POLITECNICO DI TORINO

DEPARTMENT OF CONTROL AND COMPUTER ENGINEERING

Master's Degree Course
in Data Science and Engineering

Master Thesis

Responsible Artificial Intelligence for Critical Decision-Making Support: A Healthcare Scenario

Supervisor
Prof. Elena Baralis
External Supervisor – Porini:
Dott. Luca Malinverno

Candidate
Vito Palmisano

Academic Year 2021-2022
Date 20/12/2022

Summary

Artificial intelligence is being implemented in an increasing number of areas, among which healthcare, finance and justice, with the aim of using it as decision-making support. However, high-performing models are usually not trusted by the final user, since they are not transparent. Understanding the model behaviour is fundamental, especially in critical tasks, but it is not enough to build AI systems that have the well-being of humans in the first place. To understand what AI needs to be trustworthy, the concept of responsibility in the Artificial Intelligence field is introduced. It is analyzed in detail, considering all the components needed to build a Responsible Artificial Intelligence system, from explainability to fairness, passing through accountability, security, inclusiveness and reliability. The thesis addresses the problem of the lack of interpretability of models making use of the Responsible AI Toolbox (RAIToolbox), built by Microsoft. The thesis aims also to understand and analyze the strengths and weaknesses of this toolbox, performing different experiments.

RAIToolbox leverages explanation methods to help humans during the decision-making process, inspecting the reasons behind model predictions. It makes use of post-hoc techniques to analyze the behaviour of models from both the perspective of data subgroups and individual instance predictions. The analyzes are built on the notion of cohorts, which are the combination of feature-value pairs intrinsically interpretable. They capture relevant associations, defining subgroups in the features domain. Making use of cohorts allows the tool to be used in spite of the algorithm, and virtually it can be applied to any model for supervised prediction.

Among the multiple components of the RAIToolbox, the Error Analysis one addresses the problem of identifying and analyzing data subgroups in which a model behaves differently. This allows to evaluate model fairness, identify biases and test the model via a comparison between the behaviour of the model on different cohorts and the behaviour on the overall dataset. From the perspective of individual instances, Counterfactual Analysis tool explains the prediction of any model on a specific instance by analyzing what perturbations of single features or joint feature subsets are needed to cause the model to change the prediction. Individual instances analysis, together with critical cohorts analysis, allows to better understand the results of the model Fairness Analysis.

The toolbox is tested on the COMPAS dataset, showing its effectiveness in revealing the model's behaviour at both cohort and individual instance levels. After this validation phase, the RAIToolbox is used to analyze two different scenarios' medical datasets. The former is about 101766 diabetics patients, characterized by

48 features and its target is to predict patients' readmission time to the hospital. The latter is about myocardial infarction complications and contains information about 1700 patients characterized by 123 features and the aim is to predict causes of death if any. Both datasets contain sensitive features and are used to train classification and regression models, allowing the analysis of the different tasks' models. All the performed analyzes allow an understanding of how the RAIToolbox can be used to help physicians better trust models and take responsible decisions. Moreover, the tools allow identifying the strengths and weaknesses of the datasets in the scenario where they were considered.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

AI Artificial Intelligence.

AIM Artificial Intelligence in Medicine.

ASR Automatic Speech Recognition.

CNN Convolutional Neural Networks.

CT Computed Tomography.

DL Deep Learning.

FAQs Frequently Asked Questions.

FN False Negative.

FNR False Negative Rate.

FP False Positive.

FPR False Positive Rate.

MI Myocardial Infarction.

ML Machine Learning.

MRI Magnetic Resonance Imaging.

NLP Natural Language Processing.

P&S Products and Services.

RAI Responsible Artificial Intelligence.

SDGs Sustainable Development Goals.

SVM Support Vector Machine.

TN True Negative.

TNR True Negative Rate.

TP True Positive.

TPR True Positive Rate.

XAI Explainable Artificial Intelligence.

Chapter 1

Introduction

In this chapter, there will be presented all the notions needed to understand the context where this master's thesis work is placed. The first section is dedicated to Artificial Intelligence (AI). A brief recap of AI evolution will be done and it will be seen how different definitions succeeded one to the other, changing the original aim AI seemed to have at its birth. AI sub-fields and its interactions with strictly related fields will be analyzed and, at the end of this section, there will be briefly listed some AI applications. The next section will analyze in detail the applications of AI in the medical field, starting from some historical key inventions going to what AI in Medicine (AIM) is today, to try to understand, in the end, what trends will guide the evolution of AIM. Starting from this last concept, it will be analyzed the trends that are characterizing the whole AI world: the need for trustworthiness in AI systems and the need to put humans' well-being at the centre of the evolution of AI systems. In the end, it will be defined what is the purpose of this master's thesis work, defining also what will be the steps that will be followed, from the theoretical concepts needed to understand the context this thesis work deals with, to the methods and tools used to perform our experiments.

1.1 Artificial Intelligence

In recent years, Artificial Intelligence has become an important topic in the research works of companies and universities. To understand the steps that led to invest so much time and money in AI, it could be useful to briefly analyze some key historical dates for the Computer Science field, starting from the 50s of the past century.

1.1.1 A bit of History

In 1950 Alan Turing published "Computing Machinery and Intelligence" [1]. In the paper, Turing proposes to answer the question "*Can machines think?*" and introduces the Imitation Game to determine if a computer can demonstrate the same intelligence (or the results of the same intelligence) as a human. Turing says that since the words "think" and "machine" cannot be clearly defined, it is needed to "*replace the question by another, which is closely related to it and is expressed in relatively unambiguous words*"[1]. At the end of his work, Turing is able to transform the original question because he is no longer asking whether a machine can "think", but he is asking whether a machine can act indistinguishably from the way a thinker acts [2]. Just six years after, John McCarthy coined the concept of "Artificial Intelligence", at the first-ever AI conference, held in 1956 at Dartmouth College. In 1957 Frank Rosenblatt built the Mark 1 Perceptron, the first computer based on a neural network that "*learns*" through trials and errors [3, 4]. Ten years later, Marvin Minsky and Seymour Papert published a book titled Perceptrons, which become the landmark work on neural networks [5]. Starting from these works, neural networks, which use a back-propagation algorithm to train themselves, began to be widely used in the 1980s [6]. After a period in which AI fell into disrepute, called the "AI Winter", it gradually restored its reputation in the late 1990s and early 21st century by finding specific solutions to specific problems and collaborating the more with other fields, such as statistics, economics and mathematics [6]. In 1997, Deep Blue, the IBM's chess-playing system, beat the then-world champion, Garry Kasparov, in a chess match, upsetting all chess grandmasters [7].

By 2000, solutions developed by AI researchers were widely used, although they were rarely described as "artificial intelligence" [6]. It can be observed that as, despite the term artificial intelligence being coined in 1956, only in the 21st century AI become popular, thanks to access to large amounts of data, advanced algorithms, and improvements in computing power and storage [8]. The Bloomberg journalist Jack Clark stated that 2015 was a landmark year for AI. Computers become a lot better at figuring out what's in an image and lots of companies embraced AI. But no company did it like Google. Google went from "sporadic usage" of deep learning in 2012 to apply it to more than 2,700 projects in 2015 (Fig. 1.1). Jack Clark attributes this to an increase in affordable neural networks due to a rise in cloud computing infrastructure and to an increase in research tools and datasets [9]. UNESCO stated that the amount of research into AI (measured by total publications) increased by 50% in the years 2015–2019 [10]. From Fig. 1.2, it is evident the increase in the number of publications in the field[1] [6].

[1]The annualized total for 2020 is estimated from averaged annual growth rates for prior three

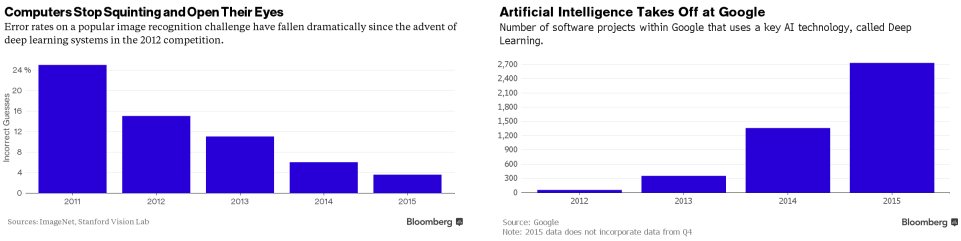


Figure 1.1: Accuracy evolution of AI systems (left) and growth of number of Google's projects making use of deep learning (right) [9].

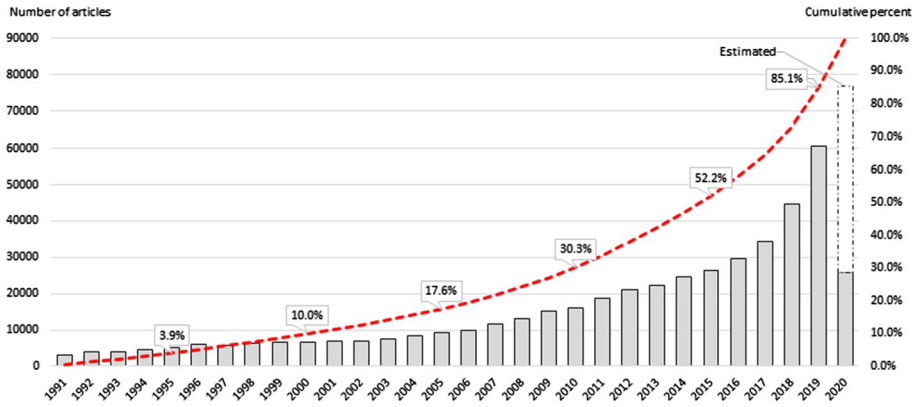


Figure 1.2: Artificial intelligence publication outputs, 1991–2020[1] [6].

1.1.2 AI Definition and Purpose

The evolution of AI over the past seven decades has been analyzed, but the definition of AI and the purpose of AI have not yet been analyzed. After the first attempt of A. Turing, in 1950, to define the AI as the ability of a machine to imitate human behaviour, S. Russell and P. Norvig, in 1995, published the book "Artificial Intelligence: A Modern Approach" [11]. In their studies they investigate four potential definitions of AI, which differentiates computer systems into four groups, based on 'thinking' vs 'acting' and 'humanly' vs 'rationally', preferring to deal with AI in terms of 'rationality' and 'acting rationally', which does not limit how intelligence can be articulated. In 2004, on top of Russell and Norvig work, John McCarthy gave the following definition of AI:

> It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of

years.

using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable" [12].

Concluding, AI uses computers and machines to simulate and improve the decision-making and problem-solving abilities of the human mind. It also enables machines to learn from their past experiences, adapt to new inputs, and carry out human-like tasks. Obviously, setting up the system and posing the proper questions still requires human intervention. In order for algorithms to learn, AI finds structure and regularities in the data [13].

AI gets the most out of data. It is important to emphasize this point. The data itself is an asset when algorithms are self-learning. Data contains the answers and to find them, the only thing needed is to use AI. Data may give a competitive advantage because its role is crucial. Even if everyone in an industry is using the same methods, the best data will prevail [13].

1.1.3 AI Taxonomy

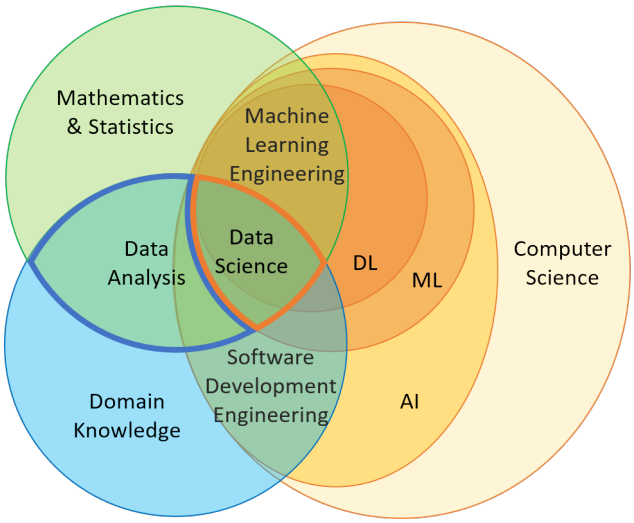


Figure 1.3: Artificial intelligence Venn's diagram is shown. It can be seen AI components and their interactions with Mathematics & Statistics and Domain Knowledge fields.

Let's go deeper in understand what are AI sub-fields and what are the most important disciplines interacting with it. From Fig. 1.3 it can be seen that AI is a field which takes land from Computer Science and articulate in the sub-fields of Machine Learning (ML) and Deep Learning (DL), which are frequently mentioned in conjunction with AI. These disciplines are comprised of AI algorithms that

typically make predictions or classifications based on input data. Machine learning has improved the quality of some expert systems, and made it easier to create them.

On the other side, AI interacts with two very important macro fields, Mathematics & Statistics and Domain Knowledge. From their interactions, other different fields arise, which operate in the world of Data and AI using different approaches and tools[2].

ML, reorganized as a separate field, started to flourish in the 1990s. The field changed its goal from achieving AI to tackling solvable problems of a practical nature. It shifted focus away from the symbolic approaches it had inherited from AI to methods and models borrowed from statistics, fuzzy logic, and probability theory [14]. An "intelligent" computer uses AI to think rationally and to perform tasks on its own, while ML is how a computer system develops its intelligence. AI is a bigger concept to create intelligent machines that can simulate behaviour, whereas, ML is an application or subset of AI that allows machines to learn from data without being programmed explicitly [15].

Now, let's focus on the differences between DL and ML. As mentioned above, both DL and ML are sub-fields of AI, and DL is a sub-field of ML. How DL and ML differ is in how each algorithm learns. DL uses deep neural networks. The "deep" in a DL algorithm refers to a neural network with more than three layers, including the input and output layers. The rise of DL has been one of the most significant breakthroughs in AI in recent years because it has reduced the manual effort involved in building AI systems. It is important to say that the wide use of DL was, in part, enabled by big data and cloud architectures, making it possible to access huge amounts of data and processing power for training AI solutions [16].

1.1.4 Applications of AI

Today, AI plays a role in everyday life, powering search engines, product recommendations and speech recognition systems. AI adds intelligence to existing products, and the models adapt when given new data. Many already used products will be improved with AI capabilities, let's think of social network ads, online shop recommendation systems and autonomous driving cars. They are the result of the application of AI and large amounts of data to improve many already existing technologies. The interactions with voice assistants and search engines are all based on DL and these products keep getting more accurate the more they are used. Going a little bit in detail, the following are some of the most common examples of the numerous, real-world applications of AI systems today.

[2]This thesis work, will not go deeper on the interaction aspects. It was decided to insert them to not lack completeness.

- Automatic Speech Recognition (ASR) or speech-to-text: It is a capability which translates human speech into a written format.
- Customer service: Online chatbots are replacing human agents along the customer journey, changing the way to think about customer engagement. Chatbots answer Frequently Asked Questions (FAQs) about topics, or provide personalized advice.
- Computer vision: This AI technology enables computers to derive meaningful information from digital images, videos, and other visual inputs, thanks, for example, to the use of Convolutional Neural Networks (CNN).
- Recommendation engines: Using past consumption behaviour data, they help to discover data trends that can be used to develop more effective cross-selling strategies.
- Fraud detection: Banks and other financial institutions can use machine learning to spot suspicious transactions. Anomaly detection can identify transactions that look atypical and deserve further investigation.

The next section elaborates on the possibilities AI finds in medicine and healthcare, the field that this thesis work want to analyze. It evaluates the opportunities AI offers in medicine but also the criticisms of using AI in a sensitive field like this one.

1.2 AI for medical purposes

The term AI is applicable to a wide range of areas in medicine, such as robotics, medical diagnosis, medical statistics, and human biology. AI in medicine (AIM) has two main branches: virtual and physical [17]. The virtual branch includes informatics approaches from information management to control of health management systems, including electronic health records, and active guidance of physicians in their treatment decisions. The physical branch is best represented by robots used to assist the elderly patient or the attending surgeon. Also embodied in this branch are targeted nanorobots, a unique new drug delivery system [17]. Only the virtual branch will be covered, since the physical one is out of the interests of this work.

To understand AIM implications, it is necessary to ask what factors are needed for successfully perform patient care. From the point of view of physician, the more you know and the more patients you treat the better patient care you can provide [18]. Usually, this happens with time, meaning physicians acquire knowledge and experience during their career. This concept of experience and knowledge is

central. The more experience and data (information analysis) there is, the better knowledge-based decisions will be. Here comes the main limitation of the human mind, because over 40 productive career years, a radiologist will look at approximately 225,000 MRI/CT exams [19], while AI can start off with this number and within a short period of time reach into the millions of scans, thus further improving its accuracy [18].

This is why an AI-driven application is able to out-perform dermatologists at correctly classifying suspicious skin lesions [20] or why AI is being trusted with tasks where experts often disagree, such as identifying pulmonary tuberculosis on chest radiographs [21, 22]. Since the advent of ML and DL, applications of AIM have expanded, creating opportunities for personalized medicine rather than algorithm-only–based medicine. Predictive models can be used for diagnosis of diseases, prediction of therapeutic response, and potentially preventative medicine in the future. AI may improve diagnostic accuracy, improve efficiency in provider workflow and clinical operations, facilitate better disease and therapeutic monitoring, and improve procedure accuracy and overall patient outcomes [23]. All these advantages let's understand why AIM research is growing rapidly. In 2016, healthcare AI projects attracted more investment than AI projects within any other sector of the global economy [24]. However, among the excitement, there is equal scepticism, with some urging caution at inflated expectations [25, 22].

Going deeper and analyzing some key periods for AIM, could help to easier understand what will be its future. It is important to notice that, as for AI diffusion and growth, also for AIM diffusion were very important two aspects: collection and sharing of data and technology improvements.

1.2.1 AIM History

Starting from the time period between 50s and 70s, despite innovations in engineering, medicine was slow to adopt AI. However, this first AI period, was an important time for digitizing data that later served as the foundation for future growth and utilization of AIM. The development of the Medical Literature Analysis and Retrieval System and the web-based search engine PubMed by the National Library of Medicine in the 1960s became an important digital resource for the later acceleration of biomedicine [26]. Clinical informatics databases and medical record systems were also first developed during this time and helped establish the foundation for future developments of AIM [23]. During the winter period, between 70s and 2000s, although the lack of general interest during this time period, collaboration among pioneers in the field of AI fostered the development of The Research Resource on Computers in Biomedicine by Saul Amarel in 1971 at Rutgers University [27]. In 1973, the Stanford University created the Medical Experimental–Artificial Intelligence in Medicine, a timeshared computer system,

and enhanced networking capabilities among clinical and biomedical researchers from several institutions [28]. Largely as a result of these collaborations, the first National Institutes of Health–sponsored AIM workshop was held at Rutgers University in 1975 [26]. These events represent the initial collaborations among the pioneers in AIM [23].

One of the first prototypes to demonstrate feasibility of applying AI to medicine was, in 1976, the development of a consultation program for glaucoma using the CASNET model, a causal–associational network able to apply information about a specific disease to individual patients and provide physicians with advice on patient management [29]. In the early 1970s, MYCIN was developed [30]. Based on patient information input by physicians and a knowledge base of about 600 rules, MYCIN could provide a list of potential bacterial pathogens and then recommend antibiotic treatment options adjusted appropriately for a patient's body weight. In 1986, DXplain, a decision support system, was released by the University of Massachusetts. This program uses inputted symptoms to generate a differential diagnosis [31]. It also serves as an electronic medical textbook, providing detailed descriptions of diseases and additional references. When first released, DXplain was able to provide information on approximately 500 diseases. Since then, it has expanded to over 2400 diseases [32]. By the late 1990s, interest in ML was renewed, particularly in the medical world, which along with the above technological developments set the stage for the modern era of AIM [23].

