Prédiction du risque des maladies endocriniennes à l'aide de la science des données et de l'intelligence artificielle explicable

Firas Ketata

Thèse de doctorat de l'établissement Université Bourgogne Franche-Comté
en co-tutelle avec l'Université de Carthage
préparée à l'Université de Franche-Comté

École doctorale n° 37
Sciences Pour l'Ingénieur et Microtechniques

Doctorat de Sciences pour l'Ingénieur

Par
Firas KETATA

Prédiction du Risque des Maladies Endocriniennes à l'aide de la Science des Données et de l'Intelligence Artificielle Explicable

Thèse présentée et soutenue à Besançon, le 2 octobre 2024

Composition du Jury :

Mounir SAYADI	Professeur à l'ENSIT, Tunis	Président
Maria DI MASCOLO	Directrice de recherche au laboratoire G-SCOP, Grenoble	Rapporteuse
Mohamed Hédi BEDOUI	Professeur en Biophysique à la Faculté de Médecine de Monastir	Rapporteur
Edith GRALL	Maître de conférences à l'Université de Technologie de Troyes	Examinatrice
Hatem BOULAHDOUR	Professeur à l'UFC et médecin au CHU, Besançon	Examinateur
Noureddine ZERHOUNI	Professeur à SUPMICROTECH-ENSMM, Besançon	Directeur de thèse
Zeina AL MASRY	Maître de conférences à SUPMICROTECH-ENSMM, Besançon	Codirectrice de thèse
Slim YACOUB	Professeur à l'INSAT, Tunis	Codirecteur de thèse

Ph.D. thesis of the University of Bourgogne Franche-Comté
under joint supervision with University of Carthage
prepared at University of Franche-Comté

Doctoral school n° 37
Engineering Sciences and Microtechniques

PhD in Engineering Sciences

By

Firas KETATA

Risk Prediction of Endocrine Diseases using Data Science and Explainable Artificial Intelligence

Thesis presented and defended in Besançon, on 2 october 2024

Composition of the jury:

Mounir SAYADI	Professor at ENSIT, Tunis	President
Maria DI MASCOLO	Research director at G-SCOP laboratory, Grenoble	Reviewer
Mohamed Hédi BEDOUI	Professor in Biophysics at the Faculty of Medicine of Monastir	Reviewer
Edith GRALL	Associate Professor at The University of Technology of Troyes	Examiner
Hatem BOULAHDOUR	University professor and Hospital practitioner at UFC and CHU, Besançon	Examiner
Noureddine ZERHOUNI	Professor at SUPMICROTECH-ENSMM, Besançon	Thesis Supervisor
Zeina AL MASRY	Associate Professor at SUPMICROTECH-ENSMM, Besançon	Thesis Co-Supervisor
Slim YACOUB	Professor at INSAT, Tunis	Thesis Co-Supervisor

Acknowledgements

La réalisation de cette thèse a été possible grâce aux efforts et aux contributions de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je tiens tout d'abord à exprimer ma profonde reconnaissance à Madame Zeina Al MASRY, mon encadrante et superviseur de thèse, pour ses efforts inlassables et sa précieuse contribution à la contextualisation et à la méthodologie des approches proposées. Son expertise et son soutien ont été essentiels à l'aboutissement de ce travail.

Je souhaite également remercier chaleureusement Pr Noureddine ZERHOUNI, directeur de thèse, pour son encadrement rigoureux, ses conseils avisés et sa capacité à orienter et affiner le travail de cette thèse. Sa direction a été un pilier dans le développement de cette recherche.

Mes remerciements vont aussi à Pr Slim YACOUB, mon deuxième directeur de thèse, pour son encadrement et son soutien constant tout au long de cette aventure académique.

Je tiens à exprimer ma reconnaissance à l'ENSMM qui m'a donné l'opportunité de participer à l'enseignement dans un poste ATER. Je remercie principalement Jean-Marc NICOD, Fabrice STHAL, Christophe VARNIER, Joël IMBAUD, Guillaume LAURENT et tous les professeurs et employés de l'ENSMM où j'ai passé mes trois années de doctorat.

Je voudrais également remercier l'équipe administrative du Laboratoire Femto-ST pour leur aide dans les démarches administratives, en particulier Isabelle GABET et Estelle PETITE. Un grand merci à l'équipe administrative de l'école doctorale, notamment Alika ROSSETTI, pour son aide précieuse dans les démarches administratives.

Je tiens à exprimer ma gratitude aux médecins partenaires, Dr Paul VALENSI, Dr Karima BEN MOHAMED et Dr Sabrina KHENSAL, pour leur collaboration et leurs perspectives médicales qui ont apporté une valeur ajoutée significative à cette thèse.

Je remercie également tous les membres du jury, Maria DI-MASCOLO, Hedi BEDOUI, Edith GRALL, Hatem BOULAHDOUR et Mounir SAYADI pour leur temps, leurs critiques constructives et leur intérêt pour mon travail.

Je remercie également Naceur SABEUR pour ses efforts soutenus tout au long de la thèse et son aide précieuse.

Un merci tout particulier à mes parents, Raouf KETATA et Kaouthar KALLEL, pour leur double rôle en tant que parents et connaisseurs en domaine. Leur soutien, leurs avis éclairés et leurs recommandations ont été inestimables.

Je remercie de tout cœur ma confidente Maha et ma chère Souha pour leur aide et leur soutien indéfectible tout au long de ce parcours.

Je n'oublie pas mon frère Mahdi et ma sœur Maryouma pour leur soutien moral et leur présence rassurante. Une pensée spéciale pour mes grands-parents, Ma et Pa, pour leur amour et leur encouragement continus.

Je souhaite aussi remercier chaleureusement mes collègues Aziz, Lobna, Valentin, Wassim, Rania, Roua et Amal pour leur aide précieuse, leur soutien continu et leurs encouragements tout au long de ce parcours.

Je tiens également à exprimer ma gratitude à tous les membres de ma famille pour leur soutien inconditionnel : Raoudha, Hajjour, Houssem Eddine, Khalil, Neila, Ammoun et Mohamed.

À toutes ces personnes, je vous exprime ma reconnaissance la plus sincère. Vous avez chacun, à votre manière, contribué à la réussite de cette thèse.

Table of contents

Résumé

Les anomalies glucidiques et le syndrome métabolique sont des pathologies complexes qui affectent une part significative de la population mondiale et impliquent des coûts élevés en termes de dépistage et de traitement. La bêta-thalassémie majeure, une maladie génétique rare entraînant une anémie sévère nécessitant des transfusions sanguines régulières, est souvent associée à des complications endocriniennes graves, notamment des anomalies glucidiques dues à une surcharge en fer causée par les transfusions. Ces complications exigent une surveillance et un traitement constants, générant ainsi des charges financières importantes pour les systèmes de santé et les patients. De même, le syndrome métabolique, caractérisé par une combinaison de troubles métaboliques et cardiovasculaires, représente un enjeu majeur de santé publique. Son dépistage précoce, notamment chez les adolescents, est crucial mais aussi coûteux, en raison de la nécessité d'effectuer des tests biologiques sophistiqués.

Face à ces défis financiers et médicaux, l'objectif principal de cette thèse est de proposer des outils de prédiction du risque des anomalies glucidiques chez les patients atteints de bêta-thalassémie majeure et du syndrome métabolique chez les adolescents, en utilisant des techniques de apprentissage automatique. Ces outils visent à aider les médecins à personnaliser le traitement et le dépistage, tout en optimisant la gestion des ressources financières et temporelles dans les systèmes de santé. En fournissant des prédictions précises, il devient possible de concentrer les efforts de dépistage et de traitement sur les patients les plus à risque, réduisant ainsi les coûts inutiles et améliorant l'efficacité globale des interventions médicales.

L'intégration des outils de apprentissage automatique dans le domaine médical pose cependant des défis, notamment en termes d'explicabilité et de fiabilité des prédictions. Les médecins doivent comprendre et évaluer ces prédictions pour les intégrer en toute confiance dans leur prise de décision clinique. Une autre difficulté réside dans le coût élevé lié à l'extraction de variables biologiques nécessaires pour établir ces prédictions, particulièrement dans le cadre du dépistage du syndrome métabolique.

Pour répondre à ces enjeux, cette thèse propose plusieurs contributions majeures. Premièrement, des modèles de prédiction ont été développés pour estimer le risque d'anomalies glucidiques chez les patients atteints de bêta-thalassémie majeure et le risque de syndrome métabolique chez les adolescents. Ces modèles fournissent aux médecins des informations essentielles sur les patients à haut et à faible risque, leur permettant de personnaliser les traitements et de cibler les dépistages avec une précision accrue. Cela contribue à une meilleure gestion des ressources médicales et financières, en focalisant les efforts sur les patients les plus vulnérables et en limitant les interventions coûteuses pour les individus à faible risque.

Deuxièmement, cette thèse aborde la question de l'explicabilité des prédictions de apprentissage automatique en intégrant des outils d'intelligence artificielle explicable. Ces outils offrent aux médecins des moyens d'évaluer la fiabilité des prédictions, en leur donnant accès à des explications claires sur les facteurs influençant les résultats des modèles. Cette transparence est cruciale pour renforcer la confiance des professionnels de santé dans les outils prédictifs et les aider à intégrer ces informations dans la prise de décision clinique.

Troisièmement, une approche a été proposée pour réduire les coûts du dépistage du syndrome métabolique en s'appuyant sur des variables cliniques plutôt que biologiques. Cette démarche permet de maintenir une précision élevée des prédictions tout en réduisant de manière significative les dépenses liées aux tests biologiques, rendant ainsi le dépistage plus accessible et moins coûteux pour les systèmes de santé.

Enfin, la thèse propose une solution pour améliorer la fiabilité de l'explicabilité fournie par les modèles de apprentissage automatique. En combinant les techniques d'intelligence artificielle explicable avec une méthode d'augmentation de données, il est possible d'obtenir des explications plus stables et robustes, ce qui renforce la confiance des médecins dans la prédiction des risques de ces maladies. Des métriques ont également été développées pour évaluer la fiabilité de cette approche, garantissant ainsi une explicabilité cohérente et fiable, indépendamment des variations dans les ensembles de données.

En conclusion, cette thèse apporte des solutions innovantes pour améliorer la gestion du risque des anomalies glucidiques chez les patients atteints de bêta-thalassémie majeure et du syndrome métabolique chez les adolescents. En combinant les outils de apprentissage automatique et l'intelligence artificielle explicable, elle permet d'optimiser l'utilisation des ressources médicales et financières tout en renforçant la confiance des médecins dans les systèmes de prédiction. Ces avancées contribuent à la personnalisation des soins et à l'amélioration de l'efficience des systèmes de santé, répondant ainsi à des enjeux médicaux et économiques majeurs.

General Introduction

Problem Statement

Medical Motivation

Endocrine glands are organs that produce chemical substances called hormones, which regulate many body functions such as metabolism, growth, reproduction, sleep, and mood [Rosol 24]. The endocrine system, composed of these glands, produces and regulates hormones and is essential for maintaining homeostasis. An endocrine disorder is a medical condition that affects the normal function of the endocrine system [Rosol 24]. Given that endocrinology naturally involves multiple organs and hormones, with widespread effects throughout the body, abnormalities in this system can lead to a wide variety of hormonal and metabolic pathologies [Kawa 21]. Some diseases are relatively common, such as obesity, glucose disorders, metabolic syndrome, thyroid disorders, reproductive disorders, and cardiovascular diseases [Kawa 21]. Consequently, the total number of patients affected by these pathologies represents a significant population, especially diabetes and thyroid disorders, which are the most widespread endocrine diseases in the world according to the World Health Organization [Biondi 19]. According to the International Diabetes Federation, 537 million adults worldwide suffer from diabetes in 2023.

Moreover, endocrine disease treatment and screening can be costly and difficult to manage, as seen in cases such as carbohydrate abnormalities in patients with beta-thalassemia major (β-TM) or metabolic syndrome (MetS) screening in adolescents.

β-TM is a rare disease caused by a deficiency or absence in the production of the beta-globin chain in red blood cells, leading to increased hemolysis (the destruction of red blood cells) and severe anemia requiring regular, lifelong transfusions of compatible, phenotyped blood [Taher 21]. The problem with these polytransfusions is that they induce iron overloads and deposits in vital organs such as the heart and liver, leading to fatal cardiac, hepatic, or endocrine complications such as carbohydrate abnormalities [Patne 18]. Without treatment, the life expectancy for these patients is no longer than 20 years. However, if properly treated, survival can extend beyond adulthood, through regular monitoring aimed at early detection of complications [He 19]. Treatment consists mainly of regular transfusions, with iron chelators added to delay the onset of post-transfusion complications. Splenomegaly (removal of the spleen) can reduce the need for transfusions but may result in serious and fatal infectious complications. Stem cell transplantation (grafting) remains the only curative treatment for these patients, but it is expensive and not always available at the necessary dose [He 19].

Major financial and time burden limitations also exist for MetS screening. MetS is characterized by various cardiovascular (including hypertension) and metabolic disorders (insulin resistance, glucose intolerance, hypercholesterolemia, and abdominal obesity). The combination of these symptoms is unanimously recognized as a major cardiovascular and metabolic risk factor, with the main complications of MetS being cardiovascular disease, diabetes and obesity [Grundy 04]. It constitutes a public health problem, with an increasing incidence over the last few years, particularly among adolescents [Weiss 04]. Several groups of experts have proposed diagnostic criteria that have given rise to various definitions of MetS to decide on prevention strategies. All the proposed definitions group

together similar metabolic and cardiovascular abnormalities, but with variable detection thresholds. These various definitions have had to be adapted for adolescents [Cook 03, De Ferranti 06, Viner 05, Weiss 04]. The prevalence of MetS in children and adolescents is increasing, in parallel with the upward trend in obesity rates. Obese children have a high probability of remaining obese later in life, increasing obesity-related diseases such as diabetes, hypertension, or heart disease. Although MetS has been extensively studied in adults, little information is known about the disease in children and adolescents. There is no consensus on the definition of MetS in children. Several definitions of MetS have been proposed in the literature in this context. The definitions of Cook, De Ferranti, Viner, Weiss, and the IDF are among the most widely used in the literature [Cook 03, De Ferranti 06, Viner 05, Weiss 04].

Despite their widespread use, traditional screening methods have several drawbacks and limitations, including screening uncertainty due to the existence of several definitions. Moreover, the data collected during MetS screening sessions requires a significant financial and time burden, especially for a large population. The main reasons for this high cost are switching to blood testing and acquiring biological variables.

Motivation and objectives

Data science and machine learning (ML) are powerful tools and have shown promising results by exploiting vast data, particularly in the medical field, as components of medical decision support. They provide interactive solutions for predicting the risk of endocrine diseases and identifying high- and low-risk individuals. This ML risk estimation is crucial for doctors with various medical needs and can provide important information for physicians.

For instance, risk prediction of carbohydrate abnormalities for patients with β-TM helps doctors personalize and manage treatment and follow-up plans for patients. Similarly, in the case of MetS, risk prediction helps doctors personalize screening and diagnosis, saving significant time and financial resources, especially for large populations. In addition, a global risk prediction that considers various definitions of MetS for adolescents is also valuable for physicians in improving screening accuracy.

However, applying ML for decision support in the medical sector poses several challenges, citing the lack of explainability and clarity of ML models and predictions for physicians. Especially in assessing the reliability of predictions, doctors need assurance that predictions are reliable and safe. This lack of explainability and confidence prevents the integration of ML-based decision support systems into the medical sector. Even when explainable ML (XAI) is applied, its reliability is questioned, particularly with changes in test and training data, as each change alters the explainability provided by XAI.

Moreover, the high cost and time involved in extracting biological and clinical variables for risk prediction make such predictions impractical and increase the difficulty of model integration into hospitals.

To achieve this, the main goal of this work is to tackle the limits identified previously by several research questions (RQ) that will be targeted in this thesis wish are presented below:

RQ1: How can we use ML by exploiting several datasets to predict the risk of the diseases discussed above?

RQ2: How can we explain the risk prediction and give physicians access to evaluate its reliability?

RQ3: How can we ensure less costly risk prediction?

RQ4: How can we improve and evaluate the reliability of the explainability provided by XAI?

The main objective of the thesis is to predict the risk of carbohydrate anomalies for patients with β-TM and the risk of MetS in adolescents. The idea is to provide information on identified high- and low-risk individuals to doctors for significant medical decision support. In response to the limitations of ML, our goal is to study the XAI to explain ML and its predictions and give doctors access to evaluate the reliability of the predictions of carbohydrate anomalies. We also aim to use XAI to select clinical rather than biological characteristics for risk prediction to reduce the financial burden of risk prediction in MetS screening. Finally, we aim to study the reliability of XAI and propose an approach to improve and evaluate this reliability.

Figure 1 illustrates several objectives of the thesis in a structured manner according to chapter divisions.

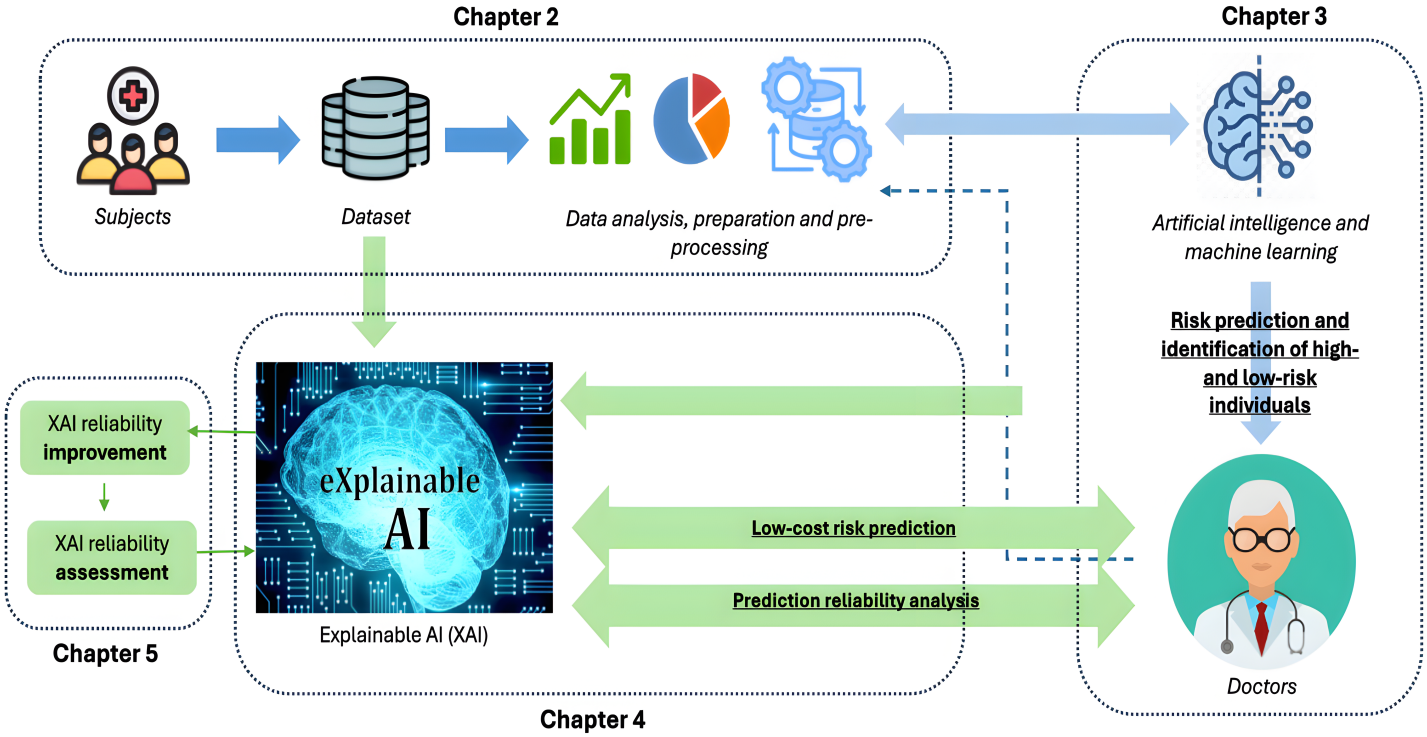


Figure 1 – Thesis objectives and chapter organization

Thesis Outline

The following parts of this thesis manuscript are organized as follows:

— **Chapter 1** is an introductory chapter that presents primarily surveys the literature within our research scope and then identifies the limitations of existing research activities. Based on these limitations, the research questions and challenges targeted in this thesis are presented. Hence, this chapter defines data science and artificial intelligence (AI). Next, studying the state of the art in a wide scope regarding the application of data science and ML for medical decision support. Finally, the research scope is refined to study activities that predict the risk of carbohydrate and MetS abnormalities, identifying limitations and defining research questions and challenges.

— **Chapter 2** is dedicated to analyzing and preparing the datasets used in the thesis. Hence, it begins by defining the statistical and visualization approaches for data analysis and pre-processing. Next, analyze and prepare the five datasets used in the thesis, each involving a discussion of the approach used to improve data quality and an argument for this choice. Finally, compare the data quality of the various datasets and discuss the difference between private and public datasets.

— **Chapter 3** presents the methodology and results for predicting the risk of carbohydrate abnormalities in patients with β-TM and the risk of MetS in adolescents using ML. Hence, it starts by defining supervised ML and the various classification models with metrics designed to evaluate this classification. Then, it presents the methodology for risk prediction with results. Finally, it presents the limitations of our approaches, which are mainly related to the lack of explainability of the prediction of carbohydrate abnormalities and the significant financial burden of MetS risk prediction.

— **Chapter 4** proposes solutions to the limitations and research questions posed in the previous chapter. The focus is on studying Explainable Machine Learning (XAI) to provide physicians with tools to evaluate the reliability of risk predictions for carbohydrate abnormalities and select features to reduce the financial and time burden of MetS screening. The chapter begins by defining XAI and exploring its several types. It then presents the methodologies and results of XAI for assessing risk prediction reliability and reducing financial burdens. While XAI has produced promising results, the reliability of its explainability has been challenged. Each time the training and test data changes, the explainability varies, making XAI validation unreliable.

— **Chapter 5** includes a response to the unreliability of XAI discussed in the previous chapter. This chapter presents a new approach for improving XAI reliability by combining it with a data augmentation technique. We then define and develop metrics to evaluate the reliability of XAI after this combination. The proposed approach is tested on two public datasets (Hypothyroidism and Diabetes) and validated on the three private datasets related to our thesis objectives (Glucidic anomalies and MetS).

Finally, we conclude this research by summarizing the contributions, discussions, limitations, and perspectives.

Figure 2 presents each chapter's thesis structure, research questions, objectives, and discussion.

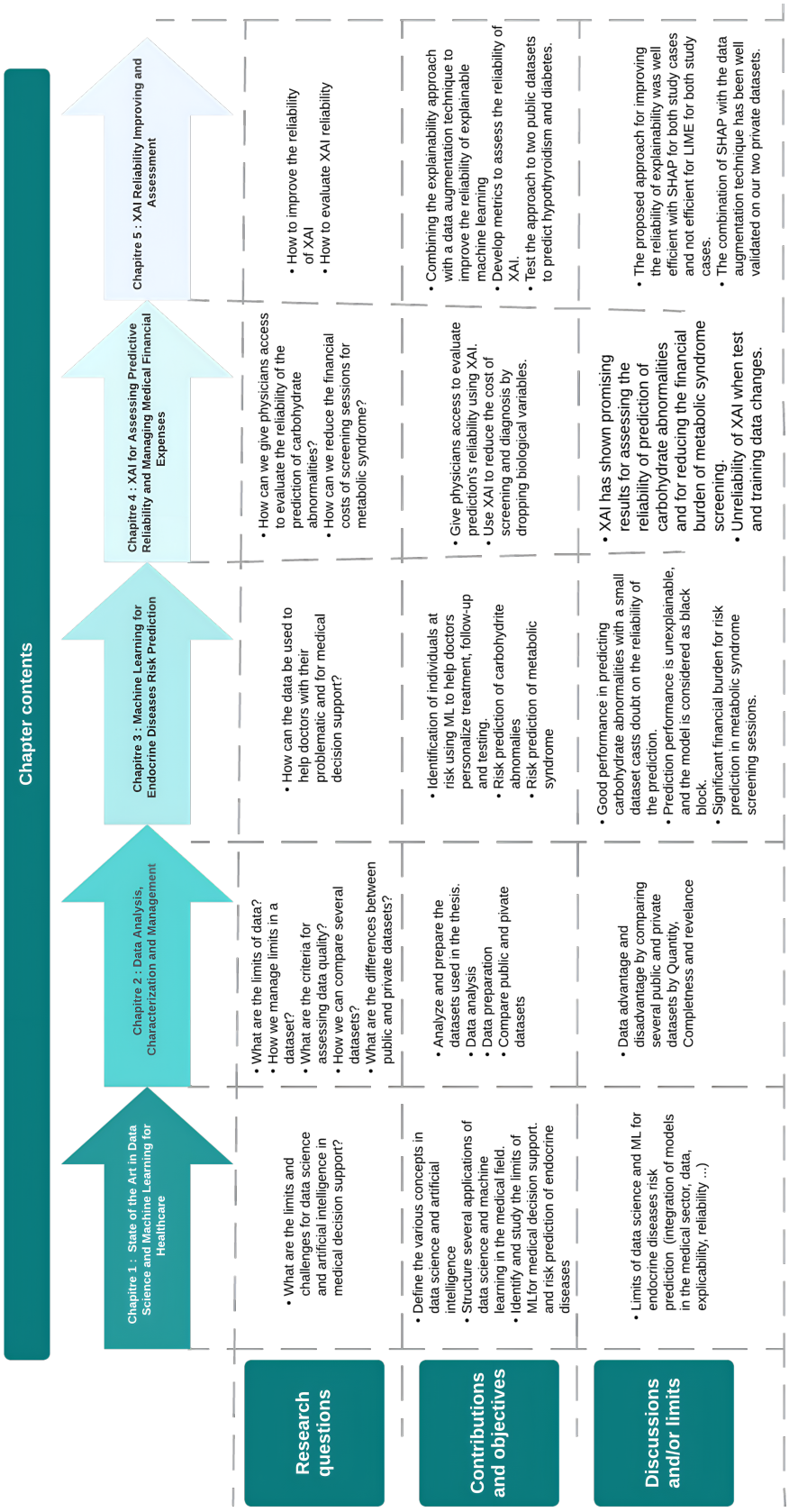


Figure 2 – Thesis outline and chapter contents

Publications

— **International Conference:**

F. Ketata, Z. Al. Masry, N. Zerhouni and S. Yacoub, "Explainable Machine Learning Approach with Augmentation for Mortality Prediction," 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET), Hammamet, Tunisia, pp. 01-06, doi:10.1109/IC_ASET58101.2023.10150509, 2023

— **Journal Papers:**

F. Ketata, Z. Al. Masry, S. Yacoub and N. Zerhouni, "A Methodology for Reliability Analysis of Explainable Machine Learning: Application to Endocrinology Diseases," *IEEE access*, doi:10.1109/ACCESS.2024.3431691, 2024

F. Ketata, S. Khensal, Z. Al. Masry, S. Yacoub, N. Zerhouni and K. Benmohamad, "Risk prediction of carbohydrate abnormalities in patients with β-Thalassemia major," *Computers in Biology and Medicine*, submitted.

K. Benmohamad, F. Ketata, Z. Al. Masry, N. Zerhouni and P. Valensi, "A Generalization and Low-Cost Approach for Metabolic Syndrome Risk Prediction in adolescents Using Explainable Machine Learning," *JOURNAL OF CLINICAL ENDOCRINOLOGY & METABOLISM*, submitted.

Chapter I

State of the Art in Data Science and Machine Learning for Healthcare

I.1 Introduction

Data science and ML have revolutionized many areas, including medicine. These disciplines offer significant potential for improving medical decision-making, especially in the endocrinology service. In this chapter, we explore the current state of the art in data science and ML research applied to endocrine diseases and aim to identify and present limits related to this field.

This chapter begins with a general overview of data science and ML in Section I.2. Then, in Section I.3, we examine how these techniques are used for screening, diagnosis, prognosis, and treatment for medical decision support. Subsequently, we address the current and potential impact of ML in our society, highlighting the specific challenges and opportunities associated with endocrine diseases in Section I.4. In the same section, we then present a review of the literature relating only to the risk prediction of the diseases targeted in the thesis. We also establish the foundational problems and limits in Section 1.5 that we seek to address in this thesis. Finally, Section I.6 concludes the chapter.

I.2 Brief overview of Data Science and Artificial Intelligence

Data science is an interdisciplinary field that involves collecting, cleaning, exploring, and modeling data to extract useful knowledge. This discipline combines concepts and techniques from statistics, mathematics, computer science, and other related fields. Its goal is to transform raw data gathered from various sources into actionable information for decision-making, problem-solving, and discovering new knowledge across a wide range of application areas [Nasution 23]. Figure I.1 summarizes the different concepts related to data science.

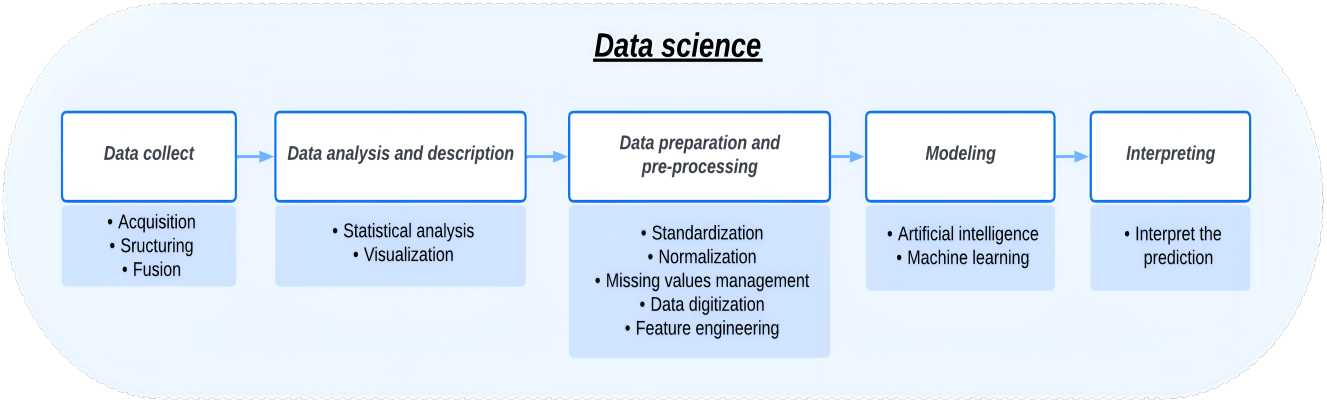


Figure I.1 – Data science process

I.2.1 Data Characterization, Analysis and Preparation

Analyzing and preparing data are essential steps to making ML work effectively. By understanding the data through analysis, we can choose the best ways to prepare it for ML. This ensures that the data is well-prepared and can be used effectively. We decide which steps in data pre-processing are necessary and which are helpful additions to improve the data for better ML results.

Before describing the various steps involved in data analysis, we discuss the various data types and sources found in the literature, specifically focusing on medical data.

Various data types are available for analysis and ML in the medical field. These data types presented in Table I.1 can be categorized into several groups based on their nature and how they are obtained.

Data type	**Data Content**	**Source**	**Objectives of ML**	**Reference**
Images data	MRI, mammography, ultrasound, thermography	Cameras	Visual diagnosis, screening, prediction	[Castiglioni 21]
Tabular data	Excel, CSV, TXT files; biological and clinical data	Screening sessions, blood tests, patient follow-up records	Risk prediction, diagnosis, personalized treatment	[Hernandez 22]
Time-series data	Sensor data, patient follow-up records	Sensors, follow-up records	Future event risk prediction	[Bock 21]

Table I.1 – Summary of Data Types, Sources, Content, and Objectives

Medical data is available from public sources such as Kaggle, the UCL ML repository, and data.gov. It can also be obtained through collaboration with hospitals to get private datasets. Private datasets are particularly interesting and often linked to real medical problems. However, they are generally raw and untreated. Therefore, it is important to consider various pre-processing and data preparation types to prepare them for exploitation.

Data Analysis and Description

Data analysis is the process of inspecting data to discover useful information. Data analysis has multiple facets and approaches, such as visualizations and statistical analyses.

Data visualization is a method for exploring and understanding the structure and relationships within medical datasets. Visualization techniques enable the graphical representation of data, facilitating the identification of patterns, trends, and anomalies. Bar and pie charts visually represent categorical data distributions, while box and violin plots display the distribution of numerical data and any outliers present. Scatter plots reveal relationships between two variables, with each data point representing a single observation. Each visualization approach offers unique insights into several aspects of the data, allowing researchers to uncover valuable information for further analysis and decision-making.

In addition to visualization approaches, statistical tools provide formal methods for analyzing and interpreting medical data. These tools allow for identifying relationships between variables, estimating model parameters, and testing hypotheses. Hypothesis tests, such as the t-test [Kim 15], ANOVA [St 89], and Chi-square test [Sharpe 15], are used to assess the significance of differences between groups or variables. Linear regression [Montgomery 21] is commonly employed to model the relationship between a dependent variable and one or more independent variables, while logistic regression is used for binary outcomes. Analysis of variance (ANOVA) assesses the variability between groups in

a dataset. Correlation analysis quantifies the strength and direction of relationships between continuous variables. Survival analysis evaluates the time until an event of interest occurs. Time series analysis [Hamilton 20] examines data collected over time to identify patterns and forecast future values. Clustering methods, such as K-means [Sinaga 20] and hierarchical clustering [Nielsen 16], group similar observations together based on their characteristics. Each statistical tool provides unique insights into the underlying patterns and relationships within the data, enabling researchers to draw meaningful conclusions and make informed decisions. Figure I.2 shows an overview of approaches to data analysis.

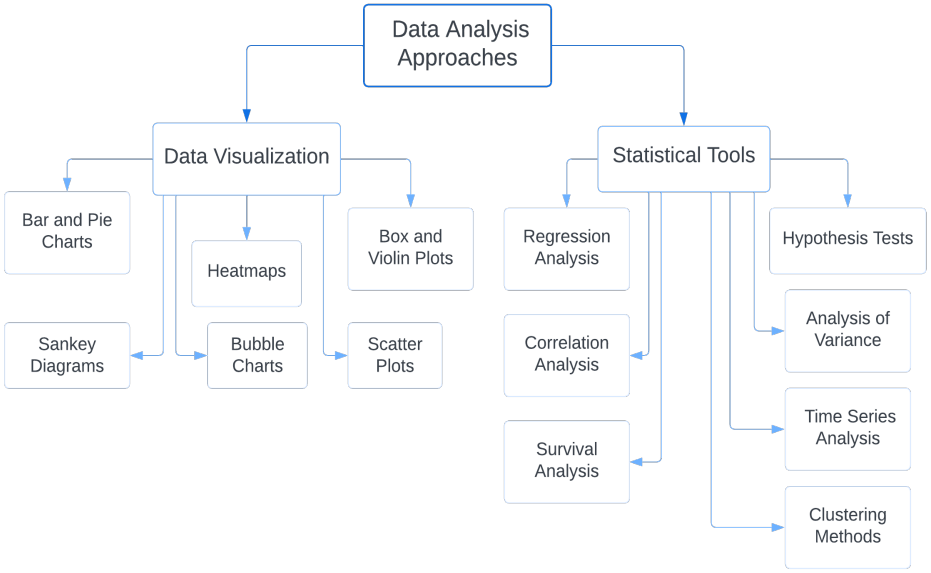


Figure I.2 – Data analysis approaches

Data Preparation and Pre-processing

Using interpretations from data analyses, we aim to identify which data pre-processing must be done on the data, ensuring that it is clean, consistent, and well-structured for subsequent steps. This part explores several approaches used to prepare and preprocess data before applying analytical techniques.

Data cleaning is a fundamental step in preparing datasets for analysis, focusing on identifying and rectifying errors, inconsistencies, and missing values. This process encompasses techniques, including handling missing values, outlier detection, removal, and data imputation. Handling missing values involves employing strategies such as deleting records with missing values, imputing missing values using statistical methods, or utilizing predictive models to estimate missing values. Outlier detection and removal are crucial in identifying and eliminating data points that significantly deviate from the rest of the dataset, potentially indicating errors or anomalies. Additionally, data imputation is employed to fill in missing values based on the available information within the dataset.

As for data transformation, which involves converting the dataset into a suitable format for analysis. Techniques for data transformation include feature scaling and feature engineering. Feature scaling standardizes the range of features in the dataset, ensuring that they have a similar scale and distribution, which is essential for many ML algorithms. Feature engineering involves creating new features or transforming existing ones to improve ML models' performance. This may include encoding categorical variables, creating interaction terms, or extracting relevant information from existing features.

— **Feature Scaling:** Feature scaling is a technique used to standardize the range of features in a dataset. It ensures that all features have a similar scale and distribution, which is crucial for many ML algorithms. Common methods of feature scaling include standardization (subtracting the mean and dividing by the standard deviation) and normalization (scaling features to a specified range, such as [0, 1]).
— **Feature Engineering:** Feature engineering involves creating new features or transforming existing features to improve the performance of ML models. This may include encoding categorical variables using one-hot or label encoding techniques, creating interaction terms by combining existing features or extracting relevant information from existing features, such as text or image data.
— **Dimensionality Reduction:** Dimensionality reduction is a technique used to reduce the number of features in a dataset while preserving as much information as possible. This is particularly useful when dealing with high-dimensional datasets, as it can help improve the performance of ML models and reduce computational complexity. Common methods of dimensionality reduction include principal component analysis (PCA) [Kurita 19] and linear discriminant analysis (LDA) [Tharwat 17].

Figure I.3 shows the overview approaches to data analysis.

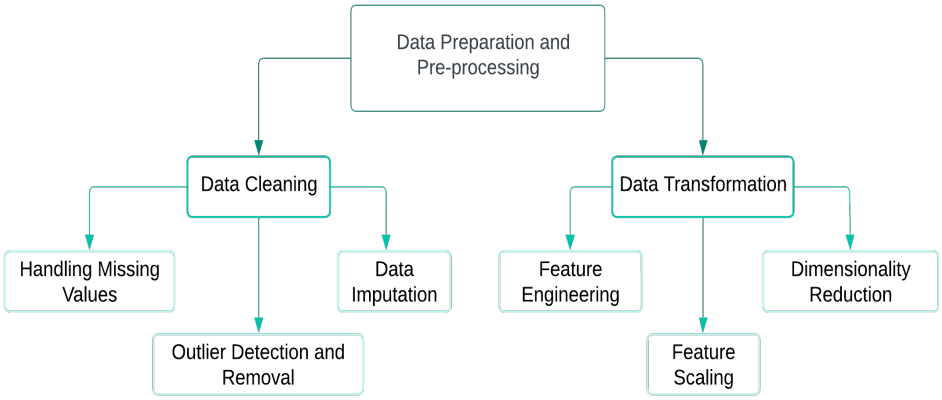


Figure I.3 – Data preparation approaches

I.2.2 Machine Learning Models Category

Once the data has been well-prepared, we move on to the stage of data exploitation to ensure AI with learning. AI and ML are foundational technologies in analyzing complex datasets and making data-driven decisions. This part overviews various ML and AI techniques, categorizing them into supervised, unsupervised, semi-supervised, and reinforcement learning methods, as shown in Figure I.4.

There are four main types of learning for AI. The most widely used type is supervised learning, which involves training a model on a labeled dataset, where the desired output is known, to make predictions or decisions. Let us explore several common techniques in supervised learning. We divide supervised ML models into three types. First, linear models are a class of algorithms that assume a linear relationship between input features and the target variable, Such as Linear Regression [James 23], Logistic Regression [Das 21a], Support Vector Machines (SVM) [Pisner 20], and Linear Discriminant Analysis (LDA) [Zhu 22]. Then, tree-based models wish to partition the input space recursively into regions, making decisions based on the input feature values. For example, the decision tree

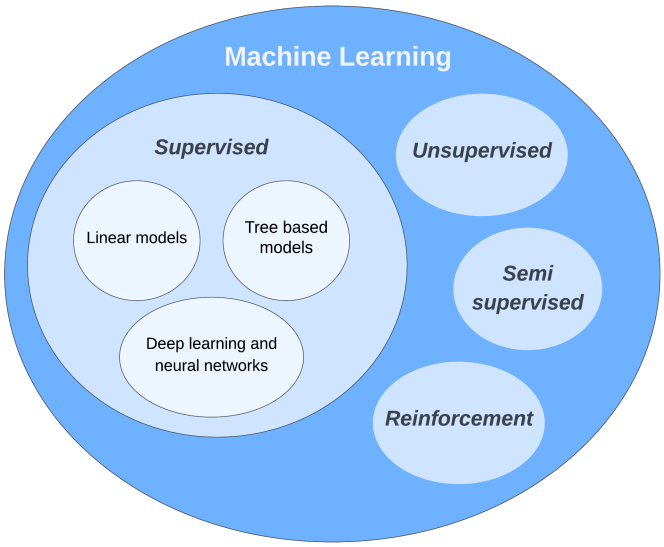


Figure I.4 – Machine learning types and models

model [Jo 21], Random forest [Rigatti 17], XGBoost [Chen 15], CatBoost [Hancock 20] and LightGBM [Rufo 21]. Finally, deep learning and neural networks are subsets of ML that utilize neural networks with multiple layers to learn complex patterns in data. Neural networks are computational models inspired by the structure and function of the human brain, composed of interconnected nodes organized in layers [LeCun 15].

While, unsupervised ML involves training models on unlabeled data to discover patterns, structures, or relationships within the data. Unlike supervised learning, unsupervised learning does not require labeled output, making it useful for clustering, dimensionality reduction, and anomaly detection [Hahne 08, Ayed 24].

