCC) with 537 patients with follow-up data. Later, blood molecular data with 965 features was also included.	Remove high correlated features, outliers, and clinical variables with no predictive power on metabolic data. Feature selection SHapley Additive exPlanations (SHAP) values.	Random Forest	DN model with only clinical data: 92% of AUROC, DN model with clinical and omics: 99% AUROC

A very similar approach was adopted by Makino et al. [38] and Dagliatti et al. [37] with their logistic regression (LR) stepwise feature selection method based on the Akaike information criterion (MLC). Stepwise feature selection is a method of selecting a subset of features by iteratively adding or removing variables. The MLC is a trade-off between model goodness and complexity, and measures the relative quality of a statistical model [57]. It can be used in stepwise feature selection to evaluate the performance of the model at each step and decide which feature to add or to remove. Although it appears similar to the RFE method, this technique trains on the selected subset of features at each step and can use either forward selection or backward elimination, whereas RFE trains on all features and removes the least important feature at each step.

Aminian et al. [41] computed the relative importance of each feature in the final model using MLC for the regression models and the Concordance index (C-Index) for the RF models. The C-Index is a metric that considers the temporal dependence associated with the model result and can be used to rank features by importance or even to analyze the global performance of the model.

Singh et al. [36] use a simpler and faster approach based on Univariate feature selection to select the most relevant

variables. These features are selected based on univariate statistical tests between the feature and the target variable, and do not take into account dependencies and relations between features.

Song et al. [42] adopted a slightly different approach, using the Gradient GBM classifier because it uses an embedded method of feature selection during model training. This allows the most important features to be selected and the model retrained using only these variables.

Table 3 shows the clinical variables that were mentioned in more than three papers as one of the most prominent variables able to give high predictive power to the model for analyzing the emergence or development of DN, and their respective importance. Two of the reviewed articles include omics data and select the molecular variables that contribute the most to the outcome of the model, thus increasing its predictive capacity.

Table 4 details the three plasma biomarkers selected by Chan et al., while Table 5 shows the five molecular variables selected by Al-Sari, (2 ketones and 3 sugar derivatives).

Table 3: MOST IMPORTANT CLINICAL VARIABLES IDENTIFIED

Papers	Feature	Meaning
[39], [43], [45], [46]	eGFR or GFR	Glomerular filtration rate (GFR) measures how well the kidneys work. eGFR is an estimate, usually calculated using the Modification of Diet in Renal Disease (MDRD) equation and the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation.
[39], [43], [45], [46]	UAlb or Alb	Albumin levels in the blood. Low levels of this protein are called hypoalbuminemia, and high levels are known as hyperalbuminemia
[37], [40], [45], [46]	HbA1c	Glycated hemoglobin (HbA1c) measures glucose levels over the past 2 to 3 months.
[39]–[41], [43]	UACR or ACR	Laboratory tests are used to detect proteinuria, the presence of protein (usually albumin) in the urine.
[40]–[42], [45]	Age	In some articles, it is the age of the patient, in others it is the age at which the patient started to be followed.
[37], [40], [41], [45]	BMI	Body Mass Index uses a person's height and weight to calculate an estimate of body fat.

Table 4: MOST IMPORTANT OMICS IDENTIFIED BY CHAN ET AL. *[43]*

Molecular feature	Meaning
TNFR1	Tumor necrosis factor receptor 1 is a protein found on the surface of cells that binds to TNF (tumor necrosis factor), a signaling molecule involved in inflammation and cell death [58].
TNFR2	Protein related in structure and function to the TNFR1 protein and also related to TNF, which plays a role in inflammation and cell death [58].
KIM1	Kidney injury molecule 1 is a protein produced in the kidney that is considered a biomarker of acute kidney injury and plays a role in the repair and regeneration of kidney cells [59].

Table 5: MOST IMPORTANT OMICS IDENTIFIED BY AL-SARI ET AL. *[46]*

Molecular feature	Meaning
3,4 dihydroxybutanoic acid	Chemical compounds are found in many foods and also produced by the human body as a byproduct of some amino acids.
2,4 dihydroxybutanoic acid	Also, a chemical compound like 3,4-dihydroxybutanoic acid with only small molecular differences.
ribitol	It is a five-carbon sugar alcohol used as sweetener. Naturally occurring compound found in small amounts on fruit and vegetables.
ribonic acid	Also found in small amounts on fruit and vegetables, but it is also a metabolic pathway intermediate and a byproduct of xylose fermentation.
myo-inositol	Six-carbon cyclic sugar alcohol. A naturally occurring compound found in some foods, particularly fruits and nuts. It is also produced by the human body as a byproduct of glucose metabolism.

C. Risk Models

This section systematizes several approaches to building a model that can predict the risk of developing diabetic nephropathy. Some approaches do not fully exploit the time factor inherent in the data (static approaches), while others manage to make better use of this factor (dynamic/temporal approaches).

1) Static approaches

Dong et al. [45] used data from non-DN patients at baseline who were followed for three years. The authors then used 408 patients who remained without DN and 408 patients who developed DN after the follow-up period. This data was used to build the model, it contains all the characteristics that the patient presented at baseline and the variable to predict is whether or not they developed the disease after the three years of follow-up. The patients were divided into training and test sets with the size of 652 and 164, respectively. Binary classification was performed using seven different ML classifiers: Light gradient boosting machine (LightGBM), eXtreme gradient boosting (XGBoost), Adaptive boosting (AdaBoost), Artificial Neural Networks (ANNs), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR). This binary classification predicts the presence or absence of DN within 3 years.

There are several other papers that have taken a similar approach and transformed the problem into a binary classification. Romero et al. [39] followed a similar strategy, but defined eight different time windows for all the 7 years of patient follow-up data. Each window corresponds to one year of data, except for the first two windows, which correspond to only 6 months each. The tree-based classifiers OneRule, J48, and RF were chosen for their simplicity, speed of classification, and user-friendly graphical presentation.

Dagliatti et al. [37] used a binary outcome variable but for three different time thresholds of 3, 5, and 7 years to predict the risk of DN. Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machines (SVMs), and Random Forest (RF) were tested.

Aminian et al. [41] used data from both surgical and non-surgical patients with T2DM. Multivariate time-to-event regression and random forest machine learning models were created to predict the 10-year risk of developing DN. The 10-year risk of morbidity and mortality was estimated for patients with and without metabolic surgery.

Sarkosh et al. [40] trained an LR-based risk score in 1907 diabetic patients, of whom 763 developed DN within five years. The outcome variable was also binary, as in the papers cited above. The authors used multivariate LR analysis to generate risk scores and divided patients into four different groups based on their respective risk of DN: low, moderate, high, and very high.

Chan et al. [43] used the same binary outcome in a train/test set of 686 patients and a validation test of 460 patients. Using clinical data and biomarkers, the authors generated risk probabilities using the final RF model and scaled the results to a continuous score between 5 and 100. The authors named the whole system IntelKidneyX. It stratified patients as follows: low risk (46%), intermediate risk (37%) and high risk (17%) of developing DN within 5 years.

Al-Sari et al. [46] and Makino et al. [38] did almost the same as the previously cited papers, but instead of defining outcome as absence or presence, it was defined as progressor or non-progressor in the Al-Sari paper and as worsening or stable in the Makino et al paper. Al-Sari et al. used data from 190 patients who had no progression of DN and 190 patients who had progression of DN during a mean follow-up of 5.4 years. He used the RF classifier to predict whether the patient would progress to DN during the follow-up period. On the other hand, Makino et al. extracted clinical features from longitudinal, textual, and structural data. LR models were trained using data from 15,422 stable patients (remaining DN stage 1) and 15,388 patients who experienced disease progression at some point (from DN stage 1 to DN stage 2-5).

Unlike the works presented above, Allen et al. [44] are able to predict 3 different outcomes, DN progression to any stage, DN progression to stages 3-5, and DN progression to stages 4-5. Three different models were created for each possible outcome, each predicting the risk of progression to DN over the next 5 years. RF and XGBoost were used as classifiers with a training and test set of 62,994 and 7,656, respectively.