In 2007, IBM created the open-domain question–answering system Watson, based on DeepQA technology, which uses NLP and various searches to analyze data over unstructured content to generate probable answers [33]. This system was more readily available for use, easier to maintain, and more cost-effective. By drawing information from a patient's electronic medical record and other electronic resources, one could apply DeepQA technology to provide evidence-based medicine responses. As such, it opened new possibilities in evidence-based clinical decision-making [33, 18]. In 2017, Bakkar et al. [34] used IBM Watson to successfully identify new RNA-binding proteins that were altered in amyotrophic lateral sclerosis. Given this momentum, along with improved computer hardware and software programs, digitalized medicine became more readily available, and AIM started to grow rapidly [23].

In the 2000s, thanks to the overcame of the limitations due to the lack of large datasets and of computing power, the adoption of CNNs made possible to analyze images and to recognize patterns with a high accuracy. They found important applications in radiology, oncology, cardiology, gastroenterology, ophthalmology, surgey and so many other medical fields. In 2017, Esteva et al. [20] trained a CNN to identify nonmelanoma and melanoma skin cancers with results indicating CNN performance comparable with experts. Weng et al. [35] shown how a CNN can be used to predict cardiovascular risk in a cohort population. AI was shown to improve

accuracy in cardiovascular risk prediction compared with the established algorithm defined by the American College of Cardiology guidelines [35, 23]. These are just some of the examples that can be reported when it comes to AIM applications.

This thesis work will deal only with the part relating to "support decision making through past knowledge" and not with images, for a matter of simplicity, since this thesis work is focused on a broad concept such as Responsible AI in the medical field, and not on the role of image analysis in the medical field.

1.2.2 AIM: Today and Tomorrow

Till now, the historical path of AIM improvements under a technological point of view was analyzed. But today's "systems thinking" about health care not only focuses on the classical interactions between patients and providers but takes into account larger-scale organizations and cycles. So, it needs to be considered also that the health care system must not be stationary but must learn from its own experiences and strive to implement continuous process improvements [17]. This process involves building or participating in an organization, which uses AI to achieve significant progress. The dynamics of individual patients needs to be captured within a larger societal ecosystem, including their responses to received medications as well as their behavioral interactions. This global care coordination allows process mapping, facilitates control, and better supports changes to the system with a demonstrated increase in response to medication, decrease of costs and more efficacious interventions [17].

AI has to be used to improve organizational performance by enabling individuals to capture, share and apply their collective knowledge to make "optimal decisions in real time". Hamet et al. [17] think that major efforts are required from academia and the information technology industry to achieve desired efficacy and minimize cost. The current status of medical records is mostly in the form of incommunicable silos of wasted information for the health system and for knowledge acquisition [17]. Laboratories and clinics need to collaborate to accelerate the implementation of electronic health records [36, 17].

A very important point, which often slows progress, is that AIM's scope is to support doctors, not to replace them. Machines lack human qualities such as empathy and compassion, and therefore patients must perceive that consultations are being led by human doctors. Furthermore, patients cannot be expected to immediately trust AI, a technology shrouded by mistrust [22]. Therefore, AIM commonly handles tasks that are essential, but limited enough in their scope so as to leave the primary responsibility of patient management with a human doctor [22].

In the end, it was said that new scientific and clinical findings should be shared through open source, and aggregated data must be displayed for open access by

physicians and scientists and made automatically available as point-of-care information. It is needed to develop cost-effective AI models and products to allow physicians, practices, and hospitals to incorporate AI into daily clinical use. Physicians should not view this as "human versus machine" but rather as a partnership in an effort to further improve clinical outcomes for the patient with diseases [23]. AI-augmented medical systems will serve to improve workflow and provide safer, more consistent and more quantitative results grounded on knowledge-based decisions [18].

The societal and ethical complexities of these applications require further reflection, proof of their medical utility, economic value, and development of interdisciplinary strategies for their wider application [17]. The road to implementing AIM is still long, fraught with various issues to be addressed along the way, from government approvals to ethical issues, as well as addressing misconceptions in the public relating to AIM [18].

If ethical standards are created, measures of success and effectiveness are developed, AI tools are made open-source and user-friendly, and proven clinical utility is achieved, the use of AIM will benefit society [17]. The concept of using AIM should be as a decision support system with the final action being from humans [18]. All these concepts are becoming important for all AI applications and not only for AIM ones. Society is asking for transparency and security and the research works seem to go in the same direction.

1.3 Future of AI

In this section, it will be analyzed what are the trends guiding investments in AI research and it will be seen that there are not only investments to improve performances or to reduce costs.

Thanks to the advantages AI can give, more and more businesses are adopting it in the short term to solve specific challenges [18]. Considering the technical point of view, Knowledge Graphs [37] are an emerging technology within AI, while NLP applications are expected to increase in sophistication. The former can encapsulate associations between pieces of information and drive upsell strategies, recommendation engines, and personalized medicine; the latter are expected to enable more intuitive interactions between humans and machines [16]. Furthermore, the advent of quantum computers could lead to a radical transformation of the way the use of AI and its results are approached.

Beyond the technological point of view, it has to be asked what it will be the role of ethics in the future of AI, how much important big data is and why domain knowledge could be crucial for the success of AI. The following is an extract from an interview made at SAS at the end of 2019, in which AI experts and data scientists

try to predict what will be AI trends for the next decade.

> "In 2019 we focused a lot on the algorithms and people understanding what AI is. [...] As we look to 2020 some of the trends I see are really about that pragmatic AI and a continued focus on transparency, ethics, and removing bias in AI systems. The trends we will see in 2020 and beyond will be specifically in how well and how accurate and how Justified AI is. [...] The whole area of how machine learning is applied in everybody's lives is going to be the next wave were you going to hear a lot of AI discussion. [...] There will be, in my opinion, a crossroads with machines making actual decisions instead of humans. The component of ethics will only evolve as we continue interesting machines to make decisions for humans. There is so much unstructured information that is barely analyzed. Images, text files, all your files, transforming all these big data into insights is going to be a huge trend. [...] It really is "Who has the data", that's who will be the King" [13].

From this interview, it can be seen how the main participant in the discussion about what the future of AI will be, is the need to "*focus on transparency, ethics, and removing bias in AI systems*". It can be noticed also a concept already discussed in the previous sections: the role and importance of Data. These concepts will be the focus of this thesis work, so let's give them a little introduction, and start to get familiar with them.

1.4 Problem Statement: The need for Human Centered AI

Let's start analyzing the concept of transparency, which strictly related to the concept of explainability. Transparency includes transparency of all AI relevant elements, from data to system and business model. Instead, explainability refers only to humans understanding the output of an algorithm, in particular, a ML one. Often a ML model is considered a black-box system (i.e. a system producing outputs without revealing any information about its internal workings. The explanations for its conclusions remain opaque or "black"). In these systems even the model programmers themselves cannot fully understand or explain how they achieved a particular result [38]. When such software programs are dealing with data as sensitive as healthcare-related ones, it can be appreciated the need to better understand how they arrived at a specific result. Such a grasp of the technology will allow physicians to use them with better trust. As said, AIM has to be a decision support system with the final action being from humans, so those

who takes the decisions must be able to understand why the algorithm gives one answer rather than another [39, 40].

Explainable AI (XAI) aims to shift the traditional black-box approach. It is used to describe an AI model, its expected impact and potential biases. It helps characterize model accuracy, fairness, transparency and outcomes in AI-powered decision-making. As stated by J. Amann et al [40] "*Omitting explainability in clinical decision support systems poses a threat to core ethical values in medicine and may have detrimental consequences for individual and public health*". In addition to that, not only clinicians but also patients, and indeed any stakeholder in healthcare, need to better understand AI when it comes to the medical field to gain the maximum benefit from this still quite new technology.

Despite what XAI allows, it is not sufficient. When a human beings make decisions, the action itself is normally connected with a direct responsibility by the agent who generated the action. You have an effect on others, and therefore, you are responsible for what you do and what you decide to do. If to take decisions is not a human being, but an artificial intelligence system, it becomes difficult and important to be able to ascribe responsibility when something goes wrong [39]. A typical example is the case of a self-driving car. If the automation system of the car causes an accident, who is responsible? To use AI safely, as a support to the activity of physicians, it is necessary it is trustworthy and validated in clinical practice. But this could not be sufficient because there could be an automation bias. Automation bias is the tendency for humans to favour machine-generated decisions, ignoring contrary data or conflicting human decisions. It leads to errors of omission and commission, where omission errors occur when a human fails to notice or disregards, the failure of the AI tool. High decision flow rates, where decisions are swiftly made on physician's examinations, and the physician is reading examinations rapidly, predispose to omission errors. Commission errors occur when the physician erroneously accepts or implements a machine's decision despite other evidence to the contrary [39]. The fact that the reasoning of a model is explainable and transparent does not imply that it is correct or fair. This is why explainability is not sufficient. To overcome these issues, in 2018, the European Commission established the High-Level Experts Group on Artificial Intelligence with the general objective to support the implementation of the European Strategy on Artificial Intelligence, including the elaboration of recommendations on future-related policy development and ethical, legal and societal issues related to AI [39]. Based on fundamental rights and ethical principles, the guidelines list seven key requirements that AI systems should meet to be trustworthy [41]. This guidelines will discussed and analyzed in the chapter dedicated to Responsible AI. The group recommended that the development, deployment and use of AI systems should adhere to the ethical principles of respect for human autonomy, prevention of harm, fairness/equity and applicability.

The solution could be to create an ethical AI, human-centred, subject to constant action control: an AI subjected to a civil liability, written in the software and for which the producers must guarantee the users so that they can use AI reasonably and with human-controlled automation. Responsible AI is needed to cover all issues caused by AI. During this work, it will be better understood what Responsible AI is and what it can be done to direct the use of AI on the Ethic path.

1.5 Purpose and Outline of the Thesis

The first sections gave some definitions and historical key notions to understand the evolution of AI and AIM. They gave also some context about the need for Responsible AI in the medical field, explaining why XAI is not sufficient to build a trustworthy environment.

This thesis work wants to go deeper, analyzing this last point, explaining in detail what Responsible AI is, how it is related whit all its components, and how important each component is, from Explainability to Accountability, through Fairness, Security, Privacy and others concepts that need to be analyzed and studied. The differences and nuances between all these elements will be underlined, with the aim, in the end, to get the full picture of what RAI is.

At this point, the RAIToolbox will be analyzed. It is an important tool built by Microsft, which allows to perform analysis of the data and of the model and to explain a black-box model.

In the end, three different open-source medical datasets having different characteristics and structures, will be presented and analyzed. Classification models will be trained based on the task datasets are built for, and the RAIToolbox will be used to perform analysis about the models and the datasets, covering all the points the tool is built for. The aim is to analyze the tool form all the points of view, trying to understand if it is able to satisfy the requirements needed to build a trustworthy and responsible AI system. Both strengths and criticisms of the tool will be analyzed, trying to figure out all the possible applications and all the possible situations may occur during its use.

1.6 Contribution

This master's thesis work comes after an internship made in Porini. Thanks to the nice experience made, we decided to continue to work together also on the thesis.

Porini is a competence centre and Microsoft's Gold Partner. It aims to support the management of medium and large companies in Italy and around the world

in the design and implementation of solutions able to drive all company functions toward digital transformation in progress. The extensive suite of Porini products based on the Microsoft platform and Azure technology is available on-premises and in the cloud and is offered worldwide both directly and through a network of qualified partners. Thanks to its team of about 180 professionals located in four countries, Porini supports its customers during the adoption and development of solutions aimed at improving decision-making and governance systems enterprise, exploiting technology and innovation to equip itself with adequate tools for the pursuit of strategic objectives.

The company was founded in Como in 1968 as a company specializing in specific solutions and consulting services for companies in the fashion, clothing, textile and retail sectors in Italy and around the world. Over the years, Porini has expanded its skills by becoming a Microsoft ISV and including in its portfolio Social CRM, Business Analytics, Artificial Intelligence, Machine Learning, IoT, Performance Management, Collaboration and Knowledge Management solutions for medium and large companies in many other sectors such as Manufacturing, Financial Services, Travel & Transportation and Health. Since 2018 Porini has become part of the DGS group: thanks to this agreement, Porini can offer the market an increasingly advanced competence centre on Microsoft's technological platforms, both nationally and internationally. Porini is Microsoft's golden partner for Analytics & Advanced Analytics and since the beginning of 2021, it has created an internal structure, the Porini Innovation & Research Center for Data Science (PIRC). PIRC focuses on innovation projects and the management of relationships with universities and institutions of research for the shared thesis or collaborative research projects. Inside PIRC I had the pleasure to have Luca Malinverno as supervisor and to work with Francesco Ghisoni, a colleague who stood by me throughout the work. The PIRC Team is working on an innovative training project that allows access to on-demand Artificial Intelligence laboratories. In addition to the existing environments, the team planned the creation of an AI laboratory dedicated to the medical and healthcare area. The thesis aim was to include me in the team dealing with the creation of this environment, to follow the training and testing phases of the algorithms that will then be made available in the laboratory, with a focus on the ethical aspects.

Figure 1.4: In order, the logo of: Porini, Porini Innovation & Research Center, Porini Education.

Chapter 2

Responsible AI

2.1 Introduction to Responsible AI

Unwanted effects of AI have received a lot of attention in recent years [42]. Many companies are beginning to worry about the possible issues of AI technologies, as they become more widely used in businesses and are questioning how to get ready to prevent unanticipated bad effects. Before AI can be used at scale in enterprises and communities, there are a number of crucial problems that need to be addressed [42]. Every AI application has potential risks that should be carefully considered and addressed. According to Ghallab [43], there are three broad categories of dangers that are prevalent in various applications of artificial intelligence, which are not independent and present technical, scientific, legal, and political challenges:

- **Safety of critical AI applications**: More and more, AI approaches are being used in safety-critical applications and fields that may have very high costs on the social, economic, or environmental fronts, like, for instance, the health sector. Given the complexity and opacity of many AI models and techniques and the intricate traceability of the hardware and software components within systems, which are becoming larger and more complex, procedures requiring informal technical descriptions and declarations of conformity to standards may not be sufficient.

- **Security and privacy of individual users**: The state of the art for digital interaction security is quite advanced, but its deployment, particularly in portable applications and connected products, is insufficient. From the perspective of specific users, transparency and understandability are equally crucial. Due to the consequences a decision support system could lead. It should be able to explain its assumptions, limitations, and criteria.

- **Social risks**: Implementing AI technologies could cause societies to face issues such as biases, economic risks, political risks, unemployment increase. In decision support systems numerous cases of gender, ethnical or seniority biases have been reported [44, 45]. These systems lacks transparency, intelligibility or/and rely on training data which is biased in hidden ways difficult to uncover and mitigate [43]. Considering political risk, the Cambridge Analytica scandal is an example [46], while High Frequency Trading (HFT) and Algorithmic pricing are some of the AI deployments that could lead to Economic risks [43]. Furthermore, technology developments are strongly suspected to be a contributing factor for the observed increase in social inequalities [43].

Some of the issues, such as the influence on liability and malicious usage, are outside the purview of private businesses and demand government intervention. Nevertheless, others must be addressed at the level of each individual corporation [42]. Experts and larger groups are debating the most of these issues, and it appears that there is broad agreement regarding their causes and potential effects. However, there is less agreement and experience regarding how to effectively address such issues in businesses that develop and employ AI, both from a technological and organizational standpoint [42, 47]. Many challenges face the practical implementation of AI for social good efforts. Additionally, in the field of fairness, accountability, and transparency of AI, decades of research has only recently begun to be more thoroughly incorporated into practical settings, and many questions remain [47].

Considering the need for governments intervention, the required measures are part of the regulatory mechanisms of society. It takes decades to fully comprehend, inform, raise awareness, and develop the social forces necessary to enforce legislation. This is because these processes have a slow reaction time. But technological progress has accelerated significantly and the difference between the two dynamics necessitates taking proactive measures. Social experiments and comprehensive research on social dangers and countermeasures are the foundation of a proactive strategy. In the end, social experimentation prior to a technical deployment lowers the gap between the dynamic of technology grow and the dynamics of social regulation [43].

Taking into considerations the effort of single companies, as of 2019, more than 20 firms (For example, Microsoft, Google, IBM, Sage, Workday, Unity Technologies, and Salesforce) have produced frameworks, principles, guidelines, and policies related to the responsible development and use of AI [47]. These governance documents typically address a set of social and ethical concerns, propose principles in response, and in some cases offer concrete reforms or internal governance strategies. The comparison of the many AI documents created by businesses, government agencies, and non-governmental organizations reveal a strong consensus in the ethical priorities of these companies [48]. The social and ethical issues that are

frequently brought up center on concerns for the welfare of the public, of customers and of employees and on concerns for algorithmic bias and fairness, transparency and explainability, trust in AI, and the dependability and safety of AI products [49]. These and other high-level RAI concepts, nevertheless, can frequently be ambiguous, host a wide range of alternative interpretations, and be challenging to apply in real-world situations [47].

A crucial issue for AI in the near future is how to convert high-level principles into practical, ethical actions. Companies working on AI should pay attention to the issue of closing the principles-to-practices gap, as should those who might purchase and use AI systems, as well as other stakeholders and the general public [47].

Fortunately, some first steps are being taken. In 2016, with the publication of "Genaral Data Protection Regulation", Europe became a model for many other countries across the world. GDPR focuses on the protection of data, regulating risk management and accountability. The aim is to set the individual dimension as the central role. Thanks to GDPR, if businesses want to use customer's information, they must first obtain a consent or permission from him. Although the GDPR is an EU regulation, it has global implications because it requires any overseas marketers, who want to connect with EU citizens, to follow its regulations [50, 51].

The European Commission did not stop to the publication of the GDPR and in 2019 published "Ethics guidelines for trustworthy AI" [41], which proposes an assessment check list for AI practitioners based on seven principles:

- Human Agency and Oversight;
- Technical Robustness and Safety;
- Privacy and Data Governance;
- Transparency;
- Diversity, Non-Discrimination and Fairness;
- Societal and Environmental Well-Being;
- Accountability.

These seven principles have to be continuously evaluated and addressed throughout the entire AI's system life cycle [41].

In this thesis work, the European Commision's seven principles are taken as the foundations for the building of Responsible Artificial Intelligence. So, in the next sections, each principle will be described in detail.

2.2 Human agency and oversight

Responsible AI places human (e.g., end-users) at the center and complies with legal requirements, stakeholder expectations, and regulatory requirements. Prior to designing and implementing responsible AI, organizations need to understand the practices that will help them drive ethics and trust of AI use [51]. Organizations must comprehend the procedures that will guide ethics and trust in the use of AI before creating and executing RAI. Companies must launch an education campaign outlining what AI is, why and how it is employed within the organization and what the obstacles are. The campaign has to present the guiding concepts, the approach, the training course, and the equipment [42]. This training program begins with those who are most closely involved in designing and developing services and products that make use of AI. Later on, training can be made available to the entire business and it might be very technical or non-technical [42].

The respect for human autonomy is the basic principle from which to start. AI systems should support human autonomy and decision-making and should be at the service of society and generate tangible benefits for people. AI systems should always stay under human control and be driven by value-based considerations. AI used in products and services should in no way lead to a negative impact on **human rights** [42]. This necessitates that AI systems support **human agency**, promote fundamental rights, and act as enablers of a democratic, prosperous, and egalitarian society while still allowing for **human oversight** [41].