As for semi-supervised ML, this is a hybrid approach combining elements of supervised and unsupervised learning. It leverages a small amount of labeled data and a larger amount of unlabeled data to improve model performance [Zhou 21b].

Finally, reinforcement learning is a type of ML that involves an agent interacting with an environment to learn a policy that maximizes cumulative rewards over time. It is inspired by behavioral psychology, where agents learn to make sequential decisions through trial and error [Wiering 12].

I.3 Data Science and Machine Learning for Medical Decision Support

ML has revolutionized the medical field, opening new horizons for medical decision-making support [Shehab 22]. By mining vast medical datasets, ML enables healthcare professionals to derive valuable insights and personalized recommendations [Varoquaux 22]. ML can detect complex patterns and recognize relationships between several medical attributes using sophisticated algorithms, which is often difficult for humans to do [Janiesch 21].

The contribution of AI to medical decision support can be dedicated to risk prediction, early detection, accurate diagnosis, or the most suitable treatment. This can provide information and instructions to the doctor to make the right decision regarding a diagnostic

result, an additional test after screening, or patient-specific treatment and follow-up for several medical services such as Endocrinology, Radiology, Cardiology, Neurology, and oncology.

I.3.1 Screening

The main challenge of ML in disease screening is the earlier and more accurate detection of various medical disorders [Zuluaga-Gomez 19]. By exploiting the vast datasets, these techniques can identify specific patterns and markers associated with specific diseases. ML can also be used to develop predictive models that assess individual risk of developing a disease using information such as medical history, genetic characteristics, lifestyle habits, and environmental risk factors [Benmohammed 22]. These predictions can help healthcare professionals identify individuals at higher risk and set up targeted screening programs, enabling early intervention and appropriate diagnosis.

Moreover, data science enables advanced analysis of large quantities of medical data, which can reveal subtle patterns and significant correlations between variables. For instance, analysis of laboratory data and vital signs can help identify early indicators of diseases such as cancer, heart disease, and endocrine disorders [Lassoued 18]. By integrating ML techniques, it is possible to develop screening models that combine various clinical and imaging data to improve the sensitivity and specificity of early diagnosis [Ma 19]. According to physicians, applying ML and data science in disease screening provides more accurate diagnostic tools, reduces errors, and improves clinical outcomes. Identifying early signals of disease and enabling early intervention help save lives and improve patient's quality of life by promoting more effective, better-targeted treatments.

I.3.2 Diagnosis

While screening primarily focuses on early predictive capabilities, diagnosis emphasizes the use of ML to provide novel insights and substantial enhancements in the accuracy and efficiency of medical diagnoses [Bohr 20]. Analyzing medical data, including symptoms, medical history, clinical test results, and imaging data, enables the identification of complex patterns and accurate classifications to be made [Bohr 20]. ML can be used to develop diagnostic models that learn from large amounts of data and recognize distinctive patterns associated with different diseases. These models can be trained to detect subtle signs or combinations of features that sometimes escape the human eye [Oh 18]. As an illustration, in the field of medical imaging, techniques enable the automated analysis of radiological images, such as X-rays, CT scans, ultrasounds, and MRIs [Lassoued 18, Ma 19, Bohr 20].

This facilitates detecting and localizing tumors, lesions, and structural anomalies, helping radiologists and clinicians formulate more accurate and rapid diagnoses. In addition, integrating ML with other medical data, such as blood test results or genetic information, enables the development of more comprehensive and holistic diagnostic models. These models can help assess the probability of the prevalence of a specific disease and guide healthcare professionals toward appropriate investigations and treatments. By improving the accuracy and speed of diagnoses, ML and data science contribute to more informed medical decision-making, helping to optimize care and reduce diagnostic errors [Esteva 21]. Ultimately, these techniques offer considerable potential to improve clinical outcomes, reduce diagnostic delays, and promote earlier and more effective treatments, leading to significant patient healthcare improvements.

I.3.3 Risk Prediction

Identifying high-risk individuals and taking appropriate preventive measures is the main objective for risk prediction using ML. Through in-depth medical data analysis, these techniques enable the development of predictive models that assess individual risk of developing a specific disease [Huang 23]. Using complex, multidimensional datasets, including information such as medical history, risk factors, genetic data, and lifestyle habits, ML algorithms can identify patterns, trends, and associations that often escape human observation [Huang 23, Barragán-Montero 21, Aggarwal 21].

These predictive models enable identifying individuals at higher risk and implementing early preventive interventions, such as lifestyle changes, regular medical monitoring, or prophylactic treatments. In addition, ML and data science can also be used to predict the progression of chronic diseases, such as diabetes or certain types of cancer, by integrating longitudinal data on patients' health [Peng 21]. This makes it possible to tailor treatment and monitoring strategies to individual risk, leading to more targeted and effective healthcare. In summary, ML and data science offer considerable potential for disease prediction, enabling early identification of at-risk individuals, personalized preventive interventions, and overall health outcomes.

I.3.4 Treatment

ML and data science significantly contribute to the disease treatment field, offering innovative possibilities for therapy optimization and clinical decision-making [Kolluri 22]. These techniques enable valuable information to be extracted from large medical datasets, including clinical data, test results, treatment responses, and genetic data [Hall 23]. Using this information, ML can be used to develop predictive models and decision support systems that help healthcare professionals choose the most effective and personalized treatments for each patient [Sarker 21].

These models can help identify sub-populations of patients who will benefit most from a specific treatment, enabling a more targeted and individualized approach. In addition, ML can also be used to optimize treatment protocols by analyzing clinical data and identifying the most effective treatment regimens [Zhang 22]. This includes adjusting drug doses, optimizing treatment schedules, and predicting potential side effects. Integrating ML with real-time data makes it possible to monitor patients' responses to treatments and adapt therapies accordingly, promoting more favorable outcomes. In brief, ML and data science open up new possibilities for disease treatment, helping select the most effective therapies, personalize treatment protocols, and improve patient clinical outcomes. These technological advances enable a more precise and individualized approach, paving the way for precision medicine and improved healthcare.

I.4 Machine Learning in our World Today and Prospects for Endocrine Diseases

Endocrine disease is a medical condition affecting the endocrine system, a network of glands that produce and release hormones into the bloodstream. Hormones are chemical messengers that regulate various bodily functions, such as metabolism, growth, reproduction, and mood.

Endocrine diseases can result from hormone imbalances, dysfunction of specific endocrine glands, or problems with hormone receptors. Common endocrine diseases include diabetes, thyroid disorders, adrenal gland disorders, pituitary gland disorders, and reproductive hormone disorders.

Figure I.5 shows the several endocrine organs and the diseases associated with each organ.

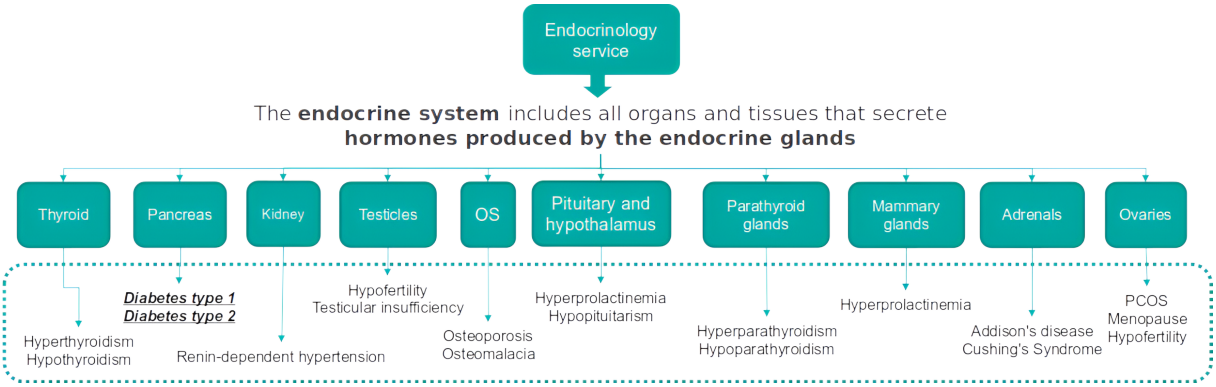


Figure I.5 – Endocrine organs and diseases

Endocrine diseases can have complex interactions with other health conditions, leading to potential complications. As a demonstration, individuals with certain endocrine disorders may have an increased risk of developing cardiovascular diseases. Hypothyroidism, for instance, has been linked to an elevated risk of cardiovascular issues.

In addition, for patients with β-TM, a major risk is presented for having endocrine diseases. The same goes for the link between obesity and MetS.

Furthermore, there is a well-established correlation between type 2 diabetes and obesity, with obesity being a significant risk factor for the development of diabetes. Similarly, both hyperthyroidism and hypothyroidism can contribute to metabolic imbalances that may lead to weight-related issues, including obesity.

A risk prediction of an endocrine disease using ML by exploiting data is important information for physicians to help manage patient complications and personalize follow-up, treatment, or diagnosis.

In this section, we will present the literature review within the scope of our research. It involves exploiting tabular biological and clinical data to predict the risk of specific endocrine diseases using ML. Risk prediction and identifying high-risk and low-risk individuals have a specific context for each disease area. Mainly, we focused on the risk prediction of carbohydrate abnormalities for patients with β-TM and the prediction of MetS for adolescents. Also, the prediction of hypothyroidism and diabetes, since we will be using two public datasets targeting these two diseases.

I.4.1 Carbohydrate Anomalies Risk Prediction in Patients with β-TM

Beta thalassemia, a hereditary condition causing hemolytic anemia due to mutations in the hemoglobin genes, is prevalent in regions around the Mediterranean. The homozygous form, termed β-TM or transfusion-dependent thalassemia, results in significant hemolysis in early childhood, necessitating transfusions, and long-term therapy to remove excess iron from the body. These transfusions cause a buildup of iron in various organs such as the heart, liver, spleen, and endocrine glands, with the liver typically exhibiting the highest levels of iron overload, followed by the pancreas and the heart

[Kattamis 22, Ali 21, Carsote 22, Sevimli 22]. Managing iron levels through ICT and ensuring treatment adherence is crucial to mitigating complications associated with iron overload and reducing morbidity and mortality. Monitoring iron levels is typically done through serum ferritin levels and magnetic resonance imaging to assess liver iron concentration and cardiac iron load using T2-weighted MRI, aiding in evaluating the effectiveness of chelation therapy or determining the risk to end organs [Sevimli 22]. Despite these measures, complications arising from iron overload are rising, including endocrine issues and disturbances in glucose metabolism [De Sanctis 23]. Diabetes, in particular, poses a significant risk as it often becomes irreversible and is associated with other life-threatening complications. Therefore, early detection of carbohydrate metabolism abnormalities preceding the onset of diabetes is crucial.

In β-TM patients, dysglycemia develops progressively, spanning from normal glucose tolerance to impaired glucose tolerance and, in some cases, diabetes mellitus, occasionally requiring insulin-dependent diabetes [De Sanctis 22]. Pre-diabetes manifests as impaired fasting glucose or impaired glucose tolerance, both detectable through standardized oral glucose tolerance tests [Care 23]. These glycemic irregularities typically surface during the second and third decades of life in β-TM patients [Gomber 18]. Following international guidelines, individuals with β-TM should undergo yearly screenings for glucose abnormalities starting at 10 [Farmakis 22].

In this context, ML models were developed for risk prediction of carbohydrate anomalies in patients with β-TM. Authors in [Yousefian 17] utilized the k-nearest neighbors (KNN) algorithm and radial basis function network (RBFN) on the "ZAFAR" database consisting of 255 Iranian patients diagnosed with β-TM and intermediate. They aimed to forecast the likelihood of diabetes in individuals with major and intermediate beta-thalassemia. The findings indicated an accuracy of 81.70% for RBFN and 69.12% for KNN. Subsequently, in [Yousefian 19] authors employed a multilayer perceptron (MLP) on the same dataset. Their results demonstrated that the application of MLP yielded an improved accuracy of 89.48% compared to the earlier models.

I.4.2 MetS Risk Prediction in Screening Sessions

MetS represents a precursor condition associated with severe ailments like type 2 diabetes and cardiovascular diseases. It manifests through increased waist circumference, elevated blood pressure, abnormal fasting blood glucose, insulin resistance, and dyslipidemia [Eckel 05]. The prevalence of MetS is escalating globally, particularly in both developed and developing nations, correlating closely with obesity and insulin resistance [Bitew 20, Reisinger 21]. Defined by the presence of multiple risk factors such as abdominal adiposity, hypertension, elevated triglycerides, reduced high-density lipoprotein cholesterol (HDL-C), and hyperglycemia, MetS predisposes individuals to various comorbidities and heightens the risk of developing T2D and cardiovascular events.

However, uncertainties persist regarding the definition and management of MetS in younger populations [Magge 17]. Despite this, numerous studies have established a link between MetS, diverse comorbidities, and the likelihood of cardiovascular complications [Liu 21, Koskinen 17]. Consequently, early screening for metabolic disturbances is crucial for assessing current complications and averting future health issues.

Yet, screening for MetS poses challenges due to the plethora of definitions utilizing different percentile thresholds. Ford and Li identified over 40 definitions for children and adolescents in 2008 [Ford 08]. These definitions often employ varying percentile thresholds, with thresholds for abdominal adiposity, hypertension, hypertriglyceridemia,

and low HDL-C varying depending on age, gender, and height. Among these, definitions established by the International Diabetes Federation (IDF) in 2007 [Zimmet 07], Cook [Cook 03], and De Ferranti [De Ferranti 06] are commonly referenced in the literature.

Diverse definitions contribute to uncertain screening when variations arise between them. Additionally, the extensive biomarker extraction in screening presents a significant challenge, along with the costs, time, and discomfort associated with large-scale screening sessions, particularly for children. Consequently, using digital tools, particularly ML, to predict MetS risk may offer valuable insights to specialists who can identify at-risk individuals during screening.

In response to the rising prevalence of MetS, non-invasive predictive studies have emerged, aiming to detect and prevent MetS at an early stage through simple means, thereby avoiding invasive measures [Huh 21, Xu 22, de Kroon 08]. These predictive models rely on external data and do not necessitate invasive procedures. They predominantly focus on anthropometric and lifestyle parameters, with some studies attempting to identify relevant factors in lifestyle-related information, considering gender as a crucial factor [Gutierrez-Esparza 21]. While certain models excel in real-life scenarios, providing high specificity but lower sensitivity, others achieve good results but are challenged by interpretation due to model complexity [Romero-Saldaña 16, Datta 19]. Notably, interpretable scoring systems based on artificial neural networks (ANNs) have been developed, enhancing transparency in risk assessment [Benmohammed 22, Shin 23].

I.4.3 Hypothyroid Risk Prediction

Thyroid disease is the most common endocrine disease worldwide, second only to diabetes, according to the World Health Organization [Biondi 19]. The most common thyroid gland disorders, hyperthyroidism and hypothyroidism, have been identified in over 110 nations worldwide, impacting 1.6 billion people. Most of these cases are in Asia, Africa, and Latin America [Alam Khan 02]. At this time, more than 25,000 emergency clinics worldwide collect information regarding patients in different configurations. However, studies use the time-consuming and expensive traditional approach [Sonuç 21] using traditional examination and quantifiable testing. According to medical professionals, early illness diagnosis and treatment are essential to halt the progression of a disease or even death. Even with much trial and error, clinical prediction is often considered a challenging task [Mir 20]. The thyroid is a small gland near the neck's base, located directly below Adam's apple [Bhaladhare 21]. Numerous bodily functions are regulated by the complex network of glands that make up the endocrine system. The thyroid gland secretes hormones that control metabolism in humans. While iodine deficiency is the most frequent cause of hypothyroidism, there are other potential causes as well [Knudsen 02]. The thyroid gland's hormones are T3, T4, and calcitonin, with T3 and T4 being the most literal forms [Garg 16].

ML is a computer science field that has recently gained popularity in medical applications. Thus, ML could be used to forecast thyroid disease early. Several benefits of an ML algorithm include high speed, self-learning, and fault tolerance to noise [Jordan 15]. Massive volumes of data (big data) that would normally be too complex or impossible to process may now be analyzed by people using ML. It is possible to forecast hypothyroidism using patient symptoms data and an ML model, a time-saving and money-efficient method to help physicians make the most appropriate decisions. Several ML algorithms available in the literature have been developed and tested for the risk prediction, detection, or diagnosis of hypothyroidism. This application has proven highly effective.

Several approaches for predicting, identifying, and categorizing thyroid diseases have been published in the literature based on the thyroid dataset obtained from the UCI Repository [Dua 17]. The effectiveness of several ML models for categorizing thyroid disease into normal, hypothyroidism, and hyperthyroidism groups was compared in [Razia 18]. The authors acquired the datasets from the ML collection of the University of California, Irvine (UCI). There are 7200 samples in the dataset and 21 attributes for each sample. According to the authors, decision tree (DT) performed 99.23% better than Support vector machine (SVM), Naive Bayes (NB), and multilinear regression (MLR). Nevertheless, little information about data preprocessing is offered to determine whether the outcomes apply to real-time datasets.

Authors in [Shankar 20] suggest using a multi-kernel SVM to categorize thyroid disorders. On UCI thyroid datasets, the multi-kernel SVM achieved 97.49% performance accuracy. The performance is increased, and feature selection is carried out by the upgraded gray wolf optimization. Authors in [Das 21b] used ML methods and specific features to perform multiclass classification. There are four classifications for hypothyroidism. With 99.81% accuracy, random forest (RF) outperformed the KNN, SVM, and DT algorithms. However, the authors did not address how well their suggested methodology for classifying thyroid diseases performed. Still, there is potential for increased performance.

A MLP methods was proposed in [Hosseinzadeh 21] for the classification of thyroid disorders. The accuracy is increased by 0.7% when the MLP is used in conjunction with a group of six networks, as opposed to using only one MLP. While 99% classification accuracy was achieved by MLP on huge dataset samples, deep learning approaches such as MLP require high computer resources to train at a faster pace. Authors in [Mishra 21] used the ML approaches of sequential minimal optimization (SMO), DT, RF, and K-star classifier to predict hypothyroid illness. For this study, a sample size of 3772 unique records is considered. RF and DT outperformed the other two approaches, with accuracy scores of 99.44% and 98.97%.

Authors in [Alyas 22] conducted a comparative study of the ML approaches DT, RF, KNN, and ANN to identify thyroid disease. To predict thyroid disease, the tests were performed using the largest dataset and considered both sampled and unsampled data. RF achieved the highest accuracy of 94.8% in its prediction. Researchers also used deep learning models to predict the classification of thyroid diseases. For example, in [Jha 22], classified thyroid diseases using a deep neural network (DNN) were developed. The UCI dataset was used to evaluate performance. The authors found that DNN could classify thyroid illness with 99.95% accuracy. However, a sizable dataset is needed to fully train the model for performance evaluation. In addition, the deep learning models require greater computer power to train.

The authors in [Chaganti 22] reported a feature engineering approach for ML and deep learning models to predict thyroid disease. The approach used forward feature selection, backward feature elimination, bidirectional feature elimination, and ML-based feature selection using extra tree classifiers. Extensive experiments show that the extra tree classifier-based feature yields the best results, with 99% accuracy and an F1 score when used with the random forest classifier.

I.4.4 Diabetes Risk Prediction

Diabetes, characterized by elevated blood sugar levels and associated organ damage such as kidney failure [Abhari 19], poses significant challenges in diagnosis and management. Computer programs have streamlined the development of IT systems based on clinical data for disease identification, particularly in cases of insufficient insulin production or utilization leading to diabetes [Allalou 16].

Predominantly, the most prevalent type of diabetes in adults is mellitus, with diagnostic criteria including pre-determined factors like impaired fasting glucose or impaired glucose tolerance as per the American Diabetes Association. Diagnosis may also rely on blood glucose levels exceeding 200 mg/dL, assessed through HbA1c, oral glucose tolerance test, or fasting plasma glucose tests [Mujumdar 19]. Juvenile type 1 diabetes, characterized by insulin dependence, arises from insufficient insulin production by beta cells, occurring across age groups from infants to adults [Andrews 95, Hasan 20, Rajendra 21, Chiu 94, Yahyaoui 19]. Improper regulation of glucose levels can lead to various complications such as heart problems, nerve disorders, and kidney-related issues [Cunningham 00].

Type 1 diabetes typically manifests before the age of 30, necessitating insulin dependence for patients. In contrast, type 2 diabetes mellitus, often termed adult diabetes, is non-insulin-dependent and results from decreased insulin secretion by β cells, exacerbated by genetic predisposition, obesity, and unhealthy lifestyles, often emerging in middle age. Type 2 diabetes can also precede gestational diabetes in females and certain ethnic populations, impacting adolescents and children as well. Gestational diabetes is anticipated based on maternal characteristics during pregnancy's later stages and biomarkers throughout gestation, with recent studies highlighting the elevated risk among pregnant women of diverse ethnic backgrounds [Dutta 18]. Modern technology has facilitated the recording of vast amounts of data, enabling the utilization of ML in disease management. Physicians analyze clinical metrics such as blood pressure and temperature iteratively, guiding treatment decisions through refined assessments [Magoulas 99]. Additionally, AI is pivotal in fuzzy-based classification and disease diagnosis using neural networks. Ensembles of artificial neural networks enhance disease diagnosis accuracy, despite the challenges posed by computer-aided comprehension [Huang 07].

In [Arunachalam 22], the SVM algorithm is proposed for diabetes classification. SVM operates in a high-dimensional space through kernel functions, with 14 diverse attributes utilized for classifying diagnosed diabetes, undiagnosed diabetes, pre-diabetes, and non-diabetes cases. Performance analysis uses receiver operating characteristic (ROC) curves and cross-validation functions, with RBF and linear kernel functions emerging as the top-performing classification schemes. Additionally, researchers in [Kononenko 01] examine early diabetes prediction using ML methods, leveraging data from various health organizations. Supervised learning algorithms are employed for classification and comparative analysis based on attributes, with a modified approach applied for feature selection and algorithm mapping. Decision tree and random forest algorithms exhibit superior performance, achieving a high specificity of over 98%. Authors in [Saru 19], an early diagnosis of diabetes is proposed using fuzzy SVM. A dataset comprising eight attributes is collected from the PID database, with data preprocessing applied to all attributes. Feature selection based on F-score optimization is performed to select the most relevant attributes, leading to the classification of diabetes using fuzzy SVM. Additionally, a ML logistic regression model is proposed for analyzing glycated hemoglobin (HbA1c) in type 2 diabetes patients based on clinical datasets.

I.5 Limits and Challenges of ML in Clinical Applications Related to Endocrine Diseases

Table I.2 summarizes the most important articles and those most closely linked to our research perimeter, with their objectives and limitations.

Targeted Disease	Reference	Objective	ML Models	Results	Limitations
Carbohydrate Anomalies	[Yousefian 17]	Diabetes Prediction in individuals with β-TM	KNN, RBFN	Accuracy 81.70% (RBFN), 69.12% (KNN)	Lack of explainability
Carbohydrate Anomalies	[Yousefian 19]	Diabetes Prediction in individuals with β-TM	MLP	Accuracy 89.48%	Lack of explainability
MetS	[Datta 19]	MetS risk prediction	Various ML algorithms; comparative analysis	High accuracy	Model complexity may hinder practical implementation
MetS	[Huh 21]	MetS risk prediction	External data-based predictive models	High specificity, lower sensitivity	Interpretation challenges due to model complexity
MetS	[Xu 22]	Early detection of MetS	Various ML approaches	Moderate sensitivity and specificity	Limited generalizability across diverse populations
MetS	[Benmohammed 22]	Develop interpretable ML models for MetS prediction	Interpretable ML models	Transparency in risk assessment	High cost prediction
MetS	[Shin 23]	MetS risk prediction	ANN	High accuracy	High cost prediction and lack of explainability
MetS	[Shin 23]	Risk prediction of MetS	DT	AUC 0.889	Lack of explainability
MetS	[Mohseni-Takalloo 24]	Risk prediction of MetS	SVM	Accuracy 78.4%	Lack of explainability
Hypothyroid	[Shankar 20]	SVM for thyroid disorder classification	SVM	Accuracy 97.49%	Lack of explainability
Hypothyroid	[Das 21b]	Thyroid diseases classification	RF, KNN, SVM, DT	RF accuracy 99.81%	Lack of explainability
Hypothyroid	[Hosseinzadeh 21]	Thyroid disorder classification	MLP	Improved accuracy	Lack of explainability
Hypothyroid	[Mishra 21]	ML for hypothyroidism prediction	DT, RF	High accuracy	Limited discussion on model interpretability
Hypothyroid	[Alyas 22]	ML for thyroid disease prediction	DT, RF, KNN, ANN	Accuracy 94.8%	Lack of explainability
Hypothyroid	[Jha 22]	DNN for thyroid disease classification	DNN	Accuracy 99.95%	Training complexity
Hypothyroid	[Alshayeji 23]	Risk prediction of hypothyroid	SMOTE	Accuracy 99.5%	Lack of explainability
Diabetes	[Saru 19]	Early diagnosis of diabetes using SVM	SVM	Good F1-Score	Lack of explainability
Diabetes	[Arunachalam 22]	SVM for diabetes classification	SVM with RBF and linear kernels	Good AUC	Lack of explainability
Diabetes	[Sonia 23]	Risk prediction of diabetes	MLP	Accuracy 97%	Lack of explainability

Table I.2 – Summary of ML studies for various diseases risk prediction

Abbreviations : SVM: Support vector machines, DNN: Deep neural network, KNN: k-nearest neighbors, RBFN: Radial basis function network, MLP: Multilayer perception, ANN: Artificial neural networks, DT: Decision tree, RF: Random forest,

Several major challenges hinder progress in integrating AI models to manage endocrine diseases. Firstly, a predictive model applicable in real clinical settings should meet specific medical requirements regarding ethics, explainability, performance, and generalization.

However, these ML models are often seen as black boxes due to their lack of explainability. This lack of clarity limits their practical use in healthcare, especially in medical decision-making. The main concern is: how can doctors trust a prediction made by AI?

Furthermore, despite advancements in AI explainability approaches, critical challenges persist regarding the reliability of the explanations, particularly when changing training and testing data. Stability, concordance, and generalization of explanations limit their utility in varied clinical environments.

In addition, the cost of data acquisition can be high for certain diseases, especially in screening sessions where a large population is targeted.

A holistic approach is necessary to overcome these challenges and fully harness the potential of ML models for managing endocrine diseases. This includes enhancing model explainability, considering the economic and practical aspects of screening and treatment, and ensuring the reliability of explanations provided by AI approaches.

Table.I.3 summarizes the limitations of the literature in predicting the risk of endocrine diseases, the research questions, and the objectives targeted in the thesis.

Table I.3 – Literature limitations and research questions addressed in the thesis

Limits	**Research questions**	**Objectives**
Lack of integration of ML models in the medical sector for carbohydrate abnormalities risk prediction	— How to make ML models ready for integration in the medical sector? — How to give confidence to ML prediction?	— Developing explainable approaches to endocrine disease risk prediction — Give physicians access to evaluate the reliability of prediction using XAI
Significant financial and temporary costs for MetS screening	— Can reducing features using XAI reduce financial and temporary expenses? — How can we reduce the cost and time of screening and treating endocrine diseases using ML ?	Exploit XAI to avoid the transition to biological features acquisition for identifying subjects at low risk and reduce financial and timing costs.
XAI reliability	— How to improve XAI reliability? — How to assess XAI reliability?	— Develop an approach to improve XAI reliability — Develop metrics to assess XAI reliability

I.6 Conclusion

In this introductory chapter, we embarked on a survey through the areas of data science and AI as they intersect with medicine, particularly endocrine diseases. We began by delving into the fundamental concepts of data analysis, preparation, and ML techniques, laying the groundwork for exploring how these methodologies can be applied in medical decision-support systems. Then, we studied the literature review focusing on endocrine diseases and specifically on risk prediction of the diseases addressed in this thesis, such as carbohydrate anomalies in β-TM patients or MetS in screening sessions. By analyzing the literature in this research area, we have identified and presented the limitations and challenges targeted in this thesis. The main limitations and proposed challenges are mainly related to integrating AI models in the medical sector, XAI, building confidence in ML, the high cost of screening and diagnosis, and the reliability of XAI. As we pursue

this thesis, we will explore these intricacies in greater depth, seeking to harness the power of data science and AI for medical advancements and critically evaluate their impact and implications for successful patient care and healthcare systems. In subsequent chapters, we will explore specific methodologies, case studies, and potential avenues for overcoming the challenges highlighted in this chapter. Our ultimate aim is to contribute to the ongoing dialogue on the role of data-driven approaches in revolutionizing the management and treatment of endocrine diseases.

Chapter II

Data Analysis, Characterization, and Management

II.1 Introduction

The previous chapter presented the limits and challenges targeted in the thesis in general terms. We aim to investigate these limitations further and define, develop, and test solutions in depth to achieve our objectives. But first, we must analyze and prepare the data, which is crucial to ensuring the success of a data-driven approach. Therefore, in this second chapter, we delve deeper into the process of data analysis, characterization, and management.

This chapter aims to provide a reliable foundation for the continuation of the thesis, establishing a thorough understanding of the data we use and the methods we employ to analyze and prepare it. By combining this knowledge with advancements in AI and ML, we aim to enhance the management of patients with endocrine diseases, thereby contributing to improved clinical outcomes and a better quality of life.

Section II.2 examines the data itself in detail, examining its source, nature, and significance, particularly from biological and clinical perspectives. We also discuss data quality, statistical analysis methods, and distribution visualization. Next, Section II.3 addresses the pivotal data preparation and preprocessing phase. This includes data transformation, handling missing and outlier values, and standardization and normalization. Additionally, we delve into feature engineering, an essential step in extracting pertinent information from raw data. Finally, Section II.4 presents our findings from the discussed concepts and techniques by analyzing and preparing datasets used in this thesis to predict endocrine disease risks. We review public and private datasets, highlighting each case's challenges and specific considerations. After analyzing and preparing the datasets, we discuss and compare the quality of several datasets at the end of the chapter.

In the sequel, we consider the dataset constituted by X as a matrix presenting the features (INPUT), and y as representing the presence or absence of the disease. In other words, the target to be predicted (OUTPUT), with f() being the prediction function.

II.2 Data Description Methodology

Data analysis and description are foundational to any data-driven research or decision-making process. These steps involve comprehensively understanding the nature, source, and characteristics of the data under investigation.

Firstly, elucidating the source and event behind the data is crucial. This involves identifying where the data originates, whether generated from clinical trials, observational studies, electronic health records, or other sources. Understanding the context in which the data was collected provides valuable insights into its reliability, biases, and potential limitations.

Secondly, categorizing and comprehending the several data types present within the dataset is essential. This may include numerical, categorical, ordinal, or time-series data. Recognizing the data types facilitates appropriate data handling techniques and statistical analyses tailored to each type.

Furthermore, interpreting the data's meaning from biological and clinical perspectives is paramount. For instance, in medical research, variables may represent physiological parameters, biomarkers, patient demographics, or clinical outcomes. Understanding the clinical relevance of these variables is essential for drawing meaningful conclusions from the data.

In addition, data quality assessment is another integral aspect of data analysis and description. This involves evaluating the data's completeness, accuracy, consistency, and timeliness. Data quality issues such as missing values, duplicates, or outliers can significantly impact the validity and reliability of subsequent analyses and interpretations.

Statistical analysis techniques to identify data patterns, trends, and associations. Descriptive statistics, inferential statistics, and hypothesis-testing methods are commonly used to summarize and draw inferences from the data.

Moreover, graphical distribution analysis techniques such as histograms, box plots, and scatter plots are employed to explore the distribution and relationships between variables visually. Visual representations aid in uncovering hidden patterns, outliers, and anomalies within the data, enhancing data understanding and interpretation.

Data analysis and description comprehensively examine the data's source, types, meanings, quality, and statistical characteristics. These foundational steps lay the groundwork for subsequent data preparation, modeling, and decision-making processes.

II.2.1 Data Types and Sources

Understanding the source and event behind the data is paramount in data analysis and description, as it provides critical context for interpreting the collected information. The events serving as data sources can vary depending on the nature of the study or research being conducted. In medical contexts, data may be gathered during screening sessions, follow-up medical visits, or the diagnostic process. Each event represents a unique opportunity to collect pertinent information regarding the health and well-being of the individuals under study.

Moreover, the types of data collected can encompass a wide range of formats, including tabular data, medical images, time series, clinical texts, and more. Each data type presents its own set of characteristics and challenges regarding analysis and interpretation. For instance, tabular data are structured in tables with columns representing several variables, often facilitating the application of ML algorithms. On the other hand, medical images require specialized techniques for processing and analysis, such as image segmentation and classification, to extract meaningful insights.

By comprehending the source and collection context of the data and the types of data utilized, researchers can better interpret the outcomes of their analyses and formulate relevant conclusions. This deep understanding of the data is essential for ensuring the validity and reliability of data-driven studies and guiding clinical decisions and public health policies.

II.2.2 Data Mining, Biological and Clinical Perspective

In medical data analysis, comprehending the significance of the variables within a dataset from biological and clinical perspectives is paramount. Each variable encapsulates crucial information about the individuals under study's physiological, pathological, or clinical aspects. Therefore, understanding these variables' biological and clinical meanings is essential for conducting meaningful analyses and drawing accurate conclusions.

From a biological perspective, variables in medical datasets often represent physiological parameters, biomarkers, genetic markers, or other biological entities relevant to the health condition being investigated. For example, variables may include blood pressure readings, cholesterol levels, genetic mutations, or biochemical markers indicative of disease

progression. Understanding the biological significance of these variables allows researchers to discern underlying biological mechanisms, pathways, and interactions contributing to disease development or progression.

On the other hand, from a clinical perspective, variables may encompass diagnostic criteria, treatment modalities, patient demographics, or clinical outcomes. These variables provide insights into the clinical manifestation of the disease, treatment efficacy, patient prognosis, and overall healthcare management. For instance, variables may include diagnostic codes, medication dosages, surgical interventions, or patient-reported symptoms. Understanding the clinical relevance of these variables enables researchers to assess disease severity, predict patient outcomes, and tailor treatment strategies to individual patient needs.

Moreover, interdisciplinary collaboration between biomedical scientists, clinicians, and data scientists is crucial in elucidating the intricate relationships between biological mechanisms and clinical outcomes. Researchers can leverage their combined expertise to uncover novel insights, develop innovative diagnostic tools, and improve patient care by bridging the gap between basic biological research and clinical practice.

To sum up, comprehending variables' biological and clinical meanings within medical datasets is fundamental for conducting robust data analyses and deriving meaningful insights. This multidimensional understanding empowers researchers to unravel the complexities of disease pathogenesis, identify prognostic markers, and advance personalized medicine approaches for improved patient outcomes.

II.2.3 Statistical Analysis

Statistical analysis is an aspect of extracting meaningful insights from medical datasets. Various techniques are employed to assess relationships, test hypotheses, and interpret the significance of findings. This section provides a detailed overview of correlation analysis, hypothesis testing, and the interpretation of p-values.

Correlation Analysis

Correlation analysis evaluates the strength and direction of the relationship between two or more variables within a dataset [Gogtay 17]. The Pearson correlation coefficient (r) and Spearman's rank correlation coefficient (ρ) are two commonly used correlation coefficients.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{II.1}$$

where x_i and y_i are the individual data points, $\bar{x}$ and $\bar{y}$ are the means of x and y respectively, and n is the number of observations.

Spearman's rank correlation coefficient (ρ) assesses the monotonic relationship between variables and is calculated based on the ranks of the data points.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{II.2}$$

Correlation analysis helps identify variables' associations but does not imply causation. Additional analyses or experimental studies may be needed to establish causal relationships.

Interpretation of p-values

In hypothesis testing, the p-value represents the probability of obtaining results as extreme as or more extreme than the observed data, assuming the null hypothesis is true. A small p-value (typically < 0.05) indicates strong evidence against the null hypothesis [Gibbons 75].

It is important to interpret p-values in the context of the chosen significance level (alpha) and consider potential sources of bias or confounding in the data.

II.2.4 Graphical Distribution Analysis

Graphical distribution analysis is an essential component of exploratory data analysis (EDA) that involves visualizing the distribution and relationships within datasets. Various graphical techniques are employed to gain insights into the underlying patterns and structures of the data.

Histograms

Histograms are graphical representations of the distribution of data values, as shown in Figure II.1, displaying the frequency of observations falling within predefined data intervals or "bins." They are effective tools for visualizing the distributional characteristics of a variable, including its central tendency and spread. Histograms help identify patterns and anomalies in the data, such as outliers or multimodal distributions.

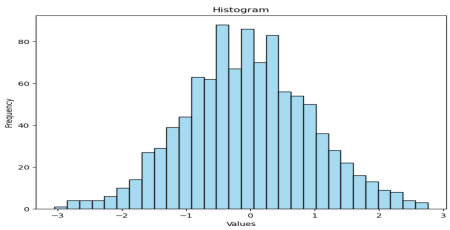


Figure II.1 – Histograms

Boxplots

Boxplots presented in Figure II.2, also known as box-and-whisker plots, offer a summary view of the data distribution by displaying key summary statistics such as quartiles, medians, and potential outliers. They are useful for detecting outliers and comparing the distributions of several variables or groups. Boxplots provide insights into the variability and skewness of the data distribution.

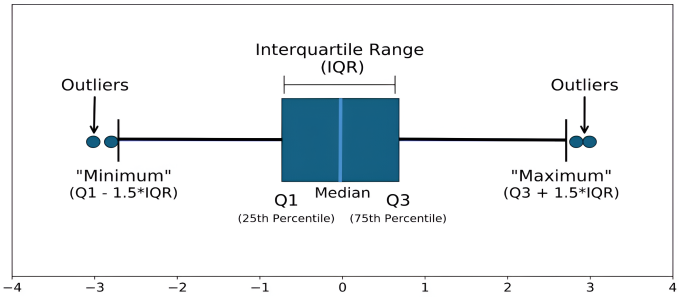


Figure II.2 – Boxplot

Density curve

Density curves, also known as probability density functions, represent the distribution of a continuous variable by showing the relative likelihood of different outcomes. When analyzing the shape of a density curve, we primarily look at its skewness, which indicates the direction and extent to which the distribution deviates from a symmetrical, normal distribution as shown in Figure II.3. A normal, or no skewed, distribution has a bell-shaped curve, where the mean, median, and mode are all aligned at the center. Positively skewed distributions, also known as right-skewed distributions, have a longer tail on the right side, indicating that there are a few exceptionally high values pulling the mean to the right of the median. Conversely, negatively skewed distributions, or left-skewed distributions, have a longer tail on the left side, meaning there are a few exceptionally low values pulling the mean to the left of the median.

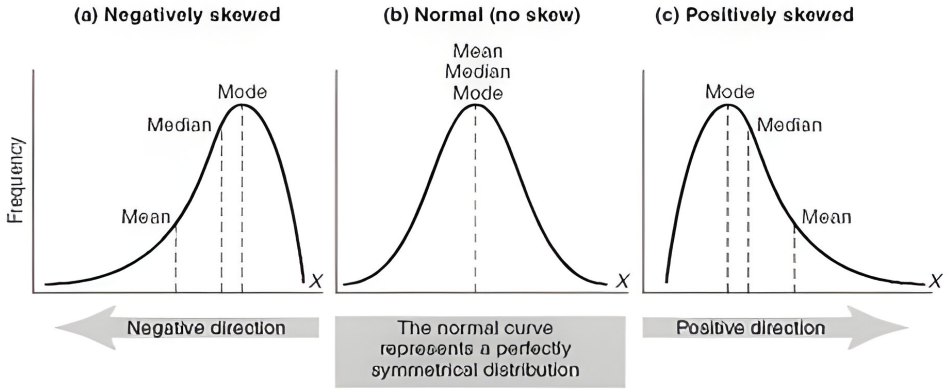


Figure II.3 – Density Curve [Chiniah 16]

Correlation Matrices

Correlation matrices display the correlation coefficients between pairs of variables in a tabular format or heatmap, as shown in Figure II.4. They allow for identifying linear relationships between variables and help detect highly correlated variables. Correlation matrices are useful for feature selection, identifying redundant variables, and understanding the overall dependency structure within the dataset.

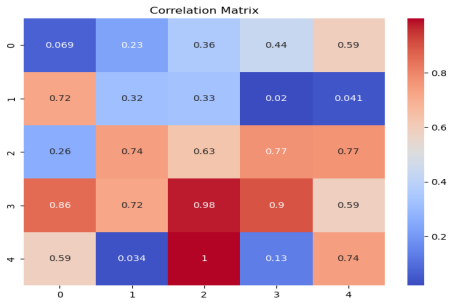


Figure II.4 – Correlation Matrices

Bar Charts

Bar charts displayed in Figure II.5 effectively represent categorical data by displaying bars of proportional lengths corresponding to the frequencies or values of the categories. They provide a visual summary of categorical variables and help compare the relative frequencies or values across several categories. Bar charts are useful for identifying patterns, trends, and outliers in categorical data.

Employing these graphical techniques can help analysts gain valuable insights into their datasets' distributional characteristics and relationships. The appropriate visualization method depends on the nature of the data and the specific research questions under investigation.

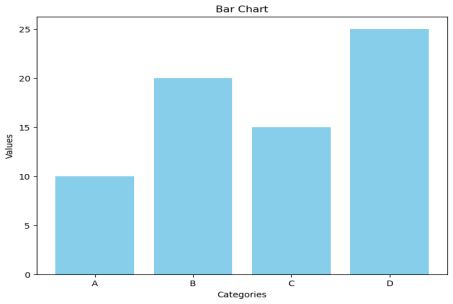


Figure II.5 – Bar Charts

Violin Plots

Violin plots combine the advantages of density curves and boxplots by displaying the data distribution as both a density curve and a boxplot simultaneously, as shown in II.6. They are useful for visually comparing data distributions across several categories or groups. Violin plots provide insights into the distributional characteristics of the data, including skewness, multimodality, and presence of outliers.