Figure 2 provides a general overview of the different approaches described above.

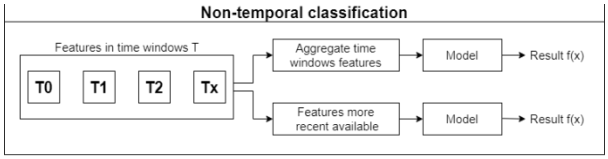

Figure 2: Non-temporal approaches.

2) Dynamic approaches

Different temporal approaches have been proposed to deal with EHR and provide risk prediction for DN. Within the remaining selected articles, the following approaches were used: stacked temporal, multitask temporal, discrete survival, and landmark boosting.

The stacked temporal technique was used in both the work of Singh et al. [36] and Song et al. [42] work. It aggregates the data within each time window and uses the data from all time windows to make a final, unique prediction. T time windows, with F features in each, result in only one time window with T multiplied by F features. One of the disadvantages of this technique is that the larger the temporal space considered, the higher the dimensionality of the data, which can lead to a large overfitting. In Figure 3, the physician appointments within each time window are aggregated to form a one-dimensional space, which is then fed into the model and a prediction is obtained.

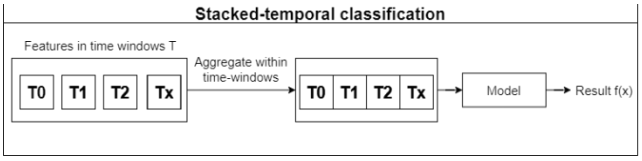

Figure 3: Stacked temporal approach.

The multitask temporal method was proposed in the paper by Sing et al. The authors decided to predict the outcome for each time window separately. Each time window must have at least five physician appointments within that time. When predicting the risk of DN for a new patient, each time window with five or more appointments is used and the final result is the average of the different results obtained in each time window. This stratification of the problem is shown in Figure 4, where it can be seen that the ML model operates independently in each time window and the result is the average of the different results obtained.

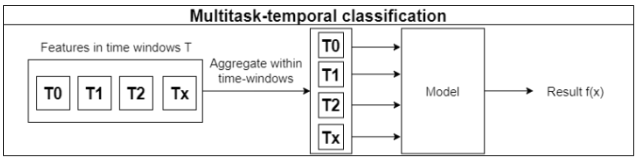

Figure 4: Multitask temporal approach.

Discrete survival and landmark boosting are two techniques mentioned in the paper by Song et al. The first makes an individual prediction in each time window, with no overlap between windows. A disadvantage of this technique is that it assumes that there is no relationship between examples in different time windows, even if they come from the same patient. This can be seen in Figure 5.

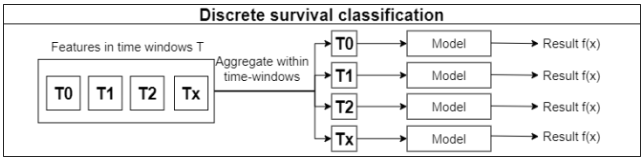

Figure 5: Discrete survival approach.

On the other hand, landmark boosting is very similar to discrete survival, but in each time window *t*, the prediction made in the previous time window $t-1$ is also considered. In effect, there is a transfer of knowledge between the time windows, making each prediction more accurate. This can be seen in the representation of the approach shown in Figure 6, where each model receives not only the features corresponding to a time window, but also the prediction made in the previous time window.

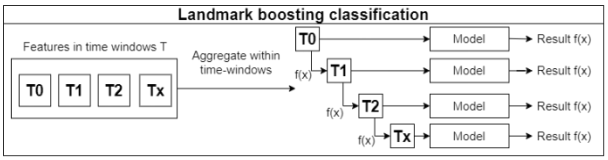

Figure 6: Landmark boosting classification.

D. Used models, interpretation, and performance.

This section discusses the type of models most commonly used to predict the onset or development of DN. It also presents the main interpretation techniques used and a comparison of performance.

Taking into account the selected papers, five different classifiers were proposed: Random Forest (RF), Logistic Regression (LR), LightGBM, GBM, and Multi-Task Logistic Regression (MTLR). From Figure 7, we can see that the method most selected was RF, followed by LR, and finally LightGBM, GBM, and MTLR, which were selected only once.

Performance is the most important individual factor that defines the classifier, but it is not the only aspect to consider. RF was the most used classifier because the decision trees that make it up can be interpreted and the final result can be explained [44]. However, as a whole, these methods are often difficult to interpret, especially when the number of decision trees is large. It has a good classification speed and can be represented graphically [39]. It is therefore a classifier with a good balance between speed, complexity, and interpretability. Logistic regression has also been proposed several times because it provides a clear interpretation of its coefficients, which are usually represented graphically by nomograms, concepts with which physicians are very familiar [37], [60]. GBM was chosen by Song et al [42] because of its robustness and effectiveness in predicting DN risk, as demonstrated in previous work. In addition, it incorporates feature selection. Multitask logistic regression was proposed by Singh et al [36] because it was appropriate for the type of solution proposed in their multitask temporal methodology. It consists of a multitask learning approach where learning is performed in parallel, and tasks are related to each other [61]. In this case, there is a learning task for each time window, and this approach is used to capture the dependency between tasks.

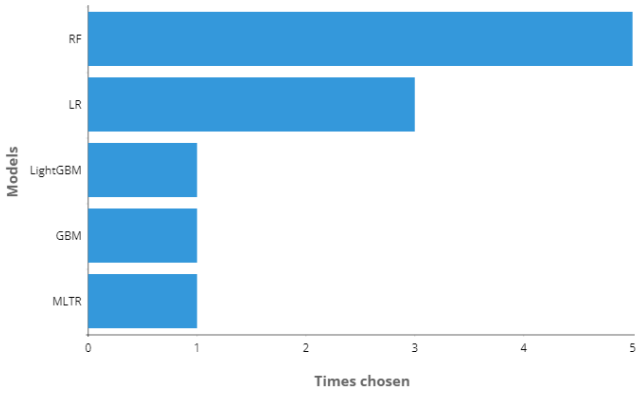

Figure 7: Most used ML classifiers in proposed methods

It is possible to identify three main techniques to interpret the results generated by the predictive model: i) SHapley Additive exPlanations - SHAP values, ii) monograms, iii) decision tree visualization. SHAP values were proposed by Lundberg et al. in 2017 to analyze model predictions [62]. It calculates the importance of each feature for a given prediction, where each feature can have a positive or negative impact on that specific prediction. The contribution of features can be local (each observation) or global (set of observations). In this particular case, local explanations aim to show the reasons that lead to a certain result generated by the model for a specific patient. On the other hand, global explanations aim to show which variables were most important for the overall predictions of the model. These are calculated by aggregating the different local explanations. Nomograms are graphical representations of LR models. They work like scoring systems, where each feature is assigned a certain number of points according to its value, and the result varies according to the number of points accumulated in the sum of the different features [63]. Finally, some of the articles used only tree-based models because they can be interpreted directly by visual inspection of the associated decision tree. RF is an ensemble of many independent trees, and the output is based on the multiple decision tree outputs. By looking at the different decision trees, it is possible to see which features are used to make predictions, the importance of each feature, and the overall patterns of predictions [64].

Some papers predict the onset of DN, some predict the worsening, and some authors predict the worsening for specific stages of the disease. In addition, there are papers where the result corresponds to only one specific time window, while others implement a different prediction for each time window, taking into account a certain number of years. This heterogeneity makes it difficult to compare their performance directly. *Table 6* provides detailed information on each of the proposed methods.