Considering **fundamental rights**, AI systems have the potential to undermine them. An evaluation of the impact on basic rights should be done in circumstances where such risks exist and any potential trade-offs between the various principles and rights have to be identified and recorded. Its necessary to made questions about how the AI system influences human users' decisions (e.g., recommended actions or choices, option presentation) and whether it might compromise human autonomy by unintentionally interfering with its capability to make decisions. For example, companies have to think about whether the AI system should inform users that a decision, content, recommendation, or outcome is the result of an algorithmic decision (e.g. are you making aware users that they are interacting with a chat bot in the case of a conversational system)? This needs to be done before the system is developed, and it should include a review of whether such risks can be minimized or justified as necessary in a free society to respect others' freedoms and rights. In addition, procedures for obtaining feedback from outside sources on AI systems that might violate fundamental rights must to be established [41].

When talking of **human agency**, it refers to the fact that users should be able to independently make well-informed choices concerning AI systems. They should be given the information and resources necessary to engage and comprehend AI

systems in a satisfying manner, as well as the ability to reasonably self-evaluate or challenge the system when appropriate. AI systems should assist people in making wiser decisions that are in line with their objectives. Since they may use sub-conscious processes, including various forms of unfair manipulation, deception, herding, and conditioning, which all have the potential to threaten individual autonomy, AI systems can occasionally be used to shape and influence human behavior through mechanisms that may be challenging to detect. The functionality of the system must be based on the general notion of user autonomy. The right to be free from decisions entirely based on automated processing is essential to this when it causes legal repercussions or other major implications for the user. To ensure human agency, AI implementers should consider the task distribution between the AI system and humans for meaningful interactions and suitable human oversight and control. They should consider whether if the AI system augments or enhances human abilities and should take precautions to avoid overconfidence in the AI system to complete tasks [41].

In the end, to prevent an AI system from undermining human autonomy or having other negative impacts, **human oversight** is needed. These governance mechanisms can all be used to achieve oversight. The following governance mechanisms can all be used to achieve oversight.

- Human-in-the-loop approach: allows for human intervention during each step of the system's decision-making.
- Human-on-the-loop approach: allows for human intervention during the system's design cycle and monitoring of the system's operation.
- Human-in-command approach: allows for human control over the AI system's overall activity and enables the decision of when and how to use the system in a given situation.

Additionally, it must be made sure that public enforcers can exercise oversight in accordance with their role. Depending on the application area and potential risk of the AI system, oversight methods may be needed to varied degrees to assist other safety and control measures. The less control a human can have over an AI system, the more rigorous testing and stricter regulations are necessary [41]. To ensure human oversight, companies have to consider the appropriate level of human control for the particular AI system and use case. They have to identify the "human in control", the circumstances or instruments for its intervention, and the degree of its control. They have also to establish systems to identify potential problems and implement any necessary step to facilitate auditing and fixing problems and also to make sure there is a stop procedure to safely stop an operation totally or partially when necessary, and give a human the next steps control [41].

2.3 Technical Robustness and Safety

Technical robustness, which is closely related to the idea of preventing harm, is an essential element for achieving Trustworthy AI. Technical robustness necessitates the development of AI systems with a risk-prevention mindset, in a way that ensures they consistently act as intended while minimizing unintended and unanticipated harm and preventing unacceptable harm [41, 52]. Humans' mental and physical integrity must always be protected and this needs to hold true also when adversarial agents try to modify the operational environment of the system. The first step in minimizing these AI risks is to create rules for risk controls with clearly defined goals, execution processes, metrics, and performance measures [51, 53].

As first, the system must be **secure** and **resilient to attack**. Assault may attack the model (model leaking), the data (data poisoning), or the supporting infrastructure (hardware and software). For AI systems to be considered secure, precautions must be taken to prevent and minimize any potential malicious actor exploitation of the system as well as any unintentional applications of the AI system. In order to protect the integrity and resilience of the AI system against potential attacks, businesses must take into account the many types and natures of vulnerabilities, such as data pollution, physical infrastructure, and cyber-attacks [41, 42].

In case of issues, consequences must be minimized, for this reason AI systems need safeguards that allow for a **fallback strategy**. This could imply that AI systems convert from a statistical to a rule-based process or that they pause to request a human operator. The system's ability to carry out its intended function without damaging the environment must be ensured. Errors and unwanted repercussions are minimized as part of this. Companies must determine whether there is a likely risk that the AI system will hurt users or other parties (including the environment or animals), and if so, what the likelihood, possible harm, impacted audience, and severity are [53]. They must also make plans for reducing or managing these risks. The extent of the risk that an AI system poses determines the type of safety precautions that are necessary. It is essential for safety measures to set thresholds and put governance procedures in place to activate fallback plans. It is also required to test these plans proactively if it can be predicted that the development process or the system itself will offer particularly high risks [41, 53].

As next, but not less important, it is crucial to determine what damage will result if the AI system predicts something incorrectly [53]. Additionally, it's crucial that the system is able to quantify how common these errors are when occasionally erroneous predictions cannot be prevented. The ability of an AI system to make accurate decisions, such as correctly classifying information into the appropriate categories, or to make accurate predictions, suggestions, or conclusions based on data or models, is referred to as **accuracy** [41]. Unintended risks from incorrect

predictions can be supported, mitigated, and corrected by a clear and well-formed development and evaluation process. When an AI system directly affects human lives, accuracy is essential, so it's essential to make sure the data being utilized is complete and updated, and if extra data is needed to boost accuracy or remove bias [53, 41].

In the end, **reproducible and reliable** outcomes from AI systems are essential to have a robust and safe system. A reliable AI system is one that performs well across a variety of inputs and contexts. This is necessary to examine an AI system and safeguard against unforeseen consequences. If an AI experiment is reproducible and displays the same behavior under the same circumstances, it is referred to as reproducibility. This makes it possible for researchers and decision-makers to precisely characterize what AI systems do. For the purposes of testing and verifying the reliability of AI, it is crucial to establish procedures in place that explicitly document and operationalize when an AI system fails and in what particular types of scenarios [41].

2.4 Privacy and Data Governance

Closely related to the idea of harm prevention is privacy, a fundamental human right that is significantly impacted by AI systems. An appropriate data governance is necessary to prevent privacy harms. It must cover data's quality and integrity, relevance to the application domain in which AI systems will be deployed, access methods, and ability to handle data in a way that respects privacy [43].

Privacy and data protection must be guaranteed by the AI systems throughout the whole lifecycle of the system. This covers both the data the user initially submitted and the data collected about him as a result of its interactions (e.g. the outputs generated by the AI system for that specific user). If the dataset contains personal data, one of the first steps is to determine what kind and how much of it there is. This is because recordings of people's behavior may enable AI systems to deduce not only people's preferences but also, for example, their gender, sexual orientation, age or religion [41, 42]. It must be verified that information about individuals won't be utilized for unfair or illegal discrimination in order for people to trust the data collection process. For this reason, it is crucial to design the AI system or train the model without using or using very little potentially sensitive or personal data. Additionally, it is crucial to implement privacy-enhancing methods like encryption, anonymization, and aggregation [42, 50, 41] and, if it is possible, to incorporate a Data Privacy Officer (DPO) during the early stages of the system's development [50].

Equally important is the **quality** of the dataset used, because it is essential for performance. Because data may contain errors, inaccuracies and social biases, it

is crucial to establish oversight systems for data acquisition, storage, processing, and use. Additionally, data's integrity must be guaranteed because feeding a malicious dataset to an AI system could modify its behavior. At each stage, including planning, training, testing, and deployment, processes and datasets must be tested and documented [50, 51]. This should also apply to AI systems that were purchased from outside sources rather than developed internally, evaluating the level of quality of the external data sources used. A constant check for hacking or compromise of datasets is also important [50, 41].

In the end, to guarantee data protection, protocols governing **access to data** should be implemented in every organization that manages the data of individuals. These protocols ought to specify who has access to data and under what conditions. This is because individuals' data should only be accessible by properly qualified people who have the know-how and necessity to do so [50]. Additionally, it is crucial to have a monitoring system to keep track of who accessed data, when, where, how, and for what reason [41].

2.5 Transparency

This criteria covers the transparency of the data, the transparency of the system and the transparency of the business models, which are the crucial elements of an AI system [41]. The transparency criteria is strongly related to the notion of explainability, because it is crucial that the organization's use of AI is transparent to the stakeholders, giving them access to information about how an AI system processes their data and arrives to certain conclusions [51].

To allow for **traceability** and a rise in transparency, the datasets and the processes that result in the AI system's decision, including the algorithms utilized and the processes for data collection and data labeling, should be **documented** to the highest standard [51]. This also holds true for decisions made by the AI system, including those that are the results of the algorithm, as well as potential alternative decisions that might result from other scenarios, such as those for different user subgroups. This makes it possible to determine the reasons why an AI decision was incorrect, which in turn may assist avoid errors in the future. Auditability and explainability, that will be analyzed in the next sections, are made easier through traceability [54, 41].

Another key factor for transparency is **communication**. Users of AI systems should not be led to think that they are dealing with human beings. They have the right to be informed that they are dealing with an AI system (e.g., through a disclaimer) [42, 54]. This requires that AI systems be immediately recognizable as such. To ensure respect for fundamental rights, it should also be possible, in appropriate circumstances, to not select this option in favor of human interaction.

In addition to this, it is important to inform AI operators or end users of the capabilities and limitations of the AI system in a manner appropriate to the use case. This could include communicating the accuracy and limitations of the AI system, such as potential or perceived risks, such as biases [54]. Equally important is to explain the function of the AI system and who or what will profit from the service or product, clearly specifying the product use scenarios [41]. The user profile should always be taken into consideration in explanations, employing the level of transparency required depending on its profile. According to the required level, employ a solution that provides local or global explanations, so that the user can request an explanation for the conclusion generated by Al [42]. Even when employing AI tools from third parties, this still holds true [42]. It has to be considered whether it is possible to comprehend how the algorithm arrived at its results, including what features and to what extent they impacted the algorithm's decisions. Obviously, this requires the inclusion of some functionality in the design phase [42].

As mentioned above, **explainability** is closely related to transparency criteria, as it is necessary for stakeholders to have access to the reasons behind the results of the AI system. Due to its importance, in the following the concept of explainability is described in detail, analyzing also three key explainability techniques that will come useful in the next chapter.

2.5.1 Explainability

Explainability refers to the capacity to explain both the technical processes of an AI system and the associated human choices, such as a system's application areas. The ability of humans to comprehend and trace an AI system's decisions is a requirement for technical explainability [51]. Unfortunately, trade-offs may be necessary between improving a system's explainability, which could decrease its accuracy, and boosting accuracy at the expense of explainability [42]. But when an AI system makes decisions that have a substantial impact on people's lives, it should be possible to ask for an adequate explanation of the AI system's decision-making process and determine the degree to which the decisions produced by the AI system can be understood, even at the cost of accuracy [41, 51]. Additionally, the rationale for the AI system's design decisions, the extent to which it influences and shapes the organization's decision-making process, and the reasons for its use should be made available, enabling an assessment of why this particular system will be used in this particular context [51]. To ensure interpatibility, it is necessary to take this goal into account from the earliest stages of AI system development, seeking to use the simplest and most interpretable model possible for the application in question, or considering whether it is possible to access the model's internal workflow or whether it is possible to check interpretability after

the model has been trained and developed [41].

In the context of Machine Learning, there are several proposals for XAI. Supervised ML models can be classified into two groups: white box models and black box models. On the one hand, white box models, such as simple decision trees or linear regressions, are models that can generate comprehensive explanations based on the model itself. On the other hand, black box models, such as complex deep learning (DL) architectures, can't provide direct explanations for the decision taken by the system in a way that is comprehensive for a human being [42]. Only the stimulus/response behavior can be accounted for when analyzing "black box" systems in order to deduce the (unknown) box's behavior [55]. The main issue is that there is a tradeoff between complexity and explainability: more complex models can potentially be more accurate, but in exchange the model is opaquer [38, 42]. To be able to use more complex models while being able to generate explanations that can be understood, there are different proposals available depending on the explanations desired [42].

Researchers have developed different algorithms to explain AI systems, that can be classified in two broad categories of explanations: self-interpretable models and post-hoc explanations. Self-interpretable models are the white box models described before. As said they can be directly read and interpreted by a human. In this case the model itself is the explanation. Post-hoc explanations are explanations, often generated by other software tools, that describe, explain, or model the algorithm to give an idea of how the algorithm works. Post-hoc explanations often can be used on algorithms without any inner knowledge of how the algorithm works, provided that it can be queried for outputs on chosen inputs [38]. More over, post-hoc explanations are grouped into two kinds: local explanations and global explanations. A local explanation explains a subset of decisions or is a per-decision explanation. A global explanation produces a model that approximates the non-interpretable model. In some cases, a global explanation can also provide local explanations by simulating them on specific inputs to provide local explanations for those individual inputs [38].

The most common type of local explanation are LIME (Local Interpretable Model-Agnostic Explainer), SHAP (SHapley Additive ex- Planations) and counterfactuals. LIME takes a decision, and by querying nearby points, builds an interpretable model (by default it is a logistic regression model) that represents the local decision, and then uses that model to perform feature explanations. SHAP, instead, provides a per-feature importance for an input on a regression problem by converting the scenario to a coalitional game from game theory and then producing the Shapley values from that game. In the end, counterfactual explanations queries the model to understand if a change in the input, corresponds to a change in the output [38].

Global explanations instead produce post-hoc explanations on the whole algorithm. Often, this involves producing a global model for an algorithm or a system [38].

An important challenge regarding XAI is that there is still not an available formalism to define a common reference for what an explanation should be. There are some criteria to consider while evaluating an explanation, but that is still not enough to set a general reference to build them [42].

2.6 Diversity, Non-Discrimination and Fairness

Inclusion and diversity must be promoted across the whole life cycle of an AI system if we want to develop trustworthy AI. This involves not only taking into account and involving all stakeholders in the process, but also guaranteeing equal access and treatment through inclusive design procedures [41]. Systems should be **user-centered**, **accessible** to anyone who wants to use AI products or services, regardless of gender, age, race, or other characteristics. It is crucial that people with disabilities, who are present in all societal groups, can access this technology [41]. The influence of the AI system on the entire prospective user audience, taking into consideration also those that might be tangentially impacted, must be carefully considered during the design, development, and deployment phases. It is also necessary to determine whether any individuals or groups may be disproportionately impacted by negative effects [53].

In addition, AI systems shouldn't take a one-size-fits-all approach, instead, they should take into account **Universal Design principles** to accommodate to the broadest range of users while adhering to the necessary accessibility guidelines. To apply these principles, the diversity within the team, the training data and the level of cultural sensitivity must be promoted when designing algorithms. The aim of "diversity-in-design" mechanism is to address problems caused by cultural biases and prejudices [53, 41]. This will make it possible for everyone to have equal access to and participation in current and future computer-mediated activities, especially assistive technologies.

Another important point to design trustworthy AI systems, is to **consult stakeholders** who may be impacted by the system during its life cycle. It is advantageous to get regular input even after deployment and establish longer-term methods for stakeholder participation, for instance by making sure that employees are informed, consulted, and involved throughout the entire process of adopting AI systems at organizations [41].

All these principles are directly related with the **fairness principle**. Datasets used by AI systems may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could

lead to unintended direct or indirect prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Identifiable and discriminatory bias should be removed in the collection phase where possible [41].

The way in which AI systems are developed may also suffer from unfair bias. This could be counteracted by putting in place oversight processes to analyse and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner [41]. It is necessary to ensure an adequate working definition of "fairness" and apply it in designing AI systems, using quantitative analysis or metrics to measure and test the applied definition of fairness. It is important to consider diversity and representativeness of users in the data, testing for specific subgroups of the population or problematic use cases, and to use available technical tools to improve the understanding of data, model and performance [42]. Fair AI seeks to ensure that the applications of AI technology lead to fair results. This means that they should not lead to discriminatory impacts on people [56] in relation to features considered sensitive, like race, ethnic origin, religion, gender, sexual orientation, disability or any other personal condition. When optimizing a machine learning algorithm, we must take into account not only the performance in terms of error optimization, but also the impact of the algorithm in the specific domain [42, 57].

Fairness assessment is strictly related with the aim of this work, so, in the following, its the metrics for its evaluation and the fairness mitigation methods are described.

2.6.1 Fairness

Since 2010, academics and business have paid a lot more attention to fairness in AI. Due in part to the fact that fairness is a sociological and ethical idea, researchers have struggled for decades to provide a single definition of it. Fairness is a difficult ideal to attain in practice because it is mostly a matter of subjectivity and fluctuates with social environment and time [53]. Because this thesis work addresses how to make decisions that are consistent with social ideals, it is adopted the concept of fairness in the context of decision-making.

> "Fairness is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" [57].

Evaluation and mitigation are the two key steps that may be taken to ensure fairness in machine learning. In the first step, the degree of bias in the model is measured and quantified (in terms of one or more criteria), and in the second step,

the model's flaws are fixed in order to decrease or eliminate the impact of bias on one or more sensitive features [42].

Considering the evaluation step, three different criteria can be employed to evaluate the fairness of a supervised ML model [42, 53]. They are described in the following.

Independence criteria (Demographic parity)

Independence criteria is sometimes referred to as demographic parity. It is achieved when the model prediction is independent of the sensitive variable, i.e., the proportion of positive samples given by the model is the same for all sensitive groups [42, 58]. Demographic parity is mathematically defined using the following set of equations. A classifier satisfies demographic parity under a distribution over (X, A, Y) if its prediction R is statistically independent of the sensitive feature A. This is equivalent to

$$\mathbb{P}[R = r|A = a] = \mathbb{P}[R = r|A = b] \quad \forall a, b \in A, \forall r \in R \tag{2.1}$$

It is important to notice that allocation harms could arise if demographic parity is not achieved. They occur when AI systems distribute opportunities, resources, or knowledge differently throughout various groups [58]. Due to the underlying presumption that resources should be distributed fairly among groups, demographic parity can be used to measure the level of allocation harms. However, utilizing demographic parity to measure fairness relies on a few assumptions, including the notions that either the dataset is an inaccurate reflection of reality or that the phenomenon being modeled is unfair despite the accuracy of the dataset [58]. These assumptions could be not true in reality. The phenomenon being modeled may not be unfair, or the dataset may accurately reflect the phenomenon. Demographic parity may not offer a relevant or useful evaluation of the fairness of a model's predictions if either assumption is false [59, 58].

Separation criteria (Equalized Odds)

[59]

The separation criteria, also known as equalized odds, is achieved when the model prediction is independent of the sensitive variable given the target variable, that is, when the TPR (true positive rate) and the FPR (false positive rate) are equal in all sensitive groups, respectively [42].

A classifier satisfies equalized odds under a distribution over (X, A, Y) if its prediction R is conditionally independent of the sensitive feature A given the label Y, and that this is equivalent to

$$\mathbb{P}[R = r|Y = y, A = a] = \mathbb{P}[R = r|Y = y, A = b] \quad y \in Y, \forall a, b \in A, \forall r \in R \tag{2.2}$$

Equalized odds requires that the TPR and the FPR be equal across groups [58, 59] (and therefore the false negative rate (FNR) and the true negative rate (TNR) are equal) for every value of the sensitive characteristics [60], i.e.

$$\mathbb{P}[R = 1|Y = 1, A = a] = \mathbb{P}[R = 1|Y = 1, A = b] \quad \forall a, b \in A \tag{2.3}$$

$$\mathbb{P}[R = 1|Y = 0, A = a] = \mathbb{P}[R = 1|Y = 0, A = b] \quad \forall a, b \in A \tag{2.4}$$

FPR's inclusion acknowledges that the costs of misclassification vary for various groups [58]. False positive predictions, for instance, might highlight existing discrepancies in outcomes between minority and majority groups when a model predicts a negative consequence that already disproportionately affects people from minority populations. By penalizing models that only outperform on majority groups, equalized odds further ensures that accuracy is high across all groups [59].