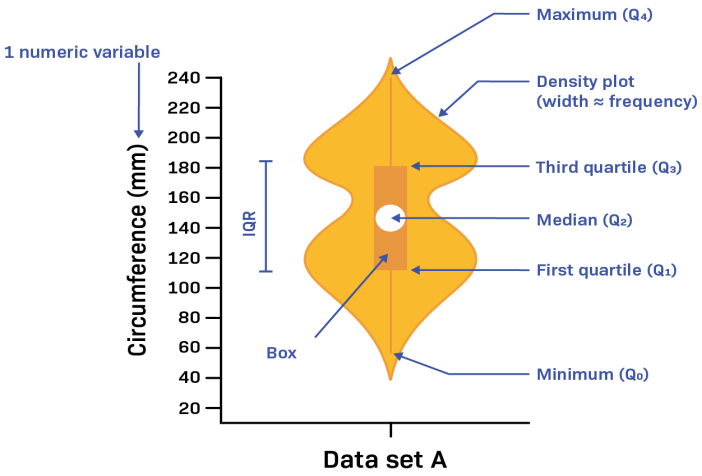


Figure II.6 – Violin Plots [Hameed 24]

II.3 Data Preparation and Pre-processing Tools

Data preparation and pre-processing are essential stages, tightly integrated with the preceding data analysis and description phase. While data analysis provides insights into the structure and characteristics of the dataset, data preparation and pre-processing are pivotal steps to refine the data for further analysis. Data preparation ensures the integrity of the dataset by addressing issues identified during analysis, such as inconsistencies, missing values, or outliers. Moreover, transformations and feature engineering techniques applied during data preparation enhance the dataset's suitability for subsequent analyses, allowing for more accurate modeling and interpretation of results. Thus, these interconnected stages collectively form a comprehensive approach to data exploration and refinement, laying the groundwork for robust and insightful data-driven insights.

II.3.1 Data Digitization

Data digitization is critical in preparing datasets for ML tasks, converting raw data into a format suitable for analysis and modeling. One common transformation technique involves encoding categorical variables into numerical representations, a necessary step as many ML algorithms require numerical input. Label encoding is a technique where each category of a categorical variable is assigned a unique numerical label [Ayed 23]. While simple to implement, label encoding may introduce ordinality where none exists, potentially leading to misinterpretation by the model. Alternatively, one-hot encoding creates binary columns for each category in a categorical variable, with a value of 1 indicating the presence of the category and 0 otherwise. This technique avoids ordinality issues and ensures each category is treated as independent, albeit at the cost of increased dimensionality. Careful consideration must be given to the choice of encoding method based on the data's nature and the ML algorithm's requirements. Data digitization may involve scaling numerical features to a standard range or normalizing them to have a mean of 0 and a standard deviation of 1, ensuring uniformity and comparability across features. These transformation techniques play a crucial role in enhancing the effectiveness of ML models by enabling them to effectively process and learn from the data.

II.3.2 Missing Values Management

Missing values in medical datasets can arise from various factors, such as errors in data collection or entry, equipment malfunction, patient non-compliance, or the absence of measurements for certain variables. These missing values can significantly affect the quality and reliability of datasets, potentially leading to biased analyses and inaccurate conclusions. The impact of missing values depends on their extent and the underlying mechanisms causing their occurrence. Ignoring or mishandling these values can distort statistical estimates, reduce statistical power, and undermine the validity of research findings. Therefore, it is crucial to understand the causes and consequences of missing values and implement appropriate strategies for their management.

A common approach to handling missing values is deletion, where observations with missing values are removed from the dataset. This includes listwise deletion, where entire observations with missing values are excluded, and pairwise deletion, where only the specific variables with missing values are excluded from analyses involving those variables. While deletion methods are straightforward to implement and can prevent bias due to imputation, they can lead to a loss of valuable information and reduced sample size, potentially affecting the generalizability and statistical power of the analysis.

Imputation methods involve replacing missing values with estimated or predicted values based on observed data. Mean imputation replaces missing values with the mean of observed values for that variable, while median imputation uses the median. Regression imputation predicts missing values using regression models based on other variables in the dataset. Imputation preserves sample size and retains valuable information, but it may introduce bias and inaccuracies if the imputation model is unspecified or if the missing data mechanism is not adequately accounted for.

Multiple imputation techniques generate several imputed datasets with plausible values for the missing data, allowing for uncertainty estimation and incorporating variability due to imputation. This approach offers more robust estimates than single-imputation methods but requires more computational resources and may be more complex to implement. Multiple imputation is particularly useful when the missing data mechanism is non-random or when there is significant uncertainty about the missing values.

Predictive mean matching (PMM) is a sophisticated imputation technique for handling missing values in datasets. Unlike simple imputation methods that replace missing values with fixed statistics such as mean, median, or mode, PMM utilizes predictive modeling to estimate and impute missing values based on the relationships observed in the data. The PMM approach involves several steps. First, a predictive model, such as regression or decision trees, is trained on the observed data without missing values. This model learns the complex relationships between the dataset's features and target variable(s). Once the model is trained, it is used to predict the missing values for the instances with missing data.

However, instead of directly using the predicted values from the model, PMM employs a matching technique to select the most similar observed instances (i.e., those without missing values) as donors for imputing the missing values. The predicted value from the model is then replaced with the observed value from the nearest neighbor, ensuring that the imputed values are plausible and consistent with the distribution of the observed data. One of PMM's key advantages is its ability to preserve the underlying distribution and variability of the data while imputing missing values. By incorporating information from similar observed instances, PMM generates more realistic imputations that reflect the true underlying relationships in the data. Additionally, PMM can handle both continuous and categorical variables, making it a versatile imputation technique suitable for various datasets.

Overall, PMM is a powerful approach for handling missing data that leverages predictive modeling and nearest neighbor matching to generate accurate and reliable imputations. Its ability to retain the structure and characteristics of the original data makes it a valuable tool for researchers and practitioners working with incomplete datasets.

II.3.3 Outlier management

Outliers are observations that significantly deviate from the rest of the data and can disproportionately impact statistical analyses. They may arise due to measurement errors, data entry mistakes, natural variability, or rare events. Outliers can distort the distributional characteristics of the data, bias parameter estimates, and affect the robustness of statistical models. Therefore, it is essential to identify and manage outliers effectively to ensure the validity and reliability of data analyses.

Various techniques exist for detecting outliers, including graphical methods, such as scatter plots and boxplots, and statistical methods, such as z-scores, Mahalanobis distance, and clustering algorithms. Graphical methods visually inspect the data for unusual observations, while statistical methods quantify the degree of deviation from the expected

values. Outliers can also be detected using ML algorithms, such as isolation forests and k-nearest neighbors, which identify observations that are significantly different from most of the data.

Once outliers are identified, several strategies can be employed to manage them. One approach is to remove outliers from the dataset by deleting them entirely or treating them as missing values. While this approach can improve the robustness of statistical analyses, it may lead to a loss of valuable information and a reduced sample size. Alternatively, outliers can be transformed using robust statistical techniques, such as winsorization or trimming, which replace extreme values with less extreme ones based on predetermined thresholds. Another strategy is to use robust statistical models that are less sensitive to outliers, such as robust regression or robust estimation techniques. These models downweight the influence of outliers, leading to more stable parameter estimates and improved model performance.

When managing outliers, it is essential to consider the underlying causes and context of the data. Outliers may contain valuable information or represent genuine phenomena that should not be disregarded hastily. Additionally, the choice of outlier management strategy should be guided by the specific objectives of the analysis and the assumptions underlying the statistical model. Sensitivity analyses and robustness checks can help assess the impact of outlier management decisions on the validity and reliability of the results.

II.3.4 Data standardization and normalization

Standardization and Normalization Types

Standardization and normalization are preprocessing techniques used to scale and transform features in a dataset to a common scale, facilitating better performance of ML models.

— **Standardization:** In standardization, also known as z-score normalization, each feature is rescaled to have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean of the feature and dividing by its standard deviation:

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma} \tag{II.3}$$

where x is the original feature value, μ is the mean of the feature, and σ is the standard deviation.

— **Normalization:** Normalization scales each feature to a range between 0 and 1. One common normalization technique is min-max scaling, which is calculated as:

$$x_{\text{normalized}} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{II.4}$$

where x is the original feature value, and min(x) and max(x) are the minimum and maximum values of the feature, respectively.

Reasons for Standardization and Normalization

Standardization and normalization of features play a crucial role in improving the performance of ML models. Improved convergence is one benefit, as many ML algorithms, particularly gradient descent-based optimization algorithms, converge faster when features are on a similar scale. Standardization and normalization ensure that features with larger magnitudes do not dominate the optimization process. Another benefit is equal weighting,

where these scaling techniques prevent features with larger scales from disproportionately influencing the model's learning process. By scaling features to a common range, each feature contributes more equally to the model's decision-making process. Additionally, standardization and normalization enhance robustness to outliers. By scaling features to a common scale, the impact of outliers and extreme values on the model's performance is reduced, making the models more resilient to such anomalies.

Models Benefitting from Standardization and Normalization

Several ML models benefit from standardization and normalization, including:

Standardization and normalization are essential for various types of ML algorithms. For linear models, such as linear regression, logistic regression, and SVM, these scaling techniques improve model performance by aiding in convergence and preventing any single feature from dominating due to its scale. Distance-based algorithms, including KNN, clustering algorithms like K-means clustering, and principal component analysis (PCA), rely heavily on distance metrics and are sensitive to feature scales. Standardization and normalization ensure that these algorithms perform optimally by maintaining a consistent feature scale. Neural networks, particularly deep learning models, also benefit from standardized or normalized inputs to facilitate better convergence during training. This is especially important for batch normalization layers within neural networks, which require inputs to be on a similar scale to function effectively.

II.3.5 Feature engineering

Feature engineering is the process of creating new meaningful features or transforming existing features in a dataset to improve the performance of ML models. It involves selecting, extracting, and modifying features to capture relevant information and patterns beneficial for predictive modeling.

Creation of Informative Features

One aspect of feature engineering involves creating new features that are highly informative concerning the target output based on existing features. This may include generating polynomial features by combining existing features through multiplication or raising them to higher powers. This allows models to capture non-linear relationships between variables. Additionally, creating interaction features by combining pairs of existing features can capture synergistic effects or interactions between variables. For example, in a medical dataset, the product of a patient's age and blood pressure might be a more informative feature than age or blood pressure alone. Constructing derived features based on domain knowledge or insights from the data is also valuable. This could involve aggregating or summarizing information from multiple variables to create new features that better represent underlying patterns or relationships.

Dimensionality Reduction

Another aspect of feature engineering involves reducing the dimensionality of the feature space to alleviate the curse of dimensionality and improve model performance. Techniques such as principal component analysis (PCA), feature selection, and feature extraction help identify and retain the most relevant features while discarding redundant or less informative ones.

Considerations and Best Practices

Feature engineering requires careful consideration of domain knowledge, data characteristics, and modeling objectives. To avoid overfitting, a balance must be struck between adding complexity and capturing relevant information. Iterative experimentation and validation are crucial for evaluating the effectiveness of feature engineering techniques and refining the feature set for optimal model performance.

II.4 Analyze and Prepare Several Datasets Used in Thesis

After defining the various aspects of data analysis and pre-processing. In this section, we present the application of these techniques to analyze, characterize, and prepare the datasets used in the thesis.

We have exploited both public and private datasets of varying size and quality. This variety serves, first and foremost, to address several medical problems and identify high- and low-risk individuals. It will also enable us to test and compare our proposed approaches presented in the following chapters on various datasets. Overall, we have exploited five tabular datasets containing biological and clinical variables. Each dataset admits a binary output presenting the disease's presence or absence.

II.4.1 Public Datasets (Hypothyroid and diabetes)

Public datasets are vital in advancing research and fostering innovation in various fields, including ML and data science. Academic institutions, research organizations, government agencies, and industry partners often make these datasets available. Platforms such as Kaggle, the UCI ML Repository, data.gov, and Google Data provide access to diverse datasets spanning multiple domains, including healthcare, finance, transportation, and social sciences. These datasets are meticulously curated, annotated, and openly shared to facilitate collaboration, reproducibility, and knowledge dissemination within the scientific community. Leveraging public datasets enables researchers, practitioners, and enthusiasts to explore real-world problems, develop and validate ML models, and gain valuable insights into complex phenomena. Additionally, public datasets serve as benchmarks for evaluating algorithm performance, benchmarking new methodologies, and addressing pressing societal challenges through data-driven approaches.

Hypothyroid Diagnosis Dataset

Data Analysis and Description of Hypothyroid Dataset : The dataset is from the UCI ML Repository [Dua 17]. It comprises 3772 subjects with 29 features, including a binary output column indicating the presence or absence of hypothyroidism. The age range of the population in the dataset spans from 1 to 95 years, with 67.9% being female and 32.1% male.

The positive class largely outweighs the negative class in the population, with 3387 cases identified as positive for hypothyroidism compared to only 291 identified as negative, as shown in Figure II.7.

The majority of individuals detected as positive for hypothyroidism are women, accounting for 67.1% of the population. In contrast, men comprise only 32.9%.

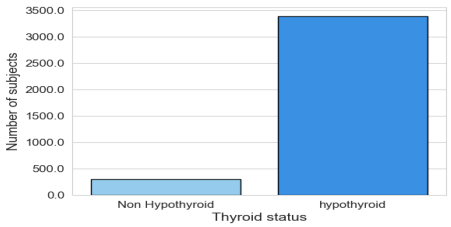


Figure II.7 – Imbalanced output classes

The age group between 55 and 75 shows the highest likelihood of developing hypothyroidism, as depicted in Figure II.8.

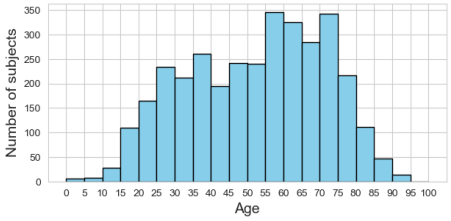


Figure II.8 – Age distribution of positive subjects

Table II.1 displays the existing features in the dataset associated with their corresponding definitions and types.

The correlation matrix in Figure II.9 indicates weak correlations between the features. However, a relatively strong correlation between the output column "Target_hypothyroid" and the TSH, T3, TT4, and FTI features suggests a significant relationship with the output.

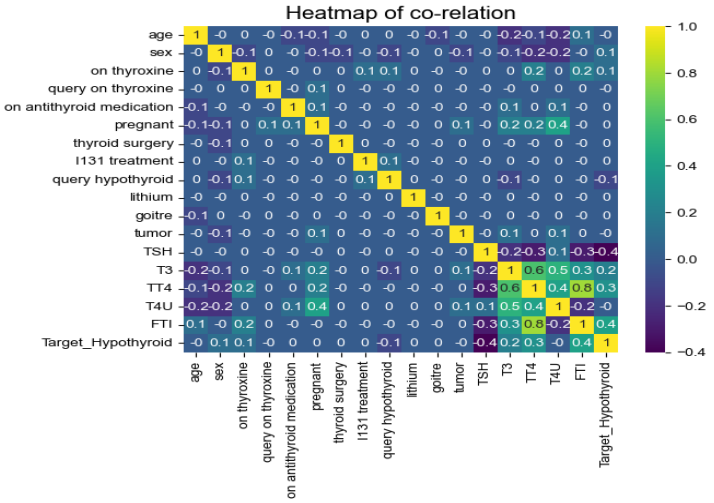


Figure II.9 – Correlation between features

The histograms in Figure II.10 reveal differences in scale among certain variables. This emphasizes the necessity of data normalization when employing linear models or neural networks for prediction.

Table II.1 – Data description and type

Features	Description	Type
age	Patient's age	Int
sex	Gender of the patient	Int
On thyroxine	Whether patient currently taking thyroxine	Bool
Query on thyroxine	Patient is questioned about use of thyroxine	Bool
On antithyroid meds	Patient currently taking antithyroid medication	Bool
pregnant	If the patient now is pregnant	Bool
Thyroid surgery	Whether the individual has had thyroid surgery	Bool
I131 treatment	If the patient has receiving I131 treatment	Bool
Query hypothyroid	Patient thinks they have developed hypothyroidism	Bool
lithium	If patient currently taking lithium	Bool
goitre	Patient have goitre	Bool
tumor	If patient diagnosed with a tumor	Bool
hypopituitary	Patient hypopituitarism	Flt
TSH	Blood test results for TSH level	Flt
T3	Blood test results for T3 level	Flt
TT4	Blood test results for TT4 level	Flt
T4U	Blood test results for T4U level	Flt
FTI	Blood test results for FTI level	Flt
TBG	Blood test results for TBG level	Flt
Target-hypothyroid	Diagnosis of hypothyroidism	Int

Abbreviations : Int: integer, Bool: Boolean, Flt: Float, TSH : Thyroid Stimulating Hormone, T3: Triiodothyronine, TT4: Total Thyroxine, T4U: Thyroxine Uptake, FTI: Free Thyroxine Index, TBG: Thyroxine Binding Globulin

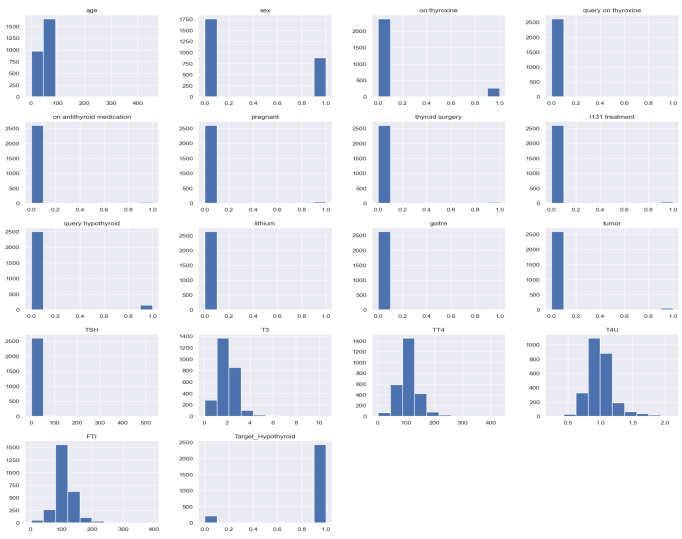


Figure II.10 – Features histogram

Unbalanced Data Manangment of Hypothyroid Dataset: The initial data preprocessing step involves converting the dataset contents into a digital format. Some columns need processing to transform them from text columns to binary form. After converting the data into a digital format, we address missing values by replacing them with average ones.

An important aspect of data preprocessing involves analyzing the output class. Based on our data analysis, we observed an imbalance between the two classes in the target column. Imbalanced data can lead to inadequate model learning, affecting its ability to accurately predict positive or negative values. Additionally, during the model's testing and validation phases, the selected data may be inconsistent, making the test phase unreliable.

Two solutions are available in this case. One technique is oversampling, which involves duplicating instances of the minority class randomly to achieve a balanced distribution between the two classes. The second proposed solution is undersampling, which involves removing instances from the majority class to achieve a balanced distribution between the two classes.

Most studies in the literature have utilized oversampling. However, this methodology often leads to overfitting and can result in biased testing and validation processes. In our study, we employ the undersampling technique. This approach reduces the dataset volume but ensures reliable and unbiased testing and validation phases.

Figure II.11 displays the data after the undersampling process. The dataset volume has been reduced from 3772 subjects to 582 subjects. Despite the size reduction, this approach ensures improved learning for the ML model and a reliable testing and validation process.

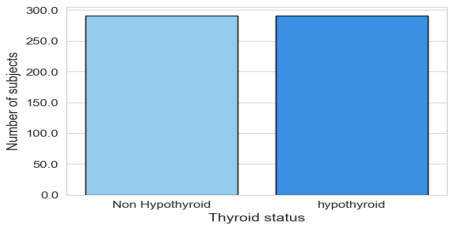


Figure II.11 – Balanced output classes

Diabetes Prediction Dataset

Data Analysis and Description of Diabetes Dataset: This dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases. Its objective is to diagnostically predict whether a patient has diabetes based on specific diagnostic measurements included in the dataset. The instances were selected with several constraints from a larger database. Specifically, all patients in this dataset are females at least 21 years old of Pima Indian heritage. The dataset consists of several medical predictor variables and one target variable (Outcome). The predictor variables are Pregnancy, Glucose, Blood Pressure, skin thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The dataset contains 768 subjects, 268 being diabetic and 500 non-diabetic, as illustrated in Fig. II.12.

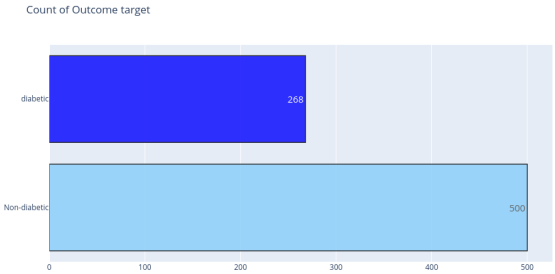


Figure II.12 – Diabetes outcome target quantity

This means that 65.1% of the population are diabetic, as shown in Figure II.13. This proportion is appropriate for training and testing the model.

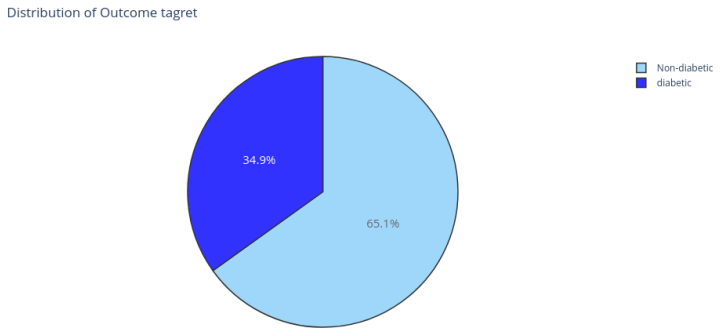


Figure II.13 – Diabetes outcome target percentage

Missing Values Management for Diabetes Dataset: Figure II.14 shows the percentage of the missing value. There are no data quality concerns for columns with a percentage of missing values of less than 10%. However, for the two characteristics of SkinThickness and Insulin, 29.56% and 48.1% of values are missing, respectively. Therefore, a PMM imputation method was developed to handle these missing values.

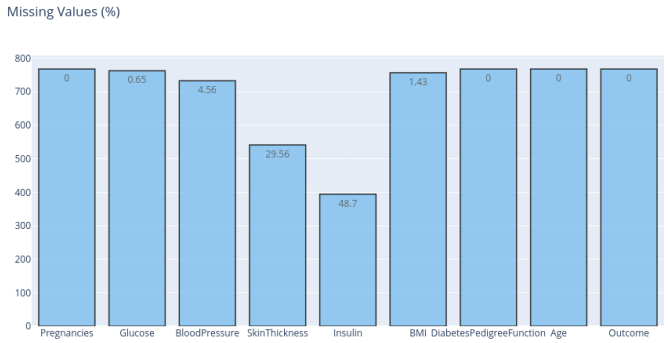


Figure II.14 – Missing Values of diabetes dataset (%)

Outlier Manangment for Diabetes dataset: The boxplot in Figure II.15 shows the features have almost the same scale. There is no need to normalize the data. Moreover, there are no outlier values to remove.

II.4.2 Private Datasets collected by doctors (Carbohydrate abnormalities and MetS)

In addition to using public datasets, our study benefits from access to private datasets collected from hospital settings. These datasets provide valuable insights into specific medical conditions and enable us to tailor our analyses to address pertinent clinical questions. In our research, we have gathered two distinct private datasets, each focusing on several medical conditions.

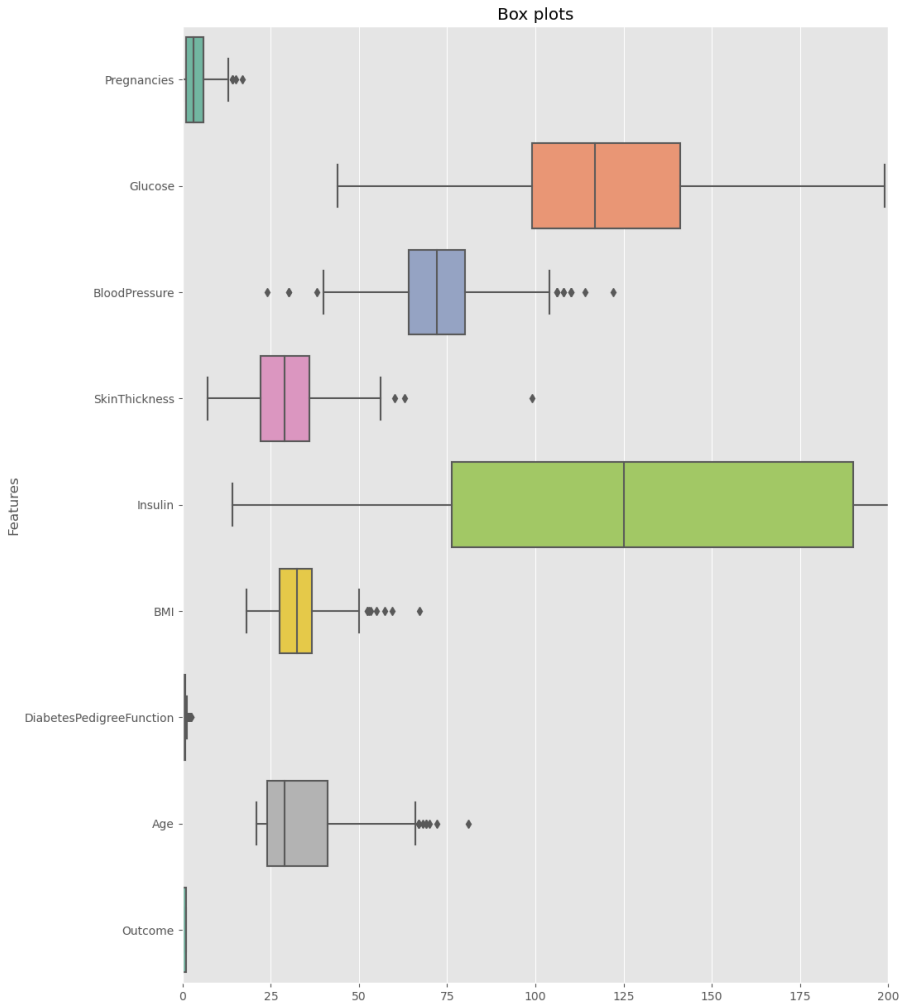


Figure II.15 – Box plot of diabetes datasets

These private hospital datasets offer unique opportunities for research and clinical insights that complement our analyses of public datasets. By combining information from diverse sources, we aim to comprehensively understand the complex interactions between genetic, clinical, and environmental factors influencing disease pathogenesis and progression. Ultimately, our research endeavors with these private datasets strive to contribute to advancements in medical knowledge, patient care, and public health initiatives.

β-TM Dataset

Data Analysis and Description of β-TM Dataset: This observational study was conducted at the Adult and Pediatric Endocrinology-Diabetology Department, Doctor Benbadis University Hospital (Constantine-Algeria). Data were obtained from a survey conducted from 2016 to 2022 among Algerian patients with β-Thalassemia Major (β-TM) receiving routine blood transfusion (TDT) from eastern Algeria. All patients were referred from the pediatric and adult hematology regional departments to assess and manage endocrine and metabolic complications of β-TM according to recently updated guidelines [Farmakis 22]. We excluded from this study all patients with other forms of thalassemia or other congenital hemoglobinopathies, β-TM patients with severe chronic illness, malnutrition, or those receiving systemic glucocorticoid treatment within the previous 4 weeks.

The data were collected through a questionnaire comprising items on socio-demographic information, personal and family medical history, transfusion history, and details regarding the date, type, and modalities of iron chelation treatment.

All patients underwent anthropometric measurements, including height (H) (cm) and weight (kg). Body mass index (BMI) (kg/m^2) was calculated. They also underwent hormonal evaluation, including assessments of somatotropic, gonadotropic, thyrotropic, and corticotropic functions, and evaluations of parathyroid gland function and vitamin D levels.

All patients in the study underwent fasting blood glucose tests after at least 8 hours of overnight fasting from the age of 10 and a standard 2-hour test after a 75-g oral glucose tolerance test (OGTT). Blood samples were sent and analyzed at the central laboratory of the University Hospital of Constantine using enzymatic methods and an automatic analyzer (ADVIA Integrated Modular System). Insulin resistance was calculated using the homeostasis model assessment of insulin resistance (HOMA-IR), as shown below:

$$HOMA - IR = \frac{Glycemia * Insulinemia}{22.5}. \tag{II.5}$$

(glycemia mmol/l; insulin mUL).

Glucose tolerance was classified into three categories based on fasting blood glucose levels: Normal fasting glucose (NFG) was defined as a glucose level below 100 mg/dL. Patients with glucose levels $\geq$ 126 mg/dL on at least two occasions, or $\geq$ 200 mg/dL after a glucose load, or randomly $\geq$ 200 mg/dL if symptoms were suggestive, were considered diabetic. Those with glucose levels between 100 and 125 mg/dL (6.1-6.9 mmol/L) were considered impaired fasting glucose (IFG). Impaired glucose tolerance (IGT) was diagnosed if blood glucose was between 140-199 mg/dL (7.8-11 mmol/L) 2 hours after a glucose load. Individuals with impaired fasting glucose and/or impaired glucose tolerance were designated as pre-diabetic according to the recommendations of "The International Network of Clinicians for Endocrinopathies in Thalassemia and Adolescent Medicine" (ICET-A) and the American Diabetes Association (ADA) [ElSayed 23, De Sanctis 16].

Blood transfusions were administered every 2-4 weeks to maintain the pre-transfusion hemoglobin level above 9 g/dL. An iron chelator (deferoxamine, deferiprone, or deferasirox) was routinely administered whenever the ferritin level exceeded 1000 ng/mL. Our population's measurement of iron overload relied solely on transfusion history and serum ferritin levels since T2* MRI is not available at the University Hospital of Constantine.

The main characteristics of the study population are shown in Table II.2. Our dataset contains 80 subjects and 22 features (After feature selection). The mean age of 80 patients (31 males, 49 females) with β-TM at the time of the study was 18,1 $\pm$ 5,9 years (range 10–30 years). There was no significant difference between the sexes. None were overweight or obese. Nine subjects (11.3 %) had a family history of type 2 diabetes.

The age of starting chelation therapy was late (9.7 $\pm$ 3.6 years), with Deferoxamine being the most commonly used iron chelator, followed by Deferasirox (56 patients, 70.0% and 20 patients, 25.0%, respectively). It should be noted that Deferoxamine became available for home use in Algeria in 2007, and only 4 out of 56 β-TM patients using Deferoxamine had an infusion pump. Adherence to chelation therapy was irregular in 59 cases (73.8%). Two subjects (2.5%) in our series received combined therapy during the study period. All poly-transfused patients in our series had post-transfusion iron overload, with a median serum ferritin level of 4600.2 $\pm$ 4332.8 ng/ml. 51 patients (63.75%) had ferritin levels above the critical threshold of 2500 ng/ml.

Table II.2 – General characteristics of the study population with β-TM according to the presence of disorders of glucose metabolism

Features	Total	Patients with NGT	Patients with IFG	Patients with IGT	Patients with diabetes	p-value
N (%)	80(100)	49(61.3)	4(5.0)	14(17.8)	15(18.7)	-
Age (yrs)	18,4 ± 5.9	15.9 ± 4.5	18.5 ± 10.2	20.9 ± 5.7	23.1 ± 5.1	≤ 0.001
Age at diagnosis of beta-TM (months)	10.8 ± 6.5	11.4 ± 6.7	6.5 ± 0.6	10.4 ± 6.6	10.7 ± 6.5	0.378
Gender						
Male	31(38.8)	16(51.6)	2(6.5)	6(19.4)	7(22.6)	
Female	49(61.3)	31(63.3)	2(4.1)	8(16.3)	8(16.3)	0.424
Adherence to chelation therapy						
Regular	21(26.3)	6(28.6)	2(9.5)	4(19)	0(0)	0.264
Not regular	59(73.8)	27(45.8)	2(3.4)	10(16.9)	15(25.4)	
Splenectomy						
Yes	59(73.8)	29(49.2)	4(6.8)	12(20.3)	14(23.7)	0.0077
No	21(26.3)	18(85.7)	0(0)	2(9.5)	1(4.8)	
Family history of thalassemia major						
Yes	46(57.5)	19(41.3)	3(6.5)	8(17.4)	8(17.4)	0.827
No	34(42.5)	14(41.2)	1(2.9)	6(17.6)	7(20.6)	
Family history of diabetes						
Yes	9(11.3)	7(77.7)	1(11.1)	0(0)	1(11.1)	0.383
No	71(88.8)	40(56.3)	3(4.2)	14(19.7)	14(19.7)	
(BMI) kg/m²	18.3±2.5	18.1±2.5	20±1.5	18.1±3.1	18.5±2	0.4477
Serum Ferritin (ng/ml)	4600.2 ± 4332.8	3102.5 ± 1865.1	4450 ± 2311.6	8215 ± 7051	5959.4 ± 4960.8	≤ 0.001
Hemoglobin before transfusion (gr/dl)	7.4 ± 0.8	7.5 ± 0.8	7.4 ± 0.7	7 ± 0.8	7.4 ± 1.1	0.1686
FPG (mg/dl)	99.44 ± 39.58	82.94 ± 13.17	107.5 ± 11.82	97.07 ± 16.56	151.2 ± 64.62	≤ 0.001
2h.post 75 g glucose (mg/dl)	143.17 ± 58.81	109.02 ± 13.09	124.67 ± 12.86	160.21 ± 27.33	244.71 ± 58.26	≤ 0.001
HOMA-IR	1.9±0.9	1.4±0.6	2.8±1.8	2.6±0.8	2.5±0.7	≤ 0.001

Data in Table II.2 are presented as n(%) or as mean ± SD, β-TM: β-Thalassemia major, HOMA: homeostasis model assessment for insulin resistance, FPG: fasting plasma glucose, NGT: normal glucose tolerance, IFG: impaired fasting glucose, IGT: impaired glucose tolerance, GHD: Growth Hormone Deficiency, ICT: Iron chelation therapy, DFO: Deferoxamine, DFX: Deferasirox, DFP: Deferiprone, HSD: hyperparathyroidism secondary to vitamin D deficiency.

Out of the 80 patients investigated, 15 (18.7%) had diabetes mellitus (DM), 4 (5.0%) had impaired fasting glucose (IFG), and 14 (17.8%) had impaired glucose tolerance (IGT). On average, patients with diabetes were older than those without diabetes (23.1 ± 5.1 years vs. 15.9 ± 4.5 years, p-value ≤ 0.05). Ferritin levels were higher among patients with diabetes compared to those with normal glucose tolerance (5959.4 ± 4960.8 ng/ml vs. 3102.5 ± 1865.1 ng/ml, p-value 0.310), but the difference was insignificant.

All our patients had at least one endocrine disorder, with hypogonadism and hypoparathyroidism being more common in patients with diabetes (14 (30.4%) vs. 6 (75.0%), p-value 0.033; 6 (40.0%) vs. 2 (13.3%), p-value ≤ 0.001, respectively), compared to patients without diabetes.

Outlier Manangment for Critical Features of β-TM Dataset: According to doctors, the HOMA-IR feature can be a very significant predictor of carbohydrate abnormalities. That is why we chose to analyze this characteristic and identify any outliers. By examining Figure II.16, which visualizes the HOMA-IR violin plot, we found that the density of this feature ranged from 0 to 3, with one relatively significant outlier greater than 5. According to the doctors, this outlier may be both logical and significant, and its presence is not attributed to a typing error. Hence, the decision was made to keep this aberrant value.

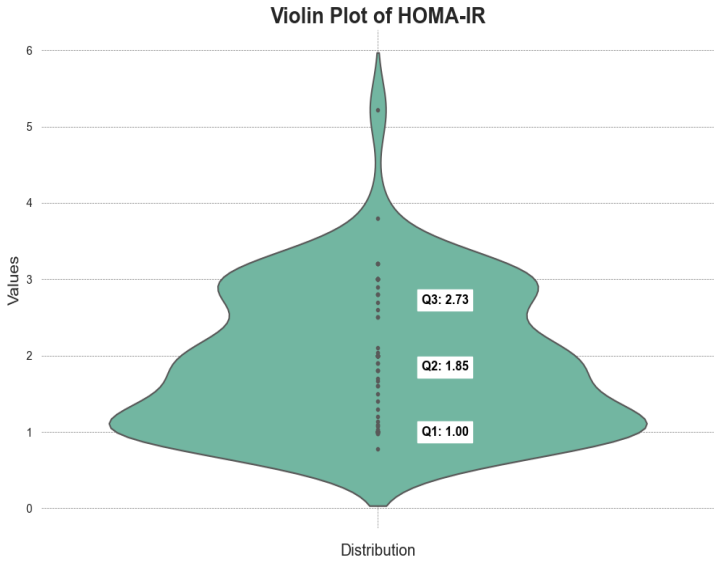


Figure II.16 – HOMA-IR Violin Plot

Missing Values Management for β-TM Dataset: Due to the limited size of the dataset, we excluded features with missing values exceeding 10% of the population [Waljee 13]. However, even after this step, 10 features still contained 1 to 3 missing values per feature. To handle these missing values without introducing bias, we imputed the median value for categorical columns to ensure they remain within existing classes, and the mean value for continuous values.

MetS Datasets

Data Analysis and Description of MetS Dataset: We used two MetS sets of research data as input. Detailed characteristics of this population and methods have been described elsewhere [Benmohammed 15, Benmohammed 11]. The first dataset (DS1) comprises 1,100 (537 boys and 563 girls) scholar adolescents [Benmohammed 15], the second dataset (DS2) includes 305 scholar adolescents (133 boys and 172 girls) [Benmohammed 11]. In the DS1 a random selection was made in three stages: the academic institution (high school and middle school); the classroom; and the students. While the DS2 included adolescents living with overweight or obesity (International Obesity Task Force criteria [Tj 00]) followed up in endocrinology department (Constantine University Hospital, Algeria). Table II.3 presents the description and type of each feature in the datasets.

All adolescents were aged 12 to 18 and were from city of Constantine (Algeria). In both datasets, the adolescents had physical examinations and blood tests. Anthropometric measurements included height (H) (cm), weight (kg), waist and hip circumferences (WC and HC) (cm).

Table II.3 – Summary of Features in MetS datasets

Features	Description	Type
Age	Age of the individual	Quantitative
Gender	Biological sex of the individual	Categorical
Weight	Body weight of the individual	Quantitative
Height	The height of the individual	Quantitative
BMI	Body Mass Index	Quantitative
WC	Waist Circumference	Quantitative
HC	Hip Circumference	Quantitative
FBG	Fasting Plasma Glucose	Quantitative
Chol	Total cholesterol levels in the blood	Quantitative
Tg	Triglycerides level in the blood	Quantitative
WC/HC	The ratio of waist to hip circumference	Quantitative
SBP	Systolic Blood Pressure	Quantitative
DBP	Diastolic Blood Pressure	Quantitative
MBP	Mean Blood Pressure	Quantitative
LDL-C	Low-Density Lipoprotein Cholesterol	Quantitative
HDL-C	High-Density Lipoprotein Cholesterol	Quantitative
TyG	Triglyceride-glucose Index	Quantitative
SM_Cook	MetS according to Cook definition	Categorical
SM_Idf	MetS according to IDF definition	Categorical
SM_Ferranti	MetS according to De Ferranti definition	Categorical

WC/HC ratio and body mass index (BMI) (kg/m^2) were calculated. Systolic and diastolic blood pressure (SBP, DBP) were also measured by the international guidelines of the National High Blood Pressure Education Program working group on high blood pressure children and adolescent populations [Program 00]. Mean blood pressure (MBP) was calculated as follows [Kodama 14]:

$$\text{MBP} = \frac{\text{SBP (mmHg)} + 2 \times \text{DBP (mmHg)}}{3}$$

Blood samples were taken after 12 hours of fasting and analyzed at the central laboratory of the Constantine University Hospital. Biological assessments included fasting plasma glucose (FPG), triglycerides (TG), total cholesterol (TC), and HDL-C measurements using enzymatic methods. Low-density lipoprotein cholesterol (LDL-C) was calculated according to the Friedewald formula. Triglyceride glucose index (TyG) was calculated according this formula:

$$\text{TyG} = \ln\left(\frac{\text{FPG (mg/dL)} \times \text{TG (mg/dL)}}{2}\right)$$

Three definitions for MetS were used: 2007 IDF [Zimmet 07], Cook [Cook 03] and De Ferranti [Magge 17] were calculated as shown in Table II.4.

The Table presented in II.5 displays the count and percentage of subjects identified as positive or negative for MetS according to three definitions. Notably, as anticipated, the prevalence of MetS is significantly higher in DS2. Furthermore, the prevalence is higher based on the De Ferranti definition than the Cook and IDF definitions in both datasets.

Missing Values Management for MetS Dataset: When analyzing the missing values in both datasets, we noticed that there were already missing values in the output columns. Managing these missing values can be delicate since the outputs are quite unbalanced, with a very small number of MetS-positive subjects. For this reason, we prefer to remove subjects with missing information on the existence of MetS to reduce the risk of impact on the ML. For the other features considered input, we chose the PMM method to cover them since these datasets admit a relatively large amount of data, which allows us to have good imputation accuracy with the PMM.

Table II.4 – Definition of the metabolic syndrome in adolescents according to the IDF, Cook et al., and De Ferranti et al.