Table 6: DETAILS AND PERFORMANCE OF PROPOSED METHODS

Proposed method	Time range	Outcome variable	Performance metrics
Random Forest [44]	5 Years	Multiclass (DN advance to any stage, DN advance to stage 3-5, and DN advance to stage 4-5)	Any stage - AUROC: 0.748, Sensitivity: 0.7, Specificity: 0.662. DN stage 3-5 - AUROC: 0.823, Sensitivity: 0.750, Specificity: 0.739. DN stage 4-5 - AUROC: 0.821, Sensitivity: 0.751, Specificity: 0.712
Random Forest [41]	10 Years	Binary target (morbidity and mortality risks)	AUC: 0.76
Random Forest [43]	5 Years	Binary	AUC: 0.77
Logistic Regression [37]	3, 5 and 7 years	Binary	3 years - Accuracy: 0.647, Sensitivity: 0.820, Specificity: 0.730 and AUC: 0.808. 5 years - Accuracy: 0.693, Sensitivity: 0.750, Specificity: 0.616, AUC: 0.734. 7 years - Accuracy: 0.686, Sensitivity: 0.714, Specificity: 0.643, AUC: 0.721.
LightGBM [45]	3 years	Binary (DN presence or absence)	Accuracy: 0.768, Sensitivity: 0.741, Specificity: 0.797 and AUC: 0.815.
Logistic Regression [38]	6 months	Binary (DN stable or aggravation)	Accuracy: 0.701 AUC: 0.743
Random Forest [46]	Non-defined	Binary (DN progression or no progression)	Accuracy: 0.96 AUC: 0.96
Random Forest [39]	8 time windows at a max of 7 years	Binary on each time window	Average Acc: 0.887
Logistic Regression [40]	5 years	Binary (DN presence or absence)	AUC: 0.758
Multitask Logistic Regression [36]	5 years with time windows of 6 months	Binary on each time window	≈ 68.3% for Threshold of 10% ≈ 71.2% for Threshold of 20%
GBM [42]	2, 3 and 4 years	Binary on each time window	2 years - AUROC: 0.830 3 years - AUROC: 0.780 4 years - AUROC: 0.820

IV. DISCUSSION

To the best of our knowledge, this is the first review in the literature to explore what is available on the use of EHR data from patients followed over a period of time to create a predictive risk DN model also for a defined period of time.

This paper can be used as a basis for further work that will be able to analyze and take full advantage of the time factor and create a system with a high predictive capacity, with full knowledge of the patient's history and what is likely to be their future.

There are several approaches in the literature for handling EHR data that are collected over time and then used to build a model to predict the risk of the onset / development of diabetic nephropathy within a given time period. This is a very heterogeneous area of research, where there is no well-defined approach to achieving the previous goal. As Fletcher points out, *heterogeneity can be, and usually is, a good thing and can be beneficial* [65].

The main findings that have emerged from this work are as follows:

- There is very little work that takes full advantage of the time factor inherent in EHR data. The works of Sing et al. [36] and Song et al. [42] are an exception. In fact, the landmark boosting method proposed in the Song et al. paper was the approach that took more advantage of the time factor. It not only predicts the risk in each time window, but also takes into account the result produced in the previous time window. Although this approach attempts to exploit the full temporal potential of EHR data, it could still be improved, as it considers all records as independent, but in fact, they are not because the patient has multiple records (appointments).
- Combining omics data with clinical data can help better predict the risk of DN over time, as confirmed in the work of Al-sari [46]. In the near future, this type of data will be linked to disease risk models because the information they contain is really valuable to increase the predictive power of the different risk models.
- Another important concern with clinical risk models is interpretability. Almost all of the proposed models were selected not only because of their good performance but also because they allow interpretation of the respective results.
- The vast majority of the selected articles were published recently (within the last 3 years), demonstrating the importance of studying existing clinical data (EHR) through longitudinal analyses, and the potential that these approaches can have in supporting patient follow-up and medical decision making.

Despite the great capabilities and improvements that these proposed models can potentially bring to medical care, the various papers reviewed have limitations, that are clearly stated by the authors. Some of the most commonly cited limitations are as follows:

- The patient sample was clinic-based rather than population-based, which means that the model was only tested on a particular dataset, extracted from the population of a particular hospital/clinic. Furthermore, in most studies, there is no external validation dataset, leading to great uncertainty about generalization to a wider population. Cabitza et al. [66] show how external validation is essential for building robust predictive models in medicine.
- Small data samples, too much missing data and missing important features. Models trained on a small amount of data can result in poor generalizability and lead to incorrect conclusions being drawn. Too much missing data can affect the consistency of the data across different visits by a given patient. This consistency is essential to build a model that can deal with the time factor and make a prediction. In addition, several papers have highlighted various missing demographic, clinical, and laboratory variables that may be essential to improve outcomes.
- Almost all of the selected papers assume that the examples are independent of each other, which is inaccurate because multiple records belonging to a single patient have been obtained. The ability to account for this inter-record dependency is key to unlocking the potential that may exist in the temporal value of EHR data and can lead to models with greater and better predictive ability. Considering this relationship, Song et al. [42] simulated some inter-record dependency by passing the prediction made in each time window to the prediction of the next time window.

Using the information obtained from the selected articles, we are now able to answer the proposed research question.

RQ: What are the most effective machine learning techniques used to construct a model that uses the temporal information in diabetic patients' EHR data to predict the development of DN or progression to higher stages?

The reviewed literature suggests that despite the potential of using ML techniques to fully exploit the temporal dimension of EHR data to predict the risk of developing or progressing to DN, this has not yet been fully achieved. However, this provides an opportunity for further research and development in this area, with the aim of achieving more effective and accurate predictions in the future. Many of the techniques used have limited use of the temporal dimension and richness of patient records available in EHR data. Approaches that use only the values available for each patient at baseline or that use statistical operations on the data to combine aggregations of different clinical visits into a single record are valid but completely ignore the temporal potential. There are also some approaches that try to make a longitudinal study of the data, but often in a somewhat incomplete way. For example, the forecasts are separated by time windows (1 year from now, 2 years from now, etc.) and in some cases these forecasts are completely independent of each other. This completely breaks with the value of time and creates a shortcut to a result that is not very different from the first approach. The Landmark Boosting approach proposed by Song et al. was able to stand out because it creates time windows and tries to establish a correlation between these windows by predicting the disease state in the current window based on the state predicted in the previous window.

In summary, all the papers included in this review were generally able to arrive at a workable risk model for the onset or development of DN using a variety of techniques. All of them have attempted, either statically or dynamically, to make partial use of the temporal factor.

V. THREATS TO VALIDITY

This section discusses all the potential threats to the validation of this work, and the various biases and weaknesses that could in any way jeopardize the results obtained.

This review uses only three different databases, and the search was done with only one query (although it included all relevant keywords). This may introduce a selection bias, meaning that our sample of studies may not be representative of the population studied. If more papers had been included, we would be more likely to have different approaches that could add value to the discussion and possibly change the conclusions drawn.

The heterogeneity of the studies also threatens the validity of this paper. The data differ in quantity, in time of collection, demographic, social and cultural characteristics of the patients, and in some cases even in the meaning of the dependent variable (outcome). Some of them had multiple disease outputs and were not specifically designed to predict the risk of DN. This results in different training and validation data between the different articles selected. They also do not have a standardized way of presenting the results. In addition, some papers omit important information, which can lead to inaccurate or inconsistent results and conclusions. This is commonly referred to as measurement bias.

The study provides a broad and consistent approach to models capable of creating a predictive model of DN using EHR data and their respective time factors. However, it is important to consider that these errors and biases may have altered or influenced the results obtained and the conclusions drawn from them.

VI. FUTURE RESEARCH DIRECTIONS

Given the small number of works that have been done in this area, there is a great need for future research to have a clearer perspective on the impact that temporal data analysis could have on medical support systems [67]. In the coming years, it is expected that there will be a huge growth in this type of work, as shown by the trends in the studies selected for this review. Therefore, the following future research directions can be outlined:

- Fully exploit the time factor: Developing strategies that take advantage of the time factor and the

dependency between different visits for the same patient, not only to obtain more data, but also to allow the algorithm to access and consider the data as a healthcare team would normally do.