The separation criteria is stricter than the independence one because it requires also that different sensitive groups have the same TPR and and FPR. This restriction is important because a model could respect the independence criteria (i.e., its predictions are independent from sensitive features), but still discriminate, classifying instances of one sensitive group more as false positive than other groups. [58, 59]. Additionally, while independence criteria assesses the allocation of resources generally, the focus of separation criteria, as shown by the positive target variable $Y = 1$, is on the distribution of resources that were actually distributed to members of that group. Separation criteria, on the other hand, rely on the target variable Y being a reliable indicator of the phenomenon being modeled, although this may not always be the case [58].

It can be considered also a relaxed version of equalized odds, the **equal opportunity**, that only considers conditional expectations with respect to positive labels, i.e., $Y = 1$. This because, in the binary case, often the outcome $Y = 1$ is the "advantaged" outcome. This metric requires equal outcomes only within the subset of records belonging to the "advantaged" class [59]. However, equal opportunity does not account for the costs of misclassification differences because it does not take into account whether FPRs are equal across groups [58].

Sufficiency criteria

The last criteria is the sufficiency one, also known as Predictive Rate Parity. It is achieved when the target variable is independent of the sensitive attribute given the model output, i.e., when the Positive Predictive Value is the same in all sensitive groups [42].

A classifier satisfies sufficency criteria under a distribution over (X, A, Y) if its target variable Y is conditionally independent of the sensitive feature A given the model output R. This is equivalent to

$$\mathbb{P}[Y = y|R = r, A = a] = \mathbb{P}[Y = y|R = r, A = b] \quad \forall y \in Y, \forall a, b \in A, r \in R \tag{2.5}$$

If this criteria is satisfied, it is the confirmation that the sensitive attributes are not needed at all for the training of the model [53].

Let's consider now the mitigation steps. In the literature, several techniques can be found and can be categorized into pre-processing, in-processing and post-processing techniques [53, 42].

- **Pre-processing**: these techniques are used to eliminate biases at the very beginning of the learning process, before the ML algorithm is trained. Utilizing pre-processing techniques requires, among other things, allowing the algorithm to alter the training data. The data can then be transformed to eliminate the bias [42].
- **In-processing**: these methods remove bias by changing the algorithms during the training phase. One techinque is to use fairness measurements as constraints or to incorporate them into the objective function [42, 58].
- **Post-processing**: these are the less intrusive methods because they don't alter the input data or the ML algorithm; they are used after the algorithm is created. This method works particularly well for reducing biases in pre-existing models or in situations when neither the training data nor the model can be changed. Using a specified function, post-processing techniques reassign the predicted labels [42].

Pre-processing or in-processing solutions are preferable in terms of performance since they apply the mitigation procedure in different phases of a typical analytics pipeline. The choice must be made to meet the specific case [42]. Analysis of these techniques is beyond the scope of this thesis work, but it was important to mention them in order to be aware of their existence.

In conclusion, the importance of fair AI has increased over the past few years. Making fair models has attracted a lot of research and the development of new solutions [53]. A unified framework for fairness in AI is required to simplify the process of adoption and implementation [42]. Although the evaluation criteria seen are commonly applied, they can not be applied to any given situation, because techniques as independence or separation measures specific fairness aspects [58]. Every day, it becomes increasingly clear that we need a single metric for this purpose. The evaluation process would be simpler to implement with the adoption of an uniform fairness technique [42]. However, it is challenging to formulate generalized definitions of fairness quantification [53].

2.7 Societal and Environmental Wellbeing

In accordance with harm prevention and fairness principles, the wider community, other sentient creatures and the environment should all be taken into account

as stakeholders throughout the life cycle of the AI system [41]. Research on AI solutions addressing global concerns, such as the Sustainable Development Goals (SDGs), should be supported. Sustainability and ecological responsibility of AI systems should also be encouraged. AI systems should ideally be employed for the benefit of all people, including future generations [53].

AI systems have the potential to assist in addressing some of the most serious social issues, but it is important to ensure that this happens in the most **ecologically friendly** and **sustainable** manner possible, minimizing the impact of the AI system's life cycle. The process of creating, deploying, and using the system, as well as its complete supply chain, should be evaluated in this regard, for instance by a critical analysis of the energy and resource consumption during training, choosing less damaging options. Support should be given to actions ensuring that the entire supply chain for AI systems is environmentally friendly [41].

In addition, the **social impact** must be taken in consideration to build human-centric AI systems. Our sense of social agency may change as a result of constant exposure to social AI systems in many spheres of our lives (including education, employment, care, and entertainment), which may also have an effect on our social connections and engagement. While AI systems can be used to improve social abilities, they can also cause those to decline [53, 42]. The physical and mental health of individuals may be impacted by this. Therefore, it is important to carefully monitor and take into account these systems' consequences. In the event that an AI system directly communicates with people, care must be taken to ensure that the AI system makes it plain that such communication is simulated and that it lacks "feeling" and "understanding" abilities. The societal effects of the AI system must also be taken into consideration (e.g. assess whether there is a risk of job loss or de-skilling of the workforce). Both companies and governments must determine the actions they will take to mitigate or restrict these risks [41, 42, 51].

Beyond evaluating how an AI system's creation, implementation, and use would affect specific people, this influence should also be evaluated from a societal standpoint, taking into account **how it will affect institutions, democracy, and society as a whole** [42, 41]. In scenarios related to the democratic process, such as political decision-making and election contexts, the employment of AI systems should be carefully considered [41].

2.8 Accountability

The need of accountability completes the set of requirements mentioned in the previous sections. Accountability relates to the extent to which humans can monitor and modify the algorithms as well as who is held responsible and culpable if issues happens [54]. Processes must be put in place in order to ensure AI systems'

responsibility and accountability: who is accountable for making sure the system is adequately tested before it is launched, who is responsible for fixing problems and who is liable for paying for consequences are crucial questions of responsibility and accountability [54].

According to a survey conducted in 2018 on StackOverflow [61] with responses from over 60000 developers, 48% of participants believed that developers who build AI systems should be responsible and take into account the system's potential consequences. Unfortunately, for any of the stated issues, there is no clear standard answer. More legislation and policy guidance are still being drafted, and the interpretation of already-existing legal and regulatory frameworks is evolving [54]. In this work, the guidelines of European commission "Ethics guidelines for trustworthy AI" are used as base to describe how accountability should be faced. According to them, auditability and risks minimization are the key point for the creation of an accountable system [41].

Enabling **auditability** means enabling the evaluation of algorithms, data, and design processes, providing traceability and logging of the operations and results of the AI systems [41]. This doesn't imply that details of intellectual property pertaining to the AI system must necessarily be made public. The evaluation of the technology by internal and external auditors, as well as the availability of such evaluation reports, can help establish its trustworthiness. In addition, AI systems should be able to be independently audited if we consider applications that affect fundamental rights, including those that are critical to security [41].

Report actions or choices that lead to a specific system outcome is not sufficinet. It is necessary to guarantee the capability of responding to the consequences of such an outcome [41, 54]. For people who may be directly or indirectly impacted, it is extremely important to **identify, evaluate, document, and minimize any potential negative effect** of AI systems. And, to minimize undesirable effects, it is important to report everything before and throughout the development, deployment, and use of AI systems [41, 54]. These evaluations must be in line with the danger that the AI systems could provide. In addition, when raising real concerns about an AI system, trade unions, whistleblowers, NGOs, or other entities must have access to the appropriate protection [41]. It is crucial to support the growth of accountability practices both within and outside of companies by offering education and training. As first, it is important to identify the employees and the branches of the team involved along the entire pipeline, teaching them the potential legal framework applicable to the AI system and establishing an "ethical AI review board" or a similar mechanism to discuss overall accountability and ethics practices [41]. Along with internal activities, think also about considering an external guidance or to bring in experts who are familiar with the work done by moral philosophers, behaviorists, sociologists, and other experts, and who can comprehend ethical questions in a holistic context and take the whole system

into account. [54]. This is a strategy found to be effective to get developers to care about software security without overwhelming them [54]. Another important point is to establish procedures for workers, distributors, consumers, and other third parties to identify potential weaknesses, risks, or biases in the AI system [41].

It has to be considered that, during the implementation, tensions between the aforementioned requirements may result in unavoidable **trade-offs**. Such trade-offs should be handled rationally and methodically [41]. This requires that the AI system's relevant interests and values have to be identified, and that, if a trade-off is necessary, it has to be acknowledged and assessed for its potential to undermine ethical principles, such as fundamental rights [41]. The creation, implementation, and usage of an AI system shouldn't go forward under circumstances where there are no discernible ethically acceptable trade-offs (e.g. in the 2020 several large tech companies decided to no longer sell facial-recognition software to law enforcement [54]). Any choice of trade-off should be adequately documented and supported by reasoning and the decision-maker must be held accountable for the way in which the trade-off is made. Additionally, the suitability of the resultant decision should be reviewed regularly to guarantee that required adjustments to the system can be made [41].

In the end, Accessible measures that guarantee adequate **redress** should be planned for when an unjust adverse impact arises. The key to ensuring trust is to put in place mechanisms to tell users and third parties about options for redress and particular focus should be given to vulnerable groups or individuals [41].

Even though legal responsibility may not always be clear, responsible companies should define who is in charge of ethical matters so that the accountability principle could be applied [54]. Of course, relying on corporate self-regulation, which frequently lacks legal binding and a clear mechanism for pursuing damages, is not the best course of action. While requests are made for government agencies to implement new regulations or adapt existing ones (such as safety standards for self-driving cars), academics and journalists may serve as guardian to spot and expose issues, such as algorithmic unfairness or biases [54].

2.9 Responsible AI recap

In the previous sections, the seven key requirements for Trustworthy AI were analyzed: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; environmental and societal well-being; accountability. The aim is to try to ensure that the AI system's respects the seven principles during the entire life cycle. Be aware that there may be fundamental conflicts between various requirements and

principles. It is important that these trade-offs and their solutions are recognized, assessed, recorded and shared, to increase the community knowledge [41].

Summing up, the path to follow is to encourage research and innovation to assist in evaluating AI systems and to help meet the standards. Companies and researchers have to share findings and raise issues with the general public and to systematically develop a new generation of AI ethics experts. It is crucial to inform stakeholders of the capabilities and constraints of the AI system, allowing them to have reasonable expectations, as well as how the requirements are implemented. This has to be done in a transparent and proactive manner, making it clear that they are interacting with an AI system. In the end, AI systems must be traceable and auditable, especially in critical contexts and circumstances, and stakeholders must be included at every stage of the AI system's development, encouraging education and training so that all interested parties are aware of and knowledgeable about trustworthy AI [41, 42, 53, 51].

As it was seen in this chapter, to ensure the implementation of those principles they have to be considered both technical and non-technical methods. In last years a lot of companies published guidelines for trustworthy AI and developed tools for error analysis, explainability of models and fairness assessment. Some examples are IBM [62], Google [63, 64] and Microsoft [65]. For this thesis work, we analyzed and used the RAIToolbox [66], a set of tools developed and made available free of charge by Microsoft. In the next section all its component are described.

2.10 RAI Toolbox

The Responsible AI Toolbox is a set of tools that offers a selection of user interfaces and libraries for model and data exploration, assessment, and learning. These interfaces and libraries enable those who are involved in the development and monitoring of AI systems to do so more responsibly and to make better data-driven decisions. Four visualization widgets are included in the Responsible-AI-Toolbox for model evaluation and decision-making:

- Responsible AI dashboard combines a number of mature Responsible AI tools from the toolbox for a Responsible AI evaluation, model debugging, and decision-making process. With the help of this dashboard, it is possible to spot model mistakes, figure out why they're happening, and take steps to prevent them.

- Error Analysis dashboard, for identifying model errors and discovering cohorts of data for which the model underperforms.

- Interpretability dashboard, for understanding model predictions [67].

- Fairness dashboard, for understanding model's fairness issues using fairness metrics across sensitive features [58].

The Responsible AI dashboard makes it simple to move between different phases of model decision-making and debugging. This personalized experience can be used to analyze the model or data holistically, go in-depth or compare cohorts of interest, explain and alter model predictions for specific occurrences, and enlighten users about business decisions and actions, among other things. The dashboard combines concepts and technology from various open-source toolkits in the following domains to achieve these capabilities:

- Error Analysis [68], which discovers data cohorts with higher error rates than the benchmark average. When the system or model underperforms for particular demographic groups or infrequently observed input conditions in the training data, these disparities may appear.
- Fairness Assessment powered by Fairlearn [58], which identifies which groups of people may be disproportionately negatively impacted by an AI system and in what ways.
- Model Interpretability powered by InterpretML [67], which explains blackbox models, helping users understand their model's global behavior, or the reasons behind individual predictions.

Error Analysis

To find cohorts with high error rates compared to the benchmark and see how the error rate is spread, utilize the Error Analysis dashboard. Visually exploring more deeply the properties of the data and models will help to identify the underlying reasons of the problems (via its embedded interpretability capabilities). For instance, Error Analysis can be utilized to find out that the model has a larger error rate for a certain cohort than the general population (for example, women with income under $50,000). Individual records from that cohort can be analyzed, to understand their feature importance values, and diagnose the contributing error factors by understanding the most significant factors responsible for this subset's inaccurate predictions [66].

Error Analysis aim is to give a clearer picture of the behaviors of the machine learning models. It can be undertaken a wide range of evaluation activities to create responsible machine learning, combining Error Analysis with Fairlearn and Interpret-Community. For improved debugging, Error Analysis can be combined with InterpretML [66].

Once the visualization dashboard is loaded, different aspects of the dataset and of the trained models can be investigated via two stages: identification and diagnosis [66].

cohorts of data with higher error rate than the overall benchmark may appear if the system or model underperforms in the training data for particular demographic groups or input situations that are infrequently observed. The way this is performed is through a Decision Tree that uses the binary tree visualization to identify cohorts with highest error rates across various variables. For each detected cohort, it can be looked into metrics like error rate, error coverage, and data representation.

Error Analysis helps debug and further explore cohorts after finding those with greater error rates. Through data exploration and model explanation, it is possible to learn more about the model or the data. The possible techniques for Error Diagnosis are:

- Data exploration, which examines feature distributions and dataset statistics. Cohort statistics can be compared to those of other cohorts or to benchmark data. Examine whether some cohorts are underrepresented or whether the distribution of their features deviates materially from the overall data.

- Global Explanation, which investigates the top K attributes that have the greatest influence on the global model explanation for a certain cohort of data. Recognize how feature values affect model prediction. Explanations might be compared to those of other cohorts or benchmarks.

- In the instance view, local explanation makes it possible to see the unprocessed data. Recognize whether each data point's forecast was accurate or inaccurate. Look for any potential problems, such as label noise or missing features. Investigate the individual conditional expectation (ICE) plots and local feature importance values (local explanation).

The algorithm used for the Error Analysis Decision Tree building is the one showed in Alg.1 [69]. It returns a list of leaves from decision tree T with error rate of at least BER + δ and error coverage at least τ.

Fairness Dashboard

Fairlearn is a Python package that empowers developers of artificial intelligence (AI) systems to assess their system's fairness and mitigate any observed unfairness issues. Fairlearn contains metrics for model assessment, enabling assessment of unfairness under several common definitions [70]. The principle metrics used in Fairlearn are Demographic parity and Equalized odds, both described in the previous Fairness section.

Algorithm 1 Failure mode generation procedure by EA RAIT paper.

Input: features: F, model: h, image cluster: C, number of features: k, tree parameters: A, error rate threshold: δ, error coverage threshold: τ
Output: leaves with high error concentration: L

$L = \emptyset$
$BER = ER(C)$
$E(x) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases} \quad \forall (x, y) \in C$
$F^* = \emptyset$
while $|F^*| < k$ **do**
    $F^* = F^* \cup \text{argmax}_{f \in F \setminus F^*} IG(E; f)$
end while
$T = train\_decision\_tree(F^*, E, A)$
for $l \in T$ **do**
    if $(ER(C_l) > BER + \delta)$ and $(EC(C_l) > \tau)$ **then**
        $L = L \cup \{l\}$
    end if
end for

Model Interpretability

InterpretML allows to explore the dataset and the model performances, understanding how model performance changes for different subsets of data and exploring model errors. It allows also to analyze dataset statistics and distributions. It gain model understanding, exploiting global and local explanations and filtering data to observe global and local feature importance [71].

The supported explainers are LIME, Mimic (for global explanations), SHAP Kernel, SHAP Tree, SHAP Deep, SHAP Linear and Permutation Feature Importance (PFI). To receive explanations in terms of the raw features before the transformation, these explanations can also be carried out by sending the feature transformation pipeline to the explainer (rather than engineered features). The explainer offers explanations in terms of engineered characteristics if this step is skipped [72].

As said before, custom cohorts can be creates or the ones generated by the Error Analysis tool can be used. The created cohorts will be visible from all of the four tabs. Once the visualization dashboard is loaded, different aspects of the dataset and of the trained model can be investigated via four tab views, comparing performances among the cohorts: Model Performance, Data Explorer, Aggregate Feature Importance and Individual Feature Importance. The model performance enables to evaluate the model by observing its performance metrics and prediction probabilities/classes/values across different cohorts. The dataset explorer, instead, allows to explore the dataset statistics by selecting different filters along the X, Y, and color axes of this tab to slice the data into different dimensions. In the end feature importance explores the top K important features that impact the overall model predictions or the single cohorts [72].

Chapter 3

Experimental Results

3.1 COMPAS dataset

In this section the RAIToolbox will be tested on the well known COMPAS dataset, to understand its potentialities and to validate the correctness of the hypothesis it is a valuable tool for the building of Responsible AI. A classification model will be trained and analyzed using the RAIToolbox to see if the latter can understand whether and where the model is discriminative.

As first, the dataset and the related problems will be described. Next, the method used for the comparison will be presented and all the possible analysis allowed by the RAIToolbox will be made and their outcomes will be described.

3.1.1 Dataset Description and previous works

Introduction to the problem

It is well known that it is difficult to deal with COMPAS, which is a dataset used to predict the recidivism of a criminal over a two-year period. It has generated a lot of discussion due to the bias present within it and to the unfairness of the models trained on it. Many works have been carried out considering this dataset. ProPublica group Larson et al. work is one of the first and most known works carried out on COMPAS dataset [44]. Other relevant works are the one carried on by Bao et al., "It's COMPASlicated", where it is analyzed how much difficult is to use COMPAS datasets in making claims about real-world outcomes [73], and chapter 10 of XAI Stories, where they are performed model explanations in search of potential racial and gender biases present in models along with potential ways to guarantee models fairness [74].

The work that will be used to validate the outcomes of the RAIToolbox is the one of Prof.ssa Elena Baralis and Dott. Eliana Pastor, which analyzed COMPAS

dataset subgroups making use of the DivExplorer tool [75].

Dataset preprocessing

There are multiple versions of the COMPAS dataset. The one used for the purposes of this work is the version called "COMPAS-scores", by Pro Publica, downloadable here [76]. It was used as starting dataset and modified in compliance with the features used in the Professor Baralis work, to have a compatible and comparable study with the one carried out by her and Dott. Eliana Pastor [75].