Definition	**Criterion**
IDF[Zimmet 07]	Abdominal adiposity (waist circumference $\geq$ 90th percentile by age & gender) and two other criteria: • Fasting glucose $\geq$ 100 mg/dL or known type 2 diabetes • SBP $\geq$ 130 mmHg or DBP $\geq$ 85 mmHg • Triglycerides $\geq$ 150 mg/dL or specific treatment • HDL-cholesterol $<$ 40 mg/dL if aged 10–16 y; $<$ 40 mg/dL in men, $<$ 50 mg/dL in women if $>$ 16 y
Cook[Cook 03]	Three or more of the following: • Waist circumference $\geq$ 90th percentile by age & gender • Fasting glucose $\geq$ 110 mg/dL • SBP or DBP $\geq$ 90th percentile by age, gender & height or treatment • Triglycerides $\geq$ 110 mg/dL • HDL-cholesterol $\leq$ 40 mg/dL
De Ferranti[Magge 17]	Three or more of the following: • Waist circumference $\geq$ 75th percentile by age & gender • Fasting glucose $\geq$ 110 mg/dL • SBP $\geq$ 90th percentile by age, gender & height • Triglycerides $\geq$ 100 mg/dL • HDL-cholesterol $<$ 45 mg/dL for boys 15–19 y, otherwise $<$ 50 mg/dL

Table II.5 – Population positive and negative for MetS

Dataset	Definition	Positive Cases	Negative Cases	Positive %	Negative %
DS1	Cook	24	1061	2.21%	97.79%
DS1	Ferranti	51	1034	4.70%	95.30%
DS1	IDF	10	1075	0.92%	99.08%
DS2	Cook	23	244	8.61%	91.39%
DS2	Ferranti	50	217	18.73%	81.27%
DS2	IDF	22	245	8.24%	91.76%

Outlier Manangment for Critical Features of MetS datasets: According to physicians, the MBP, BMI, WC, and TyG characteristics can be important predictors for MetS screening. Hence, we chose to study these variables and identify any outliers to ensure a better ML model later on. As shown in Figure II.17, we observe the presence of some outliers in the MBP, BMI, and WC characteristics. According to the physicians, these values are relatively logical and not caused by typing errors. Additionally, we note that these are more of a set of outliers rather than isolated individual values, which may be significant for the ML model. Therefore, we decided to retain these outliers.

II.4.3 Discussion

Our aim in this chapter was to analyze and prepare the datasets used in the thesis. Hence, we first defined the various data analysis and pre-processing methodologies. We then applied them to the datasets to prepare them for ML.

Now that we've prepared the data, let us discuss the quality of each dataset for comparison, considering each dataset's limitations in the following steps.

As shown in Table II.6, to compare these datasets, we have estimated evaluators such as Quantity, which considers the number of features and subjects. Additionally, completeness considers the existence of missing values, outliers, and imbalanced output data.

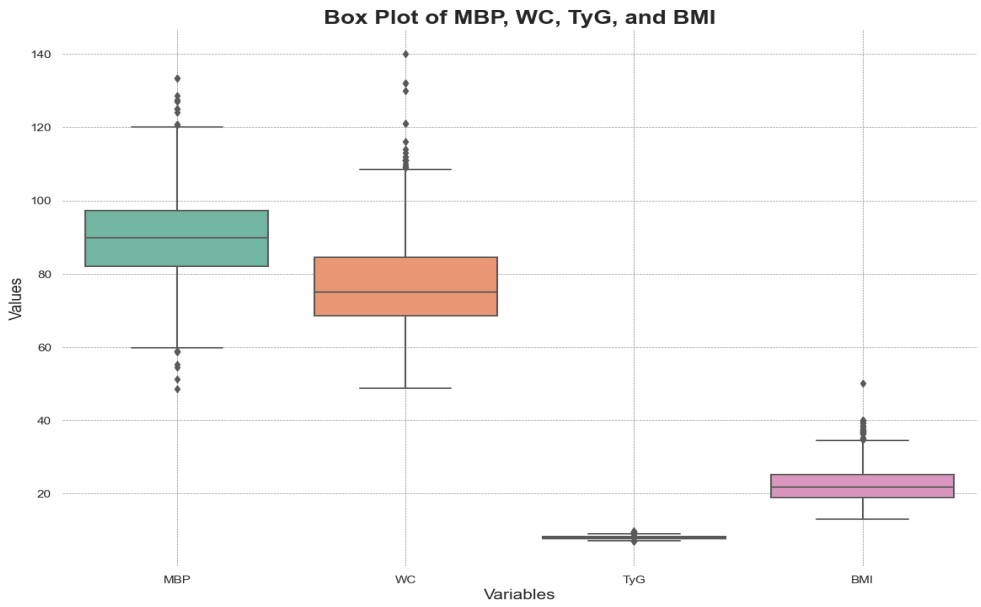


Figure II.17 – Box Plot (MBP, TyG, BMI and WC)

Finally, relevance studies the significance of features relative to the output by calculating the number of features with a high correlation and a p-value < 0.05. By studying these metrics in Table II.6, we can observe that public datasets generally have better quantity and completeness. They are often pre-processed for ML tasks. On the other hand, private datasets collected in collaboration with hospitals exhibit lower quantity and completeness but show good relevance with highly significant features.

Furthermore, public datasets are highly valuable for testing new methodologies. Primarily, they offer a substantial volume of data devoid of missing values or outliers. Additionally, since these datasets are referenced in the literature, numerous studies can be leveraged to contextualize approaches and their outcomes. Conversely, private datasets present intriguing opportunities for original research and findings, yet they may yield less reliable results in ML applications due to limited data quantity and numerous missing values. So, several limitations must be considered when testing ML or XAI in one of the datasets, especially when discussing performance.

Table II.6 – Quality comparison of several datasets

	Quantity		**Completeness**			**Relevance**	
	Features	**Subjects**	**Balanced output**	**Missing values**	**Outliers**	**Strong correlation**	**p-value $<$ 0.05**
DS_Hy	29	3772	yes	no	no	4	2
DS_D	8	864	no	yes	no	2	1
DS_B	22	80	no	yes	yes	6	5
DS_M	17	1085	yes	yes	yes	4	3
DS_MO	17	266	yes	yes	yes	4	3

Abbreviations: DS_Hy : Hypothyroid dataset, DS_D : Diabetes datasets, DS_B: β-TM dataset, DS_M : First MetS dataset, DS_MO: Second MetS dataset

II.5 Conclusion

In this chapter, we began by presenting several data analysis and preparation concepts based on statistical tools and visualization graphs. We then applied our data science knowledge and collaboration with medical specialists to analyze and prepare the datasets used in the thesis. Finally, we compared several datasets and discussed their qualities, essentially the difference between private and public datasets. Public datasets, offering ample data without missing values or outliers, are tailored for evaluating proposed novel approaches. Conversely, private datasets are distinctive because they are associated with physicians' specific challenges and characteristics strongly correlated with patient discharge. Hence, exploiting private data sets for medical decision support using ML is important. However, this data type must be well-treated and prepared for ML because of the existence of missing values, outliers, and unbalanced data. The next step will be to tackle the exploitation of these datasets by ML to predict the risk of targeted endocrine diseases to propose solutions to physicians' problems.

Chapter III

Machine Learning for Endocrine Diseases Risk Prediction

III.1 Introduction

Once the data has been prepared previously, in this chapter, we aim to use information from private datasets provided by doctors to predict the risk of carbohydrate abnormalities and MetS, assist doctors in identifying individuals at high and low risk, and offer solutions to their problems.

Hence, we start by presenting the main concepts of supervised ML in Section III.2, with detailed explanations of the several linear and tree-based models. Also, the metrics for evaluating a classification task were presented in the same section. Next, the methodologies proposed for risk prediction of the diseases targeted in this thesis and their results are presented in Section III.3. Finally, Section III.4 is given to discuss the limits and results, and Section III.5 concludes the chapter.

III.2 Supervised Models for Classification Task

In AI domain, supervised learning is a cornerstone methodology for tackling classification tasks. The goal is to assign categorical labels to input data based on their features. Supervised learning entails training a model on labeled datasets, where each data point is associated with a known outcome or class label. Within this paradigm, classification algorithms seek to discern patterns and relationships within the data, enabling the model to generalize and make accurate predictions on unseen instances.

At its core, supervised classification involves the construction of a decision boundary that delineates distinct classes within the feature space. Linear models, such as linear regression and logistic regression, serve as fundamental tools for binary and multi-class classification, leveraging linear combinations of features to delineate class boundaries. SVM extend this paradigm by identifying the optimal hyperplane that maximally separates several classes, enhancing classification performance in high-dimensional spaces.

Beyond linear methods, tree-based models offer a versatile framework for classification tasks, capable of capturing nonlinear relationships and interactions among features. Decision trees partition the feature space into hierarchical segments based on simple decision rules, culminating in a tree-like structure that facilitates intuitive interpretation. Random forests, an ensemble of decision trees, aggregate the predictions of multiple trees to enhance robustness and mitigate overfitting, making them well-suited for complex classification problems. Additionally, gradient boosting algorithms like XGBoost, CatBoost, and LightGBM iteratively refine predictive performance by sequentially fitting new models to the residuals of previous iterations, thereby boosting overall accuracy and generalization.

III.2.1 Linear Models

Linear models represent a foundational class of algorithms in supervised learning, particularly adept at tackling classification tasks. These models operate on the principle of linear relationships between input features and the target variable, aiming to delineate class boundaries through linear decision boundaries in the feature space.

Linear Regression

Linear regression is a fundamental statistical technique used for modeling the relationship between a dependent variable and one or more independent variables. This model can be adapted for binary classification by setting a threshold on the predicted values

in classification tasks. It aims to find the best-fitting linear relationship between the input features and the target variable, which is achieved by minimizing the residual sum of squares (RSS) or maximizing the likelihood function. The model parameters (coefficients) are estimated using optimization techniques such as Ordinary Least Squares (OLS) or gradient descent. The linear regression model is represented by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \tag{III.1}$$

where:

y is the predicted target variable,
β_0 is the intercept term,
$\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with each feature $x_1, x_2, \ldots, x_n$,
$x_1, x_2, \ldots, x_n$ are the input features,
ϵ represents the error term.

Linear regression relies on several assumptions, including linearity, independence, and homoscedasticity. The coefficients $\beta_1, \beta_2, \ldots, \beta_n$ represent the change in the target variable for a one-unit change in the corresponding feature, holding all other features constant, while the intercept term β_0 represents the value of the target variable when all features are zero. Its strengths include simplicity, interpretability, and fast training and prediction time for large datasets, but it is limited by assumptions of linearity, sensitivity to outliers and multicollinearity, and limited to linear decision boundaries for classification tasks.

Figure III.1 displays and summarizes the architecture of the linear regression model.

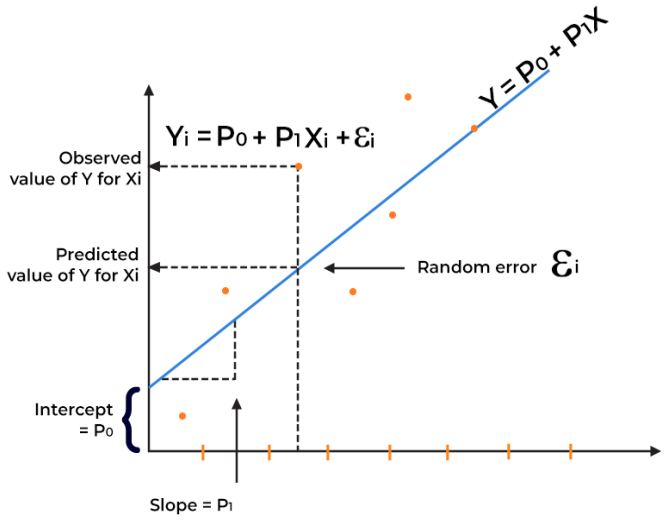


Figure III.1 – Linear regression architecture [Alyaseen 23]

Logistic Regression

Logistic regression is a statistical model used for binary classification tasks where the target variable is categorical with two possible outcomes. Despite its name, it is a classification algorithm, not a regression algorithm. Logistic regression models the probability that a given input belongs to a particular class using the logistic function (sigmoid function). The model parameters (coefficients) are estimated using optimization techniques

such as maximum likelihood estimation (MLE) or gradient descent. The logistic regression model predicts the probability that an input x belongs to class 1 ($y = 1$) using the logistic function:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}} \tag{III.2}$$

where:

$P(y = 1|x)$ is the probability of the positive class,
β_0 is the intercept term,
$\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with each feature $x_1, x_2, \ldots, x_n$,
$x_1, x_2, \ldots, x_n$ are the input features,
e is the base of the natural logarithm.

Logistic regression assumes that the relationship between the features and the log-odds of the target variable is linear. Additionally, it assumes that the observations are independent of each other. The coefficients $\beta_1, \beta_2, \ldots, \beta_n$ represent the change in the log-odds of the target variable for a one-unit change in the corresponding feature, holding all other features constant. Logistic regression provides probabilistic predictions, allowing for uncertainty estimation. Its strengths include being a simple and interpretable model and being less prone to overfitting than more complex models. However, it assumes a linear relationship between features and the log odds, which may not always hold, and it is limited to binary classification tasks.

Figure III.2 displays and summarizes the architecture of the logistic regression model.

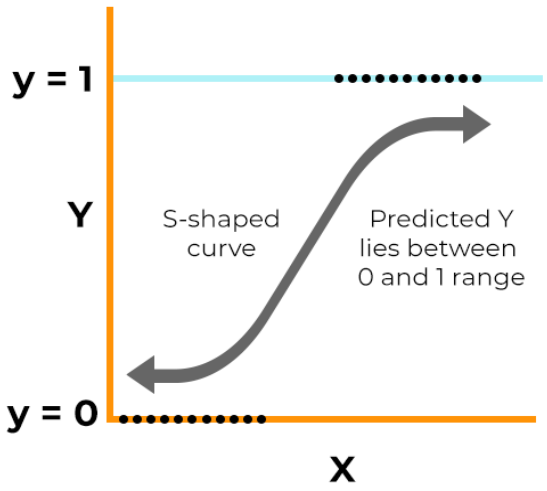


Figure III.2 – Logistic regression architecture [Baruah 24]

Support Vector Machine (SVM)

SVM is a powerful supervised learning algorithm for classification and regression tasks. It aims to find the optimal hyperplane that best separates the classes in the feature space. SVM works by finding the hyperplane that maximizes the margin, the distance between the hyperplane, and the nearest data points (support vectors) from each class, as shown in Figure III.3. This optimization problem can be solved using gradient descent or quadratic programming techniques. In a binary classification task, the decision boundary of an SVM can be represented as:

$$f(x) = \text{sign}(\mathbf{w}^T\mathbf{x} + b) \tag{III.3}$$

where:

$f(x)$ is the decision function,
$\mathbf{w}$ is the weight vector,
$\mathbf{x}$ is the input feature vector,
b is the bias term, and
$\text{sign}(\cdot)$ is the sign function.

SVM can handle non-linear decision boundaries in the input space by using the kernel trick, which implicitly maps the input features into a higher-dimensional space where a linear separation is possible. Common kernel functions include linear, polynomial, Gaussian (RBF), and sigmoid kernels. SVM assumes that the data are linearly separable or can be separated by a hyperplane with a margin. In cases where the data are not linearly separable, soft-margin SVM allows for some misclassification by introducing a penalty parameter C. The decision function of an SVM assigns a class label to each input based on which side of the hyperplane it falls on. The sign of the decision function determines the predicted class, and the function's magnitude reflects the confidence in the prediction. SVM is effective in high-dimensional spaces, versatile due to the kernel trick allowing for non-linear decision boundaries, and memory efficient as it only uses a subset of training points as support vectors. However, it is computationally intensive, especially with large datasets, requires careful selection of hyperparameters such as kernel and regularization parameter selection, and may not perform well with noisy or overlapping classes.

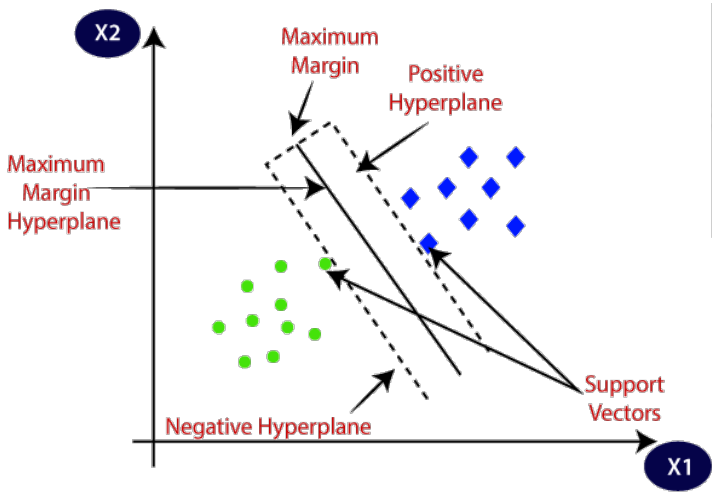


Figure III.3 – SVM architecture [Zuo 24]

III.2.2 Tree-based models

Tree-based models are a class of ML algorithms that rely on constructing a tree-like structure to make decisions. These models are widely used for classification and regression tasks. The main idea behind tree-based models is to recursively split the feature space into smaller, more homogeneous subsets until predefined criteria are met.

One of the most commonly used models in this category is the decision tree, which divides the feature space into binary segments based on decision rules derived from the data features. Each tree node represents a feature, and each branch represents a decision rule based on that feature. The tree is constructed to minimize node impurity, often measured by metrics such as the Gini index or entropy.

Tree-based models offer several advantages, including their ability to handle numerical and categorical data, interpretability, and robustness to outliers. Additionally, they can capture nonlinear relationships between input variables and the target variable.

Other tree-based models include random forests, which combine multiple decision trees to reduce overfitting and improve predictive accuracy. XGBoost, CatBoost, and LightGBM are ensemble tree-based algorithms developed to enhance the performance and efficiency of tree-based ML. These models use advanced techniques such as boosting to improve predictive accuracy while reducing computation time.

In summary, tree-based models are a powerful and versatile method for predictive modeling in various application domains. They offer a trade-off between interpretability and accuracy, making them a popular choice for many ML tasks.

Decision Tree

Decision trees are a popular and widely used supervised learning algorithm for classification and regression tasks. Their popularity stems from their simplicity, interpretability, and flexibility, which make them valuable tools for data analysis and ML. The fundamental idea behind decision trees is to build a model that resembles a tree structure, where each internal node represents a decision based on a feature, each branch represents the outcome of that decision, and each leaf node represents a final prediction or decision.

The learning process of decision trees begins with constructing this tree-like model by recursively partitioning the feature space into smaller and smaller subsets. At each step of this recursive process, the algorithm evaluates different criteria to determine the best feature and corresponding split point for dividing the data. This process continues until the data in each subset is as homogeneous as possible. The criteria for making these splits often involve measures such as Gini impurity or entropy in classification tasks and variance reduction in regression tasks. Gini impurity quantifies how mixed the classes are in the data at each node, while entropy measures the disorder or uncertainty of the data. In regression tasks, variance reduction helps identify splits that best reduce the variability of the target variable.

Building decision trees involves several key concepts. For instance, Gini impurity and entropy are used to assess the quality of splits, with lower impurity or entropy indicating a better split. Information gain, calculated from these measures, guides the choice of features and split points to maximize the effectiveness of each decision node. Additionally, pruning techniques are employed to avoid overfitting by removing branches that do not significantly improve the model's performance on unseen data.

One of the significant advantages of decision trees is their interpretability. The tree structure naturally represents decision rules that are easy for humans to follow and understand. This interpretability is further enhanced by the ability of decision trees to handle both numerical and categorical data, making them versatile for various types of problems. Through their straightforward decision-making process, decision trees provide clear insights into which features are important for making predictions.

However, decision trees are not without their challenges. They are prone to overfitting, especially when the tree is allowed to grow too deep or when the training data contains noise. Overfitting occurs when the model captures the noise in the training data rather than the underlying pattern, which can lead to poor performance on new, unseen data.

Additionally, decision trees can be unstable, as small changes in the training data might lead to different splits and, consequently, different trees. They can also be biased towards features with more levels or towards the majority class in imbalanced datasets.

To address these challenges, careful tuning of hyperparameters such as tree depth, minimum samples per leaf, and the criteria for splits is essential. Techniques such as pruning are employed to cut back the tree to avoid overfitting and improve generalization. Despite these challenges, decision trees remain a robust and flexible tool in the ML toolkit, offering a balance of simplicity and power for various analytical tasks.

In summary, decision trees are a foundational technique in ML that provides a clear and interpretable way to make predictions based on data. Their ability to handle both classification and regression tasks makes them a versatile choice for many problems. However, achieving the best performance with decision trees requires careful complexity management and a thoughtful approach to their construction and evaluation.

Random Forest: Ensemble Learning with Decision Trees

Random Forest is an ensemble learning method that leverages the power of multiple decision trees to enhance the accuracy and robustness of predictions. This approach is widely used for both classification and regression tasks due to its high performance and resistance to overfitting. At its core, Random Forest builds a collection of decision trees, each of which is trained independently on a random subset of the training data and features.

The process begins with constructing numerous decision trees, where each tree is trained on a different random sample of the data. During training, a random subset of features is considered at each node for making splits, which introduces variability among the trees. This randomness helps to reduce the correlation between the trees and thereby improves the model's generalization performance. For classification tasks, the final prediction is determined by a majority voting scheme, where the class predicted by most trees is chosen as the final outcome. In regression tasks, the model's prediction is the average of the predictions made by all the trees in the forest.

The architecture of a Random Forest model is illustrated in Figure III.4, which summarizes the process of building and combining multiple decision trees to form the ensemble model.

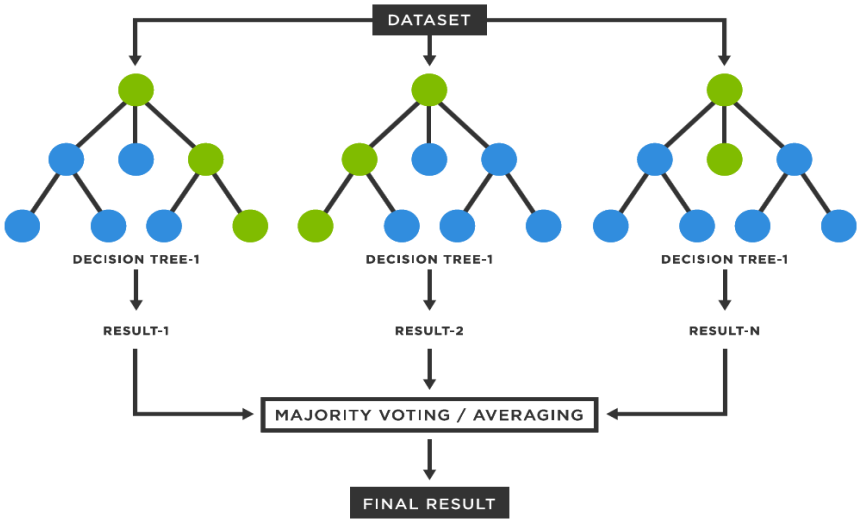


Figure III.4 – Random forest architecture [Fuster-Palà 24]

One of the primary advantages of Random Forest is its ability to achieve high accuracy across various datasets. The ensemble of diverse decision trees, combined with the mechanisms of averaging for regression and voting for classification, makes Random Forest robust to both overfitting and noise. This robustness stems from the model's ability to average out errors and variance across the many trees, which often leads to superior performance compared to individual decision trees. Additionally, Random Forest can efficiently manage large datasets with high dimensionality, which is particularly useful in real-world applications where data can be complex and voluminous.

However, Random Forest models are not without their challenges. One significant drawback is their reduced interpretability compared to single decision trees. While a single decision tree provides clear decision rules, a Random Forest consists of many trees with intricate interactions, making it harder to interpret the model's decisions. Furthermore, training and tuning Random Forest models can be computationally intensive, especially as the number of trees and the complexity of the data increase. The model also has several hyperparameters that require careful tuning, including the number of trees, the maximum depth of each tree, and the minimum number of samples required to split a node.

In conclusion, Random Forest stands out as a powerful and versatile ensemble learning method that integrates the strengths of decision trees with techniques of randomness and aggregation. This method is well-suited for a broad range of classification and regression tasks, offering high accuracy and robustness in practical applications. Despite its challenges, particularly in terms of interpretability and computational demands, Random Forest remains a popular choice for solving complex problems in ML.

XGBoost: Extreme Gradient Boosting

XGBoost, which stands for Extreme Gradient Boosting, is a highly efficient and scalable implementation of the gradient-boosting framework. Renowned for its exceptional performance and flexibility, XGBoost has gained widespread popularity in both ML competitions and real-world applications. This powerful ensemble learning method builds a strong predictive model by combining the outputs of multiple weak learners, typically decision trees.

The foundation of XGBoost lies in the gradient boosting framework, which iteratively improves the model by adding weak learners to correct the errors of the existing model. This process involves minimizing a differentiable loss function through a stage-wise approach where new trees are fitted to the negative gradient of the loss function. XGBoost distinguishes itself by incorporating both L1 and L2 regularization terms into the objective function, which helps to prevent overfitting and enhances the generalization performance of the model. The regularization terms penalize the complexity of the model parameters, encouraging simpler models that perform well on unseen data. Additionally, XGBoost employs a depth-wise tree growth strategy and prunes trees during construction to balance the trade-off between model complexity and computational efficiency. By pruning branches that do not contribute significantly to reducing the loss function, XGBoost optimizes both the accuracy and efficiency of the model.

A significant feature of XGBoost is its ability to provide insights into feature importance. The model calculates the contribution of each feature to the overall performance, ranking features based on their importance scores. This capability not only aids in feature selection but also enhances the interpretability of the model.

XGBoost is celebrated for its state-of-the-art performance across a variety of ML tasks, including classification, regression, and ranking. The algorithm supports a range of objective functions and evaluation metrics, offering flexibility to customize the model according to specific needs. Its scalability is one of its key strengths, as XGBoost can

efficiently handle large datasets with millions of samples and features. Furthermore, it supports parallel and distributed computing, which accelerates the training process on multicore CPUs and distributed computing environments.

However, XGBoost is not without its challenges. The model has numerous hyperparameters that must be carefully tuned to achieve optimal performance, which can make the model selection and tuning process quite complex. Like other ensemble methods, XGBoost models can be less interpretable than simpler models such as linear regression or decision trees, due to the intricate interactions among multiple trees. Additionally, training XGBoost models with large datasets or complex configurations can be computationally intensive, requiring substantial resources and time.

In summary, XGBoost is a powerful and versatile ML algorithm that excels in performance, scalability, and flexibility. Its advanced capabilities make it a popular choice for various predictive modeling tasks in both academic research and industrial applications.

CatBoost: Categorical Boosting

CatBoost is a gradient-boosting library developed by Yandex that is specifically designed to handle categorical variables efficiently. This model is known for its robustness, high performance, and ability to process categorical features without requiring complex preprocessing. CatBoost's design is based on an innovative algorithm incorporating novel techniques to seamlessly manage categorical features.

CatBoost's algorithm stands out for its ability to automatically handle categorical variables without manual preprocessing, such as one-hot encoding or label encoding. This process is achieved through an efficient algorithm called Ordered Boosting, which directly processes categorical features during model training. Like XGBoost and LightGBM, CatBoost follows the gradient boosting framework, aiming to progressively improve model performance by adding new weak learners. By using gradient descent, CatBoost optimizes a differentiable loss function while applying ensemble learning. The algorithm also incorporates L2 regularization to prevent overfitting and enhance the model's generalization capability. This regularization adds a penalty term to the objective function, promoting simpler and more robust models. Additionally, CatBoost is implemented in C++ to ensure increased efficiency and speed. It supports parallel computation and GPU acceleration, making it suitable for large-scale datasets and real-time applications.

One of CatBoost's main advantages is its ability to eliminate the need for manual preprocessing of categorical variables, reducing the risk of data leakage and simplifying the modeling process. Its regularization techniques and efficient handling of categorical features contribute to preventing overfitting and improving model generalization performance. In terms of predictive accuracy and computational efficiency, CatBoost proves to be a strong competitor among gradient-boosting libraries.

However, like any complex model, CatBoost presents certain challenges. Hyperparameter tuning can be time-consuming and requires substantial computational resources. Additionally, CatBoost models may be less interpretable than simpler models, especially when dealing with high-dimensional features or complex interactions between variables. Finally, CatBoost may consume more memory than traditional ML algorithms, particularly when used to process large datasets with many categorical variables.

In conclusion, CatBoost is a powerful and efficient gradient-boosting algorithm, particularly well-suited for handling categorical variables and achieving high predictive accuracy. Its ability to directly manage structured data containing categorical features makes it particularly useful for various classification and regression tasks.

LightGBM: Light Gradient Boosting Machine

LightGBM is a gradient boosting framework developed by Microsoft, designed to efficiently distribute and train large-scale datasets. This model stands out for its high performance, scalability, and flexibility, making it a suitable choice for various ML tasks. The uniqueness of LightGBM lies in its gradient-based learning algorithm that builds decision trees in a leaf-wise rather than level-wise manner, reducing the number of splits in the trees and enhancing training speed.

A key aspect of LightGBM is its leaf-wise tree growth strategy. Unlike traditional methods that grow trees level-wise, LightGBM focuses on expanding leaves to capture more complex patterns and achieve higher accuracy with fewer nodes. This approach reduces the risk of overfitting and improves the model's generalization performance. The algorithm optimizes an objective function using gradient descent and constructs ensembles of decision trees. It applies a gradient-based approach to find the best split points for both continuous and categorical features, leading to faster training and better overall performance. Additionally, LightGBM is highly scalable, capable of efficiently handling large datasets thanks to its support for distributed training across multiple CPUs and GPUs, as well as its out-of-core learning for datasets too large to fit into memory.

Among the notable advantages of LightGBM, its leaf-wise tree growth strategy and histogram-based split approach make it extremely fast compared to other gradient boosting frameworks. LightGBM can efficiently process large-scale datasets and complex models with millions of features. This framework delivers state-of-the-art performance on various ML benchmarks and real-world datasets. Its ability to capture complex patterns and directly handle categorical variables significantly contributes to its high predictive accuracy. Furthermore, LightGBM offers a wide range of hyperparameter tuning options and customization, allowing users to fine-tune the model for optimal performance. It supports custom loss functions, feature importance analysis, and early stopping techniques to improve training efficiency.

However, using LightGBM comes with certain challenges. Hyperparameter tuning can be lengthy and resource-intensive. Additionally, LightGBM may consume more memory than traditional ML algorithms, especially when training large models or managing high-dimensional feature spaces. Memory optimization techniques may be necessary to mitigate this issue. LightGBM models may also be less interpretable than simpler algorithms, such as decision trees or logistic regression. Understanding the internal workings of the model and interpreting feature importance can be difficult, particularly for complex models with numerous features.

In summary, LightGBM is a powerful and efficient gradient boosting framework that offers high performance, scalability, and considerable flexibility. It is well-suited for various ML tasks, including classification, regression, and ranking, and proves particularly effective for processing large datasets and complex models.

III.2.3 Classification Evaluation Metrics

Evaluation metrics play a crucial role in assessing the performance of classification models and quantifying their predictive accuracy. In this section, we discuss several commonly used evaluation metrics for assessing the performance of risk prediction models.

Accuracy

Accuracy measures the proportion of correctly classified instances among all the instances in the dataset. It is used to analyze the model's overall predictive ability, considering the predictive ability of both positive and negative disease subjects. The accuracy metric is calculated as the ratio of correctly predicted instances to the total number.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{III.4}$$

Precision

Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated as the ratio of true positive predictions to the sum of true positive and false positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{III.5}$$

Recall (Sensitivity)

Recall, also known as sensitivity, measures the proportion of true positive predictions among all actual positive instances in the dataset. It is calculated as the ratio of true positive predictions to the sum of true positive and false negative predictions.

$$Recall = \frac{TP}{TP + FN} \tag{III.6}$$

F1-Score

The F1-Score is the harmonic mean of precision and recall, balancing the two metrics. It is calculated as $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

Specificity

Specificity measures the proportion of true negative predictions among all actual negative instances in the dataset. It is used to analyze the predictive capacity of the model's negative subjects.

The specificity metric is calculated as the ratio of true negative predictions to the sum of true negative and false positive predictions.

$$Specificity = \frac{TN}{TN + FP} \tag{III.7}$$

AUC and ROC Curve

The Area Under the ROC Curve (AUC) measures the performance of a classification model across several thresholds. The Receiver Operating Characteristic (ROC) curve is a graphical representation of the true positive rate (sensitivity) against the false positive rate (1-specificity) for varying threshold values.

Confusion Matrix

A confusion matrix is a tabular representation of a classification model's actual versus predicted classes. It provides a comprehensive view of the model's performance by summarizing the number of true positive, true negative, false positive, and false negative predictions.

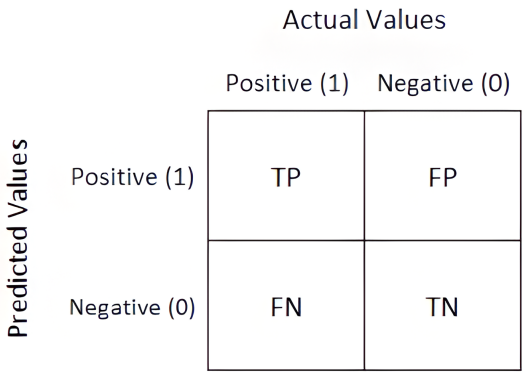


Figure III.5 – Confusion Matrix

Cross-Validation

Cross-validation is a resampling technique used to assess the performance of a predictive model by partitioning the dataset into multiple subsets. It involves training the model on a subset of the data and evaluating its performance on the remaining data, repeating this process multiple times to obtain an unbiased estimate of its performance.

Bootstrap

Bootstrap is a resampling technique used to estimate a statistic's variability by repeatedly sampling with replacement from the original dataset. It provides a robust measure of uncertainty and can be used to calculate confidence intervals for evaluation metrics.

Confidence Interval

A confidence interval is a range of values likely to contain the true value of a parameter with a certain level of confidence. It quantifies the uncertainty associated with an estimate and is commonly used to assess the reliability of evaluation metrics.

Negative Predictive Value (NPV) and Positive Predictive Value (PPV)

NPV measures the proportion of true negative predictions among all negative predictions made by the model, while PPV measures the proportion of true positive predictions among all positive predictions made by the model.

These evaluation metrics provide valuable insights into classification models' performance and help select the most appropriate model for risk prediction tasks.

C-index

The concordance index is the estimated probability of concordance between patients [Tang 19]. It is the probability that 2 patients taken at random are ordered in the same way on the outcome and the marker. This index measures the discriminating capacity of

a marker (III.8) and takes a range from 0.0 to 1.0, where a value of 1.0 indicates a perfect prediction of the risk and a value of 0.5 indicates a random prediction. It is formulated as:

$$C - index = \frac{CP + 0.5.T}{PP} \tag{III.8}$$

where CP is the concordant pairs, T the ties and PP the permissible pairs.

III.3 Risk prediction of endocrine diseases for medical decision support using ML

III.3.1 Risk prediction of carbohydrate abnormalities in patients with beta-TM

We aim to test several linear and tree-based models for predicting the risk of carbohydrate anomalies to choose the model that shows the best risk prediction, such as logistic regression (LR), SVM, random forest (RF), XG-Boost, Catboost, and LightGBM as shown in Figure III.6. To validate the risk prediction reliably, considering the limited quantity of data in the beta-TM dataset, a bootstrap approach is used to compare models by studying the confidence intervals of several metric evaluators such as accuracy, recall, precision, and f1-score. But first, Before ML prediction,we test two embedded feature selection approaches based on linear ML models to select the most significant features for the prediction. Then, both models are compared using metrics such as the c-index, AUC, precision, accuracy, and F1-score to ensure better input selection.

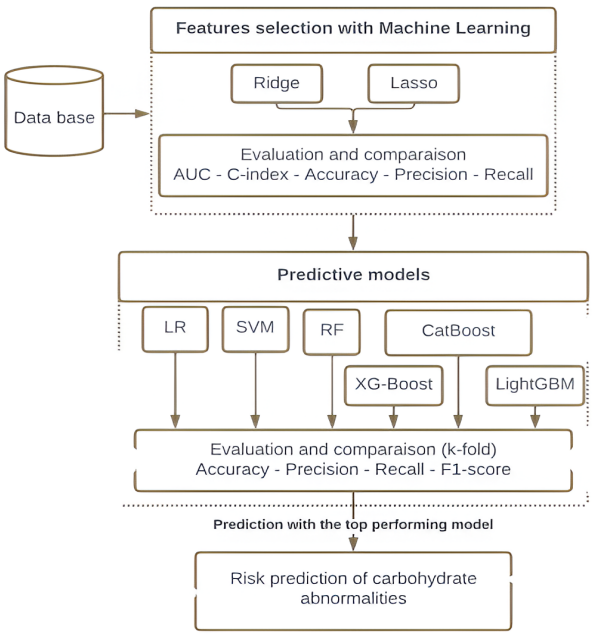


Figure III.6 – Carbohydrate abnormalities predictive models

As shown in Table III.1, the results between the two feature selection approaches were very close. Ridge regression and Lasso regression have shown an accuracy of 90%. Lasso showed a precision of 93% greater than that of Ridge 90%. On the other hand, Ridge had a recall of 84% higher than that of Lasso 81%. Hence, to choose the best model, we consult

the F1-score and the c-index which give results that encompass both the prediction of positive values 1 or negative values 0. Ridge's results are slightly superior to Lasso's with an F1-score of 87% and a c-index of 93% for Ridge and 86% and 92% for Lasso.

Table III.1 – Comparaison of Ridge and Lasso for feature selection

Metric	Ridge	Lasso
Accuracy	0.90	0.90
AUC	0.92	0.92
Precision	0.90	0.93
Recall	0.84	0.81
F1-score	0.94	0.91
c-index	0.93	0.92

The ROC of the two regressions are displayed in Fig.III.7. This shows that the AUC of Ridge of 0.94 is better than Lasso of 0.84. The AUC values confirm that ridge regression is best suited to our dataset for feature selection. Ridge regression has reduced The number of features from 45 to 22. In other words, with only 22 features, we can achieve better and faster results than with 45 features.

(a) ROC curve of Lasso (b) ROC curve of Ridge

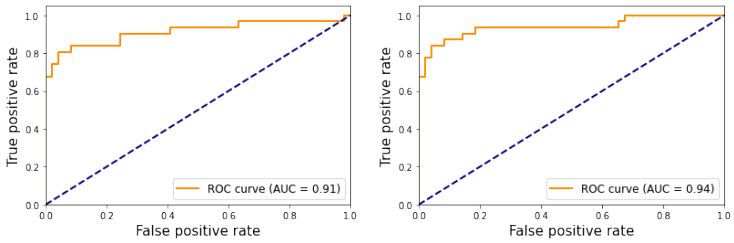


Figure III.7 – ROC curve

Several tests were executed to find the best hyperparameters of each model for predicting carbohydrate anomalies. The final hyperparameters are presented in Table III.2.

Table III.2 – Models hyperparameters for risk prediction of carbohydrate anomalies

Models	**Hyperparameters**
LR	Tolerance: 1×10^{-4}, Solver: 'lbfgs'
SVM	Kernel: 'rbf', Tolerance: 1×10^{-3}
RF	Trees: 800, Split function: "GINI"
XG-Boost	Learning rate: 0.1, Trees: 100
CatBoost	Iterations: 100, Learning rate: 0.05
LightGBM	Trees: 100, Learning rate: 0.05

The study used the Bootstrap approach to compare and validate ML models on a small dataset, with 1500 samples and 60% dataset size, to ensure reliable validation and an accurate 95% confidence interval analysis [Bouthillier 21]. The minimum and maximum

of 95% confidence intervals for each metric (Accuracy, Precision, Recall, F1-score) for several models are shown in Table III.3. Catboost displays the best (minimal:maximal) accuracy and F1_Score that is nearly equal to both LightGBM and XG-Boost. So, visually, Catboost is the best model showing a good predictive ability for both positive and negative subjects. We also see that the Recall of both LR and RF has a large confidence interval, indicating a broad range of Recall values for these two models.

Table III.3 – Min-Max intervals for each model for carbohydrate risk prediction

Models	Accuracy [min: max]	Precision [min: max]	Recall [min: max]	F1_Score [min: max]
RF	[65.1: 95.1]	[70.6: 100]	[25: 92.9]	[40: 91.7]
XG-Boost	[84.4: 97.8]	[81.2: 100]	[64.7: 100]	[78: 97.3]
CatBoost	[86: 97.8]	[80: 100]	[69: 100]	[81.1: 97.3]
LightGBM	[83: 97.9]	[75: 100]	[68.8: 100]	[77.2: 97.7]
LR	[76: 95]	[66: 100]	[53.8: 100]	[64.9: 92.9]
SVM	[77: 97.6]	[62.5: 100]	[61.1: 100]	[68.7: 96.6]

To compare models in a more trustworthy and visual way. Figure III.8 displays each model's confidence intervals for the F1_Score. The figure analysis reveals that confidence intervals for CatBoost, XG-Boost and LightGBM are less wide and admit density peaks between 0.8 and 1 of F1_Score. This contrasts with RF, LR, and SVM, where the confidence interval is relatively wide, with F1_Score peaks lower than those of the other models. In addition, The curves of CatBoost, XG-Boost and LightGBM have a normal, non-skewed distribution with a minimal error margin, compared with the other models, which have a negatively skewed curve with a larger margin of error. This means that the CatBoost, XG-Boost, and LightGBM models perform better and are more accurate in predicting carbohydrate abnormalities.