- ML with omics data: Further and better research into the impact that omics data can have on DN prediction by ML models should be explored so that it is possible to measure the impact of the respective integration. With the advent of modern biotechnologies and the great potential of ML, there is a great opportunity to bring together ML and omics data to significantly improve current systems [68].
- Apply Deep Learning (DL) techniques: Future research should focus on addressing the temporal nature of EHR data, as most traditional machine learning models are limited in their ability to handle this factor. One promising approach is the use of DL algorithms, which are well suited for detecting hidden patterns in large volumes of data and have greater flexibility and generalizability [69]. Therefore, the application of state-of-the-art DL techniques in future studies could potentially unlock the full temporal potential of EHR data and significantly improve predictive ability.

VII. CONCLUSION

This review focused on approaches that can use longitudinal data (EHR) to create ML models capable of predicting the risk of onset or development of DN. The findings suggest that the time factor inherent in the data has a clear potential to create a better predictor of DN risk. In addition, the combination of clinical and omics data can help us to achieve better results with greater credibility and generalizability. Furthermore, it is possible to test the concern of the authors of the different papers to create interpretable models whose results can be easily explained and understood by healthcare professionals.

It is important to emphasize that the studies varied in population, type and amount of data, outcome, and even purpose of the study, which may lead to limitations in the findings of this review. Further research is needed to address these limitations and to monitor how this area of temporal analysis of longitudinal data develops in the coming years.

Currently, there are only a few studies that have partially used the temporal information from EHRs to improve the accuracy of predictive ML models. However, we believe that using these temporal data will have a significant impact, especially in the detection of chronic diseases that take a long time to develop symptoms. Physicians use a patient's medical history to diagnose such diseases, and it is important for ML models to do the same. Therefore, incorporating temporal data from EHRs into ML risk prediction models has the potential to be a valuable support tool in healthcare, particularly in the diagnosis and management of chronic diseases, such as DN.

DECLARATIONS

Acknowledgements: This work is funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020.

Conflict of Interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

REFERENCES

[1] "Diabetes." https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed Oct. 29, 2022).

[2] OECD, *Health at a Glance: Europe 2020: State of Health in the EU Cycle*. Paris: Organisation for Economic Co-operation and Development, 2020. Accessed: Oct. 29, 2022. [Online]. Available: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2020_82129230-en

[3] Z. T. Bloomgarden, "Diabetes Complications," *Diabetes Care*, vol. 27, no. 6, pp. 1506–1514, Jun. 2004, doi: 10.2337/diacare.27.6.1506.

[4] P. Fioretto, I. Barzon, and M. Mauer, "Is diabetic nephropathy reversible?," *Diabetes Res. Clin. Pract.*, vol. 104, no. 3, pp. 323–328, Jun. 2014, doi: 10.1016/j.diabres.2014.01.017.

[5] P. Fioretto, I. Barzon, and M. Mauer, "Is diabetic nephropathy reversible?," *Diabetes Res. Clin. Pract.*, vol. 104, no. 3, pp. 323–328, Jun. 2014, doi: 10.1016/j.diabres.2014.01.017.

[6] "Diabetic Kidney Disease: A Report From an ADA Consensus Conference | Diabetes Care | American Diabetes Association." https://diabetesjournals.org/care/article/37/10/2864/30796/Diabetic-Kidney-Disease-A-Report-From-an-ADA (accessed Oct. 29, 2022).

[7] H. Hund, S. Gerth, D. Lossnitzer, and C. Fegeler, "Longitudinal Data Driven Study Design," in *e-Health – For Continuity of Care*, IOS Press, 2014, pp. 373–377. doi: 10.3233/978-1-61499-432-9-373.

[8] C. Ponchiardi, M. Mauer, and B. Najafian, "Temporal Profile of Diabetic Nephropathy Pathologic Changes," *Curr. Diab. Rep.*, vol. 13, no. 4, pp. 592–599, Aug. 2013, doi: 10.1007/s11892-013-0395-7.

[9] M. C. Thomas *et al.*, "Diabetic kidney disease," *Nat. Rev. Dis. Primer*, vol. 1, no. 1, Art. no. 1, Jul. 2015, doi: 10.1038/nrdp.2015.18.

[10] M. Kubat, *An Introduction to Machine Learning*. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-81935-4.

[11] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What is Machine Learning? A Primer for the Epidemiologist," *Am. J. Epidemiol.*, vol. 188, no. 12, pp. 2222–2239, Dec. 2019, doi: 10.1093/aje/kwz189.

[12] E. Brynjolfsson and T. Mitchell, "What can machine learning do? Workforce implications," *Science*, vol.

358, no. 6370, pp. 1530–1534, Dec. 2017, doi: 10.1126/science.aap8062.

[13] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[14] N. Sambyal, P. Saini, and R. Syal, "A Review of Statistical and Machine Learning Techniques for Microvascular Complications in Type 2 Diabetes," *Curr. Diabetes Rev.*, vol. 17, no. 2, pp. 143–155.

[15] J. Wong, M. Murray Horwitz, L. Zhou, and S. Toh, "Using Machine Learning to Identify Health Outcomes from Electronic Health Record Data," *Curr. Epidemiol. Rep.*, vol. 5, no. 4, pp. 331–342, Dec. 2018, doi: 10.1007/s40471-018-0165-9.

[16] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, Jan. 2017, doi: 10.1016/j.csbj.2016.12.005.

[17] K. R. Murphy and H. Aguinis, "HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results?," *J. Bus. Psychol.*, vol. 34, no. 1, pp. 1–17, Feb. 2019, doi: 10.1007/s10869-017-9524-7.

[18] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, "Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses," *FASEB J.*, vol. 22, no. 2, pp. 338–342, 2008, doi: 10.1096/fj.07-9492LSF.

[19] X. Song, L. R. Waitman, Y. Hu, A. S. L. Yu, D. Robins, and M. Liu, "Robust clinical marker identification for diabetic kidney disease with ensemble feature selection," *J. Am. Med. Inform. Assoc.*, vol. 26, no. 3, pp. 242–253, Mar. 2019, doi: 10.1093/jamia/ocy165.

[20] P. Connolly *et al.*, "Analytical validation of a multi-biomarker algorithmic test for prediction of progressive kidney function decline in patients with early-stage kidney disease," *Clin. Proteomics*, vol. 18, no. 1, p. 26, Nov. 2021, doi: 10.1186/s12014-021-09332-y.

[21] V. Singh, V. K. Asari, and R. Rajasekaran, "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease," *Diagnostics*, vol. 12, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/diagnostics12010116.

[22] "Using Machine Learning to Predict Diabetes Complications | IEEE Conference Publication | IEEE Xplore." https://ieeexplore.ieee.org/document/9677649 (accessed Dec. 04, 2022).

[23] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "A Machine Learning Approach to Predicting Diabetes Complications," *Healthcare*, vol. 9, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/healthcare9121712.

[24] S. K. David, M. Rafiullah, and K. Siddiqui, "Comparison of Different Machine Learning Techniques to Predict Diabetic Kidney Disease," *J. Healthc. Eng.*, vol. 2022, p. e7378307, Apr. 2022, doi: 10.1155/2022/7378307.

[25] M. Zuo, W. Zhang, Q. Xu, and D. Chen, "Deep Personal Multitask Prediction of Diabetes Complication with Attentive Interactions Predicting Diabetes Complications by Multitask-Learning," *J. Healthc. Eng.*, vol. 2022, p. 5129125, 2022, doi: 10.1155/2022/5129125.

[26] Y. Fan, E. Long, L. Cai, Q. Cao, X. Wu, and R. Tong, "Machine Learning Approaches to Predict Risks of Diabetic Complications and Poor Glycemic Control in Nonadherent Type 2 Diabetes," *Front. Pharmacol.*, vol. 12, p. 665951, Jun. 2021, doi: 10.3389/fphar.2021.665951.