The original version contains over 47 features. The first performed change, was to keep only the following features, to be compliant with the baseline: 'sex', 'age', 'age_cat', 'race', 'stay', 'stay_cat', 'prior_count', 'prior_count_cat', 'c_charge_degree' (it was renamed in 'charge'), 'is_recid' (the target). The second change was to create the categorical features used in the Professor work. It was created the numerical feature 'stay', computed as number of days between 'c_jail_out' and 'c_jail_in'. Next, it was discretized creating the feature 'stay_cat', having the following three possible values: '<week', '1w-3M' and '>3months' [75]. It was discretized also the feature 'prior_count', creating the feature 'prior_count_cat', having the three possible values '0', '[1-3]' and '>3'. There were explicitly leaved both categorical and numerical features concerning 'age', 'prior_count' and 'stay' to allow linear classifiers to be non-linear in this features. In Fig.3.1 it is shown the final version of the used dataset.

	sex	age	age_cat	race	priors_count	charge	is_recid	stay	stay_cat	priors_count_cat
360	Male	53	Greater than 45	Caucasian	1	M	0	1	<week	[1-3]
1781	Male	26	25 - 45	African-American	2	M	0	0	<week	[1-3]
4799	Female	30	25 - 45	African-American	10	F	1	0	<week	>3
866	Male	27	25 - 45	African-American	5	M	1	64	1w-3M	>3
251	Male	31	25 - 45	Caucasian	0	F	0	2	<week	0
...	...	...	...	...	...	...	...	...	...	...
6359	Male	39	25 - 45	Caucasian	0	F	0	1	<week	0
5759	Male	38	25 - 45	Caucasian	0	M	0	0	<week	0
5976	Male	21	Less than 25	African-American	1	F	1	41	1w-3M	[1-3]

Figure 3.1: COMPAS modified version dataset.

DivExplorer results

The work proposed by Prof.ssa Elena Baralis and Dott. Eliana Pastor, is DivExplorer, a novel approach for the complete exploration of subgroups with adequate representation in the dataset (i.e. having a support threshold s). In their work, a subgroup is characterized by an itemset which is a conjunction of attribute-value pairs, as it is in the error analysis tool of the RAIToolbox. They propose the notion

of divergence $\Delta_f(I)$ to estimate the different classification behaviour in data subgroups with respect to the overall behaviour. In addition they developed a method to understand what is the contribution of each feature to the divergence, using the notion of Shapley value (see the previous SHAP section for further details).

As said, they applied their work to the COMPAS dataset and in Fig.3.2 it is possible to see the top divergent patterns w.r.t. FPR and FNR, with a support threshold s = 0.175. The high number of priors and age lower than 25 are the terms that mostly contribute to the FPR divergence, while age greater than 45, misdemeanour charge degree and caucasian race are the terms that mostly contribute to the FNR divergence.

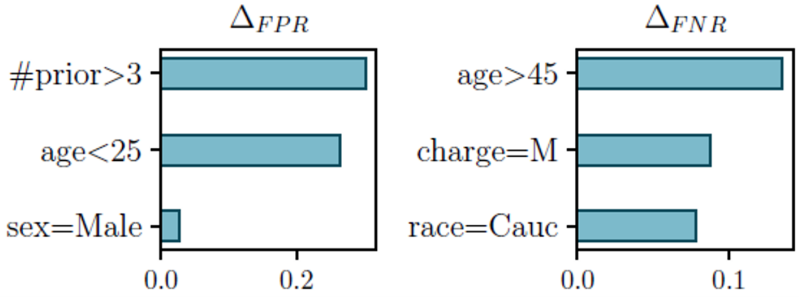


Figure 3.2: DivExplorer COMPAS results.

3.1.2 Results

We splitted the dataset in train and test sets using the proportion 80/20. The numerical features were standardized according to z-score, while for the categorial features we used one encoding. After the preprocessing steps, we trained a Support Vector Machine (SVM) Classifier, which is a black-box model. The model obtained an accuracy train score equal to 0.7040 and an accuracy test score equal to 0.6886 (error rate equal to 31.14%), seeming to not overfit. It is not a very good model (we trained also other black-box models, like random forest, but results were not different from the ones of the SVM classifier), so let's analyze it through RAIToolbox.

The first analysis we perform is made using the Error Analysis tool, which will allow us to identify critical cohorts and make analysis comparing them with the analysis made on the whole test set. We consider three features-value pairs, as in Baralis work. This is equivalent to consider the third level of the Error Analysis Tree Map. Analyzing the cohorts with the highest error rate (HER) we can understand what are the combinations of feature-value pairs where the model underperforms. Insterad, analyzing the highest error coverage (HEC) cohort, we

can understand what is the sub-group containing the higher number of misclassified instances. It is important to point out that to be the HEC cohort does not means to be a critical cohort for the model because to have an high Error Coverage does not mean to have an high Error Rate. This because EC is related also to the size of the cohort, while ER is not, i.e. the HEC cohort could be a cohort where the model does not underperform w.r.t. the whole test set. In Fig.3.3 it is possible to see the cohort with the HER, while in Fig.3.4 it is possible to see the one with the HEC.

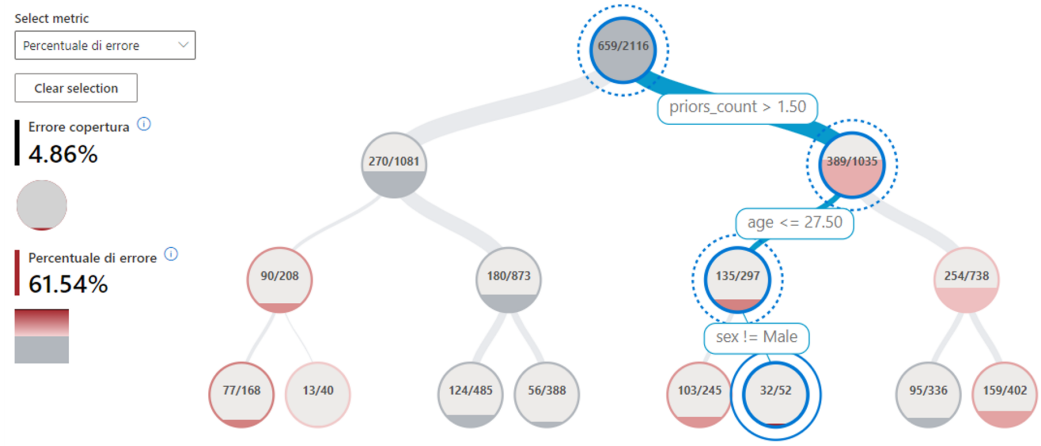


Figure 3.3: Highest Error Rate cohort for the COMPAS dataset.

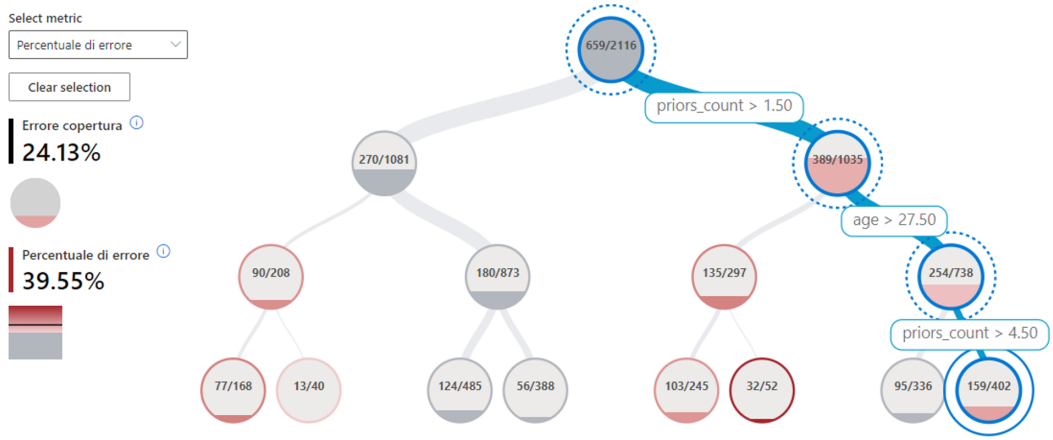


Figure 3.4: Highest Error Coverage cohort for the COMPAS dataset.

What we conclude from the Error Analysis tool applied to the COMPAS dataset, is that, when the model finds an instance belonging to the cohort with the highest

Error Rate (prior>1, age<28, sex=female), it will fail to classify it with a probability of 61.54%. W.r.t. the whole test set, the model performances inside this cohort are quite poor. But it can be observed that this result is aligned with the one obtained in the Professor Baralis work. Despite divergence can not be used inside RAIToolbox, it leads us to conclude that an high number of priors and a low age value are discriminative terms for the model. We confirm also that the sex feature gives a lower contribution for the cohort individuation, as in DivExplorer work (Fig.3.2) [75].

It could be also interesting to analyze the HEC cohort, to have a comparison and to see also how the model behaves inside this cohort. The HEC cohort (age>27, prior>4), as said, contains the highest number of misclassified instances. In this case HEC cohort contains 159 misclassified instances over a total of 659 (24.13%).

	ER	EC	y==1	$\hat{y}$==1	y - $\hat{y}$
Whole test set	31.14%	100%	36.11%	21.88%	**14.23%**
HER cohort	61.54%	4.86%	63.46%	25.00%	**38.46%**
HEC cohort	39.55%	24.13%	52.74%	54.98%	**-2.24%**

Table 3.1: Statistics and obtained result for the COMPAS dataset.

In Table 3.1 it can be seen how the model, generally, lacks in classifying instances as recidivists (y – $\hat{y}$ = 14.23%). Considering the cohort HER, this lack is evident: w.r.t. the whole test set, compared to an almost doubling of the percentage of true recidivists (63.46% vs 36.11%), the percentage of those classified as such grows very little (25.00% vs 21.88%), leading to a percentage of instances classified as recidivists very low w.r.t. the true recidivists percentage (y – $\hat{y}$ = 38.46%).

If we compare the performances over the two cohorts, with the ones over the whole test set, we can notice how the model performs more poorly over the two cohorts we analyzed, but it does not behave in the same way for both. In the cohort HEC it predicts as much recidivists as the true ones are (y – $\hat{y}$ = -2.24%), but despite this, the error rate of this cohort is higher then the whole test set one (39.55%). This discrepancy can be answered analyzing the graph in Fig.3.5, where we can see how the predictions are distributed between TP, FP, TN and FN. Despite the reduction of the FNR in the HEC cohort (from 0.628 of the whole test set, to 0.354), there is a not negligible increase of the FPR (from 0.132 to 0.442). In the cohort HER instead, we have an increase of both FPR (from 0.132 to 0.316) and FNR (from 0.628 to 0.788).

Maybe, an answer to these differences, can be founded analyzing the importance of the features-value pairs used to create these cohorts (Cohort HRC: priors_count

> 1 and sex == Female and age < 28; Cohort HER: priors_count > 4 and age > 27). In Fig.3.6 it can be seen how features importance changes considering the different cohorts.

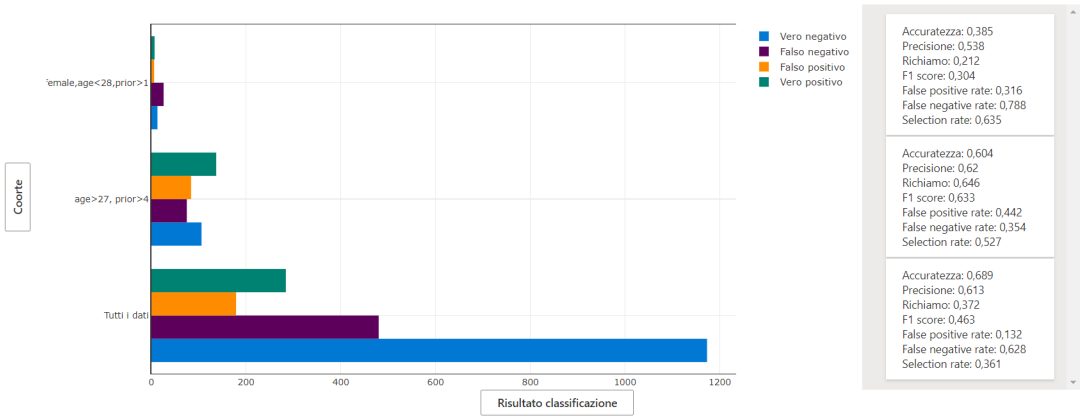


Figure 3.5: Comparison of the COMPAS number of TN, FN, TP and FP.

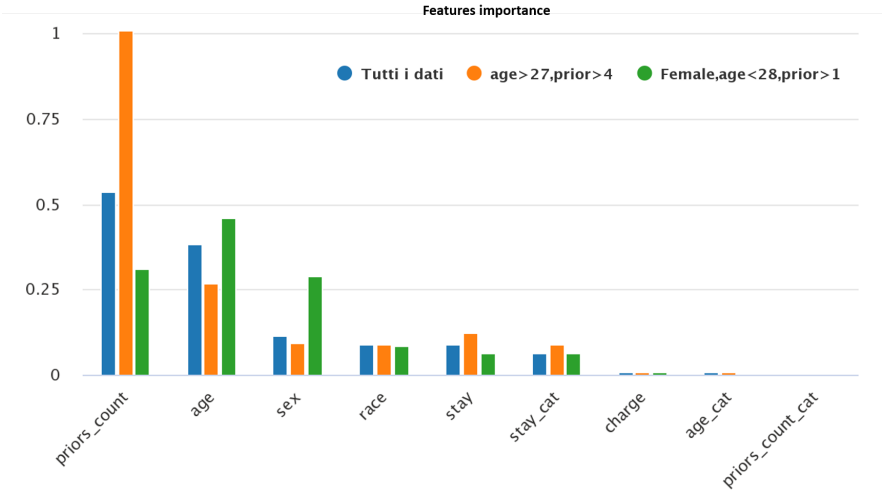


Figure 3.6: Global Feature Importance for the COMPAS dataset.

What we want to do is to analyze features intrinsically discriminative for a specific category, which are 'age', 'sex' and 'race'. Considering the cohort HER, the 'age' feature has an higher importance w.r.t. the whole test set, increasing from 0.38 to 0.46. More over, it becomes the most important feature for this cohort, due to the parallel decrease of the importance of the 'priors_count' feature. Also the 'sex' feature increases a lot its importance inside this cohort, passing from 0.11 to 0.29. We conclude that the model is biased over the 'age' and 'sex' features,

giving them a lot of importance inside the cohort with the highest error rate.

The cohort HEC, on the contrary, gives a lot of importance to the 'prior' feature (from 0.53 of the whole dataset to 1.01) and less importance to 'age' and 'sex' features (0.27 and 0.09 respectively). While the 'prior' feature is a reasonable discriminative feature (the more crimes you have committed the more likely you are to commit others), it is important to understand the reasons behind the high importance 'age' and 'sex' features have inside the model, because they could be very discriminative for the categories they represent. The 'race' feature, on the other side, have the same importance for all the cohorts, but it has too importance to be a sensitive feature.

From the two plot in Fig.3.8, it is visible how 'age' categories are equally distributed among 'sex' feature values and viceversa, but it can be noticed the underrepresentation of the Female sex and of the '>45' and '<25' age categories. From Fig.3.7 it is also evident how there is an underrepresentation of the 'asian', 'hispanic' and 'native american' race categories. It's important to understand how the model behaves over this sensitive features containing also underrepresented categories. So Let's analyze how the model behaves with them. To do this we will use the Fairness Assessment tool.

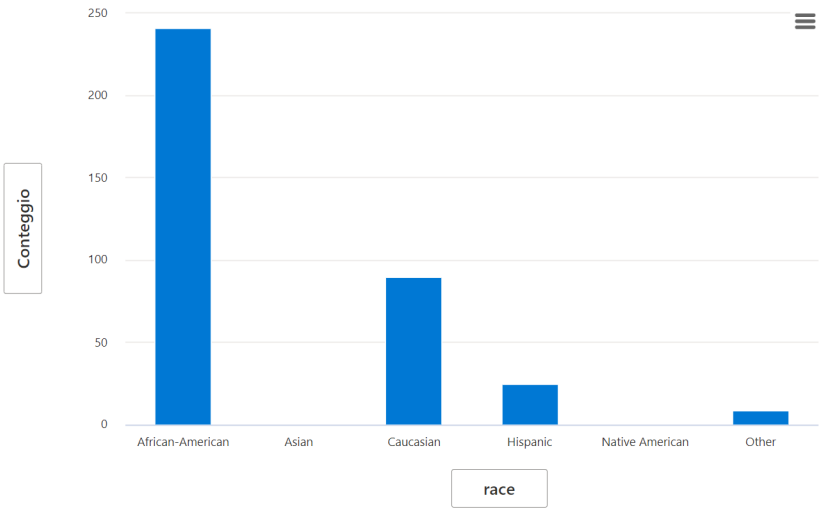


Figure 3.7: Count of the different race possible values for the COMPAS dataset.

Analyzing the 'age' feature fairness, from Fig.3.9a, it can be noticed how the model is biased over instances having age>45, wrongly classifying them as non recidivist the more w.r.t. the other two categories (FNR equal to 79% vs 61% and 59%). In addition, young people (age<25) are wrongly classified as recidivist the more then the other two categories, having a FPR equal to 23%, w.r.t. the 13% of the people having an age between 25 and 45 years and the 6.41% of people having

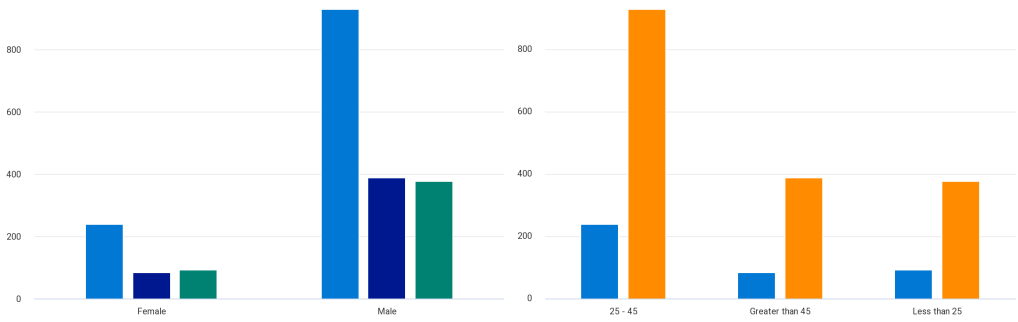


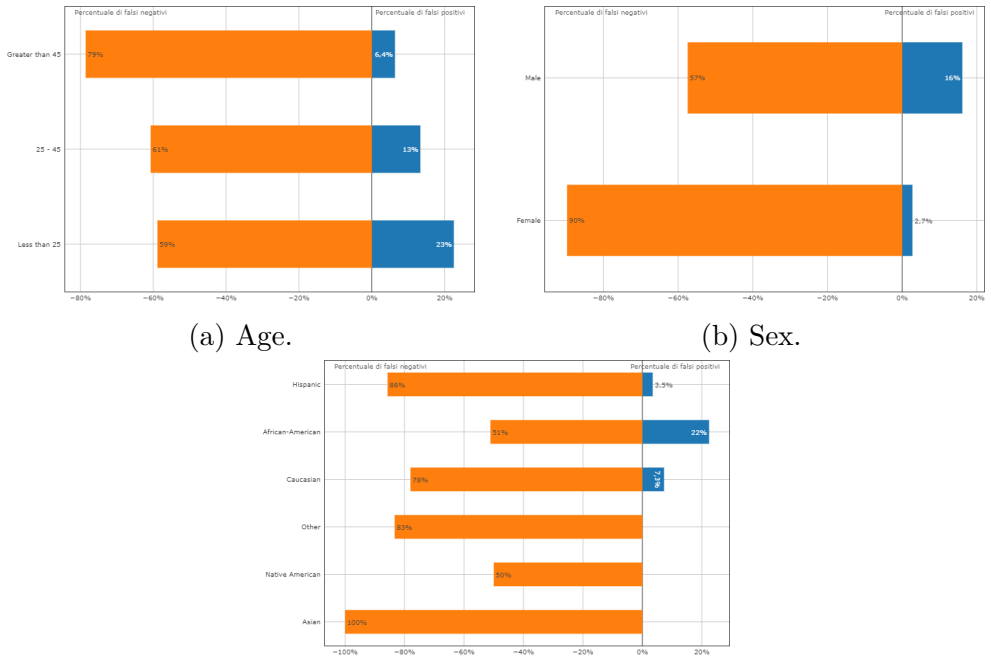
Figure 3.8: Sex vs age_cat distribution and viceversa for the COMPAS dataset.

more than 45 years.