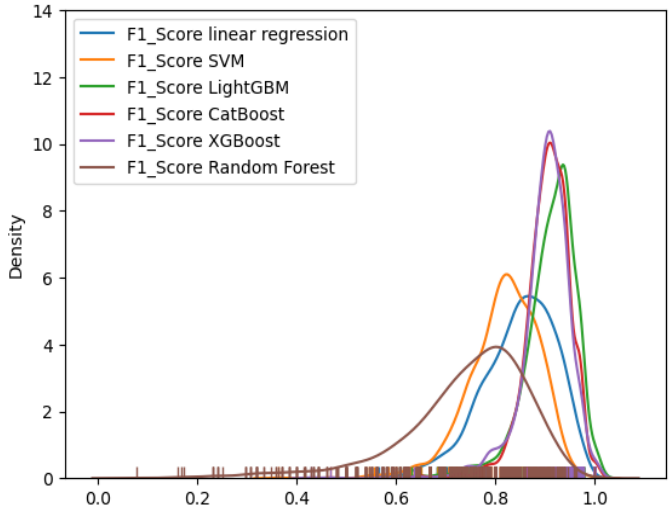


Figure III.8 – F1_Score distribution for several models of carbohydrate risk prediction

The mean values of the confidence intervals were calculated for all models with several metrics to affirm the best model for predicting carbohydrate abnormalities in patients with β-TM.

The model performance results are presented in Table III.4. Accuracy is a metric that shows the predictive capacity in a general way, taking into account the prediction of positive values 1 and negative values 0. The CatBoost showed a best accuracy of

91.9%. For recall and precision, CatBoost admits the best Recall of 84.5% and XG-Boost generated the best precision of 90.6%. This means that the XG-Boost performed better when predicting positive values. On the other hand, CatBoost was the best predictor of negative values. We now come to check the F1-Score metric, which combines both precision and recall to affirm the choice of the model. CatBoost admits the best F1-score of 91.9%, superior to XG-Boost's 91.1%. CatBoost is, therefore, the most adapted model for predicting carbohydrate abnormalities in patients with β-TM. In addition to the previous comparison, we also compared our findings in Table III.4 to three models found in the literature, which have been applied to two other datasets in beta-thalassemia. But first, let us look at the difference between data sets. The dataset in [Yousefian 17] includes 18 characteristics and 255 subjects with β-TM and thalassemia intermediate. On the other hand, our dataset contains only patients with β-TM. This may be advantageous for prediction, as they are most at risk of carbohydrate abnormalities. Our database has only 80 patients, but there is a relatively even balance of 31 positive and 49 negative subjects. There are 22 features, 6 of which are highly significant to carbohydrate abnormalities, with a p-value less than 0.05. The models tested in the literature are KNN, RBFN, and MLP, showing an accuracy of 69.1% , 81.7% , and 89.4% , respectively.

Table III.4 – Comparison between predictive models of carbohydrate abnomalies

Model	Precision-mean (%)	Recall-mean (%)	Accuracy-mean (%)	F1-Score-mean (%)	Validation	NB/PS/NG	NF	TT	NF (p-value≤0.05)
LR	83	76,9	85,5	78,9	Bootstrap	80/31/49	22	M	6
SVM	81,25	80,55	87,3	82,65	Bootstrap	80/31/49	22	M	6
RF	85,3	58,95	80,1	65,85	Bootstrap	80/31/49	22	M	6
XG-Boost	90,6	82,35	91,1	87,65	Bootstrap	80/31/49	22	M	6
CatBoost	90	84,5	91,9	89,2	Bootstrap	80/31/49	22	M	6
LightGBM	87,5	84,4	90,45	87,45	Bootstrap	80/31/49	22	M	6
KNN [Yousefian 17]	32.3	-	69.1	-	No	255/74/181	18	M/I	-
RBFN [Yousefian 17]	42.9	-	81.7	-	No	255/74/181	18	M/I	-
MLP [Yousefian 19]	61.73	81.08	89.4	-	No	255/74/181	18	M/I	-

Abbreviations: NB/PS/NG: Number of subjects/Positive target/Negative target, NF: Number of features, TT: Thalassemia type, M: Major, M/I: Major/intermediate

III.3.2 Risk prediction of MetS in screening sessions

In this part, we aim to simultaneously predict the risk of MetS through a risk-scoring calculation based on the SVM model for each MetS definition and then generalize and normalize according to several outputs of MetS screening definitions.

We have two datasets as input. The first dataset (DS1) comprises 1,100 adolescent [Benmohammed 15], while the second dataset (DS2) includes 266 adolescent [Benmohammed 11]. Each dataset has three output columns, individually indicating the presence or absence of MetS for each definition (Cook, International Diabetes Federation (IDF) and De Ferranti). Both datasets share the same features, with the distinction that subjects in DS2 are all overweight and obese. Consequently, a new multi-class column is introduced to incorporate information about overweight and obesity in this dataset. Subsequently, these

two datasets are merged to establish a generalized predictive framework for conditions correlating causally with MetS, particularly overweight and obesity. Hence, the initial step of our methodology involves merging the two databases.

Next, the second step involves ensuring MetS prediction using svm to various definitions of MetS. Building upon this predictive capability, we extract risk coefficients associated with each feature. Subsequently, we harness these risk coefficients to formulate the individualized risk function for each subject. Finally, we determine the optimal threshold for classification by leveraging the risk functions, thereby categorizing subjects into high or low-risk groups.

Finally, the third and final step involves merging the various scoring associated with each definition to formulate a novel normalized scoring system of MetS screening.

MetS Risk Scorification with SVM

The risk estimation process is presented in Algorithm 1. We will initially develop and ensure prediction using SVM model. Following this prediction, we will calculate each feature's importance or risk coefficients using the Soft Margin technique. Subsequently, each subject's risk function will be established based on the previously calculated risk coefficients. Finally, the optimal classification threshold for risk functions tied to each subject will be determined to ensure an ideal binary risk classification according to each definition of MetS.

Algorithm 1 STEP 2: Risk estimation

Step 1. Develop the SVM model and ensure prediction.

Step 2. Calculate the Risk Coefficient for each feature.

$$\min_{\mathbf{w},b,\zeta} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\zeta_i \tag{III.9}$$

subject to the constraints

$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \zeta_i \quad \text{and} \quad \zeta_i \geq 0 \quad \text{for}\, i = 1, ..., n \tag{III.10}$$

where:

- $\mathbf{w}$ is the weight vector,
- b is the bias,
- C is the regularization parameter controlling the trade-off between margin maximization and classification error minimization,
- ζ_i are the slack variables allowing for margin violations,
- y_i are the class labels,
- $\mathbf{x}_i$ are the feature vectors.

Step 3. Development of the risk function for each patient :

$$\text{Risk (patient)} = \sum_{i=1}^{n}(\text{Feature value}_i \times \text{Risk Coefficient}_i) \tag{III.11}$$

Step 4. Calculate the optimal threshold for classification (high or low risk).

Risk Normalization

In the previous step, we proposed to develop a binary risk scoring system for the MetS detection according to each definition. The current objective is to merge these scores to create a normalized global score that considers all three definitions of MetS. As illustrated in Figure III.9, a new column will be added to the database based on the three output columns from the definitions to evaluate our methodology.

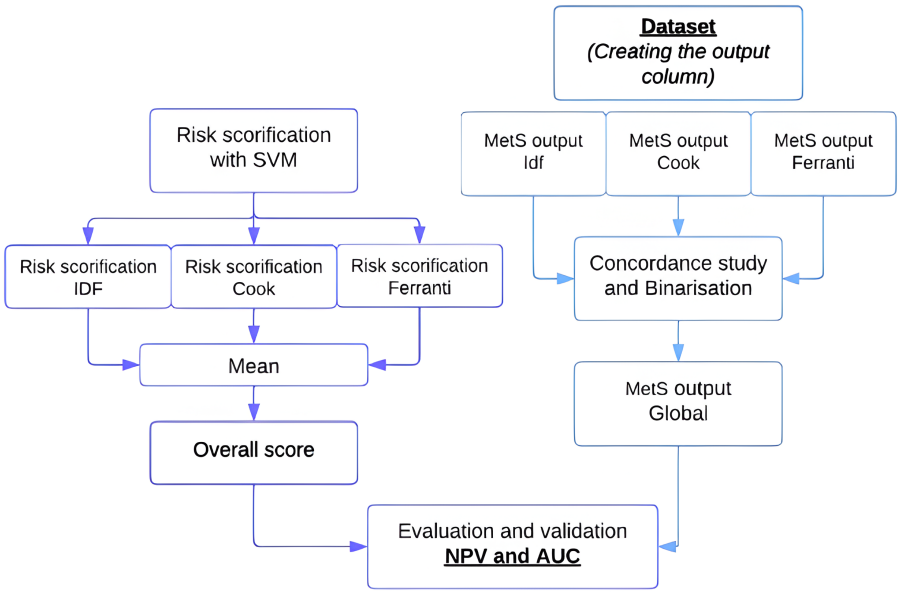


Figure III.9 – STEP 3: Risk normalization
Abbreviations : NPV: Negative Predictive Value, AUC: Area Under the Curve

Risk is presented in the form of a list for each definition. Each element of this list presents the risk of a specific subject. The idea is to merge these three lists, element by element, by the median value, to obtain a final list that considers all three definitions.

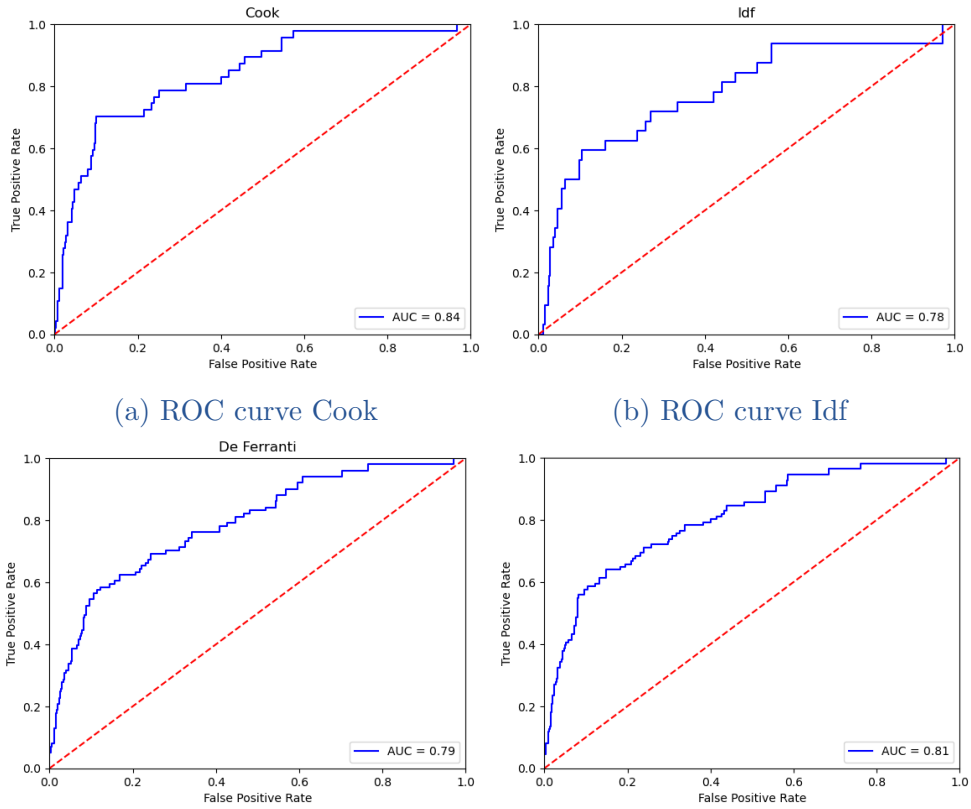
Results

To display the predictive performance of SVM for MetS screening, we chose specificity and sensitivity as metric evaluators to show the predictive ability of both positive and negative MetS values. High sensitivity, indicated by values close to 1, reflects a strong ability to predict positive outcomes (subjects with MetS). Conversely, high specificity, also close to 1, indicates a strong ability to predict negative outcomes (subjects without MetS). In Table III.5, it is observed that for the IDF definition, specificity was 0.89, while sensitivity was 0.59. This suggests our model better predicted negative outcomes than positive ones under this definition. Conversely, for the De_Ferranti definition, the model exhibited a strong predictive ability for positive outcomes, with a sensitivity of 0.87. Lastly, under the third Cook definition, the model demonstrated a strong predictive ability for positive and negative outcomes.

Table III.5 – Sensitivity, Specificity, and Cut-off Values for Several Outputs

Output	Sensitivity	Specificity	Cut-off
Output_Cook	0.70	0.89	1.01
Output_Idf	0.59	0.89	0.36
Output_Ferranti	0.58	0.87	127.85

These evaluators have a limitation in that they are influenced by the distribution of data across both classes of each output. Given that our outputs are relatively unbalanced, with more negative than positive values, the ROC curve and AUC are considered more reliable metrics for evaluating prediction accuracy. Parts (a), (b), and (c) of Figure III.10 display the three ROC curves corresponding to each definition. The AUC values for Cook, IDF, and De_Ferranti were 0.84, 0.78, and 0.79, respectively, indicating strong predictive ability for both positive and negative values.



(a) ROC curve Cook

(b) ROC curve Idf

(c) ROC curve De Ferranti

(d) ROC curve of unified risk

Figure III.10 – ROC curves

As highlighted earlier, sensitivity and specificity may not be optimal for evaluation, particularly due to the imbalance in our data. In this section, our attention is directed to the NPV, calculated at 0.70, indicating that 70% of those predicted as negative are indeed negative. Value (NPV) becomes a crucial metric in our analysis, indicating the proportion of negative predictions that are true negatives. Alongside NPV, the ROC curve and AUC provide reliable evaluation metrics. In part (d) of Fig.III.10, the AUC is 0.81, demonstrating a strong predictive ability for identifying individuals with low risk of MetS.

III.4 Discussion and limits

ML has shown promising results in predicting the risk of glucose abnormalities and MetS. However, there are several limitations, both in gaining confidence in physicians' risk prediction and discussing the cost and financial burden of collecting data for risk assessment.

In predicting carbohydrate abnormalities for patients with β-TM, a very good performance with a dataset of 80 patients raises doubts about the reliability of the prediction. This doubt makes our risk prediction not suitable for the medical field. Therefore, the question is, how can we provide enough information to physicians to assess the reliability of the prediction?

In addition, for MetS screening, the acquisition and data in the datasets require a significant financial burden, especially during screening sessions where multiple individuals are targeted. This makes a primary step like risk prediction impractical. Therefore, a second research question is how can we reduce the financial burden for MetS risk predictions and screening?

III.5 Conclusion

In this chapter, we aimed to predict carbohydrate abnormalities and MetS using ML and leveraging physician-supplied datasets. Therefore, we started by defining several concepts of supervised ML and outlining linear and decision-tree based models. We then explained various evaluation metrics, their significance, and how they can be interpreted. Subsequently, we introduced two methodologies for predicting the two targeted diseases. The results demonstrated promising performance for both diseases, showing good predictive capability for both positive and negative subjects. These strong performances were maintained even when compared to other research found in the literature. However, we also identified limitations and challenges in these predictions. While there was good predictive capacity with a small dataset of carbohydrate abnormalities, there were doubts about the predictions' reliability. Thus, the first limitation pertains to physicians' assessment of prediction reliability. Furthermore, for MetS prediction, the high cost of data acquisition and collection makes risk prediction less appealing and sometimes not feasible.

Chapter IV

XAI for Assessing Predictive Reliability and Managing Medical Financial Expenses

IV.1 Introduction

In response to the last two research questions posed in the limitations of the previous chapter, we propose in this chapter to leverage XAI to provide access to physicians to assess the reliability of predicting glucose abnormalities. Furthermore, we aim to use XAI to reduce the financial burden of predicting the risk of MetS screening sessions.

XAI has emerged as a crucial aspect of healthcare, especially in predicting and screening endocrine diseases. XAI refers to the ability of ML models to provide transparent and interpretable insights into their decision-making process. In medical practice, understanding why a model makes a particular prediction is often as important as the prediction itself, particularly when dealing with critical patient care decisions.

In this chapter, we delve into the XAI and its significance in the context of supervised models for endocrine disease risk prediction. We begin by elucidating various XAI approaches and methodologies in Section IV.2, which aim to provide a broad overview of a model's behavior and decision-making process. Subsequently, in Section IV. 3, we explore several self-, model-dependent, and global XAI approaches based on linear or tree-based models. Then, we describe several Post-Hoc and local XAI approaches focus on explaining individual predictions made by ML models. Two prominent techniques in this domain are SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). These methods offer insights into the specific factors contributing to each prediction, enabling clinicians to understand and trust model decisions at the individual patient level. Furthermore, we apply XAI techniques to interpret the predictions and to evaluate this prediction of carbohydrate abnormalities in Section IV.5, and reduce the financial cost of MetS screening in Section IV.6. By elucidating the underlying factors driving model predictions, XAI facilitates a deeper understanding of disease risk factors and aids in clinical decision-making. Lastly, Section IV.7 discusses the limitations and challenges associated with XAI, including model complexity, interpretability tradeoffs, and potential biases. Despite these challenges, integrating XAI techniques is promising for enhancing ML models' transparency, accountability, and trustworthiness in healthcare settings. Finally, Section IV.8 concludes the chapter.

IV.2 XAI Methodology

XAI is a method for explaining a model's predictions. The idea is to explain the main reasons for a specific prediction, or based on which the model has created the classification threshold.

Explainability is useful in cases where the model shows good performance, but also when the performance is mediocre. If the model has shown good performance, it is essential to verify which variables are critical for this prediction and what exactly the prediction is based on. In this case, with medical expertise, we can observe that some features are logical from a medical standpoint (such as smoking for cardiovascular diseases), indicating that the prediction is reliable. Otherwise, the prediction could be biased, based on a random correlation that lacks medical logic, thus requiring a return to the feature selection stage. Conversely, if the model has shown mediocre performance, analyzing explainability, in collaboration with medical expertise, allows for revisiting the data preparation to improve its quality.

Various methods have been developed in the field of XAI to make ML models more understandable. These methods can be categorized into different aspects based on their application, extraction techniques, and scope of explanations. This section provides a detailed overview of these aspects, divided into three main categories: whether the explanation method is model-specific or model-agnostic, how explanations are extracted, and whether explanations are for specific instances or the entire model as shown in Figure IV.1.

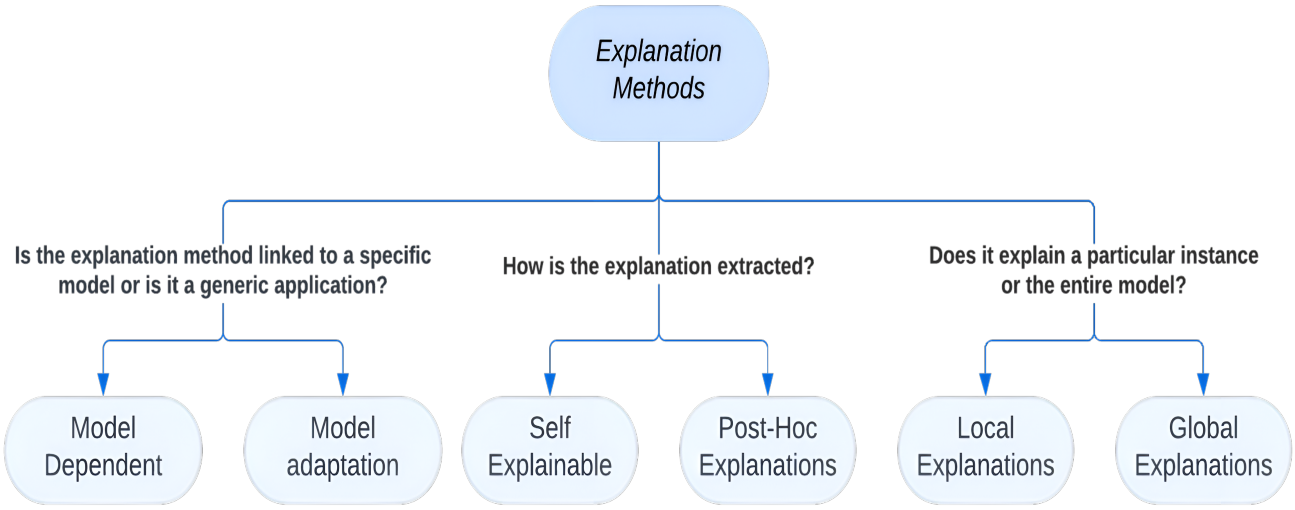


Figure IV.1 – XAI Methods

IV.2.1 Is the Explanation Method Linked to a Specific Model or Is It a Generic Application?

In the field of XAI, one of the fundamental questions is whether an explanation method is inherently tied to a specific model or if it can be applied across various models. This distinction can be divided into two primary categories: Model-Dependent methods and Model-Agnostic methods.

Model Dependent Methods are those that are designed to work with specific types of models or architectures. For example, decision trees and linear regression models are inherently interpretable because their structures are straightforward and their decision-making processes can be directly understood. These methods leverage the internal mechanisms of the model, such as the model's architecture or parameters, to generate explanations. They provide detailed insights into how individual predictions are made based on the model's internal logic. However, these explanations are closely tied to the model they are designed for and may not be easily transferable to other models.

In contrast, Model Agnostic Methods are not tied to any particular model type. They are designed to work with various models, including complex ones like deep neural networks. Techniques such as LIME and SHAP fall into this category. These methods apply general principles or algorithms to interpret the outputs of different models. Although they offer flexibility and can be used with many different models, they might not capture the nuances of the specific model's internal workings as effectively as model-dependent methods.

IV.2.2 How Is the Explanation Extracted?

Different methodologies can be used to extract explanations from AI models, broadly classified into Self-Explainable Models and Post-Hoc Explanation methods.

Self-Explainable Models are designed with inherent interpretability features. These models are built with the express goal of being understandable to humans. Examples include linear regression, decision trees, and rule-based systems. In self-explainable models, the explanation is a natural byproduct of the model's structure. For instance, decision trees provide clear paths showing how each decision was made, and linear models offer straightforward coefficients that indicate the influence of each feature on the predictions. The main advantage of these models is that they provide explanations directly and transparently as part of their operation, which can be very informative for users. However, these models may not always be capable of capturing complex relationships in the data.

In contrast, Post-Hoc Explanation Methods refer to techniques applied after the model has been trained. These methods do not alter the model itself but instead analyze the model's output to produce explanations. Examples of post-hoc methods include feature importance metrics, saliency maps, and counterfactual explanations. For instance, SHAP values can be used to decompose a model's output into contributions from each feature, and LIME can approximate the behavior of a black-box model with an interpretable surrogate model. While these methods can provide insights into the workings of complex models, they are often indirect and may not always provide a complete or accurate picture of the model's decision-making process.

IV.2.3 Does XAI Explain a Particular Instance or the Entire Model?

The scope of XAI methods' explanations can be categorized into Local Explanations and Global Explanations based on whether they target specific instances or the entire model.

Local Explanations focus on explaining individual predictions made by a model. These methods concern understanding why a model made a particular decision for a specific instance. Techniques such as LIME and individual feature attribution methods fall into this category. For example, LIME generates explanations for specific predictions by approximating the model's behavior with a simpler, interpretable model near the explained instance. Local explanations are useful for understanding individual decisions and validating model behavior case-by-case basis. However, they do not provide a comprehensive view of the model's overall decision-making processes or general patterns.

In contrast, Global Explanations aim to provide insights into the overall behavior of the model across all instances. These explanations seek to convey how the model functions as a whole, describing general patterns or rules that govern the model's predictions. Techniques for global explanations include feature importance ranking, rule extraction methods, and visualization of decision boundaries. Global explanations are useful for understanding the broader decision-making framework of the model and assessing its overall behavior, but they might lack the granularity needed for specific instances.

IV.3 Self Explainable and Model Dependent Global XAI Approaches

IV.3.1 Explaining Logistic Regression Results

Logistic regression is a widely used linear classification algorithm that models the probability of a binary outcome by fitting a logistic function to input features.

This function predicts the probability that a given instance belongs to a specific class. In logistic regression, each feature is assigned a coefficient that quantifies its impact on the log-odds of the outcome variable. A positive coefficient means that an increase in the feature's value is linked to a higher likelihood of the positive class, whereas a negative coefficient suggests the opposite. The magnitude of the coefficient indicates the strength of influence. The exponential of these coefficients represents the odds ratio, which quantifies the change in odds of the positive class for a one-unit increase in feature value. An odds ratio above 1 suggests an increase in odds, while one below 1 indicates a decrease. Additionally, logistic regression provides p-values for each coefficient, highlighting the statistical significance of the feature-outcome association; a p-value below 0.05 typically indicates significant predictive capability. The intercept term represents the log-odds of the positive class when all features are zero, adjusting the decision boundary to account for the baseline probability. To calculate the importance coefficients of features, the input features are first standardized to a mean of 0 and a standard deviation of 1, ensuring all features are comparable. The model is then trained on this standardized dataset to obtain coefficients. The importance of each feature is determined by the absolute value of its coefficient, reflecting the feature's influence magnitude, irrespective of direction. Optionally, these importance coefficients can be normalized to sum to 1, offering a relative measure of feature importance. Through these steps, logistic regression results can be interpreted, providing insight into the importance of each feature in predicting the outcome.

IV.3.2 Explaining SVM Results

SVM is a robust supervised learning algorithm primarily used for classification tasks, effectively finding the hyperplane that optimally separates the classes in the feature space. SVM identifies critical training data points, known as support vectors, which are nearest to the decision boundary and are essential for defining this boundary and classifying new data points. The margin, or the distance between the decision boundary and these closest support vectors, is maximized by SVM to enhance the classification confidence. SVM can accommodate nonlinear decision boundaries through kernel functions such as linear, polynomial, and radial basis functions (RBF), with the choice of kernel significantly impacting the model's flexibility and performance. In a linear SVM, each feature is given a coefficient, highlighting its influence on the decision boundary. The coefficients linked to the support vectors are especially critical as they directly shape the decision boundary's position and orientation, with any changes in these values potentially altering classification results. To ascertain the importance of features within SVM, the process starts with scaling the input features to uniform ranges and then training the SVM model on this scaled dataset to extract coefficients for each feature. These coefficients signify the importance of each feature in defining the decision boundary, with larger coefficients indicating a stronger influence on classification outcomes. Optionally, these coefficients

can be normalized to sum to 1, offering a relative measure of feature importance. This methodical approach allows for a detailed interpretation of SVM results, providing insights into the significance of each feature in the classification process.

IV.3.3 Decision Tree

Decision trees are versatile and interpretable ML models extensively employed for both classification and regression tasks. These models are structured with nodes, branches, and leaves, where each node represents a decision based on a feature, branches depict the outcomes of these decisions, and leaves indicate the final predicted class or value. A key aspect of decision trees is their inherent ability to measure feature importance, highlighting how frequently a feature is utilized in making decisions; features that are closer to the tree's root or appear higher in the tree hierarchy significantly influence the final prediction. Decision trees employ specific splitting criteria like Gini impurity or information gain to select the optimal feature and threshold for data segmentation at each node, aiming to enhance the purity of the resulting subsets. However, to mitigate overfitting, especially in noisy data scenarios, decision trees may require pruning techniques such as cost complexity pruning, which trims unnecessary branches to streamline the model. The interpretability of decision trees lies in their capacity to form clear and comprehensible decision rules that can be visually represented and easily understood, allowing for a straightforward explanation of how predictions are derived. To interpret decision tree models effectively, one typically begins by visualizing the tree to comprehend its structure and decision-making process. Assessing feature importance through the depth and frequency of nodes within the tree helps identify critical features, while tracing decision paths from the root to the leaves elucidates how specific predictions are made and which features are pivotal. Evaluating the splitting criteria provides additional insights into how the model partitions the feature space and selects thresholds, thus offering a comprehensive understanding of the decision-making process. Following these interpretative steps can give one a deeper insight into decision tree models and their predictive behaviors.

IV.3.4 Random Forest explanations

Random forests are advanced ensemble learning methods that leverage multiple decision trees to enhance predictive accuracy and generalization. These models build an ensemble of decision trees, each trained on a randomly selected subset of the training data and features, which collectively work to diminish overfitting and bolster robustness by integrating the predictions from multiple trees. Random forests assess feature importance by measuring how significantly each feature reduces the impurity across the nodes in the trees, considering features that yield substantial impurity reductions across the ensemble as more crucial for prediction accuracy. Bootstrap aggregation, or bagging, generates diverse data subsets for each tree, effectively minimizing variance and delivering more stable and reliable predictions. The correlation among trees within a random forest plays a critical role; less correlated trees tend to enhance the ensemble's performance by providing a variety of independent predictions. Additionally, random forests use out-of-bag (OOB) error estimation as a method to evaluate model performance, leveraging data not included in the training of each tree to provide an unbiased accuracy estimate. To interpret random forest models, one should examine the average decrease in impurity caused by each feature across all trees, consider the ensemble prediction that aggregates the outputs of all trees—typically through majority voting or a weighted average—and utilize visualization tools like tree plots or partial dependence plots to deepen under-

standing of the decision-making process. The out-of-bag error is crucial for assessing the model's generalization capabilities. Following these interpretive steps, one can thoroughly understand random forest models' workings and predictive strengths.

IV.4 Post-Hoc Explanations and Model Adaptive Local XAI Approaches

IV.4.1 SHAP (SHapley Additive exPlanations)

SHAP is a powerful method for explaining the output of ML models. It provides insights into individual predictions by indicating how each feature contributes to the model's output. The foundation of SHAP values lies in cooperative game theory, which ensures several desirable properties such as local accuracy, missingness, and consistency in its explanations.

To understand how SHAP values are calculated, we need to delve into the mathematical details of the process. The calculation involves decomposing the model's prediction for a specific instance into contributions from each feature. For a given prediction $f(x)$, where x is the input instance, SHAP values for a particular feature i are computed using the following formula:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} \Big(f(x_{S\cup\{i\}}) - f(x_S) \Big) \tag{IV.1}$$

In this equation, N represents the set of all features, and $S \subseteq N \setminus \{i\}$ denotes subsets of features that do not include feature i. The term $x_{S\cup\{i\}}$ refers to the instance with feature values from subset S along with feature i, while x_S represents the instance with feature values from subset S. The cardinality of subset S is denoted by $|S|$. This approach effectively breaks down the model's prediction into additive contributions from each feature.

The interpretation of SHAP values reveals how each feature impacts the model's output. Positive SHAP values indicate that the feature increases the prediction, whereas negative values suggest a decrease. For instance, large positive SHAP values highlight features that significantly enhance the model's output for a particular instance, while large negative SHAP values indicate features that substantially reduce the output. Conversely, SHAP values close to zero signify that the feature has minimal impact on the model's output for that instance.

SHAP's application varies depending on the types of models it is used with, and different variants of SHAP are suited for different contexts. Below is an overview of its application with various models, such as linear, tree-based, and neural networks.

— Linear SHAP: SHAP provides a straightforward and intuitive way to understand the importance of features for linear models, such as linear regression or logistic regression. Linear SHAP calculates Shapley values based on the model's coefficients, making it clear how each variable contributes to the final prediction. This is particularly useful for models where interpretability is crucial, such as economics or medical research.

— Tree SHAP: Tree SHAP is specifically designed for tree-based models, including random forests and boosting models (such as XGBoost). This SHAP variant leverages the tree structure to efficiently compute Shapley values. Tree SHAP can handle the complex interactions and non-linearities inherent in these models, providing an accurate explanation of each variable's contribution to the prediction. This helps users understand which features most influence the model's decision.
— Deep SHAP: Neural networks are often considered "black boxes" due to their complexity. Deep SHAP combines Shapley values with backpropagation techniques to explain neural network predictions. Using Deep SHAP, it is possible to visualize the contributions of individual neurons and specific layers, aiding in the interpretation of decisions made by complex models like convolutional neural networks (CNNs) or recurrent neural networks (RNNs).
— Kernel SHAP: Kernel SHAP is a more general method that can be applied to any model type, including those for which specific SHAP variants do not exist. It uses a sampling approach to estimate Shapley values and is particularly useful for models where the exact computation of Shapley values would be too computationally expensive. Kernel SHAP is flexible and can adapt to various types of data and models, although it may be less efficient for very complex or large models.

To make the most of SHAP values, various visualization techniques can be employed. Summary plots offer an overview of feature importance across multiple predictions, while dependence plots illustrate the relationship between a feature and the model output. Additionally, force plots provide a detailed view of individual SHAP values for a specific prediction, showing the contributions of each feature. Figure IV.2 displays an example of a SHAP visualization, demonstrating these concepts in practice.

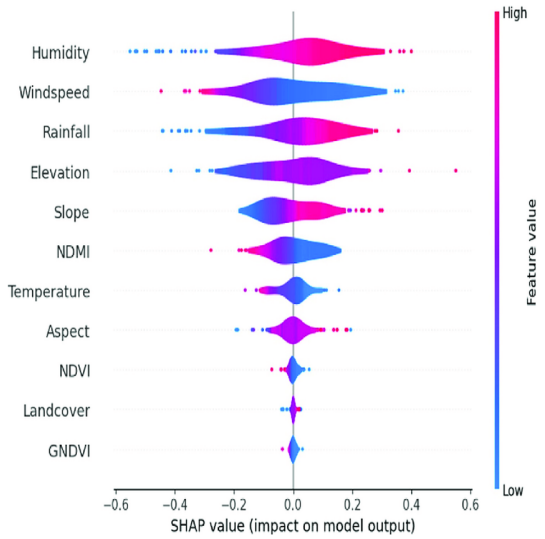


Figure IV.2 – SHAP visualization example

IV.4.2 LIME (Local Interpretable Model-agnostic Explanations)

LIME is a technique designed to explain individual predictions made by ML models. By offering locally faithful explanations, LIME helps users understand how a model arrives at specific decisions. The core idea behind LIME is to approximate complex models

with simple, interpretable surrogate models, such as linear regressions, which can be easily understood and analyzed. Importantly, LIME is model-agnostic, meaning it can be applied to any black-box model without needing to know its internal mechanics.

To generate LIME explanations, the method works by fitting a simple, interpretable model to a set of perturbed samples around the instance of interest. These perturbed samples are created by randomly modifying the feature values of the original instance, while keeping the label constant to maintain the relationship between the features and the prediction. The process for calculating LIME explanations is mathematically expressed as:

$$\hat{f}_{\text{lime}}(z) = \arg\min_{w} \sum_{j=1}^{N} \text{sim}(x, z_j)(f(z_j) - w^T z_j)^2 + \Omega(w) \tag{IV.2}$$

In this equation, $\hat{f}_{\text{lime}}(z)$ represents the prediction of the surrogate model for a perturbed sample z, while N denotes the number of perturbed samples used. The term $\text{sim}(x, z_j)$ is a similarity function that measures how close the perturbed sample z_j is to the original instance x. The coefficients of the surrogate model are represented by w, and $\Omega(w)$ denotes a regularization term that helps prevent overfitting of the simple model.

The interpretation of LIME explanations revolves around understanding how feature values influence the model's predictions. Positive coefficients in the surrogate model indicate that increasing a feature's value leads to higher predictions, whereas negative coefficients suggest a decrease in predictions with increasing feature values. Large positive coefficients highlight features that have a significant impact on the prediction for the instance, while large negative coefficients indicate features that substantially reduce the prediction. Features with coefficients close to zero have minimal impact on the prediction.

To visualize LIME explanations, several techniques can be employed. Feature importance plots are commonly used to illustrate which features most influence the model's predictions. Additionally, bar charts or heatmaps can be used to provide a detailed view of the contributions of each feature for individual instances. These visualizations offer a clearer picture of the model's behavior and decision-making process.

LIME is used in various areas, including model debugging, comparison, and trust verification. By enabling users to explore and understand the model's behavior, LIME helps identify influential features, detect potential biases, and verify the reliability of predictions. By providing interpretable explanations for complex models, LIME supports better model management and trust in ML systems.

IV.5 XAI for reliability assessment of the prediction of carbohydrate abnormalities in patients with β-TM

In the previous chapter, we tested several ML models for risk prediction of carbohydrate abnormalities. The CatBoost model showed better performance, which will satisfy the predictive capabilities expected by physicians as shown in IV.3. However, such performance with a dataset containing only 80 subjects is slightly ambiguous. Hence, this section aims to identify the most important contributors to such a prediction. Then, physicians will be offered this ranking of contributors to judge and evaluate the reliability of the prediction from a medical point of view.

If the most important characteristics have a medical logic, in this case, the prediction was based on a well-explained and reliable logic. Otherwise, if the most important features have no strong association with carbohydrate abnormalities, and the most related features admit a weak ranking, in this case, the prediction has probably been based on a false correlation between the variables, and these variables should be removed to avoid basing the ML.

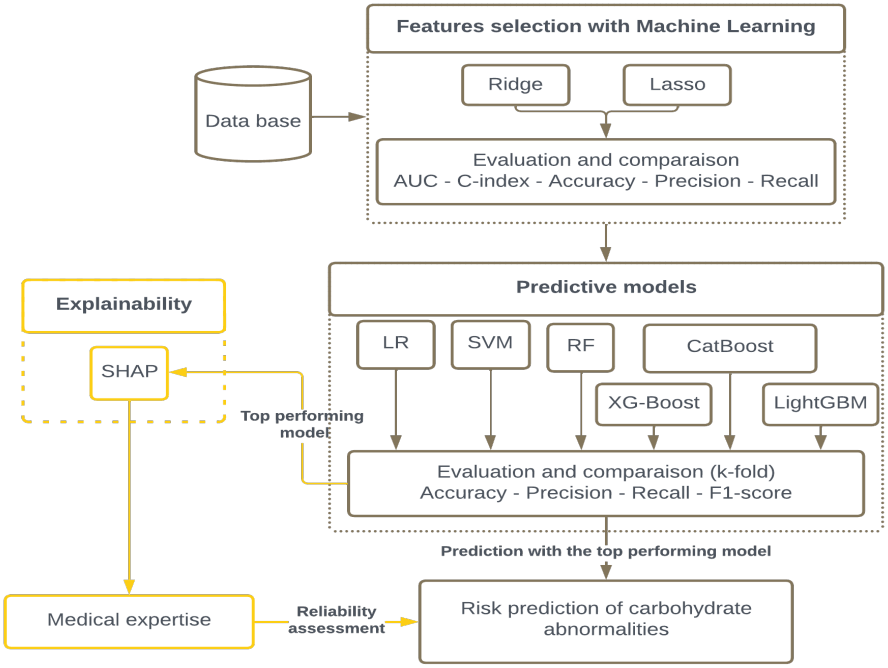


Figure IV.3 – Figure III.6 Extension: Carbohydrate Abnormalities XAI

The graph Fig.IV.4 shows the ranking of the features that contribute most to the prediction of the CatBoost model, in descending order from top to bottom. Features are shown on the y-axis and Shapley values on the x-axis. The color decay from red to blue indicates whether the value of each characteristic is large or small (red: maximum values, blue: minimum values). Each point on the graph represents the Shapley values for each specific characteristic associated with a patient. Thus, the number of points for each entry equals the number of patients. The characteristics are ranked from most important (top) to least important (bottom) in predicting carbohydrate abnormalities. When the smallest value of a characteristic admits a negative Shapley value and the largest admits a positive Shapley value, the more important this characteristic, the higher the risk of having carbohydrate abnormalities. If a characteristic has a maximum value with negative Shapley values and a minimum value with positive Shapley values, the smaller the characteristic, the higher the risk of developing carbohydrate abnormalities.

Medical expert reports: By analyzing the diagram presented in Figure IV.4, we interpret that the most important variables for predicting carbohydrate anomalies are respectively, the 2-hour post 75g glucose, FPG (Fasting Plasma Glucose), HOMA-IR, and Serum Ferritin. In previous research, doctors have identified these features as good predictors of carbohydrate diseases. So, it ensures the reliability and confidence of doctors for our prediction, as it is primarily based on these variables identified as the most crucial.

Finally, the reliability measures of XAI showed a concordance of 0.96, indicating strong agreement with the model's assessment of the features' importance. Stability of 0.35 shows moderate variation in feature impact.

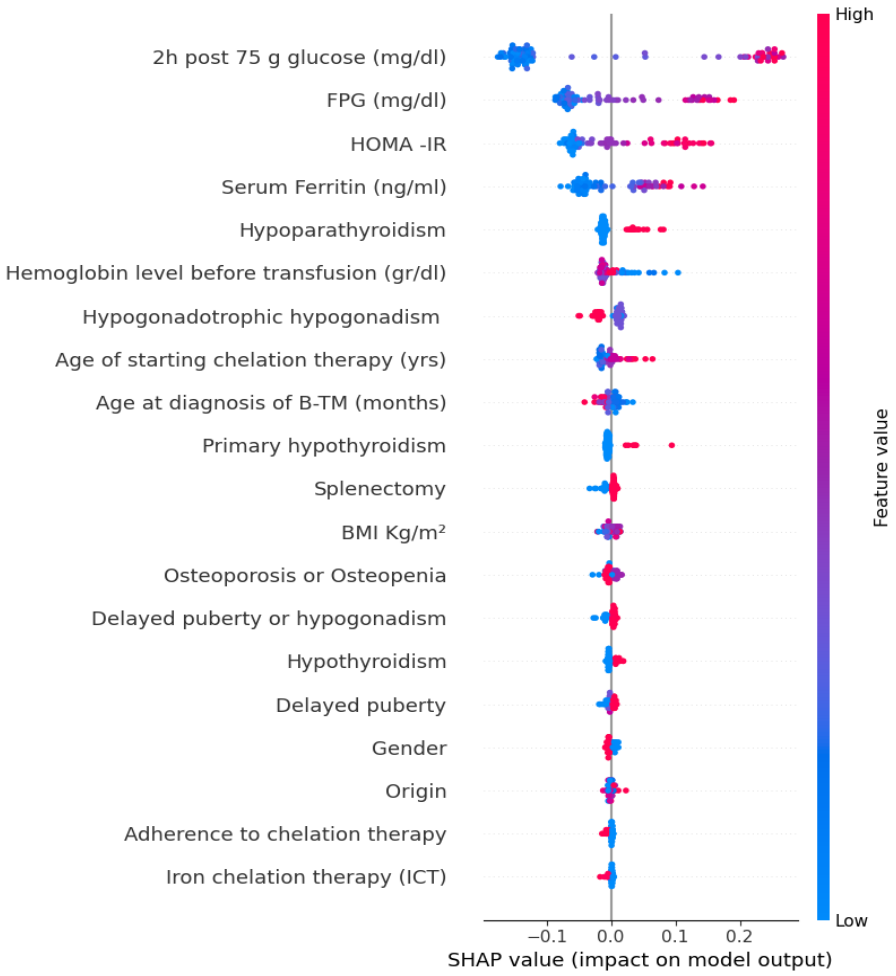


Figure IV.4 – Shapley Visualization (feature importance ranking on model output)
Abbreviations: β-TM : β-Thalassemia major, HOMA: homeostasis model assessment for insulin resistance, BMI: Body mass index FPG: fasting plasma glucose.