[27] S. Ravizza *et al.*, "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data," *Nat. Med.*, vol. 25, no. 1, pp. 57–59, Jan. 2019, doi: 10.1038/s41591-018-0239-8.

[28] N. H. Chowdhury *et al.*, "Performance Analysis of Conventional Machine Learning Algorithms for Identification of Chronic Kidney Disease in Type 1 Diabetes Mellitus Patients," *Diagn. Basel Switz.*, vol. 11, no. 12, p. 2267, Dec. 2021, doi: 10.3390/diagnostics11122267.

[29] A. L. Neves *et al.*, "Using electronic health records to develop and validate a machine-learning tool to predict type 2 diabetes outcomes: a study protocol," *BMJ Open*, vol. 11, no. 7, p. e046716, Jul. 2021, doi: 10.1136/bmjopen-2020-046716.

[30] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Sci. Rep.*, vol. 6, no. 1, Art. no. 1, May 2016, doi: 10.1038/srep26094.

[31] B. P. Swan, M. E. Mayorga, and J. S. Ivy, "The SMART Framework: Selection of Machine Learning Algorithms With ReplicaTions—A Case Study on the Microvascular Complications of Diabetes," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 2, pp. 809–817, Feb. 2022, doi: 10.1109/JBHI.2021.3094777.

[32] P. Novitski, C. M. Cohen, A. Karasik, G. Hodik, and R. Moskovitch, "Temporal patterns selection for All-Cause Mortality prediction in T2D with ANNs," *J. Biomed. Inform.*, vol. 134, p. 104198, Oct. 2022, doi: 10.1016/j.jbi.2022.104198.

[33] Q. Xu, L. Wang, and S. S. Sansgiry, "A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning," *J. Med. Artif. Intell.*, vol. 3, no. 0, Art. no. 0, Mar. 2020, doi: 10.21037/jmai.2019.10.04.

[34] T. Rahman, S. M. Farzana, and A. Z. Khanom, "Prediction of diabetes induced complications using different machine learning algorithms," Thesis, BRAC University, 2018. Accessed: Dec. 04, 2022. [Online]. Available: http://dspace.bracu.ac.bd/xmlui/handle/10361/10945

[35] W. Jiang *et al.*, "Establishment and Validation of a Risk Prediction Model for Early Diabetic Kidney

Disease Based on a Systematic Review and Meta-Analysis of 20 Cohorts," *Diabetes Care*, vol. 43, no. 4, pp. 925–933, Mar. 2020, doi: 10.2337/dc19-1897.

[36] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," *J. Biomed. Inform.*, vol. 53, pp. 220–228, Feb. 2015, doi: 10.1016/j.jbi.2014.11.005.

[37] A. Dagliati *et al.*, "Machine Learning Methods to Predict Diabetes Complications," *J. Diabetes Sci. Technol.*, vol. 12, no. 2, pp. 295–302, Mar. 2018, doi: 10.1177/1932296817706375.

[38] M. Makino *et al.*, "Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning," *Sci. Rep.*, vol. 9, p. 11862, Aug. 2019, doi: 10.1038/s41598-019-48263-5.

[39] V. Rodriguez-Romero, R. F. Bergstrom, B. S. Decker, G. Lahu, M. Vakilynejad, and R. R. Bies, "Prediction of Nephropathy in Type 2 Diabetes: An Analysis of the ACCORD Trial Applying Machine Learning Techniques," *Clin. Transl. Sci.*, vol. 12, no. 5, pp. 519–528, 2019, doi: 10.1111/cts.12647.

[40] S. M. Hosseini Sarkhosh, M. Hemmatabadi, and A. Esteghamati, "Development and validation of a risk score for diabetic kidney disease prediction in type 2 diabetes patients: a machine learning approach," *J. Endocrinol. Invest.*, Sep. 2022, doi: 10.1007/s40618-022-01919-y.

[41] A. Aminian *et al.*, "Predicting 10-Year Risk of End-Organ Complications of Type 2 Diabetes With and Without Metabolic Surgery: A Machine Learning Approach," *Diabetes Care*, vol. 43, no. 4, pp. 852–859, Feb. 2020, doi: 10.2337/dc19-2057.

[42] X. Song, L. R. Waitman, A. S. Yu, D. C. Robbins, Y. Hu, and M. Liu, "Longitudinal Risk Prediction of Chronic Kidney Disease in Diabetic Patients Using a Temporal-Enhanced Gradient Boosting Machine: Retrospective Cohort Study," *JMIR Med. Inform.*, vol. 8, no. 1, p. e15510, Jan. 2020, doi: 10.2196/15510.

[43] L. Chan *et al.*, "Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease," *Diabetologia*, vol. 64, no. 7, pp. 1504–1515, Jul. 2021, doi: 10.1007/s00125-021-05444-0.

[44] A. Allen *et al.*, "Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus," *BMJ Open Diabetes Res. Care*, vol. 10, no. 1, p. e002560, Jan. 2022, doi: 10.1136/bmjdrc-2021-002560.

[45] Z. Dong *et al.*, "Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records," *J. Transl. Med.*, vol. 20, no. 1, p. 143, Mar. 2022, doi: 10.1186/s12967-022-03339-1.

[46] N. Al-Sari *et al.*, "Precision diagnostic approach to predict 5-year risk for microvascular complications in type 1 diabetes," *eBioMedicine*, vol. 80, Jun. 2022, doi: 10.1016/j.ebiom.2022.104032.

[47] M. R. Cowie *et al.*, "Electronic health records to facilitate clinical research," *Clin. Res. Cardiol.*, vol. 106, no. 1, pp. 1–9, Jan. 2017, doi: 10.1007/s00392-016-1025-6.

[48] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, and S. A. Tsaftaris, "Causal machine learning for healthcare and precision medicine," *R. Soc. Open Sci.*, vol. 9, no. 8, p. 220638, doi: 10.1098/rsos.220638.

[49] C. M. Micheel *et al.*, *Omics-Based Clinical Discovery: Science, Technology, and Applications*. National Academies Press (US), 2012. Accessed: Jan. 07, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK202165/

[50] A. Holzinger, B. Haibe-Kains, and I. Jurisica, "Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 46, no. 13, pp. 2722–2730, Dec. 2019, doi: 10.1007/s00259-019-04382-9.

[51] K. Jayawardana *et al.*, "Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information," *Int. J. Cancer*, vol. 136, no. 4, pp. 863–874, 2015, doi: 10.1002/ijc.29047.

[52] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer," *Clin. Cancer Res.*, vol. 24, no. 6, pp. 1248–1259, Mar. 2018, doi: 10.1158/1078-0432.CCR-17-0853.

[53] D. Tong *et al.*, "Improving prediction performance of colon cancer prognosis based on the integration of clinical and multi-omics data," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 22, Feb. 2020, doi: 10.1186/s12911-020-1043-1.

[54] R. De Bin, W. Sauerbrei, and A.-L. Boulesteix, "Investigating the prediction ability of survival models based on both clinical and omics data: two case studies," *Stat. Med.*, vol. 33, no. 30, pp. 5310–5329, 2014, doi: 10.1002/sim.6246.

[55] A. Cambiaghi, M. Ferrario, and M. Masseroli, "Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration," *Brief. Bioinform.*, vol. 18, no. 3, pp. 498–510, May 2017, doi: 10.1093/bib/bbw031.

[56] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018, doi: 10.1016/j.neucom.2017.11.077.

[57] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *WIREs Comput. Stat.*, vol. 11, no. 3, p. e1460, 2019, doi: 10.1002/wics.1460.

[58] H. Wajant and D. Siegmund, "TNFR1 and TNFR2 in the Control of the Life and Death Balance of Macrophages," *Front. Cell Dev. Biol.*, vol. 7, 2019,

Accessed: Jan. 14, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcell.2019.00091

[59] D. M. Tanase *et al.*, "The Predictive Role of the Biomarker Kidney Molecule-1 (KIM-1) in Acute Kidney Injury (AKI) Cisplatin-Induced Nephrotoxicity," *Int. J. Mol. Sci.*, vol. 20, no. 20, Art. no. 20, Jan. 2019, doi: 10.3390/ijms20205238.