Instead, if we consider the 'sex' feature fairness assessment, in Fig.3.9b, we can see how the model classifies the more men as recidivists, despite the fact they are not. Instead women are the more classified wrongly as non recidivists. We can see the gap comparing the FNRs and the FPRs. Women have a FNR equal to 90%, while men a FNR equal to 57%. Considering FPR instead, women have a rate equal to 2.7% and men equal to 16%.

In the end, we performed the fairness assessment over the race feature. From Fig.3.9c it is possible to see how the model is not fair. It should be considered that the "Asian," "Hispanic," and "Native American" categories are underrepresented, preventing the model from having enough instances, belonging to them, from which to learn. This leads the fairness assessment results for these categories to not be quite significant.

In conclusion, the Error Analysis tool found that the feature-value pairs used to identify the HER cohort are quite similar to the ones characterizing the subgroup with the highest FPR divergence found in prof. Baralis work, despite the two different approaches used. The RAIToolbox allowed us to analyze the cohorts of interests, and we discovered the model behaves differently for the two cohorts and also for the whole test set. We analyzed the features characterizing the two cohorts. 'Sex', 'race' and 'age' features are sensitive features and contains underrepresented categories. They all have a high importance in the model, so we performed a fairness analysis, finding out that the model is unfair considering all this features.



(a) Age.

(b) Sex.

(c) Race.

Figure 3.9: Fairness assessment of the Age, Sex and Race features of the COMPAS dataset.

3.2 Diabetes Medical Dataset

In the previous section we shown how RAIToolbox has great potential and allows to identify and analyze critical cohorts. After its validation, in this sections, an application in the medical field will be performed, based also on the methodology followed during the COMPAS analysis. In this case, two different tasks will be performed: binary classification and multiclass calssification.

3.2.1 Dataset Description

The analyzed dataset is the Diabetes one, created by Strack et al. [77]. In order to evaluate historical trends of diabetes care in patients admitted to US hospitals and to inform future strategies that can improve patient safety, a sizable clinical database was assembled and its contents were analyzed. The readmission probability of a patient after discharge and its dependence on other clinical parameters that might be gathered during hospitalization were the main points of attention. The dataset represents clinical treatment provided over a ten-year period (1999–2008) at 130 US hospitals and integrated delivery networks [77, 78].

The dataset can be downloaded from the UCI repository [79]. It contains 101,766 hospitalization cases of patients with diabetes, represented by 55 features. All the instances satisfy the following conditions, as stated by :

- it is an inpatient encounter (a hospital admission);;
- it is a diabetic encounter, meaning that any type of diabetes was diagnosed during the encounter;
- the length of stay was between 1 and 14 days;
- laboratory tests were conducted during the encounter;
- medications were conducted during the encounter.

The dataset contains sensitive features, as patient race, sex and age. To get an idea of the dataset, other features are diagnosis, physician medical specialty, admission type, HbA1c test result, time in hospital, diabetic medications, number of medications, number of laboratory tests performed, and so on (the whole list of features with their description can be founded in the Strack et al. work [77]). Notice that the age attribute values are not natural numbers, but it is encoded as a 10-level ordinal variable according to 10 age intervals provided in the initial data table.

The aim of the analysis of this dataset is to predict readmissions ("readmitted" variable). As said previously, in this place two different tasks will be carried out. During the former task, we will perform a binary classification, trying to predict readmission within 30 days after discharge from the hospital, encoded as 1, 0 otherwise (as performed in Strack et al. work [77]). During the latter task, what we will do is to predict readmitted variable value within the three possible values in the dataset. Readmission is encoded in 3 levels in this case, with 0 value corresponding to "No" (absence of recorded readmission), 1 to "<30 days" and 2 to ">30 days".

As is typical for any real-world data, the initial database contains inaccurate, redundant, and noisy information. Numerous features have a significant percentage of missing values, making it impossible to directly treat them. The underlined features are weight (with 97% of the values missing), payer code (40% of missing values), and medical specialization (47%). All the three attributes are excluded from further analysis because they contain too much missing values. After this cleaning step, it can be founded that there are 3713 instances containing missing values. All these missing values belongs to the features 'race', 'diag_1', 'diag_2' and 'diag_3'. We decide do drop these instances.

As done in the COMPAS case, the dataset is splitted in training and test sets, according to the ration 80%/20%. All the numerical features are standardized

using z-score, while, for categorical features, the OneHot Encoding technique is applied.

At this point, the first time we performed this process, we decided to proceed with the training phase of the model and with the analysis through the RAIToolbox, but after the model was trained, the RAIToolbox raises an error. There were categorical feature-values present in the training set but not in the test one and, for this reason, it was unable to perform analysis. So we found here the first problem of the RAIToolbox: it is unable do deal with categories which have too few instances having a specific categorical value, raising an error when a categorical feature value is not present in the test set. What we done so, was to find those features causing this problem and to delete them.

After all these cleaning steps, the dataset used for the analysis contains 40 features representing patient and hospital outcomes, 78441 hospitalizations for diabetes patients train set and 19611 hospitalizations for test set. Standardization and onehot encoding are performed again and a SVM classifier is trained over the training dataset. In the next section, we show the results and the analysis made with the RAIToolbox.

3.2.2 Results

Binary Classification

In this subsection, we want to train a binary classification model, having as target the fact that the patient was readmitted within 30 days from the discharge (1) or not (0). We trained a SVM classifier, obtaining the following accuracy scores: accuracy train = 0.8860, accuracy test = 0.8925. It seems to not overfit during the training phase.

As before, the first analysis we perform is the error analysis one. Fig.3.10 shows us how the model fails to correctly classify instances belonging to the HER cohort (number_inpatient $>$ 3, num_medications $>$ 13) with a probability of 32.24%. The error covered by this cohort is of 7.49%.

Considering the HEC cohort, it contains 34.23% of misclassified instances (722 over 2109). But, as we said analyzing the COMPAS dataset, to have an high Error Coverage does not mean to be a critical cohort for the model. In fact, HEC cohort has a lower ER w.r.t. the whole test set (6.50% vs 10.75%). This cohort is the combination of these features-values pairs: number_inpatient $<=$ 1.50; diag_1 != 157 | 162 | 198 | 202 | 250.41 | 250.42 | 250.7 | 276 | 287 | 288 | 296 | 298 | 403 | 428 | 434 | 440 | 507 | 531 | 550 | 562 | 564 | 572 | 584 | 593 | 608 | 727 | 787 | 790 | 820 | 824 | 852 | V58; discharge_disposition_id != 2 | 22 | 28 | 3 | 5.

As before, our aim is to understand the reasons behind the different model behaviours in the different cohorts, analyzing the relevance sensitive features have for the classification.

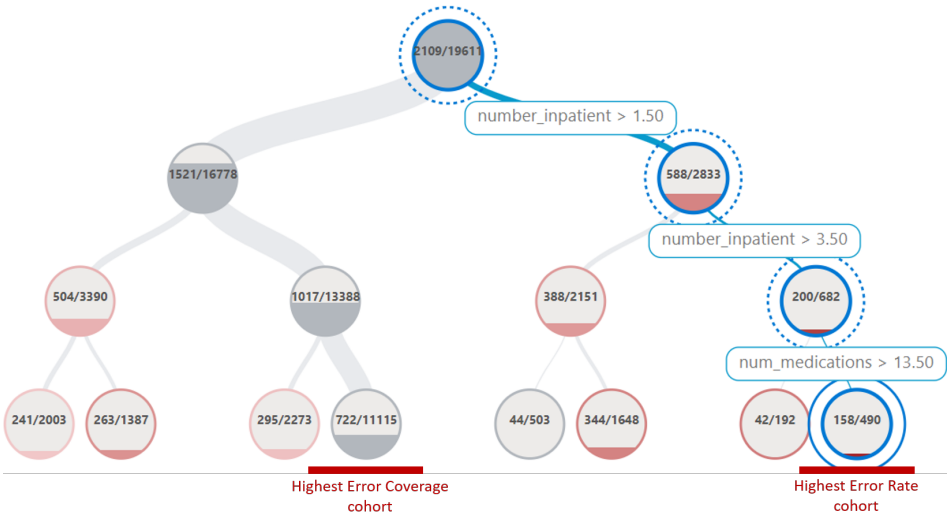


Figure 3.10: HER and HEC cohorts for the Diabetes dataset binary task.

In Fig.3.11 it is possible to see the performances of the model considering the different cohorts. In all the cohorts the model classifies all instances as 0 (i.e. not readmitted in the next 30 days). In other words, the error rate of the model is equal to 10.75% because almost all instances belonging to class 1 (10.73% of the total) are wrongly classified.

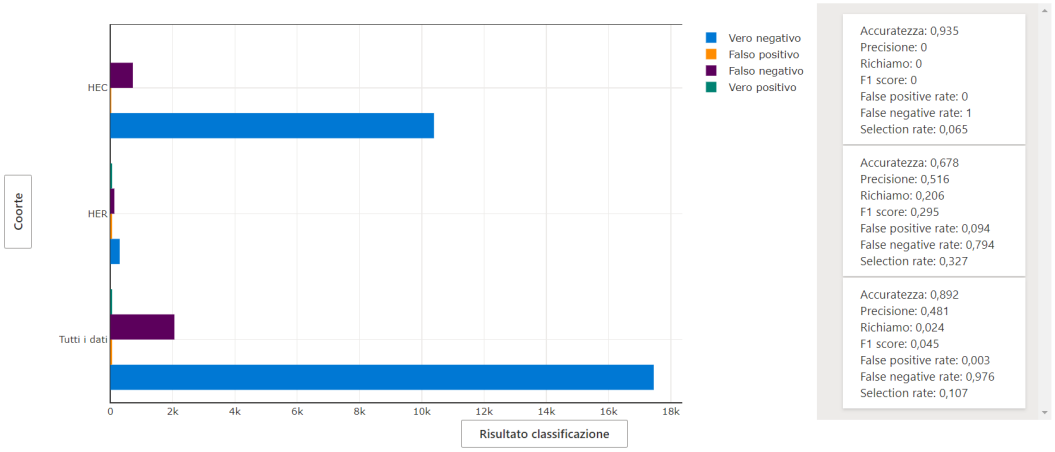


Figure 3.11: Comparison of the Diabetes number of TN, FN, TP and FP.

In Table 3.2 it can be seen what we already said. In all the cohorts, the error rate is equal to the percentage of patients belonging to the true class. But, stated this, it means that, among the 13.06% of patients labeled as 1, the half of them

are misclassified. This is confirmed also by the FPR in Fig.3.11, where the HER cohort has a FPR equal to 0.094, while the whole test set has a FPR thirty times lower. On the other side the FNR is lower in the HER cohort. Instead, considering the HEC cohort, FPR and FNR are quite similar to the ones of the whole test set (here all the instances are classified as non readmitted within 30 days).

	ER	EC	y==1	$\hat{y}$==1	y - $\hat{y}$
Whole test set	10.75%	100%	10.73%	0.53%	**10.20%**
HER cohort	32.24%	7.49%	32.65%	13.06%	**19.59**%
HEC cohort	6.50%	34.23%	6.48%	0.009%	**6.48**%

Table 3.2: Statistics and obtained result for the Diabetes dataset.

Now that we analyzed the cohorts performances and we have seen the distributions of the sensitive features devided by model result, let's analyzed the feature importance for each cohort. In Fig.3.12 there are shown the most 20 features sorted by the improtance for the whole test set. The importance of the first two features by importance ('discharge_disposition_id' and 'number_impatient') slightly changes for the HEC cohort, reducing from a value around 0.25 to 0.20. What is impactful is the importance the 'number_impatient' feature has for the HER cohort ('number_inpatient' is the number of visits of the patient in the year preceding the encounter).

If we analyze the most important features, we can see how they are not sensitive features. The most important sensitive feature is the 'age', which is in the sixth place, while 'race' and 'gender' features are not present between the most twenty important features.

A tool we did not used during the COMPAS analysis, is the local explanation one. As described in the RAIToolbox section, it allows to understand the importance of each feature for each single instance. In Fig.3.13 it is possible to see the importance of each age value for each instance. It can be noticed how, according to the previous analysis, the importance of the different values have a quite similar distribution for the different cohorts.

In Fig.3.14 it is possible to see the model results over the sensitive features, for each feature value. It can be seen from these graphs that the model does not appear to be distorted, due to the proportion of TP to FP, which appears to be uniform. To get confirmation of the above, the best thing to do is to perform fairness assessment.

From Fig.3.15 it is possible to confirm what we argued in the previous paragraph: the model is not biased among sensitive features. But this conclusion could be reached in the earlier stages of this analysis. Considering the whole test set, the model labels as 1 only the 0.53% of the patients, leading itself to have a very

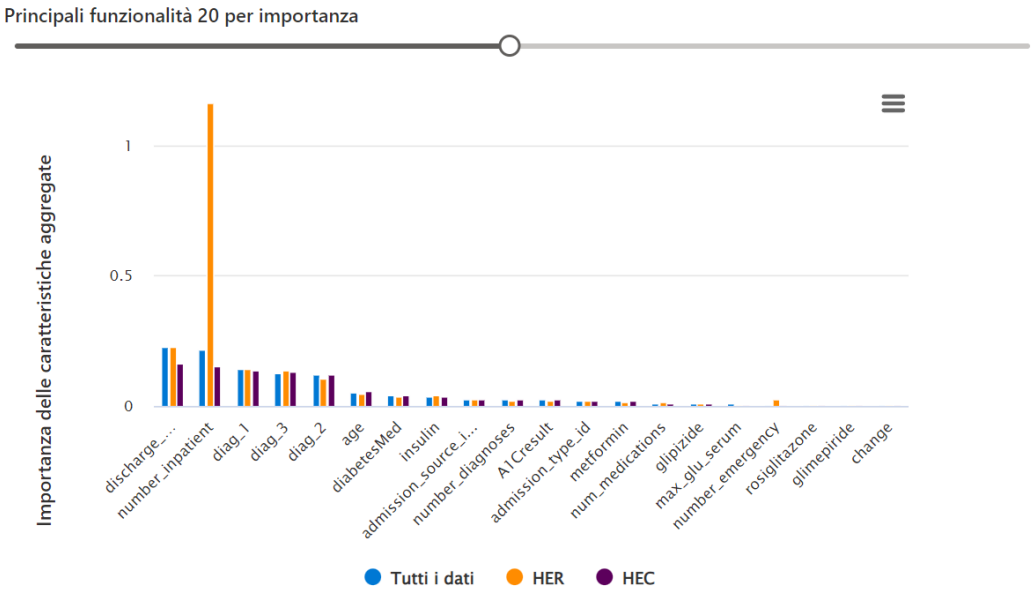
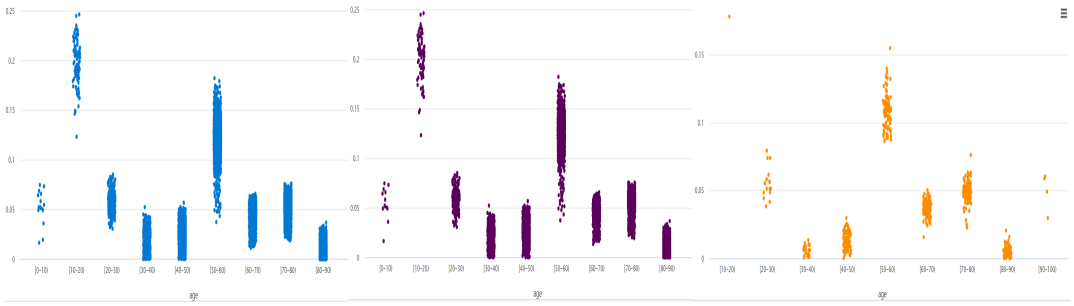


Figure 3.12: Global Feature Importance for the Diabetes dataset binary task.



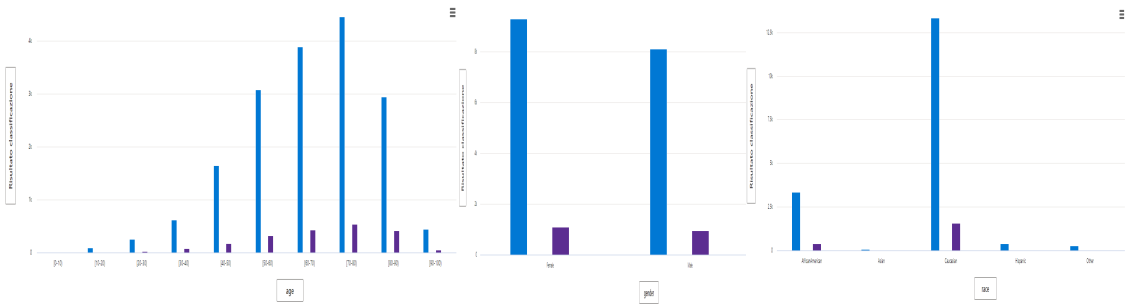
(a) Whole test set. (b) HEC cohort. (c) HER cohort.

Figure 3.13: Local instance feature importance for the AGE attribute.

high FNR, equal to 0.976. This implies that it is high probably that all sensitive features values will have an high FNR quite close to the one of the whole test set. What it could be interesting to do, is to perform fairness assessments for each single cohort. Unfortunately this can not be done automatically inside the RAIToolbox Dashboards.

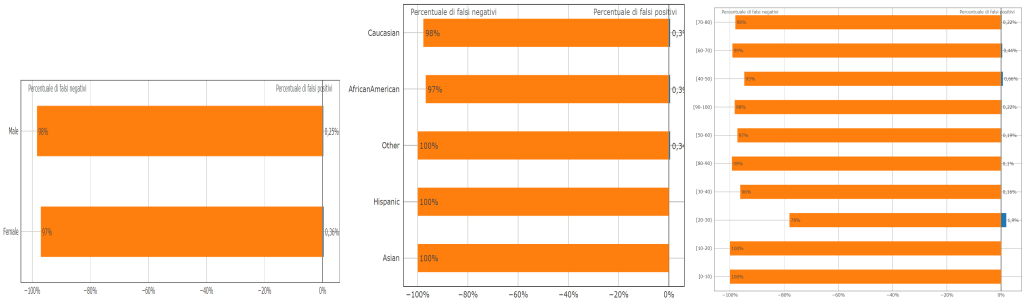
Multiclass Classification

Now, what we want to do, is to train a multi-classification model, using the classes 0 (absence of recorded readmission), 1 (readmitted within 30 days) and 2 (readmitted after 30 days). We train another SVM classifier and obtained the a



(a) Age. (b) Sex. (c) Race.

Figure 3.14: Model results for Age, Sex and Race features. The number of TN, FN, TP and FP is showed.



(a) Gender (b) Race (c) Age

Figure 3.15: Fairness assessment of the Gender, Race and Age features of the Diabetes dataset.

train accuracy equal to 0.5912, and a test accuracy equal to 0.5821. They are not good results, so, again, the first analysis we perform is the Error one. In Fig.3.16 it is shown the Error Analysis Tree, highlighting the HER and the HEC cohorts. They have, respectively, an error rate of 52.48% and 43.10%, and an error coverage of 24.54% 51.80%. Three quarters of the error is contained in these two cohorts!

It could be very interesting to analyze the model performances, but, trying to figure out them, the best we managed to obtain are the results shown in Fig.3.17. The only in details analysis we can perform, are the ones related to the original test dataset and the feature importance plot.