IV.6 XAI for a Low-cost Risk Prediction of MetS in Screening Sessions

The datasets collected during MetS screening sessions consist of biological and clinical variables. According to physicians, extracting biological variables for a large population during a MetS screening session requires a significant financial and time burden. To reduce these financial burdens, we have considered limiting the use of biological characteristics and favoring the use of clinical variables. Our idea is to identify low-risk individuals by exploiting only clinical variables, thus avoiding the need for blood tests to extract biological characteristics for this population. Therefore, we have ensured that blood tests are used only for at-risk people, considering the three MetS definitions (IDF, Cook, or Ferrenti). The aim is to exploit the XAI to select the most important clinical characteristics to create the MetS risk function and identify low-risk individuals.

However, the obvious question arises regarding XAI feature selection and data fusion as shown in Figure IV.5. Is it better for prediction to merge the data and then select the features, or is it better to select the features and merge them?

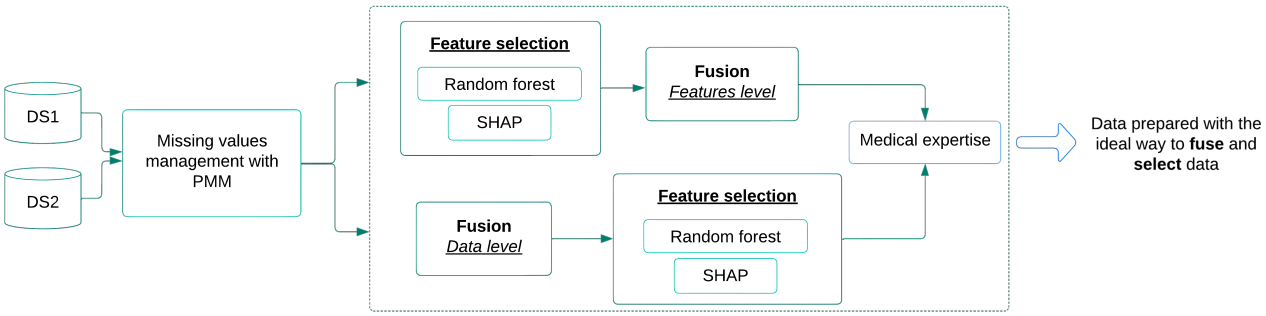


Figure IV.5 – Data fusion and selection

Feature selection and data fusion: Data fusion is the process of combining data from multiple sources to create a more complete and accurate representation of a situation or phenomenon. By merging information from different sources, the objective of data fusion is to improve the overall quality of the data. It is a very important method, especially for the generalization of ML-based approaches.

A state-of-the-art review was conducted to examine the various types of data fusion. Data fusion is divided into three main types: data-level fusion, feature-level fusion, and decision-level fusion [Khaleghi 13, Ayed 15, Jing 18]. Data-level fusion involves merging raw data before preprocessing and feature selection. The idea behind this type of fusion is to exploit the maximum amount of data for preparation and selection. On the other hand, feature-level fusion involves merging data after preparation and selection. Finally, decision-level fusion combines the different classification results by using multiple datasets or multiple outputs.

In our study, we will compare the two fusion approaches (data-level and feature-level). The first approach involves selecting features (with SHAP and random forest) from each dataset and then merging the selected features. Conversely, the second approach involves merging the data into a single dataset (DS_Merged) and then exploiting this dataset for feature selection.

Finally, decision-level fusion will be studied and applied to combine the classification results of the three definitions of MetS screening to estimate a unified risk that considers the three different definitions.

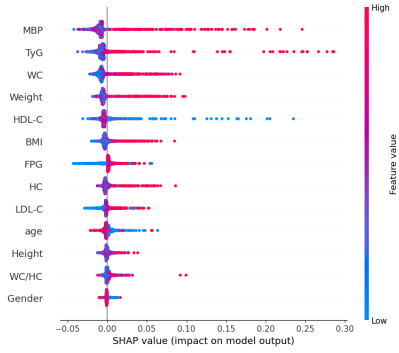
Feature Selection Before Data Fusion (feature-level): Table IV.1 shows the most important features for each data set related to the three definitions of MetS.

Table IV.1 – Feature Importance Ranking Before Data Fusion for Several MetS definitions and Datasets

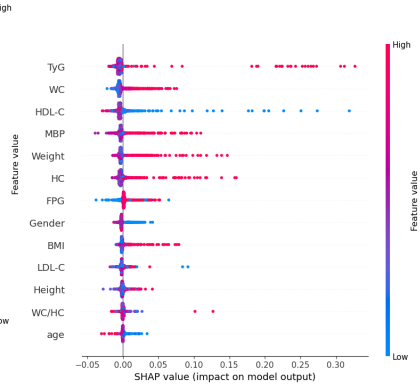
Database_Output	1st	2nd	3rd	4th
DS1_Cook	MBP	Tg	WC	HDL-C
DS1_Idf	HDL-C	WC	MBP	FBG
DS1_Ferranti	MBP	Tg	HDL-C	TyG
DS2_Cook	HDL-C	FBG	TyG	Tg
DS2_Idf	FBG	TyG	HDL-C	MBP
DS2_Ferranti	HDL-C	Tg	TyG	MBP

The goal is to identify the key characteristics to consider when selecting variables from both datasets and the three definitions. Therefore, the frequency of variable matches deemed important in both datasets and all three definitions is detailed in section (d) of Fig.IV.6. It is observed that HDL-C is a crucial variable across all definitions and datasets, with a matching frequency of 6. Additionally, MBP is identified as an important variable 5 times. Furthermore, a review of Table IV.1 reveals that MBP and HDL-C consistently rank as the most significant variables in the entire dataset.

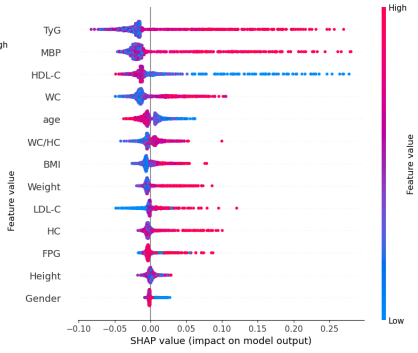
Feature Selection After Data Fusion (data-level): Parts (a), (b), and (c) of Figure IV.6 illustrate the most significant features following data fusion for each definition. The y-axis represents features, while the x-axis represents Shapley values. The color gradient from red to blue indicates the value of each feature, with red denoting the highest values and blue denoting the lowest values. Each point on the graph represents the Shapley value for a specific attribute linked to a patient. The features on the graph are arranged in descending order of importance.



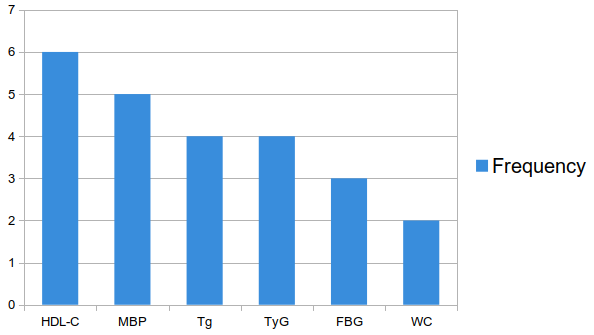
(a) SHAP values related to Cook after data fusion



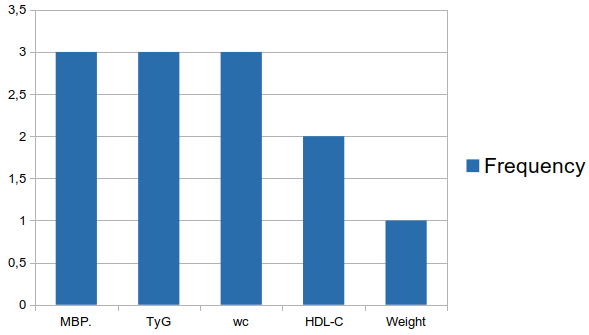
(b) SHAP values related to Idf after data fusion



(c) SHAP values related to De Ferranti after data fusion



(d) Most important features before data fusion



(e) Most important features after data fusion

Figure IV.6 – Feature selection results

Table IV.2 shows the most important characteristics using the merged database and according to the three definitions.

In part (e) of Fig.IV.6, the frequency of occurrence of the most crucial variables is depicted. By referencing Table IV.2 and part (e) of Fig.IV.6, it is evident that MBP, TyG, and WC emerge as the most significant variables.

Table IV.2 – Feature Importance Ranking for After Data Fusion

Dataset	1st	2nd	3rd	4th
DS_Merged_Cook	MBP	TyG	WC	Weight
DS_Merged_Idf	TyG	WC	HDL-C	MBP
DS_Merged_Ferranti	TyG	MBP	HDL-C	WC

Comparison and Assessment: The comparison between the variables selected by the first and second approaches highlights the impact of changing the data utilized by the random forest model on feature selection. Before data fusion, the most critical features were MBP and HDL-C, whereas after data fusion, the emphasis shifted to MBP, TyG, and WC.

Given the medical expertise and the goal of identifying low-risk individuals with clinical features favored over biological ones to streamline MetS screening and make it more cost-effective, the decision to select MBP and WC for MetS risk prediction aligns with the objective of minimizing the need for extensive blood tests in a large population during screening. This strategic selection of features can enhance the efficiency and affordability of the screening process for MetS.

Medical expert reports: Indeed, according to the doctors, the features identified as important for the prediction show a strong correlation with the presence of metabolic syndrome. This confirms the reliability of our data fusion and feature selection.

For feature selection, the doctors preferred to choose clinical features rather than biological ones to avoid requiring blood tests for patients predicted to be negative for MetS by the AI. Using only clinical variables significantly reduces the individual and social burden of MetS screening. The features to be selected, according to the doctors' recommendations, are WC and MBP.

Identifying Low-risk Individuals: The risk prediction performances using only MBP and WC for the three MetS definitions are summarized in Table IV.3.

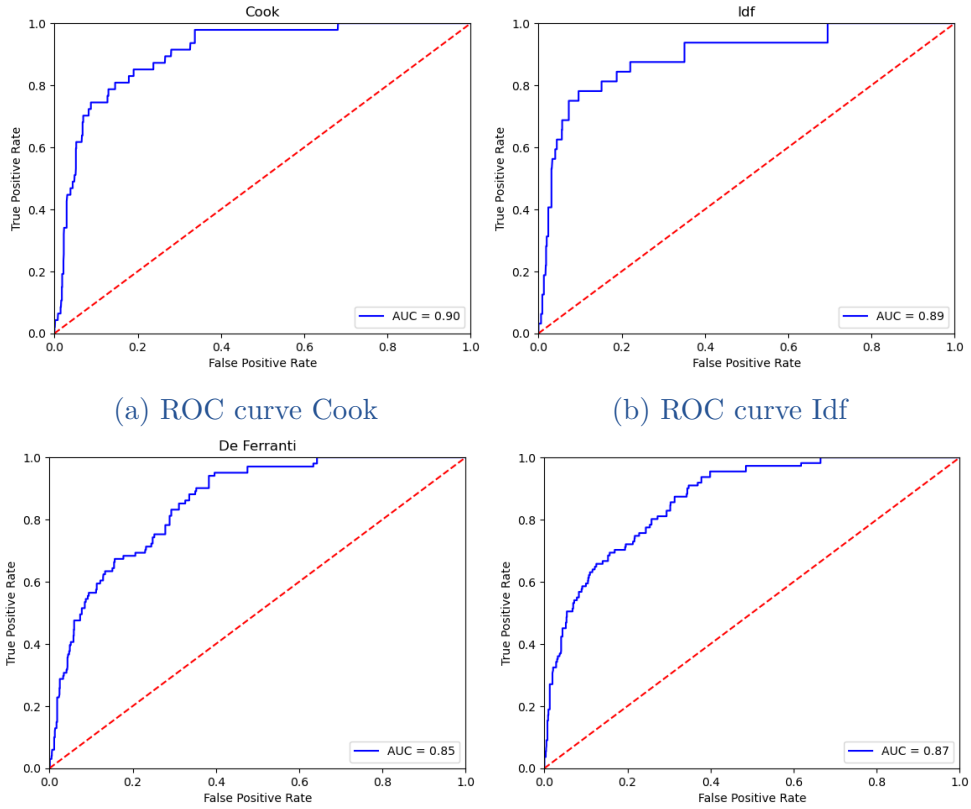
For the IDF definition, the model achieved a specificity of 0.9 and a sensitivity of 0.78. This indicates that based on this definition, the model better predicted positive values than negative ones. In the case of the De_Ferranti definition, the model exhibited a strong predictive ability for negative values, with a sensitivity of 0.94. Lastly, for the Cook definition, the model demonstrated good predictive ability for positive and negative subjects, indicating a balanced performance across the two categories. These results provide valuable insights into the model's predictive capabilities based on the selected features (MBP and WC) for each MetS definition, highlighting its strengths and areas for improvement in risk prediction.

Table IV.3 – Sensitivity, Specificity, and Cut-off Values for Several Outputs using Only MBP and WC

Output	Sensitivity	Specificity	Cut-off
Output_Cook	0.80	0.85	0.059
Output_Idf	0.78	0.90	0.031
Output_Ferranti	0.94	0.61	0.142

Evaluators like sensitivity and specificity are limited by their sensitivity to class distribution imbalance, particularly when one class significantly outweighs the other. In such cases, the ROC curve and AUC (Area Under the Curve) are more reliable metrics for assessing predictions.

The ROC curves for the three MetS definitions (Cook, IDF, and De_Ferranti) are illustrated in parts (a), (b), and (c) of Figure IV.7. The corresponding AUC values for Cook, IDF, and De_Ferranti were 0.90, 0.89, and 0.85, respectively. These AUC values indicate that the model exhibits good predictive ability for both positive and negative values across the several definitions, providing a comprehensive assessment of the model's performance in handling the class imbalance and making predictions for MetS risk.



(a) ROC curve Cook

(b) ROC curve Idf

(c) ROC curve De Ferranti

(d) ROC curve of unified risk

Figure IV.7 – ROC curves

Focusing on individuals predicted as negative for MetS across all definitions makes the Negative Predictive Value (NPV) a crucial metric for evaluating the model's performance. An NPV of 0.878 indicates that 87.8% of the population predicted as negative for MetS are truly negative, highlighting the model's ability to correctly identify individuals without MetS.

Furthermore, the ROC curve depicted in part (d) of Fig.IV.7 showcases the model's performance in predicting unified MetS risk. The corresponding AUC of 87% signifies a strong predictive capacity for the overall risk assessment, underscoring the model's effectiveness in accurately identifying individuals without MetS across several definitions.

By leveraging NPV, ROC curves, and AUC, the evaluation process provides a comprehensive and reliable assessment of the model's predictive capabilities, particularly in correctly identifying individuals without MetS, which is crucial for effective screening and risk assessment strategies.

IV.7 XAI Limits

The XAI has effectively explained the risk prediction, ensuring practitioners' confidence in this prediction. In other words, the variables on which ML was based are medically logical according to reliable references. This makes the prediction confident and ready to be integrated into healthcare.

Furthermore, XAI has provided promising results for evaluating the reliability of predicting carbohydrate anomalies. Additionally, using SHAP has been decisive in eliminating biological variables for MetS screening when predicting the low-risk subjects, significantly reducing individual and social financial burdens by eliminating blood tests for this population.

However, the reliability of explanations provided by XAI approaches remains a critical challenge [Yeh 19, Kumar 20, Schwartzenberg 20, Marx 23]. These challenges include the difficulty of valid and relevant explanations for new training and testing data. The explanation changes each time the test and training data selection is changed, especially for small datasets [Ketata 23]. Hence, there is ambiguity regarding generalization and the robustness of explainability. In addition, XAI approaches may admit a lack of consistency, where explanations can vary considerably for similar instances, raising concerns regarding their stability. Furthermore, there may be a difference between the importance of features evaluated by the model and the importance reflected in the explanations, calling into question the concordance of the explanations.

These limitations push our research towards a very important question regarding the reliability of XAI. How can we improve the reliability of XAI? How can we evaluate the reliability of XAI?

IV.8 Conclusion

In conclusion, this chapter has demonstrated the pivotal role of XAI in assessing the reliability of predictive models for endocrine disease risk prediction and managing financial medical expenses. The application of XAI techniques to predict carbohydrate abnormalities in β-TM patients and for risk prediction in MetS screening sessions has shown promising results in improving diagnostic accuracy and reducing financial burdens. However, the reliability of XAI explanations remains a challenge, particularly concerning their generalization, consistency, and concordance with model evaluations.

Chapter V
XAI Reliability Improving and Assessment

V.1 Introduction

At the end of the previous chapter, we presented XAI's limitations and challenges, especially regarding reliability. In this chapter, we propose approaches to improving XAI's reliability and then develop metrics to evaluate the reliability of this improvement.

However, an ML-based prediction provided to practitioners is generally considered opaque information. When this prediction is accompanied by explainability, it becomes clearer. Indeed, explainability can vary when the data selected for testing and training changes. This variability leads to a lack of confidence in the reliability of explainability validation, highlighting the need for improvement and thorough reliability studies. Therefore, we propose to provide practitioners with predictions that include more reliable explainability, validated by the combination with a data sampling technique and assessed using metrics that evaluate the reliability of the provided explainability.

Our study aimed to develop a structured approach to improve and assess the reliability of explanations provided by XAI approaches in healthcare. The main contributions of this study are as follows:

- Combining the explainable ML approach with a data sampling method to improve the reliability of the explainability.
- We define and develop metrics to assess the reliability of explainability after combining it with a data sampling technique.
- Develop a global metric for reliability assessment of XAI.

Our ultimate goal is to increase practitioners' confidence in ML by exploiting tabular datasets (Biological and Clinical) to predict the risk of abnormalities for each subject, assisting physicians in personalizing the screening or treatment of diseases. Consequently, several case studies are proposed to test and validate our approach on various datasets and ensure its applicability and effectiveness in real-world scenarios. We first test the proposed methodology on two public datasets (hypothyroidism and diabetes), which are generally well-treated and high-quality. We include thyroid disorders and diabetes because, according to the World Health Organization, they are the most common endocrine diseases worldwide. In these two case studies, our generic approach will be assisted by two XAI applicability approaches (SHAP and LIME) as these two XAI approaches are the most used in the literature in the context of endocrine disease prediction. Moreover, these approaches are relatively generic, can be applied to several ML models, and are not limited by a specific model. Then, we propose to apply and validate our approach in the two private datasets in relation with the medical issues ($\beta - TM$ and MetS).

Through this research, we aim not only to improve the understanding of the decisions made by ML models in healthcare, but also to facilitate wider adoption of these technologies by providing reliable, actionable explanations to healthcare professionals.

We begin this chapter by presenting research works close to our context in Section V.2. Then, we present our approach in Sections V.3 and V.4. Next, the tests and application of the proposed approach in two case studies for predicting the risk of hypothyroidism and diabetes are presented in Section V.5. Then, validate the approach for $\beta - TM$ and MetS in Section V.6 and discuss the limits and perspectives in Section V.7. Finally, the conclusion of the chapter in Section V.8.

V.2 Related Works

To assess XAI's performance, various subjective and objective metrics have been developed [Coroama 23]. Subjective metrics rely on user feedback regarding clarity, transparency, and satisfaction, while objective metrics utilize mathematical and statistical tools to evaluate the reliability of data-driven explainability. The range of metrics employed is contingent on the nature of the data and the specific field of application.

In partnership with an expert, XAI can achieve optimal accuracy. A recent study by the author of [Rosenfeld 21] explored advanced imaging techniques in radiology for disease detection, demonstrating that the XAI agent achieved a remarkable 99.5% accuracy rate. Providing medical professionals with clear explanations of classification errors is crucial in critical decision-making domains. The study introduced four criteria for evaluating the explanations provided by the agent: performance disparity between the agent's model and the explanation's logic, the number of rules, the number of features utilized in constructing the explanation, and stability. These criteria underscore the shortcomings of current research, which often oversimplify the logic of initial models without considering legal, ethical, or safety implications. They offer the advantage of being independent of the task or employing the XAI algorithm.

Within the realm of recommendation systems, there is a tendency to interchange specific terms, as highlighted by Tintarev and Masthoff [Doshi-Velez 17]. They stress the importance of user transparency, persuasiveness, scrutability, effectiveness, satisfaction, efficiency, and trust, alongside traditional accuracy measures like precision and recall. To evaluate the efficacy of explanations within these systems, they introduced metrics such as transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and user satisfaction. These metrics are designed to ensure that users comprehend the rationale behind recommendations, enabling them to rectify the system when necessary and fostering trust through clear and effective explanations.

Researchers have demonstrated that system design can influence perceived trustworthiness. Trust is evaluated by authors in [Fogg 01] through methods such as surveys or by assessing user engagement indicators like login frequency or sales. The concept of persuasiveness involves motivating users to make purchases or try products, as gauged by their responses to explanations. Effectiveness allows users to eliminate unsuitable choices through informative explanations, while efficiency, particularly relevant in chat systems, measures the speed of task completion, often quantified by the number of explanations required. User satisfaction, reflecting the system's utility and user-friendliness, is determined through user feedback metrics.

Doshi-Velez and Kim introduced in [Doshi-Velez 17] various concepts related to the quality of system explanations. They discussed how explanations can empower users to make corrections (actionability and correctability), establish causal links between inputs and outputs [Holzinger 19], and provide comprehensive system descriptions (completeness). They also addressed the ease of understanding explanations (comprehensibility [Askira-Gelman 98]), the selection of relevant features (faithfulness), and the alignment of explanations with expert knowledge (justifiability). Other aspects such as explanation consistency across similar inputs (robustness [Alvarez-Melis 18]), the ability to scrutinize unsuccessful training instances (scrutability), and the focus on essential explanatory features (simplicity) were also explored. Additional considerations included sensitivity to input variations, explanation stability, and truthfulness (soundness), with a discussion on the prioritization of completeness versus soundness [Kulesza 13]. The authors noted

that while concepts like transparency, interactivity, and security are commonly discussed in XAI, formal definitions and practical applications are still lacking, underscoring the interdisciplinary nature of this field.

Authors in [Hsiao 21] introduced seven cognitive metrics: explanation quality, user satisfaction, user engagement and curiosity, trust in the system, user comprehension, performance, and system usabilityFactors like explanation stability, robustness of the classification model, and computational requirements of explanation methods are crucial for time series classificationctive models of behavior, and utilizing factor analysis to explain behavior based on feature significance. These metrics are predominantly subjective and necessitate user feedback for accurate assessment. The categorization of evaluation methods targets specific user groups, including AI novices, domain experts, and AI professionals. Key interpretability metrics focus on the user's cognitive model, the utility and impact of explanations, trust in the system, and overall performance in human-AI collaborative tasks in [Mohseni 21].

Computational evaluation encompasses measures such as explanation accuracy, which closely links model consistency, explanation reliability, and model trustworthiness, independent of human-centered studies. For time series classification, factors like explanation stability, robustness of the classification model, and computational requirements of explanation methods are crucial. The authors in [Nguyen 20b] analyzed the effectiveness of saliency maps in providing explanations that pinpoint critical components for predictions within time series data. A truly informative explanation highlights parts crucial for accurate prediction. The stability and robustness of these explanations are tested through repeated trials, assessing their resilience to changes and the computational resources needed for generating such explanations.

The authors of [Zhou 21a] discussed explainability in AI as a combination of interpretability (how understandable explanations are to humans) and fidelity (how accurately explanations reflect the model's behavior). They argued against the feasibility of universal computation metrics for evaluating XAI methods due to factors like the subjective nature of explanations, varying contexts, dependencies on users and models, and specific types of explanations required. Objective evaluation metrics were categorized into model-based, attribution-based, and example-based explanations. Model-based explanations involve using or creating models to elucidate ML algorithms, with metrics like model size, interaction strength, or complexity. Attribution-based metrics focus on feature significance or ranking, employing metrics such as monotonicity or sensitivity. Example-based explanations utilize similar instances from the dataset, with metrics like non-representativeness and diversity [Nguyen 20a].

Arrieta in [Arrieta 20] proposed the development of specific evaluation metrics for future enhancements, focusing on the quality, utility, and satisfaction derived from explanations, improving the audience's mental model, and assessing the impact of explanations on model performance and user trust. Tools like goodness checklists, satisfaction scales, and computational measures were mentioned to assess explainer fidelity and reliability.

The study also explored using Bayesian Networks in various applications, highlighting BayLime as an enhancement of the LIME technique to address instability and enhance consistency and robustness through Bayesian reasoning [Zhao 21]. The importance of explaining Bayesian networks, particularly in legal contexts, was acknowledged [Vlek 16].

A review of explanation methods for Bayesian networks outlined essential properties such as content, communication, and adaptation to user needs in [Lacave 02]. Emphasis was placed on explaining the knowledge base, reasoning process, and evidence-

supporting conclusions. Communication involved presenting explanations, including format and probability expression. Adaptation refers to tailoring explanations to the user's knowledge level and information requirements.

In 2024, authors in [Arreche 24] proposed an end-to-end framework to evaluate XAI methods for network intrusion detection. This framework evaluates global and local scopes and analyzes metrics like descriptive accuracy, sparsity, stability, efficiency, robustness, and completeness.

Table V.1 presents a compilation of various metrics introduced in academic literature, detailing the goals of each study and the specific applications involving data and models.

Table V.1 – XAI Evaluation Metrics Summary

Reference	Metric	Metric Definition	Study Objective	Metric Type	Data	Model
[Rosenfeld 21]	Performance (D), Number of rules (R), Number of features (F), Stability (S)	Difference in performance between the agent's model and the explanation's logic.	Demonstrate the effectiveness of XAI in enhancing diagnostic accuracy in radiology.	Objective	Images	Neural Networks
[Tintarev 07]	transparency, persuasiveness, scrutability, effectiveness, satisfaction, efficiency, trust	Metrics evaluate how well a system communicates its decision-making process, convinces users, and allows for inspection.	Assess the role of clarity and user control in recommendation systems acceptance.	Subjective	-	-
[Fogg 01]	Login Frequency, Sales	Indicators of how often users log in and the volume of sales, reflecting engagement and trust.	Investigate how system design influences user trust and engagement metrics.	Obj/Subj	-	-
[Doshi-Velez 17]	Actionability, Correctability, Causality, faithfulness, justifiability, robustness, soundness.	Focus on the user's ability to act upon, correct explanations, and understand cause-effect relationships.	Evaluate how well XAI systems enable user interaction and understanding through explanations.	Obj/Subj	-	-
[Hsiao 21]	Explanation Quality, User Satisfaction, productivity, usability/interaction.	Cognitive metrics that gauge the clarity, helpfulness, and satisfaction levels of explanations from the user's perspective.	Identify and measure cognitive metrics reflecting user interaction with XAI systems.	Subjective	-	-
[Nguyen 20a]	Explanation Stability, Robustness	Evaluate how consistent and resilient the explanations are to changes and noise.	Explore the consistency and resilience of explanations in time series classification.	Objective	Time Series	-
[Zhou 21a]	Model Size, Complexity, Monotonicity, Complexity.	Assessments of the explanatory model's size, intricacy, and the predictability of feature importance.	Assess the interpretability and accuracy of explanations across several models and data types.	Objective	Images, Tabular	-
[Arrieta 20]	Goodness Checklist, Satisfaction Scale	Tools to evaluate the effectiveness, adequacy, and satisfaction with the explanations provided.	Develop and refine metrics for evaluating explanation effectiveness and user satisfaction.	Subjective	-	-
[Lacave 02]	Explanation Focus, Explanation Level.	Criteria for determining the scope, depth, and approachability of explanations in Bayesian networks.	Review and define the necessary properties of explanations in Bayesian networks for user understanding.	Obj/Subj	-	Bayesian Nets
[Arreche 24]	Accuracy, sparsity, stability, efficiency, robustness, and completeness	End-to-end framework.	evaluate both global and local scopes of XAI for network intrusion detection.	Objective	-	-

Our study examines how alterations in training and testing data impact the feature ranking generated by XAI. Consequently, our objective is to integrate the XAI methodology with a data sampling technique to enhance the validation process of XAI results. Subsequently, we devised a comprehensive metric centered on generalizability, concordance, and stability to assess the dependability of the XAI and its amalgamation with k-fold validation.

V.3 XAI Reliability Improvement

The challenge with explainable ML is that whenever the selection of test and training data for prediction is changed, the explainability of the prediction outcomes leads to a novel order of feature contributions [Ketata 23]. To overcome this problem and improve the reliability of the explainability, we propose to combine the extraction of importance coefficients by XAI approaches with the k-fold technique. Subsequently, metrics were developed to study the generalization, concordance, and stability of the combination of XAI and k-fold as shown in Figure V.1.

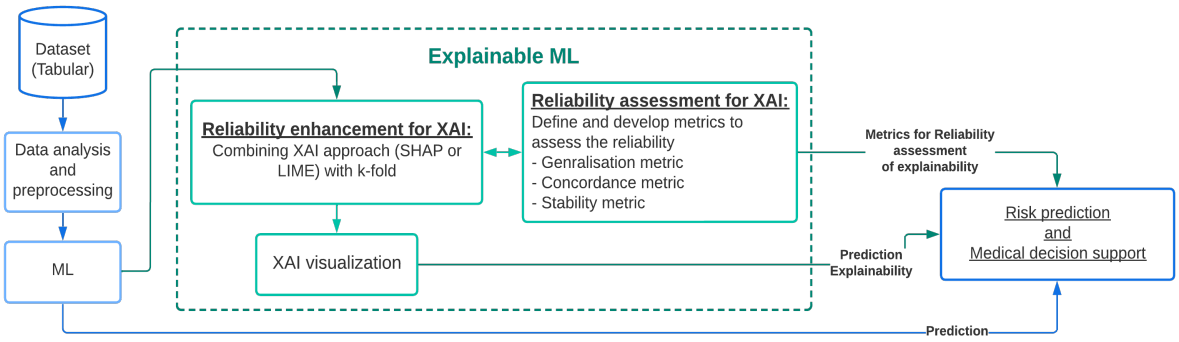


Figure V.1 – Process of the proposed methodology for XAI improvement and evaluation

V.3.1 K-fold Technique Definition

The k-fold cross-validation method is a commonly used technique to assess the performance of ML models. It involves dividing the original dataset into k equal-sized subsets (or "folds") as shown in Figure V.2.

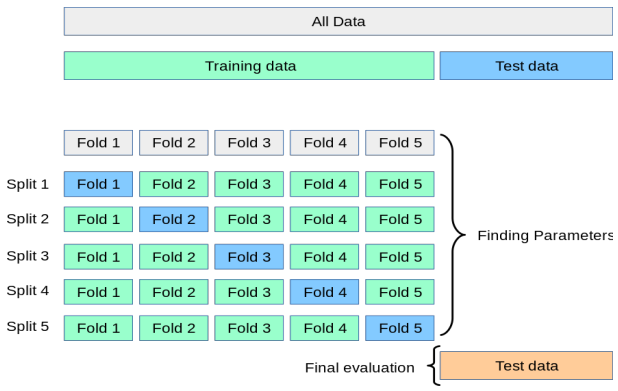


Figure V.2 – k-fold technique [DA SILVA 23]

The original dataset is divided into k equal-sized subsets (folds). Each fold contains an approximately equal distribution of the data. The model is trained k times, with each iteration using a different fold as the validation set and the remaining $k-1$ folds as the training set. For example, in the first iteration, the first fold is used as the validation set and the other folds are used for training. In the second iteration, the second fold is used as the validation set, and so on.

The main advantages of k-fold cross-validation are reduced variance and better data utilization. Common choices for the value of k include 5-fold and 10-fold cross-validation, although other values can also be used depending on the dataset size and available computational resources. However, smaller values of k may result in higher variance in performance estimates, while larger values may increase computational burden.

V.3.2 Combining XAI Approaches with k-fold Technique

For this purpose, data were divided into k samples or files using the k-fold technique. For the first iteration, one of the k samples was chosen as the validation set, with the remaining k-1 samples serving as the training set for model learning. Then, for each iteration, the data file selected as validation data is used for training, and one of the files selected previously for training is used for validation. We concatenate the feature importance coefficients for each iteration in the XAI list. Eventually, we obtain a list of feature importance coefficients divided into multiple sub-lists, each resulting from a prediction made using distinct training and test datasets, as shown in Figure V.3. As a result, visualizing the explainability of each variable's significance to prediction is more generalizable and reliable for validating the feature importance ranking.

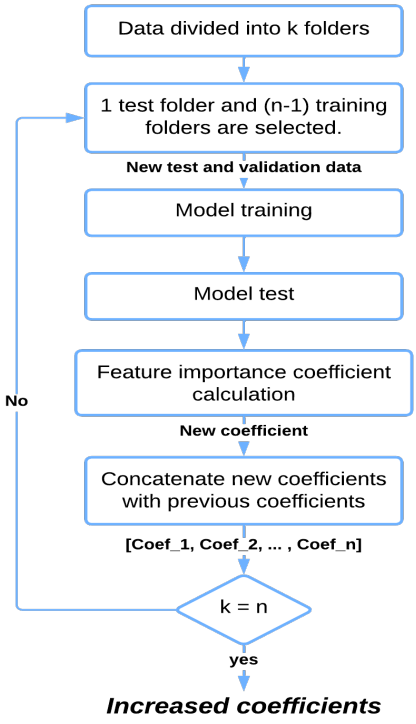


Figure V.3 – XAI with k-fold

V.4 Metrics for Assessing XAI Reliability

We have now come to present the metrics for the reliability assessment of XAI.

Generalizability Metric

The objective of this part is to identify the value of k in k-fold, where the ranking of the best feature contributors remains unchanged. The idea is to calculate the augmented XAI coefficients for several values of k. Subsequently, we computed the average absolute of the coefficients for each feature. Thus far, we have a separate list for each value of k, that contains the average importance coefficient for each feature. Each list provides a distinct rating of features for each k. Subsequently, Spearman correlation similarity analysis was conducted to examine the generalizability or variability of the feature ranks across various k values. Ultimately, a matrix was generated to display the correlation among all k values and construct a curve to assess the variation between each pair of consecutive k values. We propose studying the generalizability metric to first choose the optimum k value in the k-fold approach and to study whether the feature importance ranking provided by XAI changes when the data selected for testing or training changes. In other words, we aim to check whether the final feature ranking is generalizable and stable within the dataset. This metric is based on calculating the similarity (correlation). Therefore, it was between 0 and 1. A value of 1 indicates maximum generalizability. However, a value of zero indicates zero similarity.

The algorithm 2 describes the computing process of similarity between features importance ranking.

Concordance Metric

This metric evaluates the correlation between the feature importance provided by XAI, named XAI_coefficients and those provided by the predictive model explanation (impurity-based importance for the predictive model) [Stassin 23]. A high correlation indicates that the explanations provided by the XAI agree with the intrinsic importance of the features according to the model, which is a reliable and significant indicator of explainability. We propose to study concordance to assess the degree to which the final ranking of the most important features provided by XAI is directly correlated with the importance of these features in the kernel of the model in the ML process.

The concordance equation is presented below (V.4).

$$\text{Concordance} = \Phi(\mathbf{I}_{\text{Model}}, \mathbf{I}_{\text{XAI}}) \tag{V.4}$$

where Φ represents the Pearson correlation function, $\mathbf{I}_{\text{model}}$ is the vector of feature importance as assessed by the model, and $\mathbf{I}_{\text{XAI}}$ is the vector of mean feature importance derived from XAI.

The concordance was also between 0 and 1. A higher concordance indicates good reliability.

Stability Metric

This measure assesses the extent to which the explanations provided by XAI are consistent for similar instances [Munoz 23]. A small distance close to zero means that similar instances receive similar explanations, indicating good stability in how the model

Algorithm 2 Generalizability analysis of feature sorting based on k values

Step 1. Increase the XAI coefficient extraction for each k-value

Step 2. Compute the average absolute values of the XAI coefficients for each feature represented by the following formula.

$$M = \frac{1}{m} \sum_{i=0}^{m-1} |XAI\_coefficients_i| \tag{V.1}$$

where m is the number of subjects

Step 3. Generate a list of values, including the mean absolute values of the feature importance for each k-value :

$$L_j = [M_{0j}, M_{1j}, ..., M_{(n-1)j}] \tag{V.2}$$

for j in [1,...,n], where n is the max value of k

Step 4. The similarity between the ranked lists for all k values was calculated using the Spearman correlation presented by the following formula:

$$Generalisation = \rho(\mathbf{L}_\mathrm{j}, \mathbf{L}_\mathrm{j+1}) \tag{V.3}$$

where ρ refers to Spearman correlation.

Step 5. Display the Generalisation_metric between feature ranks for each k value.

assigns importance to features. This metric is dedicated to assessing the stability of the explainability provided by XAI for two similar instances. In other words, it is a useful metric for evaluating the certainty of its explainability. The stability metric is an important assessment in conjunction with the generalizability metric because unstable and uncertain explainability can affect the similarity between feature rankings as a function of k values, thus causing disorder in the generalizability metric.

The equation for stability is presented below (V.5).

$$\text{Stability} = \frac{1}{N} \sum_{k=1}^{N} d(\mathbf{S}_{k1}, \mathbf{S}_{k2}) \tag{V.5}$$

where N is the number of pairs of similar instances examined, d represents euclidean distance function , and $\mathbf{S}_{k1}$ and $\mathbf{S}_{k2}$ are the XAI value vectors for the k pair of similar instances.

The euclidean distance is calculated by the following formula (V.6) :

$$d(\mathbf{S1}, \mathbf{S2}) = \sqrt{\sum_{i=1}^{n} (S1_i - S2_i)^2} \tag{V.6}$$

where **S1** and **S2** are two vectors of the XAI values for compared instances, and $S1_i$, $S2_i$ are the corresponding components in these vectors.

Global Metric for Reliability Assessment

The final metric we have devised integrates generalization, stability, and concordance considerations. A heightened metric value (equal to 1) for generalizability or concordance signifies strong reliability. Conversely, a lower stability metric approaching zero suggests robust reliability. Therefore, the ultimate reliability metric is determined as the product of generalizability and concordance, subtracted from 1 minus stability, as illustrated in Equation (V.7). Consequently, the reliability_metric also falls within the range of 1 to 0. A reliability value nearing 1 denotes optimal XAI reliability.

$$\text{Reliability} = \mathit{Concordance} \cdot \mathit{Generalizability} \cdot (1 - \mathit{Stability}) \tag{V.7}$$

V.5 Test the Proposed Methodology in Several Public Datasets for Endocrine Risk Prediction

V.5.1 Hypothyroidism Risk Prediction for a Low-cost Diagnosis

Risk Prediction of Hypothyroidism Using Random Forest

We conducted various tests to optimize the hyperparameters of the random forest model. The optimal hyperparameters were set as 300 trees in the forest, "GINI" as the split quality measure, and a maximum depth of 10. To ensure a robust evaluation of risk prediction, we assessed the metrics using a 10-fold cross-validation (k=10). The random forest model exhibited performance metrics of 99.1% accuracy, 99.5% precision, 98.8% recall, and 99.1% F1-score.

SHAP Reliability Improvement and Assessment

To improve the reliability of explainable ML, in this section, we analyze the variability in the sorting feature importance for every k value. The correlation matrix presented in Figure V.4 illustrates the degree of similarity in the feature ranks across the various k values. When the value of k is less than 5, there is a limited correlation between k-values. The rankings of the features showed major changes. This implies a lack of generalization by only applying SHAP without K-fold or even with small k-values. Starting from k = 27, we observed that the ranks have nearly identical similarity, which is equal to one.

To effectively visualize the similarity, Figure V.5 shows the correlation between consecutive k values to assess the consistency of the rankings. Figure V.5 demonstrates that starting from k=27, there is a strong correlation between rankings, with an approximate value of 1. Therefore, the ranking remained mostly unchanged.

Based on this analysis, we can confidently state that utilizing a combination of Shapley and k-fold with a value of k=27 is the most reliable method for studying the most important features in the dataset, with a generalization metric equal to 1. The ultimate ranking of feature importance is displayed in Figure V.6 in descending order from top to bottom.

We begin by explaining how to analyze the graph. Features are shown on the y-axis, and Shapley values on the x-axis. The color blue and red indicates whether the characteristic values for each subject are at minimum (blue) or maximum (red).