[60] S. Jiang *et al.*, "Prognostic nomogram and score to predict renal survival of patients with biopsy-proven diabetic nephropathy," *Diabetes Res. Clin. Pract.*, vol. 155, p. 107809, Sep. 2019, doi: 10.1016/j.diabres.2019.107809.

[61] K.-H. Thung and C.-Y. Wee, "A brief review on multi-task learning," *Multimed. Tools Appl.*, vol. 77, no. 22, pp. 29705–29725, Nov. 2018, doi: 10.1007/s11042-018-6463-x.

[62] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Jun. 13, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[63] N. Salkind, *Encyclopedia of Research Design*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2010. doi: 10.4135/9781412961288.

[64] X. Zhao, Y. Wu, D. L. Lee, and W. Cui, "iForest: Interpreting Random Forests via Visual Analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 407–416, Jan. 2019, doi: 10.1109/TVCG.2018.2864475.

[65] J. Fletcher, "What is heterogeneity and is it important?," *BMJ*, vol. 334, no. 7584, pp. 94–96, Jan. 2007, doi: 10.1136/bmj.39057.406644.68.

[66] F. Cabitza *et al.*, "The importance of being external. methodological insights for the external validation of machine learning models in medicine," *Comput. Methods Programs Biomed.*, vol. 208, p. 106288, Sep. 2021, doi: 10.1016/j.cmpb.2021.106288.

[67] I. Bica, A. M. Alaa, C. Lambert, and M. van der Schaar, "From Real-World Patient Data to Individualized Treatment Effects Using Machine Learning: Current and Future Methods to Address Underlying Challenges," *Clin. Pharmacol. Ther.*, vol. 109, no. 1, pp. 87–100, 2021, doi: 10.1002/cpt.1907.

[68] R. Li, L. Li, Y. Xu, and J. Yang, "Machine learning meets omics: applications and perspectives," *Brief. Bioinform.*, vol. 23, no. 1, pp. 1–22, Jan. 2022, doi: 10.1093/bib/bbab460.

[69] F. Xie *et al.*, "Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies," *J. Biomed. Inform.*, vol. 126, p. 103980, Feb. 2022, doi: 10.1016/j.jbi.2021.103980.

Appendix B – hyperparameters of ML models

This chapter presents some of the most important hyperparameters associated with the ML models trained in both approach A and B. Table 7.1 displays all the default hyperparameters initially used in the training process. Additionally, Table 7.2 outlines the hyperparameter configurations after tuning the models in approach A, while Table 7.3 presents the tuned hyperparameters for approach B.

Table 7.1: Default hyperparameters of ML models before tuning in approaches A and B.

Model	Hyperparameters	Values
GBM	Loss	Logloss
	Learning rate	0.1
	N_estimators	100
	Criterion	friend_mse
	Max_depth	3
Catboost	Learning rate	0.03
	Depth	6
	N_estimators	10
	Evaluation metric	Logloss
	Score function	Cosine
MLP	Hidden layers size	[100]
	Activation function	Relu
	Solver	Adam
	Learning rate	Constant
	Max iterations	200
LR	Penalty	L2
	C	1.0
	Solver	Lbfgs
	Max iterations	100
LightGBM	Boosting type	Gbdt
	Num leaves	31
	Max depth	-1
	Learning rate	0.1
	N_estimators	100
Adaboost	Base estimator	Decision Tree
	N_estimators	50
	Learning rate	1.0
	Algorithm	SAMME.R

Table 7.2: Hyperparameters of ML models after tuning in approach A.

Model	Hyperparameters	Values
GBM	Loss	Logloss
	Learning rate	0.1
	N_estimators	100
	Criterion	friend_mse
	Max_depth	3
Catboost	Learning rate	0.03
	Depth	4

	N_estimators	190
	Evaluation metric	Logloss
	Score function	Cosine
MLP	Hidden layers size	[50, 50]
	Activation function	logistic
	Solver	Adam
	Learning rate	Constant
	Max iterations	500
LightGBM	Boosting type	Gbdt
	Num leaves	20
	Max depth	-1
	Learning rate	0.05
	N_estimators	80
Adaboost	Base estimator	Decision Tree
	N_estimators	190
	Learning rate	0.4
	Algorithm	SAMME.R

Table 7.3: Hyperparameters of ML models after tuning in approach B.

Model	**Hyperparameters**	**Values**
GBM	Loss	Logloss
	Learning rate	0.1
	N_estimators	100
	Criterion	friend_mse
	Max_depth	3
Catboost	Learning rate	0.03
	Depth	2
	N_estimators	110
	Evaluation metric	Logloss
	Score function	Cosine
MLP	Hidden layers size	[50]
	Activation function	Relu
	Solver	Adam
	Learning rate	Adaptive
	Max iterations	500
LR	Penalty	L2
	C	6.246
	Solver	Lbfgs
	Max iterations	1000
LightGBM	Boosting type	Gbdt
	Num leaves	31
	Max depth	-1
	Learning rate	0.1
	N_estimators	100

Appendix C – ML models analysis

This chapter presents the supplementary material used in the analysis of the models trained and selected on approach B to assist the analysis of the results presented in Chapter 5.3. Appendix C.1 presents all the plots concerning the feature importance of the various models. The MLP algorithm was excluded from the analysis since there's no scikit-learn method to determine feature importance for it. Consequently, as PyCaret utilizes scikit-learn algorithms, feature ranking for this method couldn't be displayed. Appendix C.2 presents the results obtained in the statistical analysis performed between the ML models using the McNemar's statistical test.