Concerning the analysis about the data, we can see how they are distributed among the cohorts, but we can not assess if instances belonging to class 1 are the more classified as 0 or 2. This is a very easy task, that can be solved plotting a simple confusion matrix. In addition, fairness assessment can not be performed as

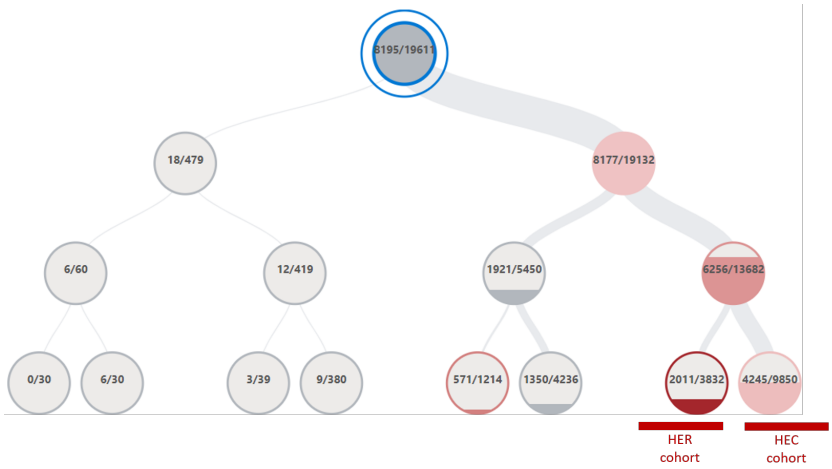


Figure 3.16: HER and HEC cohorts for the Diabetes dataset multiclass task.

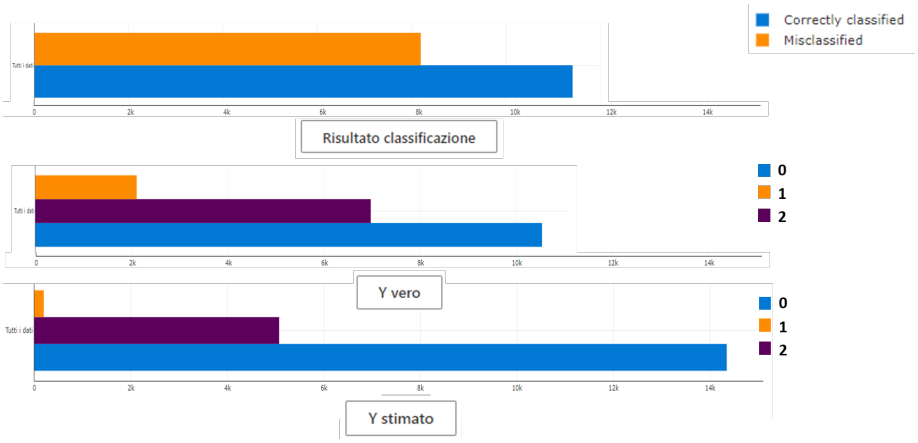


Figure 3.17: Results of the Diabetes dataset multiclass task.

for the binary classification task. All these shortcomings lead RAIToolbox to be not yet usable for the analysis of multiclass classification models, or at least, not to an extent comparable to that of binary models.

3.3 Myocardial Infarction Medical Dataset

Now that we analyzed two different datasets through the RAIToolbox and get more familiar with it, try to get the maximum from it, applying it to the Myocardial Infarction Medical Dataset. In this section we will perform only the binary classification, due to the low effectiveness and to the limitations that the RAIToolbox has for the analysis of multiclass calssification models.

3.3.1 Dataset Description

The disease known as MI has spread widely over the past 50 years, making it one of the most critical issues in contemporary medicine [78]. All nations continue to experience a high incidence of MI. Patients with MI encounter a variety of clinical progressions. There are two types of MI, with or without consequences that impact the long-term prognosis [78]. Meanwhile, complications that aggravate the course of the disease and potentially cause mortality occur in around half of patients during the acute and subacute phases. Even a professional cannot always predict when these issues will arise. In this context, predicting MI problems could lead to better outcomes by enabling the implementation of the required preventive measures [78].

The database was collected from 1992 through 1995, in the Krasnoyarsk Interdistrict Clinical Hospital (Russia). It can be founded and downloaded at [80]. At the same page it is provided the detailed description of the variables with the corresponding descriptive statistics.

The dataset contains information about 1,700 patients characterized by 111 features describing the clinical phenotypes and 12 features representing possible complications of the MI disease (123 features in total). In this case, the only sensitive features are the 'SEX' and the 'AGE' ones.

For our purposes, we will use only the 111 features describing the clinical phenotypes, while we will keep only 1 of the other 12 representing the possible complications. For the Binary classification task we decided to keep the target variable 'ZSN', having value 1 if chronic heart failure happend, 0 otherwise [80].

During the cleaning phase, all columns having more than 25% of missing values where removed (seven columns). Analyzing the number of instances having no missing values, we found out that they are only 544 instances, so we decide to remove only instances having more than ten missing values and to keep all the other, filling null values according to the feature type (we used the mean for the numerical features, while all the categorical feature where filled with the target "unknown" plus the name of the feature).

The dataset was splitted also here according to the 80%/20% proportion and

during the preprocessing step, onehot encoding was performed for categorical features, while the numerical one were standardized using the z-score standardizer.

3.3.2 Results

The model trained is also here a SVM classifier and it obtained the following accuracy scores: 0.8422 over the training set and 0.8067 over the test set. It is important to notice that our target, 'ZSN', is unbalanced, with 77% of instances belonging to the class 0 and only 23% to the class 1, so the obtained result is not impressive.

Considering Error Analysis, Fig.3.18 shows us the HER cohort, while Fig.3.19 shows one cohort we will call T (Test cohort). Here we decided to not analyze the HEC cohort, but choose to analyze the T cohort due to its trade-off between error rate and error coverage scores.

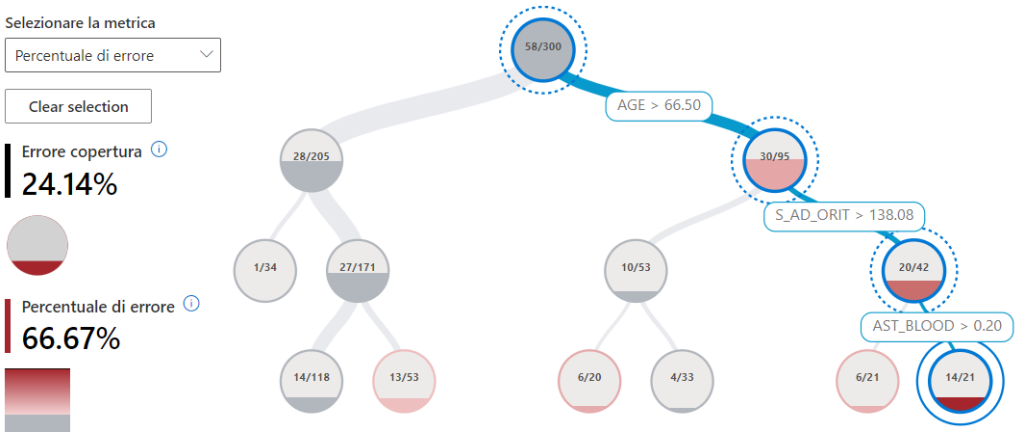


Figure 3.18: Highest Error Rate cohort for the MI dataset.

HER cohort is the combination of the feature-value pairs 'AGE' > 66, 'S_AD_ORIT' > 138.08 (Systolic blood pressure according to intensive care unit) and 'AST_BLOOD' > 0.20 (Serum AsAT content). It has an error rate equal to 66.67% (quite higher than the one of the whole test set) and covers the 24.14% of the errors. The T cohort, instead, is the combination of the feature-value pairs $49 <$ 'AGE' < 66 and 'AST_BLOOD' > 0.27. The error rate of this cohort is 24.53%, while it covers 22.41% of errors.

From Table 3.3 it can be seen how the error is generate principally by the fact that the model tends to not predict patients as at risk of chronic heart failure. We can see also how the behaviour of the model is different for the different cohorts. W.r.t. to whole test set, despite the increase of patients at risk in the T cohort, the model decreases the percentage of patients classified as such, getting an error

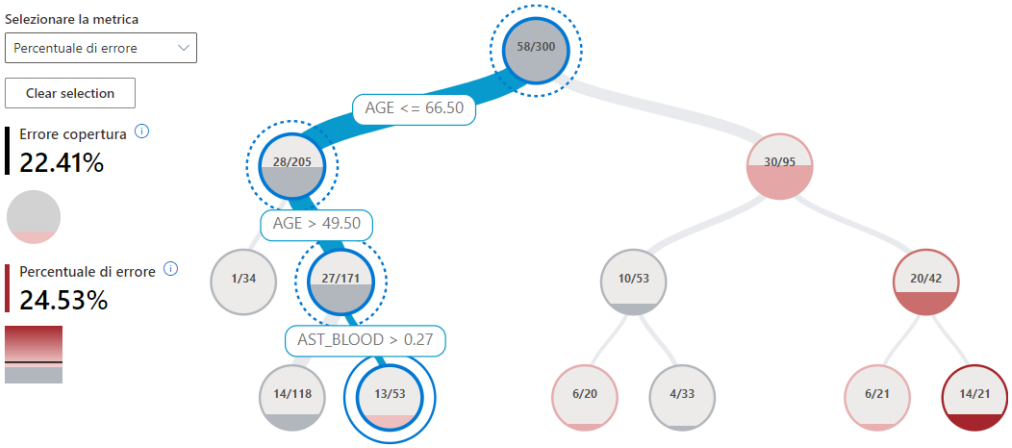


Figure 3.19: T cohort for the MI dataset.

rate slightly higher. Instead, considering the HER cohort, a doubling of positive instances is matched by a doubling of instances classified as such, but the ER triplicates. These observations are summed up and improved by the graph in Fig.3.20. The differences among the FPRs and FNRs of the cohorts are quite exhaustive. The model behaves differently based on the set of features-values pairs. Let's try to understand what are the most important features for the model and also if they are different based on the cohort.

	ER	EC	y==1	$\hat{y}$==1	y - $\hat{y}$
Whole test set	19.33%	100%	23.33%	16.67%	**6.67%**
HER cohort	66.67%	24.14%	52.38%	33.33%	**19.05%**
T cohort	24.53%	22.14%	28.30%	11.32%	**16.98%**

Table 3.3: Statistics and obtained result for the MI dataset.

In Fig.3.21 it is shown the importance of the most 25 important feature. The most important feature is the 'ZSN_A', which is strictly related with our target variable, because it represents the presence of chronic Heart failure (HF) in the anamnesis. This features, however, becomes less important if we consider the T cohort, becoming the sixth feature for importance. The second most important feature for the whole test set is 'NA_R_3_n', which represents the use of opioid drugs in the ICU in the third day of the hospital period. This feature is the most important for the T cohort, while it is less important for the HER cohort. In the and, it can be seen as the sensitive feature 'AGE' is quite important for the classification outcome, being the third feature for importance for the whole test set, while it is the first for the HER cohort. Considering the task, it was not so

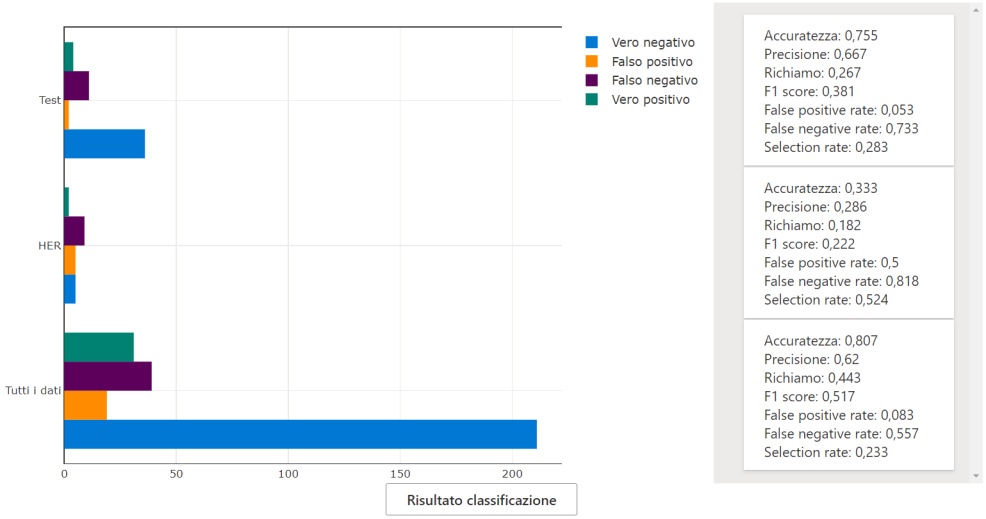


Figure 3.20: Comparison of the MI number of TN, FN, TP and FP.

strange that the 'AGE' has a very high importance. On the other side, the 'SEX' feature is the 21st for importance, which is high in ranking if we consider that they are 103 in total.

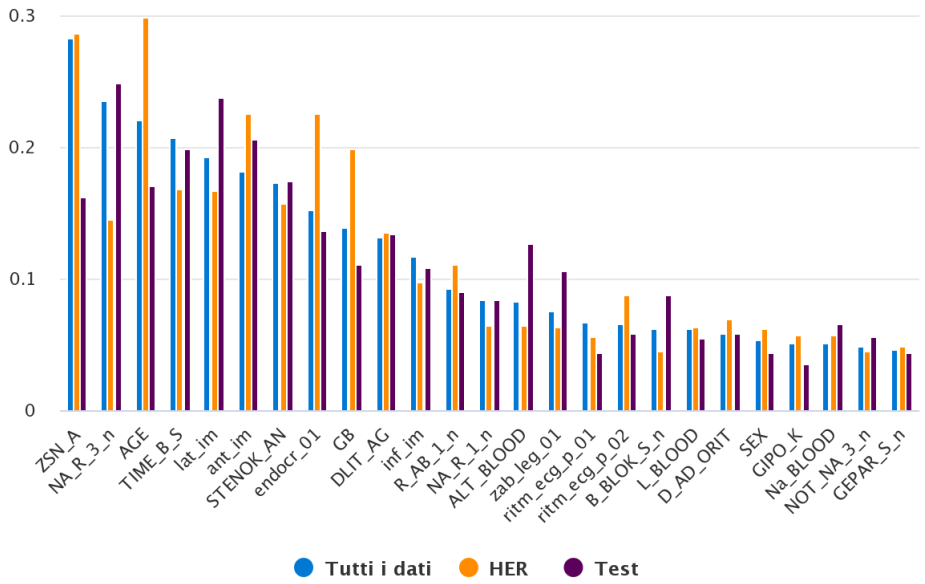
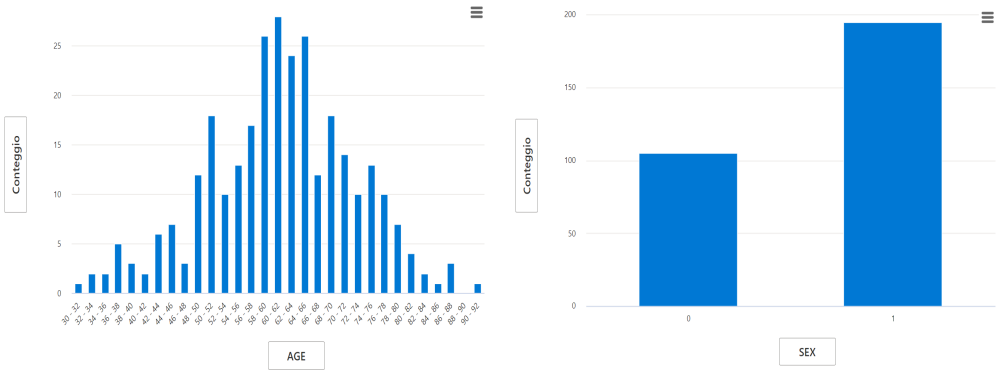


Figure 3.21: Global Feature Importance for the MI dataset.

We stated that the sensitive features are quite important for the prediction of

the model, so let's see if they are balanced and if they are also balanced with respect to the target. In Fig.3.22 it can be seen how the 'SEX' feature is unbalanced and the 'AGE' feature has the most of the instances belonging to values between 50 and 80 years. From Fig.3.23 it can be seen also that there is an unbalance between the age of men and the age of women. In the end, the results for each category value are plotted in Fig.3.24 and in Fig.3.25 the probability to be classified as at risk of chronic heart failure are shown. All these plots shows us how the model classifies the more women as at risk, maybe due to the fact the 'SEX' feature is unbalanced, and the older a patient is, the higher is the likelihood of classifying him or her at risk.



(a) Count of the different AGE possible values for the MI dataset.

(b) Count of the different SEX possible values for the MI dataset.

Figure 3.22: AGE and SEX value count.

The fact that the dataset is unbalanced toward men and the feature "AGE" could be the actual representation of reality or, on otherwise, it could be caused by a bias in the dataset, but we have no control over it. The only way to get an answer is through further studies on the subject. Instead, what we can check, is that the higher probability for women to be classified as at risk fit well the dataset or not. The same for the 'AGE' feature. In Fig.3.26 it is shown the fairness assessment for the 'SEX' feature. It can be seen how the fact that women where classified with an higher probability as 1, is not totally correct, because we have a FPR of the 17% for women, w.r.t. the 4.4% we have for men. On the other side, men are the more classified as negative, despite the fact the belong to the positive class. Unfortunately, when we tried to perform the fairness analysis for the 'AGE' feature, the RAIToolbox did not work. The fact that the 'AGE' feature is a numerical attribute causes problems with the Fairness tool. It allows to select this feature for fairness assessment, but it does not work.

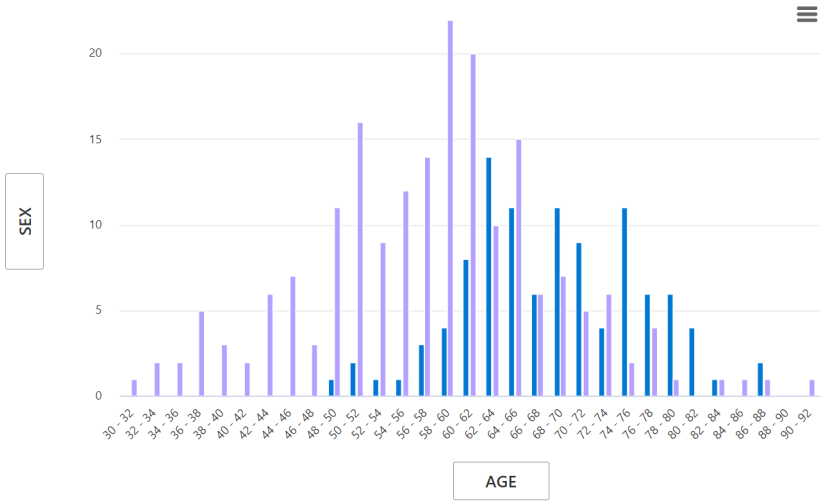
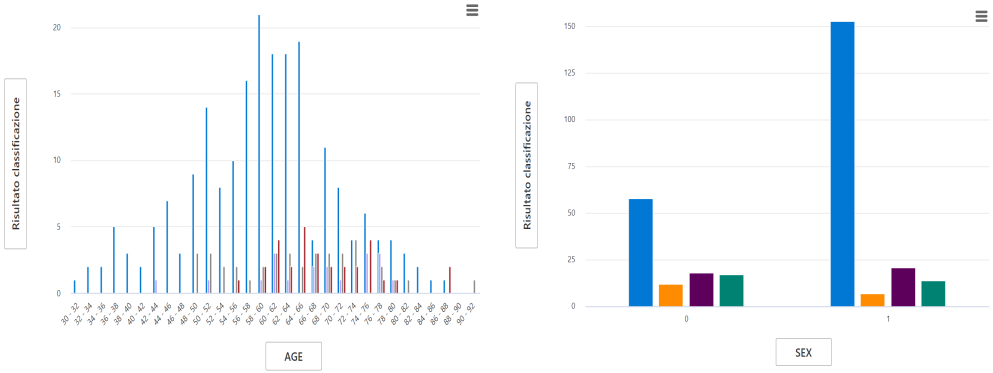
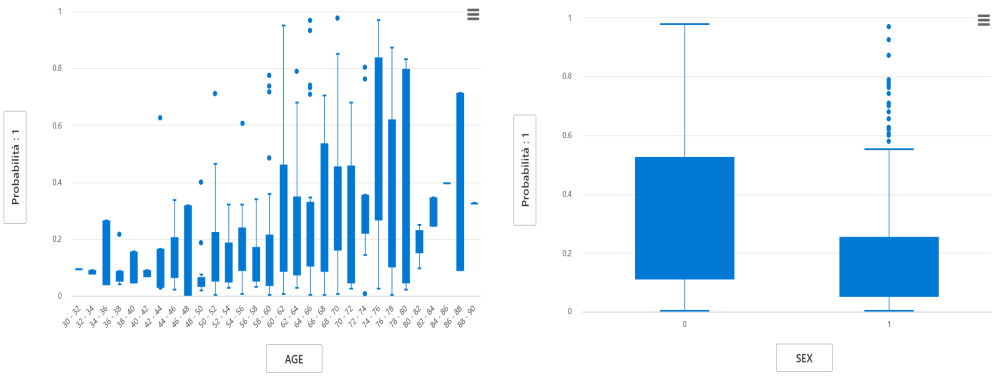


Figure 3.23: Distribution of the AGE feature values for each SEX.



(a) AGE. (b) SEX.

Figure 3.24: Model results for AGE and SEX features. The number of TN, FN, TP and FP is showed.



(a) AGE. (b) SEX.

Figure 3.25: Probability to be classified as at risk for each feature value of AGE and SEX.

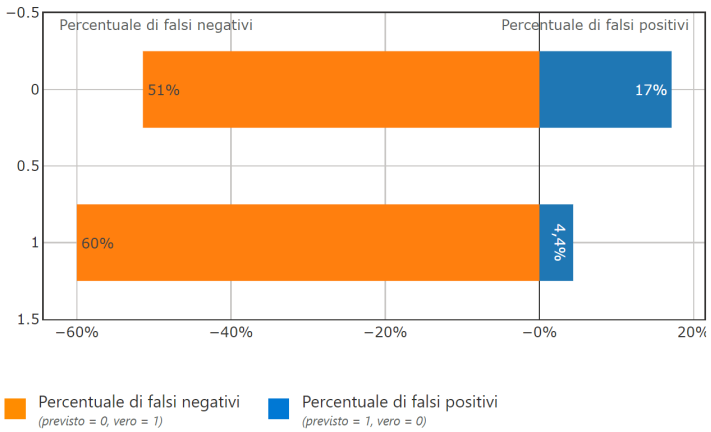


Figure 3.26: Fairness assessment of the SEX feature of the MI dataset.

Chapter 4

Conclusions

This thesis work started analyzing the history of Artificial Intelligence and giving the needed definitions. The next step was to go deeper, analyzing what AI in Medicine is and what is its history. From here, the work started to deal with the problem of the need of trustworthiness in the models created by AI. The concepts of Responsible AI and all its elements were analyzed, starting from the seven principles promoted by the European Union [41]: Human Agency and Oversight; Technical Robustness and Safety; Privacy and Data Governance; Transparency; Diversity, Non-Discrimination and Fairness; Societal and Environmental Well-Being; Accountability. After the studying of all these elements, the RAIToolbox was introduced and described. They were described its strengths and weaknesses and, in the end, it was applied to practical use cases.

As first, the aim was to validate the RAIToolbox on the COMPAS dataset, using, as evaluation comparison, the DivExplorer results obtained by Professor Elena Baralis and Dott. Eliana Pastor [75]. The Error Analysis tool identified the highest error rate cohort as having the same more important features of the highest FPR divergence subgroup of the DivExplorer work. These allowed us to state that the RAIToolbox could be a valuable instrument for error analysis.

After its validation, it was used to analyze two different medical datasets. The former was the Diabetes dataset [79, 77], which aim is to predict if the diabetic patient will be readmitted within 30 days from the discharge or not. The results of the trained model were not satisfying, but the RAIToolbox allowed to perform an in depth analysis of both dataset and model. It allowed to understand what are the most important features at all the levels, from the whole test set, to the specific cohorts and, in the end, for the single instances. It was founded out that the models behave in a very similar way for the HEC cohort and the whole test set and that it is not unfair in general. Unfortunately, it is not yet possible to perform directly fairness analysis for the single different cohorts. After the evaluation of the model trained for the binary classification task, the RAIToolbox was used to

try to perform an evalutaion of a model trained over a multiclassification task. The results were quite poor, showing the weaknesses the RAIToolbox has when it is used for multiclassification. It does not show a confusion matrix, but only how much instances, for each class, where wrongly or correctly classified. This visualization does not allow to understand to what wrong classes instances are assigned.

The last analyzed model was trained using the Myocardial Infarction Complications Dataset [80, 78], which aim is to predict one of the twelve possible complications outputs. We decided to use 'ZSN' as target, representing chronic heart failure as a binary task. The model did not performe very well, but this allowed us to making the most of the RAIToolbox's potential. We analyzed two different cohorts, the highest error rate one and the cohort having the highest trade-off between error rate and error coverage. We founded out that the model behaves differently for the different cohorts. Going deeper, we also founded out that the importance of the features changes for the different cohorts and also that the 'AGE' feature has a very high importance. Also the 'SEX' feature has a relative high importance, so we decided to analyze the distribution of the data based on these features and we founded out that the dataset is quite unbalanced with respect to both the features. In the end the fairness assessment was performed, and the result was that the model is slightly unfair considering the 'SEX' feature. Unfortunatly here it was founded out another limit of the RAIToolbox: it is unable to perform the fairness assessment over the numerical feature 'AGE'.

In conclusion, thanks to the RAIToolbox we were able to understand which are the most critical cohorts. This allowed us to start from them to analyze their features, finding their critical issues. We were able to go deeper and deeper into the data and model analysis. The ability to go into detail during the analysis is a great strength for the RAIToolbx. But to have further visualizations and graphs could be useful inside the RAIToolbox.

4.1 Future works

The path to a Responsible AI use is still unpaved and a lot of work can be done to implement techniques to enhance explainability and fairness assessment. Some possible future works could be the analysis of other tools made by other researchers or by other companies. In last years a lot of guidelines were wrote for RAI building and implementation, like the one made by Google [63, 64] and the one wrote by IBM [62].

Bibliography

[1] A. M. TURING, "I.—COMPUTING MACHINERY AND INTELLIGENCE," *Mind*, vol. LIX, no. 236, pp. 433–460, 10 1950. [Online]. Available: https://doi.org/10.1093/mind/LIX.236.433

[2] S. Harnad and P. Scherzer, "First, scale up to the robotic turing test, then worry about feeling," *Artificial Intelligence in Medicine*, vol. 44, no. 2, pp. 83–89, 2008, artificial Consciousness. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S093336570800122X

[3] Wikipedia contributors, "Perceptron," August 2022, accessed: 2020-10-13. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Perceptron&oldid=1106416414

[4] Wikipedia contributors, "Frank rosenblatt," Jul. 2022, accessed: NA-NA-NA. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Frank_Rosenblatt&oldid=1097344029

[5] Wikipedia contributors, "Perceptrons (book)," January 2022, accessed: 2020-10-13. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Perceptrons_(book)&oldid=1063970182

[6] N. Liu, P. Shapira, and X. Yue, "Tracking developments in artificial intelligence research: constructing and applying a new search strategy," *Scientometrics*, pp. 3176–3177, 2021. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s11192-021-03868-4.pdf

[7] C. Krauthammer, "Be afraid - the meaning of deep blue's victory." May 1997, accessed: 2020-10-12. [Online]. Available: https://web.archive.org/web/20170228073706/http://www.weeklystandard.com/be-afraid/article/9802

[8] S. London, G. Bradski, A. Coates, L. Deng, and M. Shah, "Ask the ai experts: What's driving today's progress in ai?" July 2017, accessed: 2020-10-12. [Online]. Available: https://web.archive.org/web/20180413190018/https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/ask-the-ai-experts-whats-driving-todays-progress-in-ai

[9] J. Clark, "Why 2015 was a breakthrough year in artificial intelligence," December 2015, accessed: 2020-10-12. [Online]. Available: https://web.archive.org/web/20161123053855/https://www.bloomberg.com/news/articles/

2015-12-08/why-2015-was-a-breakthrough-year-in-artificial-intelligence

[10] UNESCO, "Unesco science report: the race against time for smarter development," June 2021, accessed: 2020-10-12. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000377433/PDF/377433eng.pdf.multi

[11] S. Russell and P. Norvig, *Artificial intelligence: A modern approach*, 3rd ed. Pearson, 2009.

[12] J. McCarthy, "What is artificial intelligence," *Stanford*, 2004. [Online]. Available: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

[13] "Artificial intelligence," June 2022, accessed: 2022-10-14. [Online]. Available: https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html

[14] P. Langley, "The changing science of machine learning," *Mach. Learn.*, vol. 82, no. 3, pp. 275–279, 2011.

[15] "Difference between artificial intelligence and machine learning," accessed: 2022-11-25. [Online]. Available: https://www.javatpoint.com/difference-between-artificial-intelligence-and-machine-learning

[16] IBM Cloud Education, "Artificial intelligence (ai)," June 2020, accessed: 2020-10-09. [Online]. Available: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence#toc-types-of-a-q56lfpGa

[17] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, pp. S36–S40, 2017.

[18] Y. Mintz and R. Brodie, "Introduction to artificial intelligence in medicine," *Minimally Invasive Therapy & Allied Technologies*, vol. 28, no. 2, pp. 73–81, 2019.

[19] H. Yokota, M. Goto, C. Bamba, M. Kiba, and K. Yamada, "Reading efficiency can be improved by minor modification of assigned duties; a pilot study on a small team of general radiologists," *Japanese journal of radiology*, vol. 35, no. 5, pp. 262–268, 2017.

[20] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[21] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[22] V. H. Buch, I. Ahmed, and M. Maruthappu, "Artificial intelligence in medicine: current trends and future possibilities," *British Journal of General Practice*, vol. 68, no. 668, pp. 143–144, 2018.

[23] V. Kaul, S. Enslin, and S. A. Gross, "History of artificial intelligence in medicine," *Gastrointestinal endoscopy*, vol. 92, no. 4, pp. 807–812, 2020.

[24] CB Insights Research, "Healthcare remains the hottest

ai category for deals." April 2017, accessed: 2022-10-14. [Online]. Available: https://www.cbinsights.com/research/artificial-intelligence-healthcare-startups-investors/#:~:text=In%20the%20private%20market%2C%20AI,across%20270%20deals%20since%202012.

[25] J. H. Chen and S. M. Asch, "Machine learning and prediction in medicine—beyond the peak of inflated expectations," *The New England journal of medicine*, vol. 376, no. 26, p. 2507, 2017.

[26] C. A. Kulikowski, "Beginnings of artificial intelligence in medicine (aim): computational artifice assisting scientific inquiry and clinical art–with reflections on present aim challenges," *Yearbook of medical informatics*, vol. 28, no. 01, pp. 249–256, 2019.

[27] S. Amarel, "The history of artificial intelligence at rutgers," *AI Magazine*, vol. 6, no. 3, pp. 192–192, 1985.

[28] C. Kulikowski, "An opening chapter of the first generation of artificial intelligence in medicine: the first rutgers aim workshop, june 1975," *Yearbook of medical informatics*, vol. 24, no. 01, pp. 227–233, 2015.

[29] S. Weiss, C. A. Kulikowski, and A. Safir, "Glaucoma consultation by computer," *Computers in Biology and Medicine*, vol. 8, no. 1, pp. 25–40, 1978.

[30] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, "Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system," *Computers and biomedical research*, vol. 8, no. 4, pp. 303–320, 1975.

[31] P. Malik, M. Pathania, V. K. Rathaur *et al.*, "Overview of artificial intelligence in medicine," *Journal of family medicine and primary care*, vol. 8, no. 7, p. 2328, 2019.

[32] The Massachusetts General Hospital Laboratory of Computer Science., "Using decision support to help explain clinical manifestations of disease." accessed: 2022-10-16. [Online]. Available: http://www.mghlcs.org/projects/dxplain

[33] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller, "Watson: beyond jeopardy!" *Artificial Intelligence*, vol. 199, pp. 93–105, 2013.

[34] N. Bakkar, T. Kovalik, I. Lorenzini, S. Spangler, A. Lacoste, K. Sponaugle, P. Ferrante, E. Argentinis, R. Sattler, and R. Bowser, "Artificial intelligence in neurodegenerative disease research: use of ibm watson to identify additional rna-binding proteins altered in amyotrophic lateral sclerosis," *Acta neuropathologica*, vol. 135, no. 2, pp. 227–247, 2018.

[35] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PloS one*, vol. 12, no. 4, p. e0174944, 2017.

[36] C. Castaneda, K. Nalley, C. Mannion, P. Bhattacharyya, P. Blake, A. Pecora, A. Goy, and K. S. Suh, "Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine," *Journal of clinical*

bioinformatics, vol. 5, no. 1, pp. 1–16, 2015.

[37] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs." *SEMANTiCS (Posters, Demos, SuCCESS)*, vol. 48, no. 1-4, p. 2, 2016.

[38] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and M. A. Przybocki, "Four principles of explainable artificial intelligence," *Gaithersburg, Maryland*, 2020.

[39] E. Neri, F. Coppola, V. Miele, C. Bibbolino, and R. Grassi, "Artificial intelligence: Who is responsible for the diagnosis?" pp. 517–521, 2020.

[40] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–9, 2020.

[41] European High-Level Expert Group on AI, "Ethics guidelines for trustworthy ai," April 2019, accessed: 2020-10-17. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[42] R. Benjamins, A. Barbado, and D. Sierra, "Responsible ai by design in practice," *arXiv preprint arXiv:1909.12838*, 2019.

[43] M. Ghallab, "Responsible ai: requirements and challenges," *AI Perspectives*, vol. 1, no. 1, pp. 1–7, 2019.

[44] J. Angwin, J. Larson, L. Kirchner, and S. Mattu, "Machine bias," May 2016, accessed: 2022-11-4. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[45] J. L. Skeem and C. T. Lowenkamp, "Risk, race, & recidivism: Predictive bias and disparate impact.(2016)," *Criminology*, vol. 54, p. 680, 2016.

[46] P. Nemitz, "Constitutional democracy and technology in the age of artificial intelligence," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180089, 2018.

[47] D. Schiff, B. Rakova, A. Ayesh, A. Fanti, and M. Lennon, "Principles to practices for responsible ai: closing the gap," *arXiv preprint arXiv:2006.04707*, 2020.

[48] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," *Berkman Klein Center Research Publication*, no. 2020-1, 2020.

[49] D. Schiff, J. Biddle, J. Borenstein, and K. Laas, "What's next for ai ethics, policy, and governance? a global overview," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 153–158.

[50] EU, "General data protection regulation," 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[51] Y. Wang, M. Xiong, and H. Olya, "Toward an understanding of responsible artificial intelligence practices," in *Proceedings of the 53rd hawaii international conference on system sciences*. Hawaii International Conference on System

Sciences (HICSS), 2020, pp. 4962–4971.

[52] A. Askell, M. Brundage, and G. Hadfield, "The role of cooperation in responsible ai development," *arXiv preprint arXiv:1907.04534*, 2019.

[53] L. Cheng, K. R. Varshney, and H. Liu, "Socially responsible ai algorithms: Issues, purposes, and challenges," *Journal of Artificial Intelligence Research*, vol. 71, pp. 1137–1181, 2021.

[54] C. Kästner, "Transparency and accountability in ml-enabled systems," Jan. 2022, accessed: 2022-11-01. [Online]. Available: https://ckaestne.medium.com/transparency-and-accountability-in-ml-enabled-systems-f8ed0b6fd183

[55] Wikipedia contributors, "Black box," Oct. 2022, accessed: 22-11-2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Black_box&oldid=1118576430

[56] J. Stoyanovich, S. Abiteboul, and G. Miklau, "Data, responsibly: Fairness, neutrality and transparency in data analysis," in *International Conference on Extending Database Technology*, 2016.

[57] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[58] "Common fairness metrics — fairlearn 0.8.0 documentation," accessed: 2022-12-1. [Online]. Available: https://fairlearn.org/v0.8/user_guide/assessment/common_fairness_metrics.html

[59] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[60] Wikipedia contributors, "Fairness (machine learning)," Dec. 2022, accessed: 10-11-2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Fairness_(machine_learning)&oldid=1125375384

[61] "Stack overflow developer survey 2018," accessed: 2022-15-11. [Online]. Available: https://insights.stackoverflow.com/survey/2018/

[62] "AI ethics," https://www.ibm.com/artificial-intelligence/ethics, accessed: 2022-11-5.

[63] "Responsible ai practices," https://ai.google/responsibilities/responsible-ai-practices/, accessed: 2022-11-5.

[64] "Building responsible ai for everyone," https://ai.google/responsibilities/, accessed: 2022-11-5.

[65] "Responsible AI," https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6, accessed: 2022-11-5.

[66] "responsible-ai-toolbox," accessed: 2022-11-5. [Online]. Available: https://github.com/microsoft/responsible-ai-toolbox

[67] "InterpretML," accessed: 2022-12-3. [Online]. Available: https://interpret.ml/

[68] "Error analysis," https://erroranalysis.ai/, accessed: 2022-11-5.

[69] S. Singla, B. Nushi, S. Shah, E. Kamar, and E. Horvitz, "Understanding failures of deep networks via robust feature extraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 853–12 862.
[70] "Fairlearn," https://fairlearn.org/, accessed: 2022-11-5.
[71] "Interpretml documentation," accessed: 2022-11-3. [Online]. Available: https://interpret.ml/docs/intro.html
[72] "Welcome to interpret-community's documentation! — interpret-community 0.27.0 documentation," accessed: 2022-11-3. [Online]. Available: https://interpret-community.readthedocs.io/en/latest/index.html
[73] M. Bao, A. Zhou, S. Zottola, B. Brubach, S. Desmarais, A. Horowitz, K. Lum, and S. Venkatasubramanian, "It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks," 2021. [Online]. Available: https://arxiv.org/abs/2106.05498
[74] "Chapter 10 story COMPAS: recidivism reloaded," https://pbiecek.github.io/xai_stories/story-compas.html, Oct. 2020, accessed: 2022-11-4.
[75] ProPublica, "Divexplorer project," https://divexplorer.github.io/, accessed: 2022-11-4.
[76] ProPublica, "compas-analysis: Data and analysis for 'machine bias'," accessed: 2022-11-4. [Online]. Available: https://github.com/propublica/compas-analysis
[77] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records," *BioMed research international*, vol. 2014, 2014.
[78] S. E. Golovenkin, J. Bac, A. Chervov, E. M. Mirkes, Y. V. Orlova, E. Barillot, A. N. Gorban, and A. Zinovyev, "Trajectories, bifurcations, and pseudo-time in large clinical datasets: Applications to myocardial infarction and diabetes data," *GigaScience*, vol. 9, no. 11, p. giaa128, 2020.
[79] "UCI machine learning repository: Diabetes 130-US hospitals for years 1999-2008 data set," https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008, accessed: 2022-11-5.
[80] "UCI machine learning repository: Myocardial infarction complications data set," https://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications, accessed: 2022-11-5.