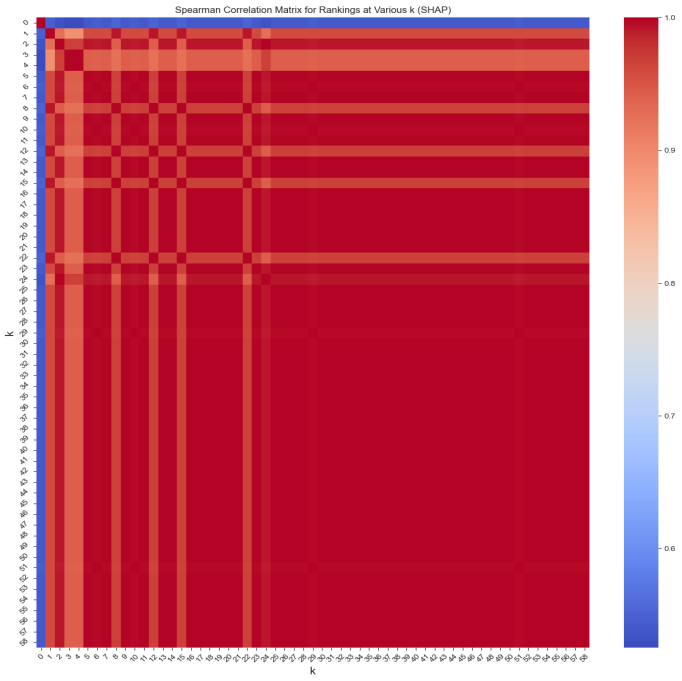


Figure V.4 – Correlation matrix between feature rankings: SHAP for hypothyroid

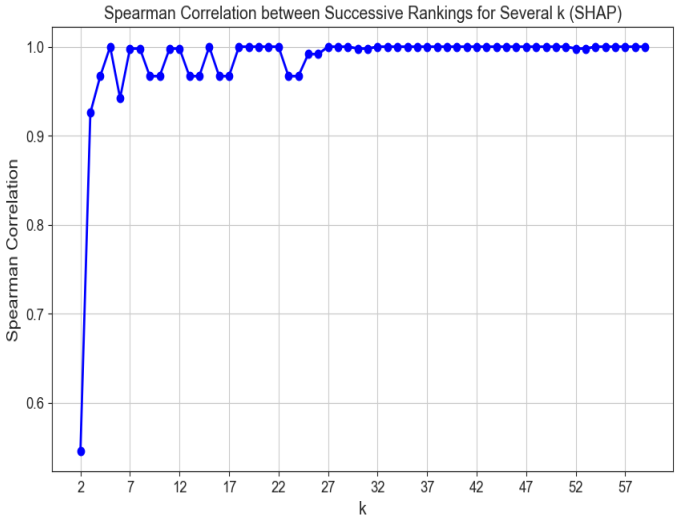


Figure V.5 – Correlation between characteristic rankings of successive k values : SHAP for hypothyroid

Each point on the graph represents the Shapley value for each specific characteristic associated with a patient. Therefore, the number of points for each entry was equal to the number of patients. The characteristics were ranked from the most important (top) to the least important (bottom) in predicting hypothyroidism. If the biggest values (in red) of a variable admit positive Shapley values, it means that the bigger the variable, the greater the risk of having the anomaly (Output = 1). Alternatively, if the smallest values of a variable admit positive shapley values, the smaller that variable is, the greater the risk of having the anomaly.

As shown in Figure V.6, TSH was the most important feature for the risk prediction of hypothyroidism. We can see the difference between the TSH SHAP values and the SHAP values of other features. The final ranking of the top contributors is highly reasonable from a medical perspective. This means that our forecast and risk assessment outcomes are dependable and are not influenced by random chance or skewed data.

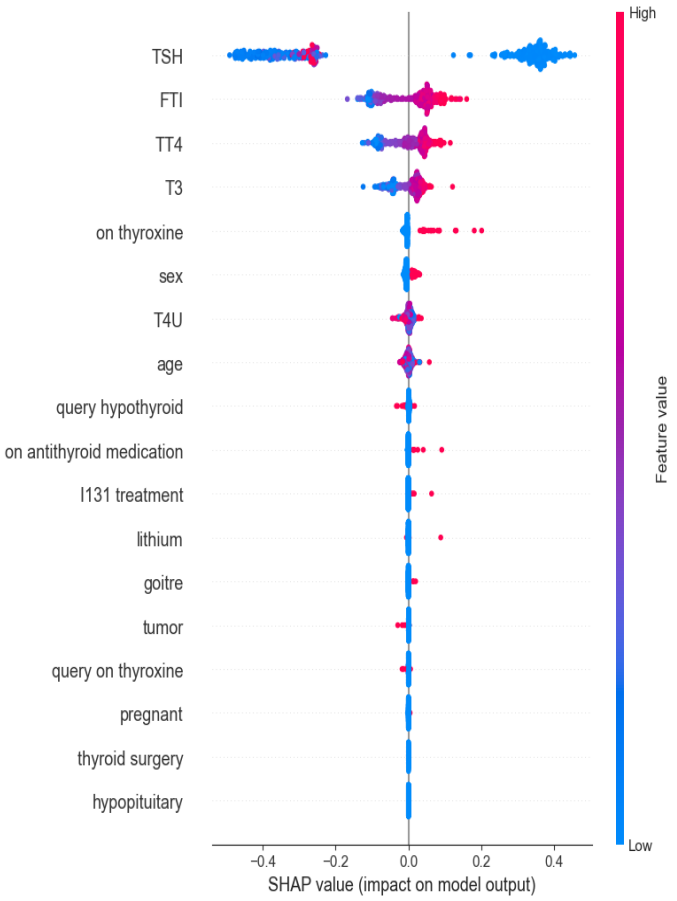


Figure V.6 – SHAP with k-fold after study k-value (k=27) : SHAP for hypothyroid

We computed the developed metrics to assess the reliability of SHAP after combining it with the k-fold. The preceding curve shows that generalizability is perfect and equal to one. The stability of the similarity between the several feature classifications was perfectly correlated. In addition, SHAP with k-fold has a very good concordance of 0.994 and a good stability of 0.087. Therefore, the overall reliability metric is 0.91, which is good reliability.

LIME Reliability Improvement and Assessment

In this section, we display the same reliability assessment methodology with the same graphs presented in the previous section with SHAP and k-fold to study the combination of LIME and k-fold.

Figure V.7 shows that by combining LIME and k-fold for k values below 17, the correlation is very weak. This demonstrates the lack of generalizability of LIME without the application of k-fold, even with low k values. However, for k values above 17, the correlation is much higher and very close to 1.

By analyzing the variation of similiraties in Figure V.8, we can see that, unlike SHAP, the generalization metric is well perturbed and not stable. This means that the feature ranking was not similar for several k-values.

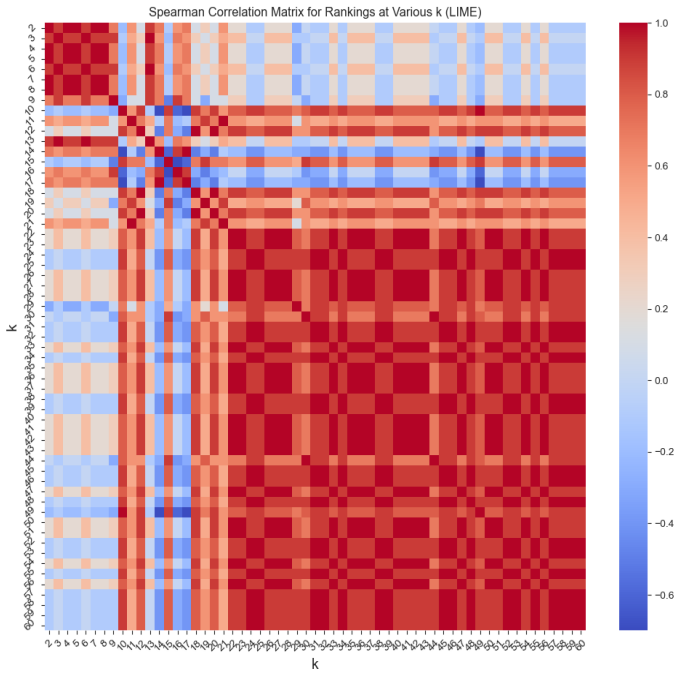


Figure V.7 – Correlation matrix between feature rankings: LIME for hypothyroid

However, we can see a trend towards stability with a correlation between 0.75 and 1 from k = 23. Hence, the generalization metric has an average value of 0.875. The concordance was relatively good at 0.81, and the stability was very good at 0.017. This means the LIME and k-fold approach were stable and certain of its explainability. On the other hand, it is non-generalizable and not sufficiently correlated with the internal interpretability of random forest. This provided an overall reliability metric of 0.69, affirming a lack of LIME and k-fold reliability for the hypothyroid case study.

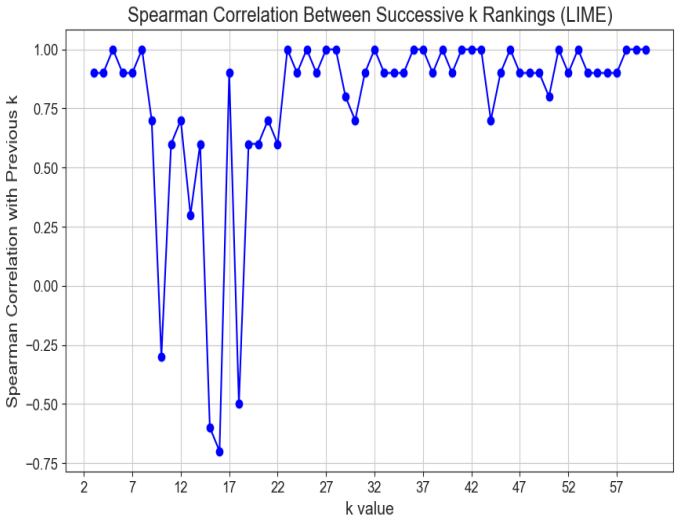


Figure V.8 – Correlation between characteristic rankings of successive k values : LIME for hypothyroid

Figure V.9 shows the ranking of the feature importance after combining LIME and k-fold. The TSH variable was ranked as the most important variable, as shown by LIME and SHAP. However, there is a change in the importance of the other variables, which leads us to compare the reliability of the two explainability approaches.

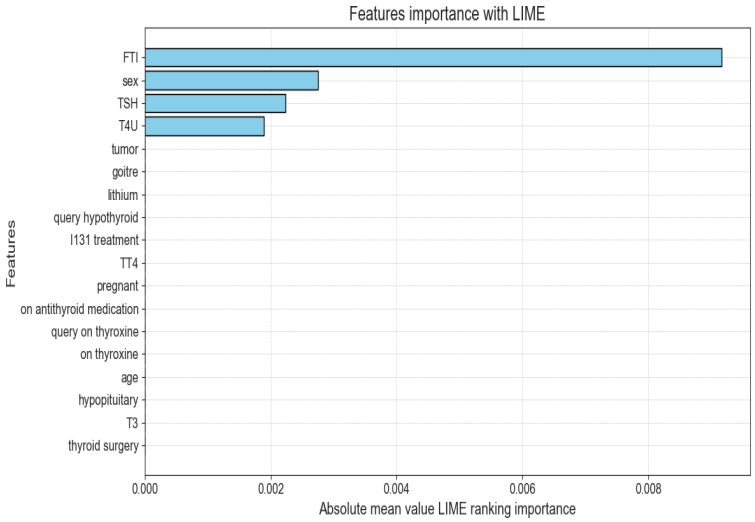


Figure V.9 – LIME with k-fold after study k-value (k=27) : LIME for hypothyroid

V.5.2 Diabetes Risk Prediction

Risk Prediction of Diabetes Using Random Forest

To provide a reliable risk prediction evaluation, we applied a k-fold cross-validation (k=10) to show the metric raters. Random forest showed a moderate performance in predicting diabetes risk. Accuracy 76.8%, precision 69.0%, recall 59.5% and F1-score 62.8%.

SHAP Reliability Improvement and Assessment

The correlation matrix presented in Figure V.10 shows that for k less than 8, the correlation is lower and perturbed than that for k greater than 9, which is a strong correlation. Hence, the ranking of features after the combination of SHAP and k-fold was stable quickly and perfectly from k = 10.

Moreover, the curve in Figure V.11 shows that from k=10, the correlation is perfectly equal to 1. Even for k values below 9, the correlation was strong between 0.975 and 1.

After combining SHAP and k-fold, Figure V.12 shows that glucose was the most important variable for predicting diabetes risk. In addition, all other variables were important for prediction, particularly BMI and age.

Finally, the generalizability of SHAP and k-fold was perfectly equal to 1, proving an identical feature importance ranking. The concordance of this combination and the random forest explanation was also significant at 0.98, with a good stability of 0.01. Hence, a very good global reliability score of 0.97 for the combination of SHAP and k-fold for the diabetes case study.

LIME Reliability Improvement and Assessment

The correlation matrix presented in Figure V.13 shows a weak and skewed correlation for the combination of LIME and k-fold, demonstrating the lack of generalizability of LIME in this case study.

The same is true for the correlation curve shown in Figure V.14, which confirms that the LIME and k-fold combination is not stabilized. Hence, the LIME and k-fold combination lack generalizability. Therefore, the generalizability metric is the mean value of all correlations. This mean value was equal to 0.7.

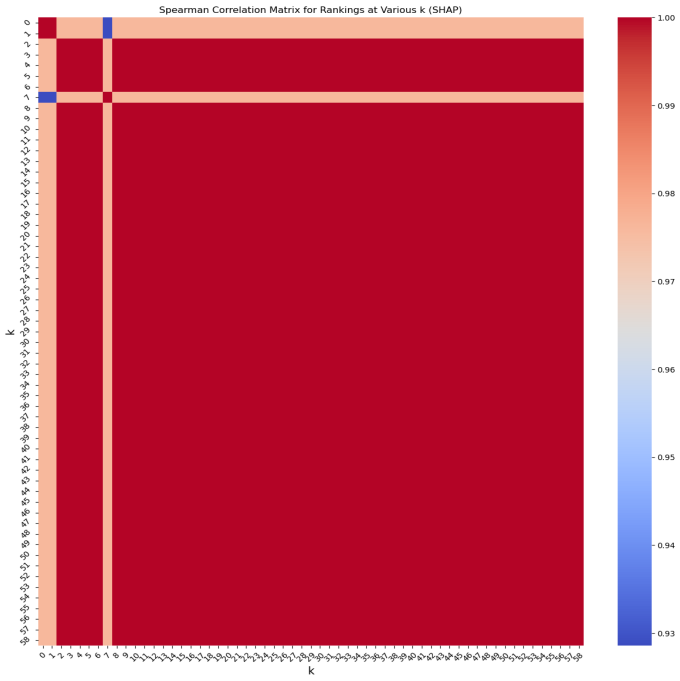


Figure V.10 – Correlation matrix between feature rankings: SHAP for diabetes

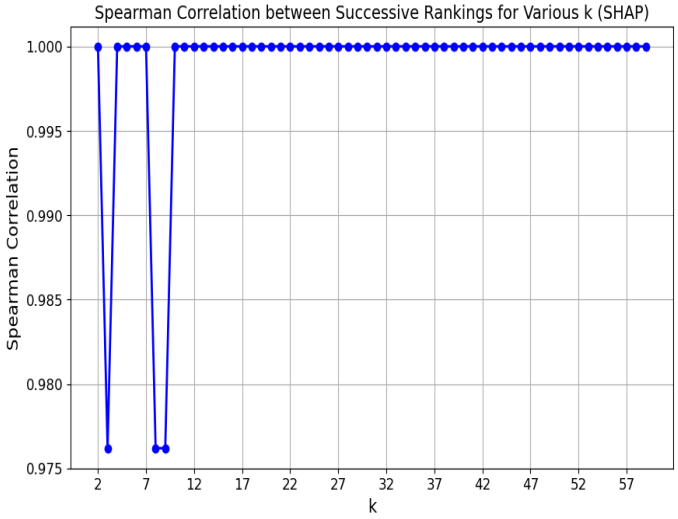


Figure V.11 – Correlation between characteristic rankings of successive k values : SHAP for diabetes

Figure V.15 shows the feature rankings for LIME, which differ from those for SHAP. This confirms the need to study the reliability of both approaches.

The concordance between LIME plus k-fold and Random Forest explainability was very low, equal to 0.27 with a stability of 0.01. This means a high degree of surety in the explainability provided by LIME and k-fold, but not sufficiently correlated with the internal interpretability of random forest. In addition, LIME and k-fold showed poor and unstable generalizability, which shows the lack of similarity between the different rankings of the most important features, leading to poor reliability. Hence, the global reliability score is 0.187.

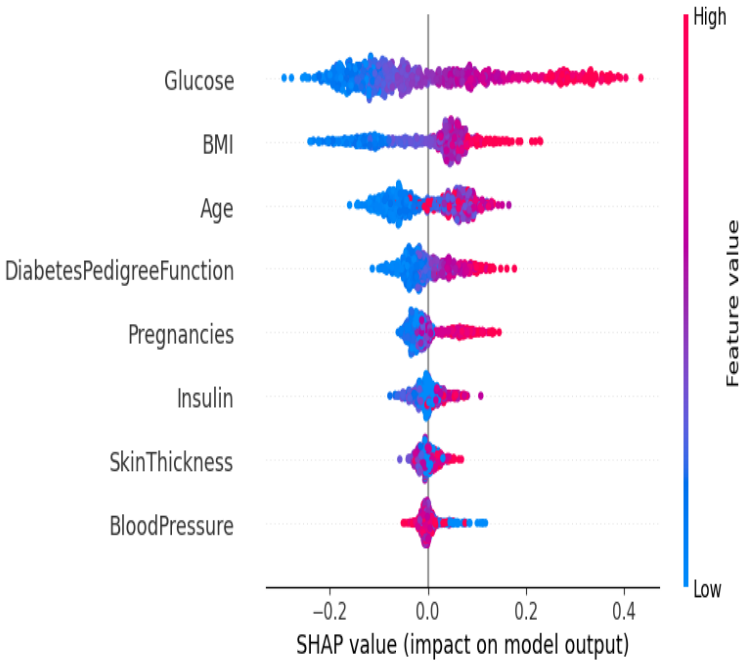


Figure V.12 – SHAP with k-fold after study k-value (k=10) : SHAP for diabetes

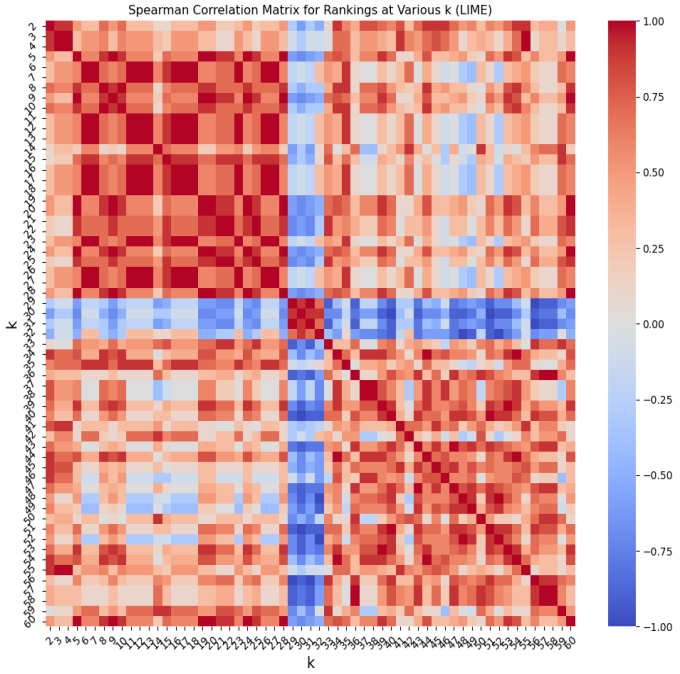


Figure V.13 – Correlation matrix between feature rankings: LIME for diabetes

V.5.3 Discussion

In this study, we propose to combine the k-fold technique with the SHAP and LIME approaches. Subsequently, we developed metrics to evaluate the concordance, generalization, stability, and overall reliability of this combination. We then tested and applied these to two different datasets to predict hypothyroidism and diabetes. Table V.2 summarizes the results obtained for both case studies. First, we noticed a difference in the ranking of the most important features between SHAP, kfold and LIME, kfold. This adds to the obligation to evaluate and compare the reliability of these two combinations.

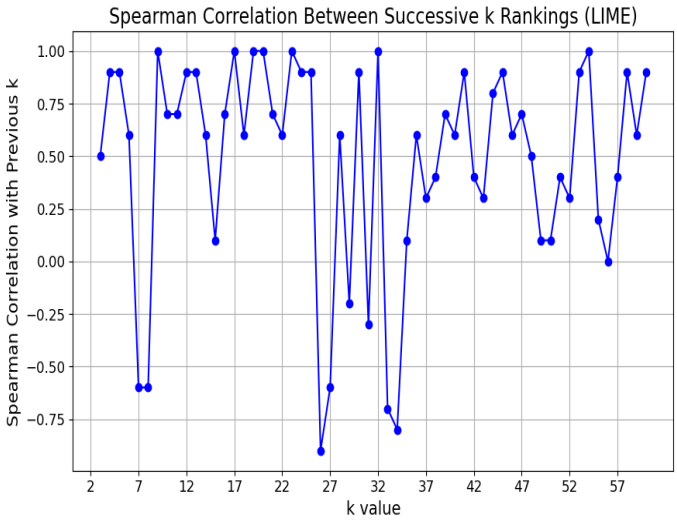


Figure V.14 – Correlation between characteristic rankings of successive k values : LIME for diabetes

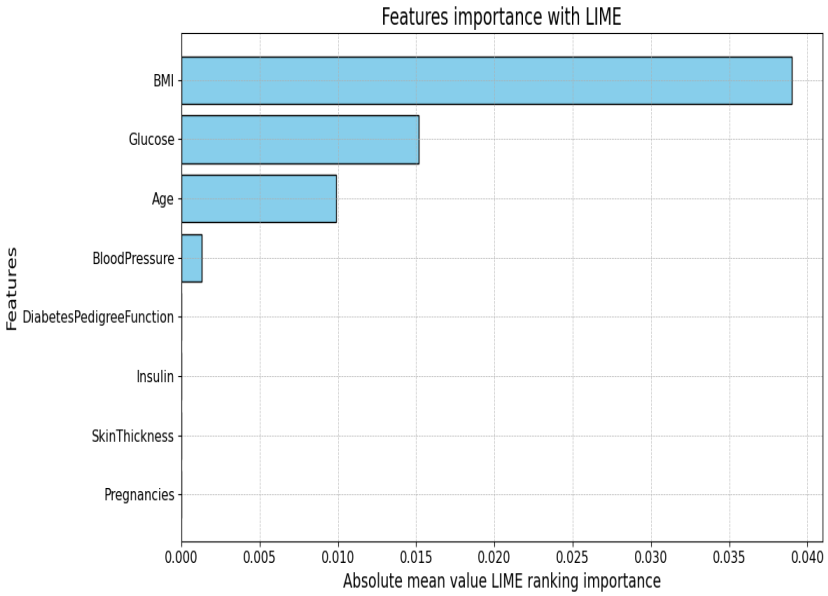


Figure V.15 – LIME with k-fold after study k-value (k=32) : LIME for diabetes

This can be observed from Table V.2, that the SHAP approach with k-fold was more generalizable and less influenced by changes in the test and training data for both case studies. This combination achieved feature ranking stability more quickly in the second case study, which may be explained by either the lower number of features or the higher amount of data in the second case study compared with the first. However, for the combination of LIME and k-fold, feature ranking did not achieve stability for either case study, indicating a lack of generalizability and a significant influence on changing the selected test and training data. Additionally, the SHAP and k-fold combinations showed strong concordance with the internal explainability of Random Forest at 0.994 and 0.98 respectively in both case studies, ensuring the convergence of SHAP explainability even when combined with k-fold. In contrast, the LIME plus k-fold approach showed poor concordance, especially for the second case study at 0.27, indicating that combining k-fold with LIME was not effective or reliable. Both approaches provide stable explanations for the two case studies. The good stability of the LIME with k-fold indicates the relative

suitability of this approach. The good stability of LIME with k-fold, poor concordance, and poor generalizability raise doubts about the effectiveness of the LIME with k-fold. In contrast, the SHAP and k-fold combination demonstrated excellent stability, concordance, and generalizability. Finally, the overall reliability score effectively demonstrates that the explanations provided by the SHAP and k-fold combination were reliable at 0.91 and 0.97 for both hypothyroidism and diabetes predictions. Moreover, without this combination, the explainability of SHAP remains influenced by the change in training and test data presented when the k variable is weak in the correlation curves. This supports the usefulness of our idea of combining SHAP and k-fold. In contrast, the LIME and k-fold combination showed mediocre scores, particularly for predicting the risk of diabetes, with a reliability score of 0.18.

Table V.2 – Summary of XAI reliability assessment

Metrics	**Thyroid prediction**		**Diabetes prediction**	
	SHAP	LIME	SHAP	LIME
Generalization	1	0.875	1	0.7
Concordance	0.994	0.81	0.98	0.27
Stability	0.087	0.017	0.01	0.01
Overall reliability	0.91	0.69	0.97	0.18

Finally, in this study, we proposed not only to share an identification of at-risk individuals that is ambiguous for practitioners but also to provide a reliable explanation of this identification combined with k-fold and metrics to assess this reliability and its degree of certainty. A reliable XAI such as SHAP combined with k-fold and metrics that address its reliability, such as generalizability, concordance, and overall reliability, can increase physicians' confidence in the prediction and its explanatory power. This may lead to greater integration of ML models for risk prediction in hospitals, particularly in the context of endocrine diseases.

V.6 Validation of the Proposed Methodology in Several Private Datasets for Endocrine Risk Prediction

After testing the proposed approach on two study cases using two explicability frameworks. In this part, we aim to apply and validate the combination of SHAP with k-fold and the various proposed metrics on real-case studies using private data sets, in response to doctors' problems.

V.6.1 β-TM

Figures V.16 and V.17 show that from k=3 the stability of the similarity between feature rankings is almost perfect. A small disturbance is apparent, but the correlation between 098 and 1 remains strong. There is a small disturbance that prevents perfect stability. This may be due to the small amount of data or to the missing data management.

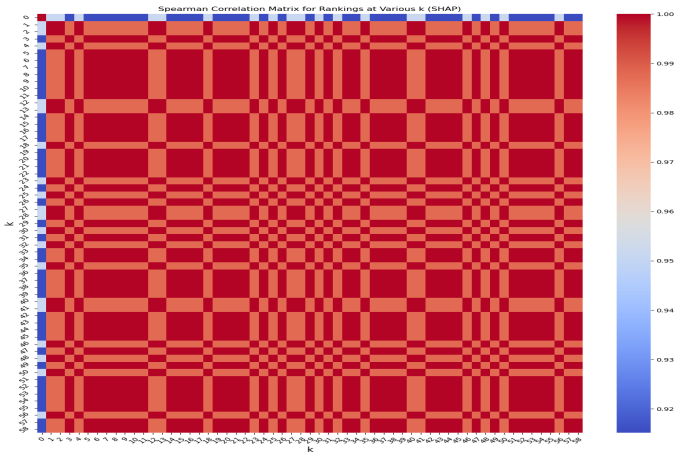


Figure V.16 – SHAP with k-fold for Beta Thalassemia correlation matrix

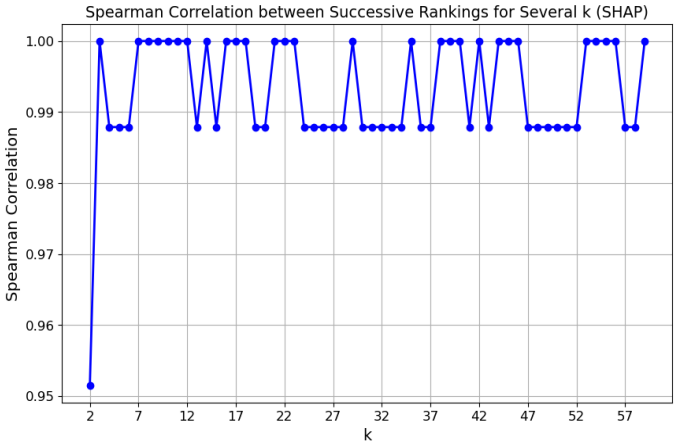


Figure V.17 – SHAP with k-fold for Beta Thalassemia Correlation curve

The generability metric is calculated as the average of the correlations in the event of non-stability. Hence, generalizability is 0.99. There is a good concordance of 0.97 but poor stability of 0.3. This poor stability is probably due to the lack of data for the beta-thal dataset.

Figure V.18 shows the final ranking of the most important features with k=3 with a reliability of 0.67.

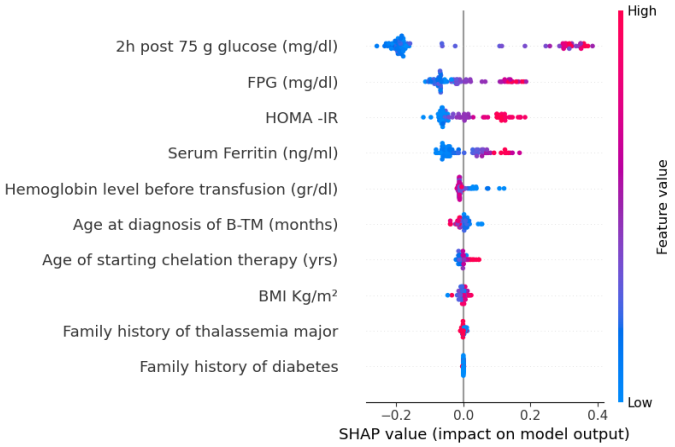


Figure V.18 – SHAP with k-fold features importances ranking for Beta Thalassemia

V.6.2 MetS

Figures V.19 and V.20 show that from k=18 the stability of the similarity between feature rankings is perfect and equal to 1. This means perfect similarity between the ranking of characteristics for different training and test data.

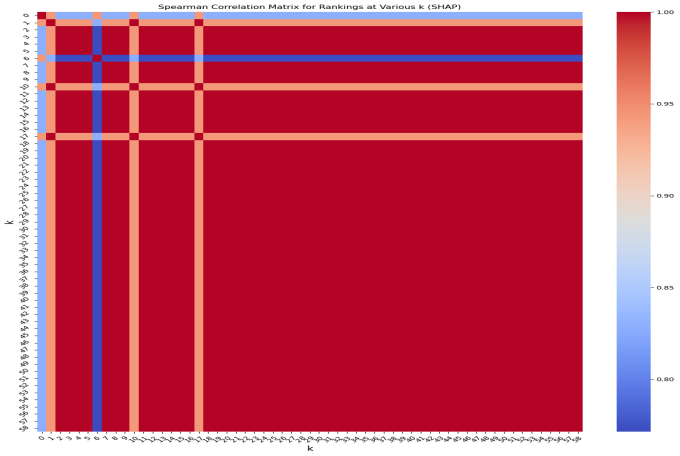


Figure V.19 – SHAP with k-fold for MetS correlation matrix

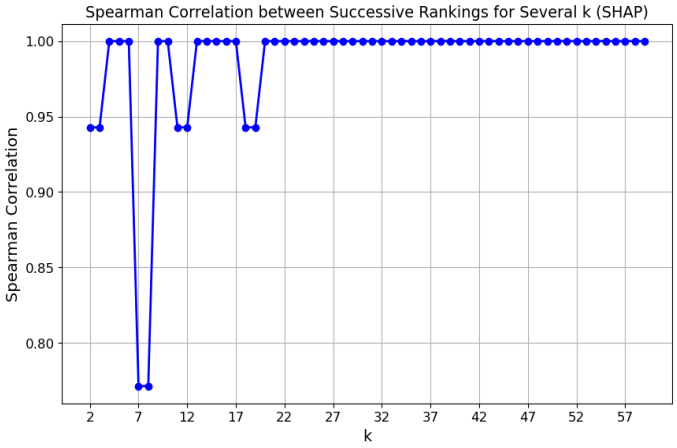


Figure V.20 – SHAP with k-fold for MetS correlation curve

The generalizability metric is therefore perfect and equal to 1. A good concordance of 0.8 and a stability of 0.001. The final reliability of SHAP with k-fold is therefore 0.72. The generalizability metric achieved permanent stability in this case study, unlike the first case study for predicting carbohydrate abnormalities. This can be explained by the Mets dataset admitting more subjects and fewer features. Hence, it is more obvious that there should be a permanent classification. Unlike the carbohydrate anomalies dataset, which has more features and less data.

V.7 Limits and perspectives

In this work, we proposed an approach to improve and evaluate the reliability of XAI validation in response to a specific limitation related to explainability change provided by XAI when training or test data is changed. This specific and limited scope is embedded within several other XAI limitations in the literature.

Our proposed approach to enhance and evaluate the reliability of XAI is limited to basic ML models that exploit tabular data. It may not be applicable to deep learning models and to other types of data, such as images. Our study tested this methodology in only two case studies using the Random Forest model. Our findings indicated that the combination of SHAP and k-fold validation is reliable, based on the evaluation metrics developed in this study.

However, the combination of LIME and k-fold validation showed poor performance. This lack of reliability could be attributed to the fact that LIME is typically used as a local explainer. It may also be caused by the uncertainty of the AI model. Since it is a frequent problem, many models in the literature are likely over-optimistic due to leakage and over-fitting. Our focus in this article was more on the reliability of XAI, and we hope in future projects to study the impact of model reliability and optimization on XAI reliability. Another assumption is presented on the limited performance of LIME with k-fold about the data quantity. Based on our analyses, a high number of features may affect the stability of the XAI feature ranking, especially when the number of subjects is relatively small. However, a high number of subjects with a low number of features may increase the stability of feature ranking as a function of k values. This prompted us to investigate this combination further using several ML models and case studies in the future. In addition, in our research, we have proposed a combination of the k-fold technique with SHAP and LIME approaches, and we envisage testing other data sampling techniques in place of k-fold to discuss the reliability of the validation produced by this combination.

Combining the XAI approach with external validation may also lead to reliable XAI validation[Riley 24]. Therefore, it looks interesting to test this combination for reliable validation of the XAI.

Ultimately, XAI aims to ensure and strengthen confidence in predictions by identifying the basis for these predictions. This explainability is crucial in the medical field, as it is a sensitive area where a prediction can recommend specific treatments or tests. However, a thorough study of the reliability of XAI is also essential to reinforce this confidence. Evaluating and improving the reliability of XAI is, therefore, a priority, motivating us to delve deeper into this field in our future research.

V.8 Conclusion

In conclusion, this chapter proposes a novel approach to improving the reliability of XAI by combining XAI techniques with the k-fold technique. By leveraging k-fold cross-validation, we aim to address the challenge of varying explainability outcomes when the selection of test and training data is changed. Integrating k-fold with XAI approaches, such as SHAP and LIME, offers a promising avenue to enhance the generalization, concordance, and stability of explainability in predictive models.

Through the development of metrics to assess the reliability of this combined approach, we have provided insights into its performance in predicting hypothyroidism and diabetes. Our findings highlight the importance of evaluating the generalizability and concordance of feature rankings between several XAI techniques and the k-fold technique.

Notably, the combination of SHAP and k-fold demonstrated superior generalizability, stability, and concordance performance, indicating its reliability in providing interpretable explanations for predictive models. In contrast, the combination of LIME and k-fold showed limitations in generalizability and concordance, underscoring the importance of careful consideration when integrating XAI techniques with cross-validation methods.

Finally, we have validated the combination of SHAP with k-fold on our private datasets. The reliability of this combination was very good for both study cases, with some disruption of generalizability probably due to the difference in data quality.

Overall, our proposed methodology offers a systematic approach to improve the reliability of XAI in predictive modeling for endocrine disease risk prediction. By enhancing the transparency and interpretability of ML models, we contribute to building trust and confidence in their application in clinical decision-making processes. Further research is warranted to explore additional methods and techniques for enhancing the reliability of XAI and advancing its utility in healthcare settings.

Chapter VI
Conclusion and perspectives

VI.1 Conclusion

In response to physicians' pressing need for improved screening and treatment personalization, this thesis presents novel approaches for predicting the risk of carbohydrate abnormalities and MetS. The goal is to identify individuals at high and low risk, thereby enhancing treatment personalization and screening efficiency.

In the first chapter, we reviewed our research domain's current state of the art, identifying key limitations and challenges. The second chapter delved into data analysis and characterization, discussing the quality of both public and private datasets. This analysis laid the groundwork for leveraging these data to predict risks and support medical decision-making. Two distinct approaches were developed to predict glycemic anomalies and MetS, demonstrating promising results in distinguishing high- and low-risk individuals. However, the practical applicability of these models in clinical settings remains constrained by their lack of explainability, with many physicians perceiving these ML models as opaque black boxes. Despite achieving good performance with small datasets, ensuring prediction reliability remains a significant challenge. Additionally, the high cost of extracting biomarkers for MetS prediction limits the practicality of this approach for widespread screening.

In response to these challenges, Chapter 4 introduced two innovative approaches to enhance prediction reliability and increase physician confidence using XAI. The first approach utilized XAI to assess prediction reliability, while the second focused on using clinical variables instead of biological ones to reduce MetS screening costs. Our results indicated a notable improvement in the reliability of glycemic anomaly predictions and a significant reduction in the financial burden of MetS screening without compromising predictive accuracy. Nevertheless, XAI has limitations, such as variability in explainability when training and test datasets are modified.

To address these limitations, Chapter 5 proposes an advanced approach to improving and evaluating XAI reliability. This approach combines XAI with a data augmentation technique and defines specific metrics to assess the reliability of this combination. We tested this method on two case studies, demonstrating that the integration of SHAP with k-fold cross-validation yields promising results, thereby enhancing confidence in XAI's reliability.

VI.2 Contributions

To sum up, we recall below the main contributions of the thesis:

— **Risk prediction of carbohydrate abnormalities and MetS:** We developed methods to identify individuals at high and low risk of carbohydrate abnormalities in patients with β-TM and the risk of MetS in adolescents during screening sessions. These contributions are crucial as they aid physicians in personalizing treatment and screening plans based on the risk levels of their patients, improving patient outcomes, and optimizing the management of financial resources within healthcare systems. By focusing on high-risk patients, we ensure that they receive the necessary attention and interventions. We also allow for a more efficient allocation of medical and financial resources by potentially reducing unnecessary interventions for low-risk individuals.

— **XAI to assess prediction reliability of carbohydrate abnormalities:** We incorporated XAI to provide doctors with insights into the reliability of predictions regarding carbohydrate abnormalities. By analyzing the main features contributing to these predictions, physicians can understand the reasoning behind the model's outputs. This transparency builds trust in the AI system, allowing doctors to make more informed decisions and better communicate with their patients about the risks and recommended treatments.

— **Reduce the financial cost of MetS screening using XAI:** Through using XAI, we demonstrated that it is possible to predict the risk of MetS using only clinical variables, excluding the need for more expensive biological tests. This finding has significant financial implications, as it allows for cost-effective screening processes. By focusing on clinically available data, healthcare providers can maintain accuracy in their predictions while significantly reducing the overall costs associated with MetS screening.

— **Improve and evaluate XAI reliability:** To enhance the reliability of XAI, we combined it with a data sampling technique. This approach improves the validation of explainability by considering changes in the data used for training and testing the models. By doing so, we ensure that the explanations provided by the AI are consistent and robust across different datasets, thereby improving the general trustworthiness of the system. Then we developed a comprehensive set of metrics to assess the reliability of XAI systems. These metrics include stability, which measures how consistent the explanations are over different runs; concordance, which assesses the agreement between different explainability methods; generalizability, which evaluates how well the explanations apply to new, unseen data; and overall reliability, which provides a holistic view of the system's trustworthiness. These metrics offer a robust framework for evaluating and improving the reliability of AI explanations in clinical settings.

VI.3 Limits and Perspectives

Despite the significant contributions of this thesis, several limitations have been identified, along with avenues for future work.

Since it is a frequent problem, many models in the literature are likely overoptimistic due to leakage and overfitting. Our focus in this research was more on XAI reliability, and we hope to study the impact of model reliability and optimization on XAI reliability in future projects. In addition, in our research, we have proposed a combination of the k-fold technique with SHAP and LIME approaches, and we envisage testing other data sampling techniques in place of k-fold to discuss the reliability of the validation produced by this combination. Combining the XAI approach with external validation may also lead to reliable XAI validation. Therefore, it looks interesting to test this combination for reliable validation of the XAI.

Our proposed approach to enhance and evaluate the reliability of XAI is limited to basic ML models that exploit tabular data. It may not apply to deep learning models and other data types like images. Using the Random Forest model, our XAI reliability study was tested in only two case studies. Our findings indicated that the combination of SHAP and k-fold validation is reliable, based on the evaluation metrics developed in this study. However, the combination of LIME and k-fold validation showed poor performance. This lack of reliability could be attributed to the fact that LIME is typically used as a local explainer and may also be caused by the uncertainty of the AI model. Another

assumption is presented on the limited performance of LIME with k-fold about the data quantity. Based on our analyses, a high number of features may affect the stability of the XAI feature ranking, especially when the number of subjects is relatively small. However, a high number of subjects with a low number of features may increase the stability of feature ranking as a function of k values.

The small database for predicting glycemic anomalies in patients with β-TM was also an important limitation. We plan to collect more data and explore other models, such as deep learning, to enhance ML reliability. This would allow for better generalization of results and improved risk prediction capability among these patients. We also plan to exploit deep learning models to analyze the predictive capacity of both carbohydrate abnormalities and mortality together, following a strong causal relationship between these two outputs as a result of physician interpretations.

Only three definitions of MetS were tested for standardized risk prediction of MetS, indicating a limitation in the variety of definitions considered. To improve the generalizability of risk prediction, we aim to gather data on other MetS definitions, providing a more comprehensive understanding and adaptability to several patient groups.

In conclusion, while this work proposed an approach to improve and evaluate the reliability of XAI validation in response to a specific limitation related to explainability change provided by XAI when training or test data is changed, this specific and limited scope is embedded within several other XAI limitations in the literature. Ensuring and strengthening confidence in predictions by identifying the basis for these predictions is crucial, especially in the medical field, where a prediction can recommend specific treatments or tests. Evaluating and improving the reliability of XAI is, therefore, a priority, motivating us to delve deeper into this field in our future research. Ultimately, we want to deploy the models developed with the XAI reliability study on an electronic board and conduct real-world tests on the proposed approach in hospitals.