Appendix C.1 – Feature importance

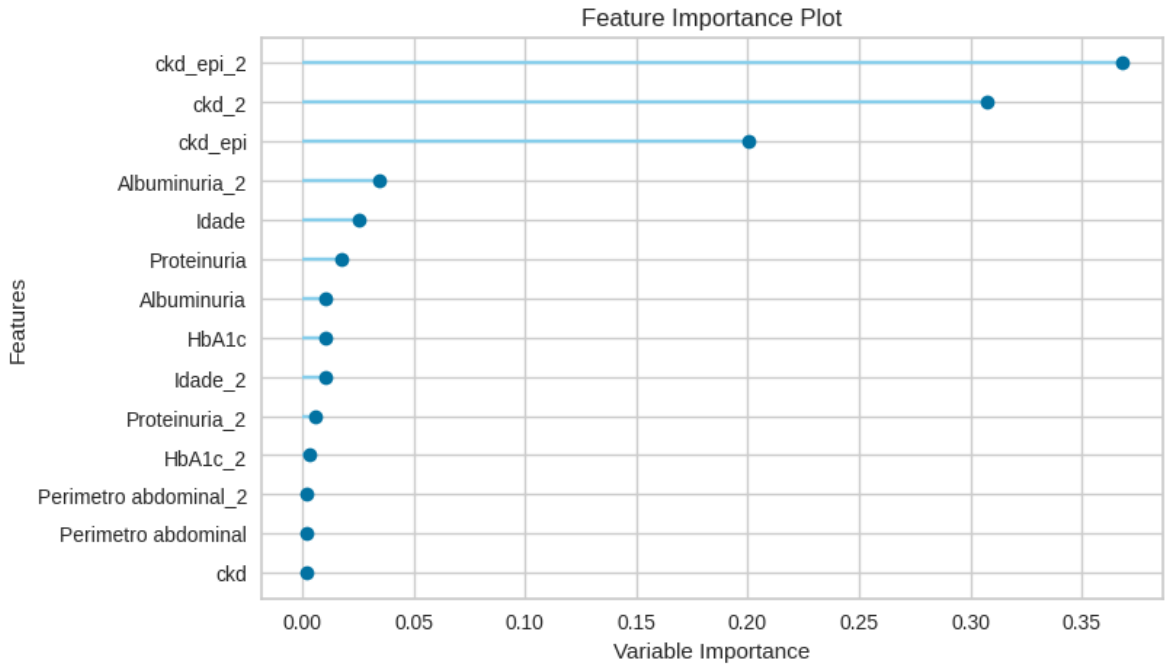

Figure 7.1: Feature importance on GBM model

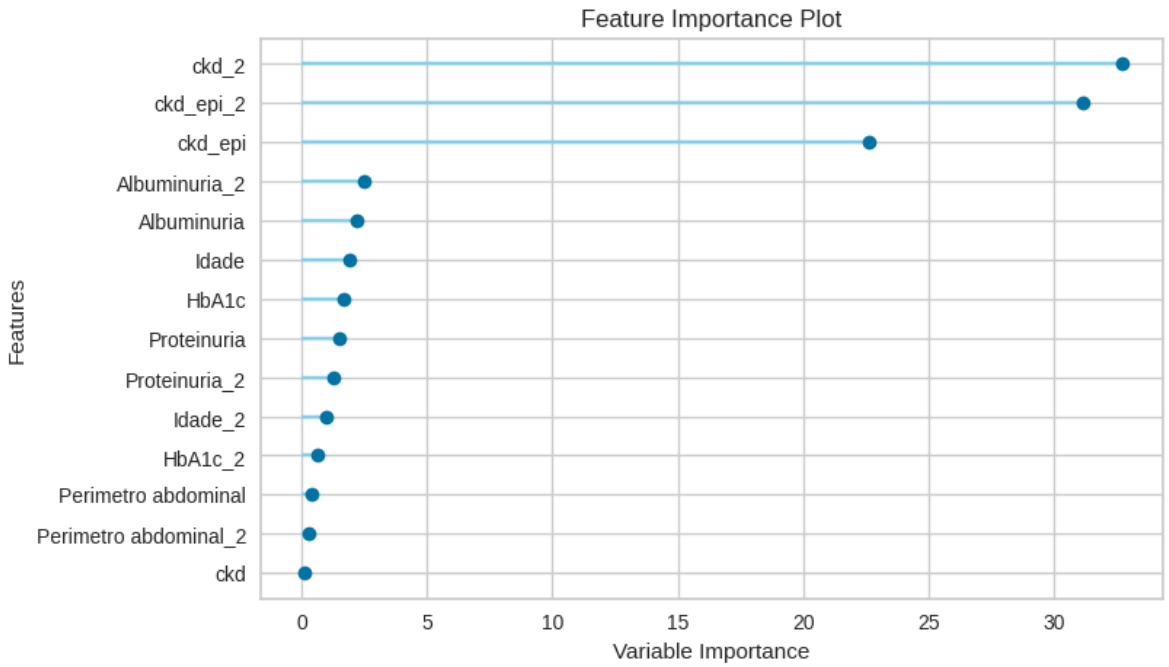

Figure 7.2: Feature importance on Catboost model

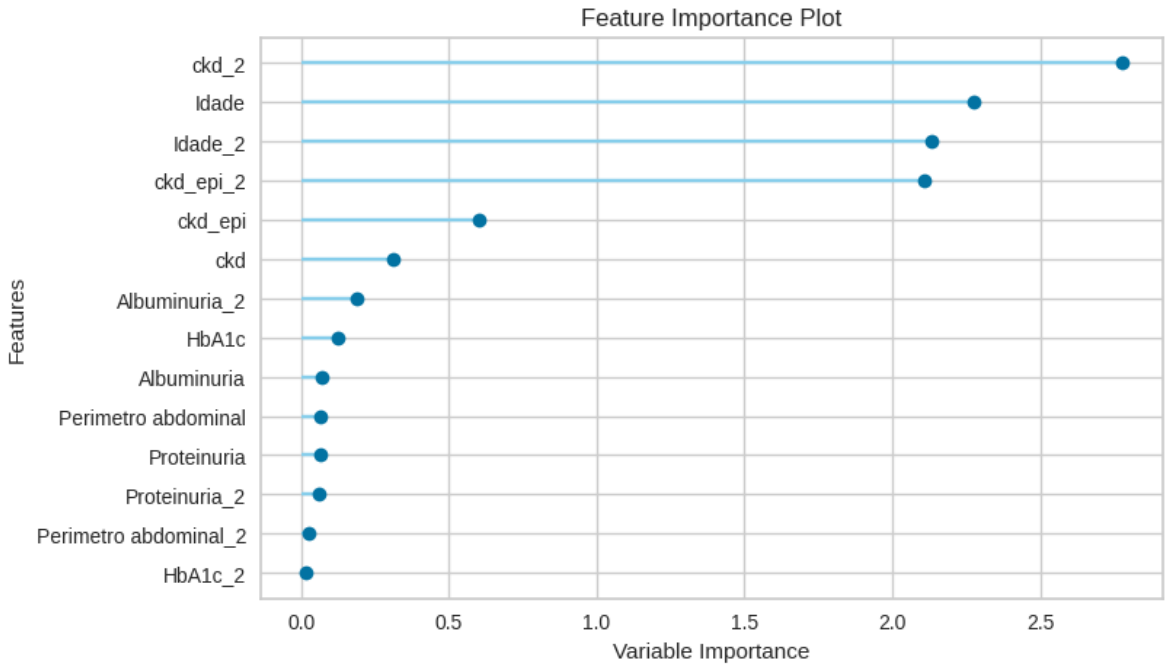

Figure 7.3: Feature importance on LR model

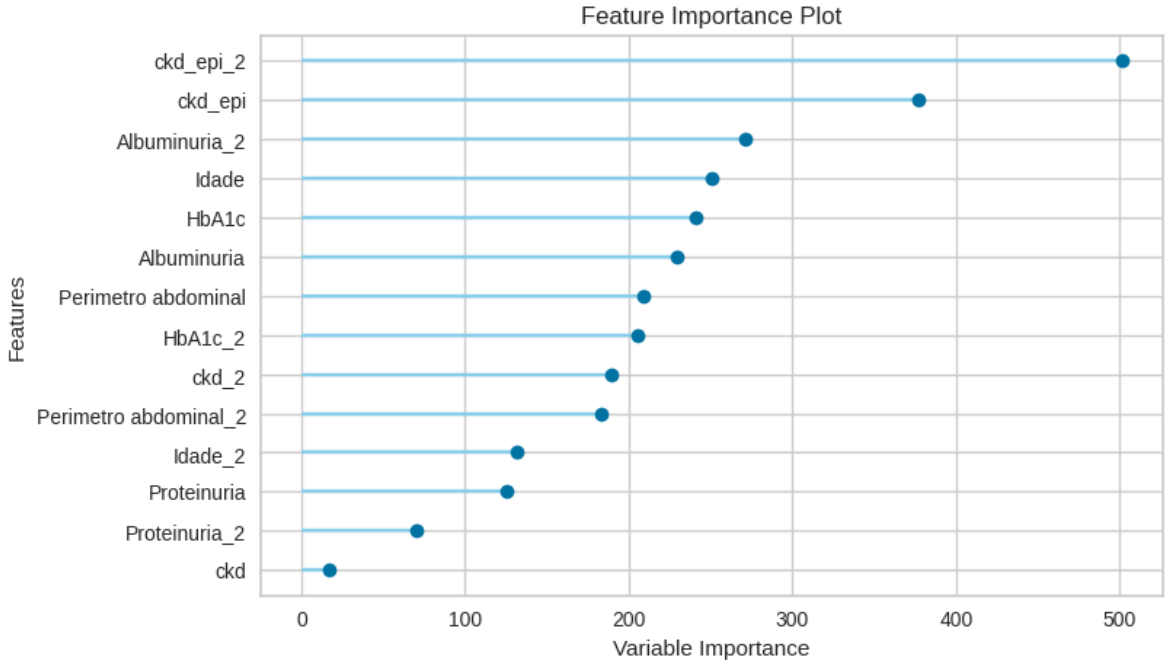

Figure 7.4: Feature importance on LightGBM model

Appendix C.2 – Statistical Significance

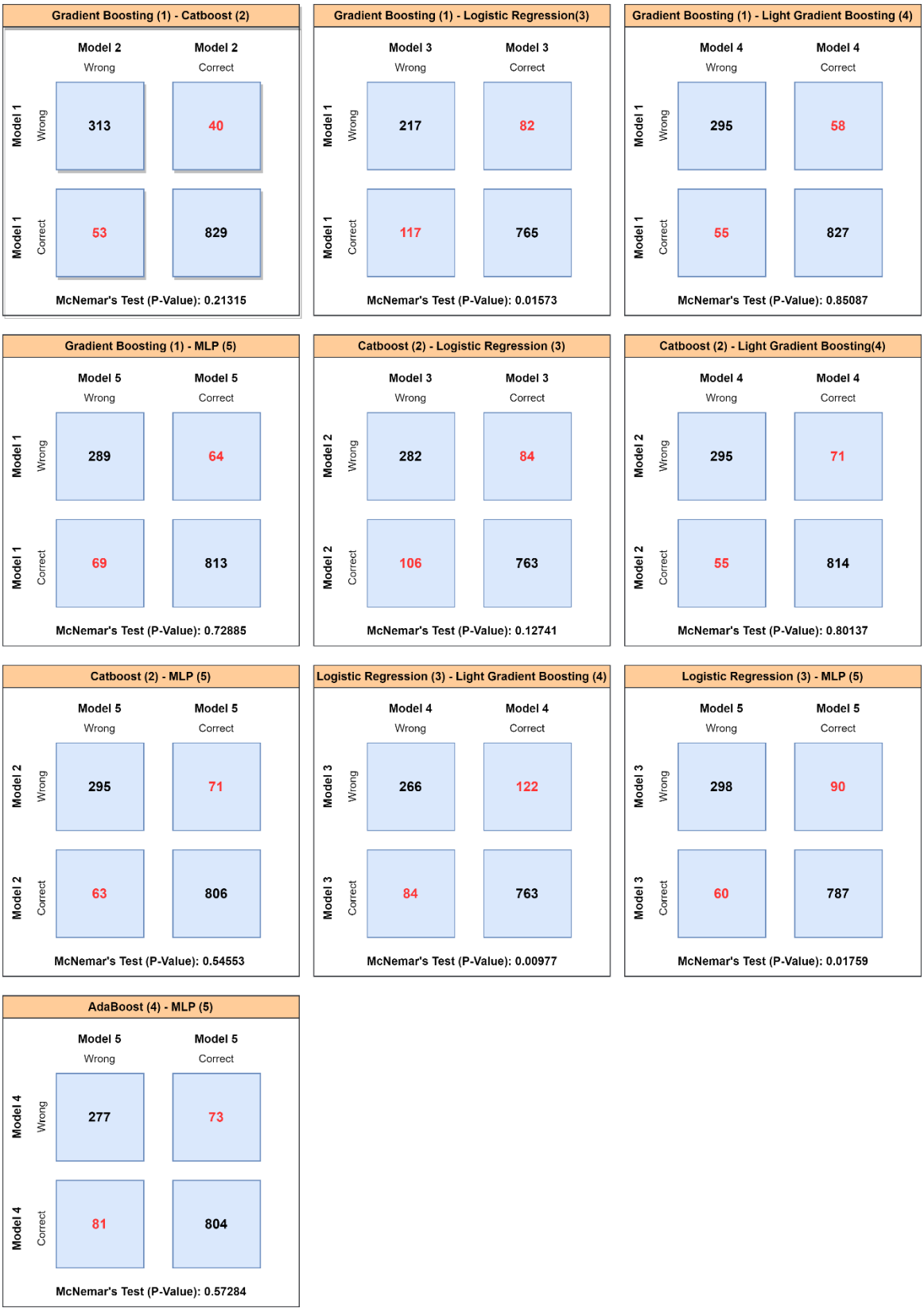

Figure 7.5: Statistical significance of ML models' performance using McNemar's test.

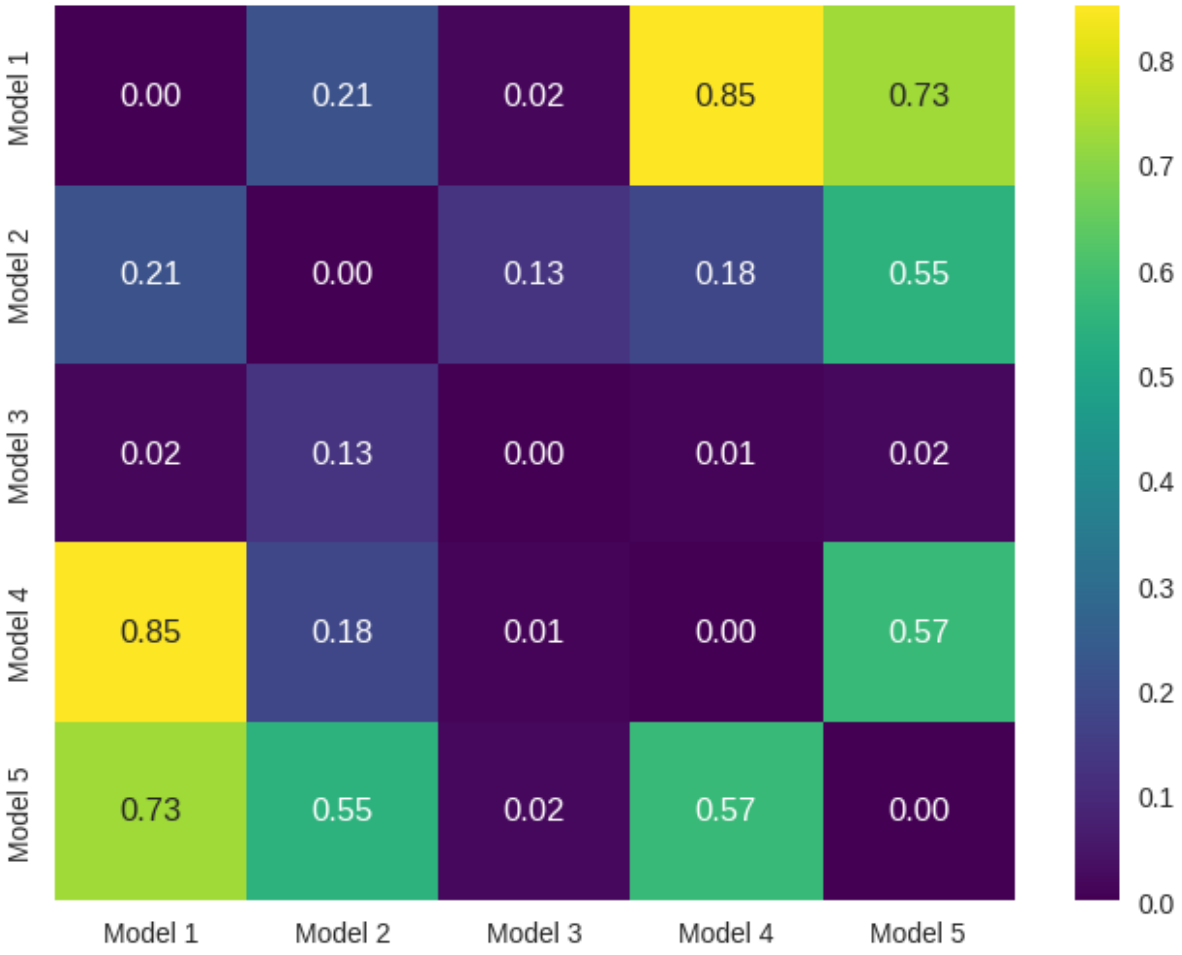

Figure 7.6: P-value between the different trained ML models

Instituto Superior
de Engenharia
Politécnico de Coimbra