Bibliography

[Abhari 19] S. Abhari, S. R. N. Kalhori, M. Ebrahimi, H. Hasannejadasl & A. Garavand. *Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods*. Healthcare informatics research, vol. 25, no. 4, page 248, 2019.

[Aggarwal 21] R. Aggarwal, V. Sounderajah, G. Martin, D. S. Ting, A. Karthikesalingam, D. King, H. Ashrafian & A. Darzi. *Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis*. NPJ digital medicine, vol. 4, no. 1, page 65, 2021.

[Alam Khan 02] V. Alam Khan, M. A. Khan & S. Akhtar. *Thyroid disorders, etiology and prevalence*. J Med Sci, vol. 2, no. 2, pages 89–94, 2002.

[Ali 21] S. Ali, S. Mumtaz, H. A. Shakir, M. Khan, H. M. Tahir, S. Mumtaz, T. A. Mughal, A. Hassan, S. A. R. Kazmi, Sadia*et al. Current status of beta-thalassemia and its treatment strategies*. Molecular Genetics & Genomic Medicine, vol. 9, no. 12, page e1788, 2021.

[Allalou 16] A. Allalou, A. Nalla, K. J. Prentice, Y. Liu, M. Zhang, F. F. Dai, X. Ning, L. R. Osborne, B. J. Cox, E. P. Gunderson*et al. A predictive metabolic signature for the transition from gestational diabetes mellitus to type 2 diabetes*. Diabetes, vol. 65, no. 9, pages 2529–2539, 2016.

[Alshayeji 23] M. H. Alshayeji. *Early thyroid risk prediction by data mining and ensemble classifiers*. Machine Learning and Knowledge Extraction, vol. 5, no. 3, pages 1195–1213, 2023.

[Alvarez-Melis 18] D. Alvarez-Melis & T. S. Jaakkola. *On the robustness of interpretability methods*. arXiv preprint arXiv:1806.08049, 2018.

[Alyas 22] T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, A. Ahmad*et al. Empirical method for thyroid disease classification using a machine learning approach*. BioMed Research International, vol. 2022, 2022.

[Alyaseen 23] A. Alyaseen, A. Poddar, N. Kumar, S. Tajjour, C. V. S. R. Prasad, H. Alahmad & P. Sihag. *High-performance self-compacting concrete with recycled coarse aggregate: Soft-computing analysis of compressive strength*. Journal of Building Engineering, vol. 77, page 107527, 2023.

[Andrews 95] R. Andrews, J. Diederich & A. B. Tickle. *Survey and critique of techniques for extracting rules from trained artificial neural networks*. Knowledge-based systems, vol. 8, no. 6, pages 373–389, 1995.

[Arreche 24] O. Arreche, T. R. Guntur, J. W. Roberts & M. Abdallah. *E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection*. IEEE Access, 2024.

[Arrieta 20] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjaminset al. *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.* Information fusion, vol. 58, pages 82–115, 2020.

[Arunachalam 22] P. Arunachalam, N. Janakiraman, J. Rashid, J. Kim, S. Samanta, U. Naseem, A. K. Sivaraman & A. Balasundaram. *Effective Classification of Synovial Sarcoma Cancer Using Structure Features and Support Vectors.* Computers, Materials & Continua, vol. 72, no. 2, 2022.

[Askira-Gelman 98] I. Askira-Gelman. *Knowledge discovery: comprehensibility of the results.* In Proceedings of the thirty-first Hawaii international conference on system sciences, volume 5, pages 247–255. IEEE, 1998.

[Ayed 15] S. B. Ayed, H. Trichili & A. M. Alimi. *Data fusion architectures: A survey and comparison.* In 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA), pages 277–282. IEEE, 2015.

[Ayed 23] M. B. Ayed, M. Soualhi, N. Mairot, S. Giampiccolo, R. Ketata & N. Zerhouni. *Explainable prediction of machine-tool breakdowns based on combination of natural language processing and classifiers.* In Intelligent systems conference, pages 105–121. Springer, 2023.

[Ayed 24] M. B. Ayed, M. Soualhi, R. Ketata, N. Mairot, S. Giampiccolo & N. Zerhouni. *A Data-Driven Methodology to Assess Raw Materials Impact on Manufacturing Systems Breakdowns.* International Journal of Prognostics and Health Management, vol. 15, no. 1, 2024.

[Barragán-Montero 21] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfmanet al. *Artificial intelligence and machine learning for medical imaging: A technology review.* Physica Medica, vol. 83, pages 242–256, 2021.

[Baruah 24] P. Baruah & B. Sarma. *Customer Churn Prediction Using Ensemble Techniques And Algorithms.* Educational Administration: Theory and Practice, vol. 30, no. 6, pages 3427–3436, 2024.

[Benmohammed 11] K. Benmohammed, M. T. Nguyen, S. Khensal, P. Valensi & A. Lezzar. *Arterial hypertension in overweight and obese Algerian adolescents: role of abdominal adiposity.* Diabetes & metabolism, vol. 37, no. 4, pages 291–297, 2011.

[Benmohammed 15] K. Benmohammed, P. Valensi, M. Benlatreche, M. Nguyen, F. Benmohammed, J. Pariès, S. Khensal, C. Benlatreche & A. Lezzar. *Anthropometric markers for detection of the metabolic syndrome in adolescents.* Diabetes & metabolism, vol. 41, no. 2, pages 138–144, 2015.

[Benmohammed 22] K. Benmohammed, P. Valensi, N. Omri, Z. Al Masry & N. Zerhouni. *Metabolic syndrome screening in adolescents: New scores AI_METS based on artificial intelligence techniques*. Nutrition, Metabolism and Cardiovascular Diseases, vol. 32, no. 12, pages 2890–2899, 2022.

[Bhaladhare 21] V. Bhaladhare, N. B. Chouragade, D. Balpande, A. Bhande, R. S. Ambad & N. Bankar. *Ayurvedic management of hypothyroidism*. NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal| NVEO, pages 1440–1447, 2021.

[Biondi 19] B. Biondi, G. J. Kahaly & R. P. Robertson. *Thyroid dysfunction and diabetes mellitus: two closely associated disorders*. Endocrine reviews, vol. 40, no. 3, pages 789–824, 2019.

[Bitew 20] Z. W. Bitew, A. Alemu, E. G. Ayele, Z. Tenaw, A. Alebel & T. Worku. *Metabolic syndrome among children and adolescents in low and middle income countries: a systematic review and meta-analysis*. Diabetology & metabolic syndrome, vol. 12, pages 1–23, 2020.

[Bock 21] C. Bock, M. Moor, C. R. Jutzeler & K. Borgwardt. *Machine learning for biomedical time series classification: from shapelets to deep learning*. Artificial Neural Networks, pages 33–71, 2021.

[Bohr 20] A. Bohr & K. Memarzadeh. *The rise of artificial intelligence in healthcare applications*. In Artificial Intelligence in healthcare, pages 25–60. Elsevier, 2020.

[Bouthillier 21] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti*et al. Accounting for variance in machine learning benchmarks*. Proceedings of Machine Learning and Systems, vol. 3, pages 747–769, 2021.

[Care 23] D. Care. *2. Classification and Diagnosis of Diabetes: Standards of Care in*. Diabetes Care, vol. 46, page S19, 2023.

[Carsote 22] M. Carsote, C. Vasiliu, A. I. Trandafir, S. E. Albu, M.-C. Dumitrascu, A. Popa, C. Mehedintu, R.-C. Petca, A. Petca & F. Sandru. *New entity—thalassemic endocrine disease: major beta-thalassemia and endocrine involvement*. Diagnostics, vol. 12, no. 8, page 1921, 2022.

[Castiglioni 21] I. Castiglioni, L. Rundo, M. Codari, G. Di Leo, C. Salvatore, M. Interlenghi, F. Gallivanone, A. Cozzi, N. C. D'Amico & F. Sardanelli. *AI applications to medical images: From machine learning to deep learning*. Physica medica, vol. 83, pages 9–24, 2021.

[Chaganti 22] R. Chaganti, F. Rustam, I. De La Torre Díez, J. L. V. Mazón, C. L. Rodríguez & I. Ashraf. *Thyroid disease prediction using selective features and machine learning techniques*. Cancers, vol. 14, no. 16, page 3914, 2022.

[Chen 15] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou*et al. Xgboost: extreme gradient boosting*. R package version 0.4-2, vol. 1, no. 4, pages 1–4, 2015.

[Chiniah 16] P. Chiniah. *The one-step ahead Time-varying density forecast window of fat-tailed Value-at-risk models*. PhD thesis, University of Mauritius, 2016.

[Chiu 94] S. L. Chiu. *Fuzzy model identification based on cluster estimation*. Journal of Intelligent & fuzzy systems, vol. 2, no. 3, pages 267–278, 1994.

[Cook 03] S. Cook, M. Weitzman, P. Auinger, M. Nguyen & W. H. Dietz. *Prevalence of a metabolic syndrome phenotype in adolescents: findings from the third National Health and Nutrition Examination Survey, 1988-1994*. Archives of pediatrics & adolescent medicine, vol. 157, no. 8, pages 821–827, 2003.

[Coroama 23] L. Coroama & A. Groza. *Evaluation metrics for Explainable Artificial Intelligence techniques: State of the Art Review and Challenges*. Applied System Innovation, 2023.

[Cunningham 00] P. Cunningham, J. Carney & S. Jacob. *Stability problems with artificial neural networks and the ensemble solution*. Artificial Intelligence in medicine, vol. 20, no. 3, pages 217–225, 2000.

[DA SILVA 23] G. B. DA SILVA, V. R. BOTELHO, C. D. L. Becker, C. Viccari & T. A. Pianoschi. *Modeling of the mass attenuation coefficients of X ray beams using deep neural networks (DNN) and NIST database*. Brazilian Journal of Radiation Sciences, vol. 11, no. 1A (Suppl.), pages 1–20, 2023.

[Das 21a] A. Das. *Logistic regression*. In Encyclopedia of Quality of Life and Well-Being Research, pages 1–2. Springer, 2021.

[Das 21b] R. Das, S. Saraswat, D. Chandel, S. Karan & J. S. Kirar. *An ai driven approach for multiclass hypothyroidism classification*. In International Conference on Advanced Network Technologies and Intelligent Computing, pages 319–327. Springer, 2021.

[Datta 19] S. Datta, A. Schraplau, H. F. Da Cruz, J. P. Sachs, F. Mayer & E. Böttinger. *A machine learning approach for non-invasive diagnosis of metabolic syndrome*. In 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pages 933–940. IEEE, 2019.

[De Ferranti 06] S. D. De Ferranti, K. Gauvreau, D. S. Ludwig, J. W. Newburger & N. Rifai. *Inflammation and changes in metabolic syndrome abnormalities in US adolescents: findings from the 1988–1994 and 1999–2000 National Health and Nutrition Examination Surveys*. Clinical chemistry, vol. 52, no. 7, pages 1325–1330, 2006.

[de Kroon 08] M. L. de Kroon, C. M. Renders, E. C. Kuipers, J. P. van Wouwe, S. van Buuren, G. A. de Jonge & R. A. Hirasing. *Identifying metabolic syndrome without blood tests in young adults—The Terneuzen Birth Cohort*. The European Journal of Public Health, vol. 18, no. 6, pages 656–660, 2008.

[De Sanctis 16] V. De Sanctis, A. T. Soliman, H. Elsedfy, S. A. Yaarubi, N. Skordis, D. Khater, M. El Kholy, I. Stoeva, B. Fiscina, M. Angastiniotis*et al. The ICET-A recommendations for the diagnosis and management of disturbances of glucose homeostasis in thalassemia major patients*. Mediterranean Journal of Hematology and Infectious Diseases, vol. 8, no. 1, 2016.

[De Sanctis 22] V. De Sanctis, S. Daar, A. T. Soliman, P. Tzoulis, M. Karimi, S. Di Maio & C. Kattamis. *Screening for glucose dysregulation in β-thalassemia major (β-TM): An update of current evidences and personal experience.* Acta Bio Medica: Atenei Parmensis, vol. 93, no. 1, 2022.

[De Sanctis 23] V. De Sanctis, A. T. Soliman, S. Daar, P. Tzoulis, S. Di Maio & C. Kattamis. *Glucose Homeostasis and Assessment of β-Cell Function by 3-hour Oral Glucose Tolerance (OGTT) in Patients with β-Thalassemia Major with Serum Ferritin below 1,000 ng/dL: Results from a Single ICET-A Centre.* Mediterranean Journal of Hematology and Infectious Diseases, vol. 15, no. 1, 2023.

[Doshi-Velez 17] F. Doshi-Velez & B. Kim. *Towards a rigorous science of interpretable machine learning.* arXiv preprint arXiv:1702.08608, 2017.

[Dua 17] D. Dua, C. Graff*et al. UCI machine learning repository, 2017.* URL http://archive. ics. uci. edu/ml, vol. 7, no. 1, page 62, 2017.

[Dutta 18] D. Dutta, D. Paul & P. Ghosh. *Analysing feature importances for diabetes prediction using machine learning.* In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pages 924–928. IEEE, 2018.

[Eckel 05] R. H. Eckel, S. M. Grundy & P. Z. Zimmet. *The metabolic syndrome.* The lancet, vol. 365, no. 9468, pages 1415–1428, 2005.

[ElSayed 23] N. A. ElSayed, G. Aleppo, V. R. Aroda, R. R. Bannuru, F. M. Brown, D. Bruemmer, B. S. Collins, J. L. Gaglia, M. E. Hilliard, D. Isaacs*et al. 2. Classification and diagnosis of diabetes: standards of care in diabetes—2023.* Diabetes care, vol. 46, no. Supplement_1, pages 19–40, 2023.

[Esteva 21] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean & R. Socher. *Deep learning-enabled medical computer vision.* NPJ digital medicine, vol. 4, no. 1, page 5, 2021.

[Farmakis 22] D. Farmakis, J. Porter, A. Taher, M. D. Cappellini, M. Angastiniotis & A. Eleftheriou. *2021 Thalassaemia International Federation guidelines for the management of transfusion-dependent thalassemia.* Hemasphere, vol. 6, no. 8, 2022.

[Fogg 01] B. J. Fogg, J. Marshall, T. Kameda, J. Solomon, A. Rangnekar, J. Boyd & B. Brown. *Web credibility research: a method for online experiments and early study results.* In CHI'01 extended abstracts on Human factors in computing systems, pages 295–296, 2001.

[Ford 08] E. S. Ford & C. Li. *Defining the metabolic syndrome in children and adolescents: will the real definition please stand up?* The Journal of pediatrics, vol. 152, no. 2, pages 160–164, 2008.

[Fuster-Palà 24] A. Fuster-Palà, F. Luna-Perejón & M. Domínguez-Morales. *Disease screening using Artificial Intelligence.* 2024.

[Garg 16] M. Garg, N. Mahalle & K. H. Kumar. *Laboratory evaluation of thyroid function: Dilemmas and pitfalls.* Medical Journal of Dr. DY Patil University, vol. 9, no. 4, pages 430–436, 2016.

[Gibbons 75] J. D. Gibbons & J. W. Pratt. *P-values: interpretation and methodology.* The American Statistician, vol. 29, no. 1, pages 20–25, 1975.

[Gogtay 17] N. J. Gogtay & U. M. Thatte. *Principles of correlation analysis.* Journal of the Association of Physicians of India, vol. 65, no. 3, pages 78–81, 2017.

[Gomber 18] S. Gomber, A. Bagaria, S. V. Madhu & P. Dewan. *Glucose homeostasis markers in beta-thalassemia.* Journal of pediatric hematology/oncology, vol. 40, no. 7, pages 508–510, 2018.

[Grundy 04] S. M. Grundy, H. B. Brewer Jr, J. I. Cleeman, S. C. Smith Jr & C. Lenfant. *Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition.* Circulation, vol. 109, no. 3, pages 433–438, 2004.

[Gutierrez-Esparza 21] G. O. Gutierrez-Esparza, T. A. Ramirez-delReal, M. Martinez-Garcia, O. Infante Vázquez, M. Vallejo & J. Hernandez-Torruco. *Machine and deep learning applied to predict metabolic syndrome without a blood screening.* Applied Sciences, vol. 11, no. 10, page 4334, 2021.

[Hahne 08] F. Hahne, W. Huber, R. Gentleman, S. Falcon, R. Gentleman & V. Carey. *Unsupervised machine learning.* Bioconductor case studies, pages 137–157, 2008.

[Hall 23] J. L. Hall, S. Honeycutt, N. Gonzalez, A. O'Donnell-Luria, C. O. Taylor, L. Stevens, A. A. Philippakis & M. C. Schatz. *National Human Genome Research Institute Genomic Data Science Analysis, Visualization, and Informatics Lab-Space: Reaching out to Clinicians.* Circulation. Genomic and precision medicine, vol. 16, no. 3, pages 275–276, 2023.

[Hameed 24] M. A. Hameed, I. Kaaya, M. Al-Jbori, Q. Matti, R. Scheer & R. Gottschalg. *Analysis and Prediction of the Performance and Reliability of PV Modules installed in harsh climates: Case study Iraq.* Renewable Energy, vol. 228, page 120577, 2024.

[Hamilton 20] J. D. Hamilton. Time series analysis. Princeton university press, 2020.

[Hancock 20] J. T. Hancock & T. M. Khoshgoftaar. *CatBoost for big data: an interdisciplinary review.* Journal of big data, vol. 7, no. 1, page 94, 2020.

[Hasan 20] M. K. Hasan, M. A. Alam, D. Das, E. Hossain & M. Hasan. *Diabetes prediction using ensembling of different machine learning classifiers.* IEEE Access, vol. 8, pages 76516–76531, 2020.

[He 19] L.-N. He, W. Chen, Y. Yang, Y.-J. Xie, Z.-Y. Xiong, D.-Y. Chen, D. Lu, N.-Q. Liu, Y.-H. Yang, X.-F. Sun*et al. Elevated prevalence of abnormal glucose metabolism and other endocrine disorders in patients with-thalassemia major: a meta-analysis.* BioMed research international, vol. 2019, 2019.

[Hernandez 22] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla & D. Rankin. *Synthetic data generation for tabular health records: A systematic review.* Neurocomputing, vol. 493, pages 28–45, 2022.

[Holzinger 19] A. Holzinger, G. Langs, H. Denk, K. Zatloukal & H. Müller. *Causability and explainability of artificial intelligence in medicine*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, page e1312, 2019.

[Hosseinzadeh 21] M. Hosseinzadeh, O. H. Ahmed, M. Y. Ghafour, F. Safara, H. K. Hama, S. Ali, B. Vo & H.-S. Chiang. *A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things*. The Journal of Supercomputing, vol. 77, pages 3616–3637, 2021.

[Hsiao 21] J. H.-w. Hsiao, H. H. T. Ngai, L. Qiu, Y. Yang & C. C. Cao. *Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI)*. arXiv preprint arXiv:2108.01737, 2021.

[Huang 07] Y. Huang, P. McCullagh, N. Black & R. Harper. *Feature selection and classification model construction on type 2 diabetic patients' data*. Artificial intelligence in medicine, vol. 41, no. 3, pages 251–262, 2007.

[Huang 23] S.-C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung & A. S. Chaudhari. *Self-supervised learning for medical image classification: a systematic review and implementation guidelines*. NPJ Digital Medicine, vol. 6, no. 1, page 74, 2023.

[Huh 21] J. H. Huh, D. R. Kang, J. Y. Kim, K. K. Koh*et al.* *Metabolic syndrome fact sheet 2021: executive report*. CardioMetabolic Syndrome Journal, vol. 1, no. 2, pages 125–134, 2021.

[James 23] G. James, D. Witten, T. Hastie, R. Tibshirani & J. Taylor. *Linear regression*. In An introduction to statistical learning: With applications in python, pages 69–134. Springer, 2023.

[Janiesch 21] C. Janiesch, P. Zschech & K. Heinrich. *Machine learning and deep learning*. Electronic Markets, vol. 31, no. 3, pages 685–695, 2021.

[Jha 22] R. Jha, V. Bhattacharjee & A. Mustafi. *Increasing the prediction accuracy for thyroid disease: a step towards better health for society*. Wireless Personal Communications, vol. 122, no. 2, pages 1921–1938, 2022.

[Jing 18] X. Jing, Z. Yan & W. Pedrycz. *Security data collection and data analytics in the internet: A survey*. IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pages 586–618, 2018.

[Jo 21] T. Jo. *Machine learning foundations*. Machine Learning Foundations. Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-65900-4, 2021.

[Jordan 15] M. I. Jordan & T. M. Mitchell. *Machine learning: Trends, perspectives, and prospects*. Science, vol. 349, no. 6245, pages 255–260, 2015.

[Kattamis 22] A. Kattamis, J. L. Kwiatkowski & Y. Aydinok. *Thalassaemia*. The lancet, vol. 399, no. 10343, pages 2310–2324, 2022.

[Kawa 21] I. A. Kawa, Q. Fatima, S. A. Mir, H. Jeelani, S. Manzoor, F. Rashid*et al.* *Endocrine disrupting chemical Bisphenol A and its potential effects on female health*. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 15, no. 3, pages 803–811, 2021.

[Ketata 23] F. Ketata, Z. Al Masry, N. Zerhouni & S. Yacoub. *Explainable Machine Learning Approach with Augmentation for Mortality Prediction*. In 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET), pages 01–06. IEEE, 2023.

[Khaleghi 13] B. Khaleghi, A. Khamis, F. O. Karray & S. N. Razavi. *Multisensor data fusion: A review of the state-of-the-art*. Information fusion, vol. 14, no. 1, pages 28–44, 2013.

[Kim 15] T. K. Kim. *T test as a parametric statistic*. Korean journal of anesthesiology, vol. 68, no. 6, page 540, 2015.

[Knudsen 02] N. Knudsen, P. Laurberg, H. Perrild, I. Bülow, L. Ovesen & T. Jørgensen. *Risk factors for goiter and thyroid nodules*. Thyroid, vol. 12, no. 10, pages 879–888, 2002.

[Kodama 14] S. Kodama, C. Horikawa, K. Fujihara, S. Yoshizawa, Y. Yachi, S. Tanaka, N. Ohara, S. Matsunaga, T. Yamada, O. Hanyu*et al.* *Meta-analysis of the quantitative relation between pulse pressure and mean arterial pressure and cardiovascular risk in patients with diabetes mellitus*. The American journal of cardiology, vol. 113, no. 6, pages 1058–1065, 2014.

[Kolluri 22] S. Kolluri, J. Lin, R. Liu, Y. Zhang & W. Zhang. *Machine learning and artificial intelligence in pharmaceutical research and development: a review*. The AAPS journal, vol. 24, pages 1–10, 2022.

[Kononenko 01] I. Kononenko. *Machine learning for medical diagnosis: history, state of the art and perspective*. Artificial Intelligence in medicine, vol. 23, no. 1, pages 89–109, 2001.

[Koskinen 17] J. Koskinen, C. G. Magnussen, A. Sinaiko, J. Woo, E. Urbina, D. R. Jacobs Jr, J. Steinberger, R. Prineas, M. A. Sabin, T. Burns*et al.* *Childhood age and associations between childhood metabolic syndrome and adult risk for metabolic syndrome, type 2 diabetes mellitus and carotid intima media thickness: the international childhood cardiovascular cohort consortium*. Journal of the American Heart Association, vol. 6, no. 8, page e005632, 2017.

[Kulesza 13] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan & W.-K. Wong. *Too much, too little, or just right? Ways explanations impact end users' mental models*. In 2013 IEEE Symposium on visual languages and human centric computing, pages 3–10. IEEE, 2013.

[Kumar 20] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger & S. Friedler. *Problems with Shapley-value-based explanations as feature importance measures*. In International conference on machine learning, pages 5491–5500. PMLR, 2020.

[Kurita 19] T. Kurita. *Principal component analysis (PCA)*. Computer Vision: A Reference Guide, pages 1–4, 2019.

[Lacave 02] C. Lacave & F. J. Díez. *A review of explanation methods for Bayesian networks*. The Knowledge Engineering Review, vol. 17, no. 2, pages 107–127, 2002.

[Lassoued 18] H. Lassoued, R. Ketata & S. Yacoub. *ECG decision support system based on feedforward neural networks*. International Journal on Smart Sensing and Intelligent Systems, vol. 11, no. 1, pages 1–15, 2018.

[LeCun 15] Y. LeCun, Y. Bengio & G. Hinton. *Deep learning*. nature, vol. 521, no. 7553, pages 436–444, 2015.

[Liu 21] C. Liu, S. Wu & X. Pan. *Clustering of cardio-metabolic risk factors and pre-diabetes among US adolescents*. Scientific Reports, vol. 11, no. 1, page 5015, 2021.

[Ma 19] J. Ma, P. Shang, C. Lu, S. Meraghni, K. Benaggoune, J. Zuluaga, N. Zerhouni, C. Devalland & Z. Al Masry. *A portable breast cancer detection system based on smartphone with infrared camera*. Vibroengineering Procedia, vol. 26, pages 57–63, 2019.

[Magge 17] S. N. Magge, E. Goodman, S. C. Armstrong, S. Daniels, M. Corkins, S. de Ferranti, N. H. Golden, J. H. Kim, S. J. Schwarzenberg, I. N. Sills*et al.* *The metabolic syndrome in children and adolescents: shifting the focus to cardiometabolic risk factor clustering*. Pediatrics, vol. 140, no. 2, 2017.

[Magoulas 99] G. D. Magoulas & A. Prentza. *Machine learning in medical applications*. In Advanced course on artificial intelligence, pages 300–307. Springer, 1999.

[Marx 23] C. Marx, Y. Park, H. Hasson, Y. Wang, S. Ermon & L. Huan. *But are you sure? an uncertainty-aware perspective on explainable ai*. In International Conference on Artificial Intelligence and Statistics, pages 7375–7391. PMLR, 2023.

[Mir 20] Y. Mir & S. Mittal. *Thyroid disease pprediction using two tier eensemble classifier*. Int. J. Adv. Sci. Technol, vol. 29, no. 06, pages 4460–4471, 2020.

[Mishra 21] S. Mishra, Y. Tadesse, A. Dash, L. Jena & P. Ranjan. *Thyroid disorder analysis using random forest classifier*. In Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2, pages 385–390. Springer, 2021.

[Mohseni-Takalloo 24] S. Mohseni-Takalloo, H. Mozaffari-Khosravi, H. Mohseni, M. Mirzaei & M. Hosseinzadeh. *Metabolic syndrome prediction using non-invasive and dietary parameters based on a support vector machine*. Nutrition, Metabolism and Cardiovascular Diseases, vol. 34, no. 1, pages 126–135, 2024.

[Mohseni 21] S. Mohseni, N. Zarei & E. D. Ragan. *A multidisciplinary survey and framework for design and evaluation of explainable AI systems*. ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 11, no. 3-4, pages 1–45, 2021.

[Montgomery 21] D. C. Montgomery, E. A. Peck & G. G. Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.

[Mujumdar 19] A. Mujumdar & V. Vaidehi. *Diabetes prediction using machine learning algorithms*. Procedia Computer Science, vol. 165, pages 292–299, 2019.

[Munoz 23] C. Munoz, K. da Costa, B. Modenesi & A. Koshiyama. *Evaluating explainability for machine learning predictions using model-agnostic metrics*. arXiv preprint arXiv:2302.12094, 2023.

[Nasution 23] M. K. Nasution, R. Syah & M. Elveny. *What is data science*. In Data Science with Semantic Technologies, pages 1–25. CRC Press, 2023.

[Nguyen 20a] A.-p. Nguyen & M. R. Martínez. *On quantitative aspects of model interpretability*. arXiv preprint arXiv:2007.07584, 2020.

[Nguyen 20b] T. T. Nguyen, T. Le Nguyen & G. Ifrim. *A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification*. In Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6, pages 77–94. Springer, 2020.

[Nielsen 16] F. Nielsen. Introduction to hpc with mpi for data science. Springer, 2016.

[Oh 18] S. L. Oh, E. Y. Ng, R. San Tan & U. R. Acharya. *Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats*. Computers in biology and medicine, vol. 102, pages 278–287, 2018.

[Patne 18] A. B. Patne, P. J. Hisalkar, S. B. Gaikwad & V. R. Bhagwat. *Effect of Blood Transfusions on Oxidant/Antioxidants Balance in Beta Thalassaemia Maio Patients*. J Clin Diag Res, vol. 12, no. 5, 2018.

[Peng 21] J. Peng, E. C. Jury, P. Dönnes & C. Ciurtin. *Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges*. Frontiers in pharmacology, vol. 12, page 720694, 2021.

[Pisner 20] D. A. Pisner & D. M. Schnyer. *Support vector machine*. In Machine learning, pages 101–121. Elsevier, 2020.

[Program 00] N. H. B. P. E. Program*et al.* *Report of the national high blood pressure education program working group on high blood pressure in pregnancy*. American journal of obstetrics and gynecology, vol. 183, no. 1, pages s1–s22, 2000.

[Rajendra 21] P. Rajendra & S. Latifi. *Prediction of diabetes using logistic regression and ensemble techniques*. Computer Methods and Programs in Biomedicine Update, vol. 1, page 100032, 2021.

[Razia 18] S. Razia, P. S. Prathyusha, N. V. Krishna & N. S. Sumana. *A Comparative study of machine learning algorithms on thyroid disease prediction*. Int. J. Eng. Technol, vol. 7, no. 2.8, pages 315–319, 2018.

[Reisinger 21] C. Reisinger, B. N. Nkeh-Chungag, P. M. Fredriksen & N. Goswami. *The prevalence of pediatric metabolic syndrome—A critical look on the discrepancies between definitions and its clinical importance*. International Journal of Obesity, vol. 45, no. 1, pages 12–24, 2021.

[Rigatti 17] S. J. Rigatti. *Random forest*. Journal of Insurance Medicine, vol. 47, no. 1, pages 31–39, 2017.

[Riley 24] R. D. Riley, L. Archer, K. I. Snell, J. Ensor, P. Dhiman, G. P. Martin, L. J. Bonnett & G. S. Collins. *Evaluation of clinical prediction models (part 2): how to undertake an external validation study*. bmj, vol. 384, 2024.

[Romero-Saldaña 16] M. Romero-Saldaña, F. J. Fuentes-Jiménez, M. Vaquero-Abellán, C. Álvarez-Fernández, G. Molina-Recio & J. López-Miranda. *New non-invasive method for early detection of metabolic syndrome in the working population*. European Journal of Cardiovascular Nursing, vol. 15, no. 7, pages 549–558, 2016.

[Rosenfeld 21] A. Rosenfeld. *Better metrics for evaluating explainable artificial intelligence*. In Proceedings of the 20th international conference on autonomous agents and multiagent systems, pages 45–50, 2021.

[Rosol 24] T. J. Rosol, A. Brändli-Baiocco, M. J. Hoenerhoff & J. L. Vahle. *Endocrine system*. Haschek and Rousseaux's handbook of toxicologic pathology, pages 517–631, 2024.

[Rufo 21] D. D. Rufo, T. G. Debelee, A. Ibenthal & W. G. Negera. *Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)*. Diagnostics, vol. 11, no. 9, page 1714, 2021.

[Sarker 21] I. H. Sarker. *Machine learning: Algorithms, real-world applications and research directions*. SN computer science, vol. 2, no. 3, page 160, 2021.

[Saru 19] S. Saru & S. Subashree. *Analysis and prediction of diabetes using machine learning*. International journal of emerging technology and innovative engineering, vol. 5, no. 4, 2019.

[Schwartzenberg 20] C. Schwartzenberg, T. van Engers & Y. Li. *The fidelity of global surrogates in interpretable Machine Learning*. BNAIC/BeneLearn, vol. 2020, page 269, 2020.

[Sevimli 22] C. Sevimli, Y. Yilmaz, Z. Bayramoglu, R. G. Comert, N. Gul, M. Dursun & Z. Karakas. *Pancreatic MR imaging and endocrine complications in patients with beta-thalassemia: a single-center experience*. Clinical and Experimental Medicine, pages 1–7, 2022.

[Shankar 20] K. Shankar, S. Lakshmanaprabu, D. Gupta, A. Maseleno & V. H. C. De Albuquerque. *Optimal feature-based multi-kernel SVM approach for thyroid disease classification*. The journal of supercomputing, vol. 76, pages 1128–1143, 2020.

[Sharpe 15] D. Sharpe. *Your chi-square test is statistically significant: now what?*. Practical assessment, research & evaluation, vol. 20, no. 8, page n8, 2015.

[Shehab 22] M. Shehab, L. Abualigah, Q. Shambour, M. A. Abu-Hashem, M. K. Y. Shambour, A. I. Alsalibi & A. H. Gandomi. *Machine learning in medical applications: A review of state-of-the-art methods*. Computers in Biology and Medicine, vol. 145, page 105458, 2022.

[Shin 23] H. Shin, S. Shim & S. Oh. *Machine learning-based predictive model for prevention of metabolic syndrome*. PLoS One, vol. 18, no. 6, page e0286635, 2023.

[Sinaga 20] K. P. Sinaga & M.-S. Yang. *Unsupervised K-means clustering algorithm*. IEEE access, vol. 8, pages 80716–80727, 2020.

[Sonia 23] J. J. Sonia, P. Jayachandran, A. Q. Md, S. Mohan, A. K. Sivaraman & K. F. Tee. *Machine-learning-based diabetes mellitus risk prediction using multi-layer neural network no-prop algorithm*. Diagnostics, vol. 13, no. 4, page 723, 2023.

[Sonuç 21] E. Sonuç*et al.* *Thyroid disease classification using machine learning algorithms*. In Journal of Physics: Conference Series, page 012140. IOP Publishing, 2021.

[St 89] L. St, S. Wold*et al.* *Analysis of variance (ANOVA)*. Chemometrics and intelligent laboratory systems, vol. 6, no. 4, pages 259–272, 1989.

[Stassin 23] S. Stassin, A. Englebert, G. Nanfack, J. Albert, N. Versbraegen, G. Peiffer, M. Doh, N. Riche, B. Frenay & C. De Vleeschouwer. *An experimental investigation into the evaluation of explainability methods*. arXiv preprint arXiv:2305.16361, 2023.

[Taher 21] A. T. Taher, K. M. Musallam & M. D. Cappellini. *β-Thalassemias*. New England Journal of Medicine, vol. 384, no. 8, pages 727–743, 2021.

[Tang 19] X. Tang, Z. Ma, Q. Hu & W. Tang. *A real-time arrhythmia heartbeats classification algorithm using parallel delta modulations and rotated linear-kernel support vector machines*. IEEE Transactions on Biomedical Engineering, vol. 67, no. 4, pages 978–986, 2019.

[Tharwat 17] A. Tharwat, T. Gaber, A. Ibrahim & A. E. Hassanien. *Linear discriminant analysis: A detailed tutorial*. AI communications, vol. 30, no. 2, pages 169–190, 2017.

[Tintarev 07] N. Tintarev & J. Masthoff. *A survey of explanations in recommender systems*. In 2007 IEEE 23rd international conference on data engineering workshop, pages 801–810. IEEE, 2007.

[Tj 00] C. Tj. *Establishing a standard definition for child overweight and obesity worldwide: international survey*. Bmj, vol. 320, pages 1–6, 2000.

[Varoquaux 22] G. Varoquaux & V. Cheplygina. *Machine learning for medical imaging: methodological failures and recommendations for the future*. NPJ digital medicine, vol. 5, no. 1, page 48, 2022.

[Viner 05] R. Viner, T. Segal, E. Lichtarowicz-Krynska & P. Hindmarsh. *Prevalence of the insulin resistance syndrome in obesity*. Archives of disease in childhood, vol. 90, no. 1, pages 10–14, 2005.

[Vlek 16] C. S. Vlek, H. Prakken, S. Renooij & B. Verheij. *A method for explaining Bayesian networks for legal evidence with scenarios*. Artificial Intelligence and Law, vol. 24, pages 285–324, 2016.

[Waljee 13] A. K. Waljee, A. Mukherjee, A. G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu & P. D. Higgins. *Comparison of imputation methods for missing laboratory data in medicine*. BMJ open, vol. 3, no. 8, 2013.

[Weiss 04] R. Weiss, J. Dziura, T. S. Burgert, W. V. Tamborlane, S. E. Taksali, C. W. Yeckel, K. Allen, M. Lopes, M. Savoye, J. Morrison*et al.* *Obesity and the metabolic syndrome in children and adolescents*. New England journal of medicine, vol. 350, no. 23, pages 2362–2374, 2004.

[Wiering 12] M. A. Wiering & M. Van Otterlo. *Reinforcement learning*. Adaptation, learning, and optimization, vol. 12, no. 3, page 729, 2012.

[Xu 22] Q. Xu, L. Wang, J. Ming, H. Cao, T. Liu, X. Yu, Y. Bai, S. Liang, R. Hu, L. Wang*et al. Using noninvasive anthropometric indices to develop and validate a predictive model for metabolic syndrome in Chinese adults: A nationwide study*. BMC Endocrine Disorders, vol. 22, no. 1, page 53, 2022.

[Yahyaoui 19] A. Yahyaoui, A. Jamil, J. Rasheed & M. Yesiltepe. *A decision support system for diabetes prediction using machine learning and deep learning techniques*. In 2019 1st International informatics and software engineering conference (UBMYK), pages 1–4. IEEE, 2019.

[Yeh 19] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye & P. K. Ravikumar. *On the (in) fidelity and sensitivity of explanations*. Advances in neural information processing systems, vol. 32, 2019.

[Yousefian 17] F. Yousefian, T. Banirostam & A. AzarKeivan. *Prediction of Mellitus Diabetes in Patients with Beta-thalassemia using Radial Basis Network, and k-Nearest Neighbor based on Zafar Thalassemia Datasets*. Diabetes, vol. 19, page 20, 2017.

[Yousefian 19] F. Yousefian, T. Banirostam & A. Azarkeivan. *Predicting the Risk of Diabetes in Iranian Patients with β-Thalassemia Major/Intermedia Based on Artificial Neural Network*. Signal Processing and Renewable Energy, vol. 3, no. 4, pages 23–33, 2019.

[Zhang 22] J. Zhang, C. Cai, G. Kim, Y. Wang & W. Chen. *Composition design of high-entropy alloys with deep sets learning*. npj Computational Materials, vol. 8, no. 1, page 89, 2022.

[Zhao 21] X. Zhao, W. Huang, X. Huang, V. Robu & D. Flynn. *Baylime: Bayesian local interpretable model-agnostic explanations*. In Uncertainty in artificial intelligence, pages 887–896. PMLR, 2021.

[Zhou 21a] J. Zhou, A. H. Gandomi, F. Chen & A. Holzinger. *Evaluating the quality of machine learning explanations: A survey on methods and metrics*. Electronics, vol. 10, no. 5, page 593, 2021.

[Zhou 21b] Z.-H. Zhou & Z.-H. Zhou. *Semi-supervised learning*. Machine Learning, pages 315–341, 2021.

[Zhu 22] F. Zhu, J. Gao, J. Yang & N. Ye. *Neighborhood linear discriminant analysis*. Pattern Recognition, vol. 123, page 108422, 2022.

[Zimmet 07] P. Zimmet, K. G. M. Alberti, F. Kaufman, N. Tajima, M. Silink, S. Arslanian, G. Wong, P. Bennett, J. Shaw, S. Caprio*et al. The metabolic syndrome in children and adolescents–an IDF consensus report*. Pediatric diabetes, vol. 8, no. 5, pages 299–306, 2007.

[Zuluaga-Gomez 19] J. Zuluaga-Gomez, N. Zerhouni, Z. Al Masry, C. Devalland & C. Varnier. *A survey of breast cancer screening techniques: thermography and electrical impedance tomography*. Journal of medical engineering & technology, vol. 43, no. 5, pages 305–322, 2019.

[Zuo 24] H. Zuo. *Visual Design of Digital Display Based on Virtual Reality Technology with Improved SVM Algorithm*. EAI Endorsed Transactions on Scalable Information Systems, vol. 11, no. 5, 2024.

List of Figures

List of Tables

Titre : Prédiction du Risque des Maladies Endocriniennes à l'aide de la Science des Données et de l'Intelligence Artificielle Explicable

Mots clefs: Science de Données, Intelligence Artificielle, Apprentissage Automatique Explicable, Prédiction du Risque, Aide à la Décision Médicale, Maladies Endocriniennes.

Résumé : L'objectif de cette thèse est de prédire le risque de maladies endocriniennes à l'aide de la science des données et de l'apprentissage automatique. L'idée est d'exploiter cette identification de risque pour aider les médecins à gérer les ressources financières et personnaliser le traitement des anomalies glucidiques chez les patients atteints de bêta-thalassémie majeure, ainsi que pour le dépistage du syndrome métabolique chez les adolescents. Une étude d'explicabilité des prédictions a été développée dans cette thèse pour évaluer la fiabilité de la prédiction des anomalies glucidiques et pour réduire les coûts financiers associés au dépistage du syndrome métabolique. Enfin, en réponse aux limites constatées de l'apprentissage automatique explicable, nous proposons une approche visant à améliorer et évaluer cette explicabilité, que nous testons sur différents jeux de données.

Title : Risk Prediction of Endocrine Diseases using Data Science and Explainable Artificial Intelligence

Keywords : Data Science, Artificial Intelligence, Explicable Machine Learning, Risk Prediction, Medical Decision Support, Endocrine Diseases.

Abstract : This thesis aims to predict the risk of endocrine diseases using data science and machine learning. The aim is to leverage this risk identification to assist doctors in managing financial resources, personalizing the treatment of carbohydrate anomalies in patients with beta-thalassemia major, and screening for metabolic syndrome in adolescents. An explainability study of the predictions was developed in this thesis to evaluate the reliability of predicting glucose anomalies and to reduce the financial burden associated with screening for metabolic syndrome. Finally, in response to the observed limitations of explainable machine learning, we propose an approach to improve and evaluate this explainability, which we test on several datasets.

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon