11-17-2023 10:00 AM

An Exploration of Visual Analytic Techniques for XAI: Applications in Clinical Decision Support

Mozhgan Salimiparsa,

Supervisor: Sedig, Kamran, *The University of Western Ontario*
Co-Supervisor: Lizotte, Daniel J., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science

Recommended Citation

Salimiparsa, Mozhgan, "An Exploration of Visual Analytic Techniques for XAI: Applications in Clinical Decision Support" (2023). *Electronic Thesis and Dissertation Repository*. 9848.
https://ir.lib.uwo.ca/etd/9848

Abstract

Artificial Intelligence (AI) systems exhibit considerable potential in providing decision support across various domains. In this context, the methodology of eXplainable AI (XAI) becomes crucial, as it aims to enhance the transparency and comprehensibility of AI models' decision-making processes. However, after a review of XAI methods and their application in clinical decision support, there exist notable gaps within the XAI methodology, particularly concerning the effective communication of explanations to users.

This thesis aims to bridge these existing gaps by presenting in Chapter 3 a framework designed to communicate AI-generated explanations effectively to end-users. This is particularly pertinent in fields like healthcare, where the successful implementation of AI decision support hinges on the ability to convey actionable insights to medical professionals.

Building upon this framework, subsequent chapters illustrate how visualization and visual analytics can be used with XAI in the context of clinical decision support. Chapter 4 introduces a visual analytic tool designed for ranking and triaging patients in the intensive care unit (ICU). Leveraging various XAI methods, the tool enables healthcare professionals to understand how the ranking model functions and how individual patients are prioritized. Through interactivity, users can explore influencing factors, evaluate alternate scenarios, and make informed decisions for optimal patient care.

The pivotal role of transparency and comprehensibility within machine learning models is explored in Chapter 5. Leveraging the power of explainable AI techniques and visualization, it investigates the factors contributing to model performance and errors. Furthermore, it investigates scenarios in which the model outperforms, offering potential to enhance user trust by shedding light on the model's strengths and capabilities.

Recognizing the ethical concerns associated with predictive models in health, Chapter 6 considers potential bias and discrimination in ranking systems. By using the proposed visual analytic tool, users can assess the fairness and equity of the system, promoting equal treatment. This research emphasizes the need for unbiased decision-making in healthcare.

Having developed the framework and illustrated ways of combining XAI with visual analytics in the service of clinical decision support, the thesis concludes by identifying important future directions of research in this area.

Keywords

eXplainable AI (XAI), Interpretability, Black Box, Transparency, XAI Methods, Visual Analytics (VA), Visualization, Interaction, Human Computer Interaction (HCI), Ranking, Clinical Decision Support Systems, CDSS, Machine Learning, Healthcare, Fairness, Bias and Discrimination, AI adoption

Summary for Lay Audience

Our study centers on utilizing the potential of Artificial Intelligence (AI) to enhance decision-making in various fields, with a special focus on healthcare. We address the challenge of interpreting AI systems' complex decision-making through the methodology of eXplainable AI (XAI). Our focus lies in bridging the gap between complex AI insights and user comprehension with the eventual goal of fostering trust and informed choices. To achieve this comprehension, we propose a framework that combines Visual Analytics (VA) and XAI methodologies, creating a more intuitive way to communicate AI-generated insights to users. Through case studies, we demonstrate how this combined approach could be used to enhance transparency in AI decisions, especially in healthcare scenarios.

The thesis comprises 7 chapters. The first chapter discusses motivation and provides an overview of the thesis structure. The second chapter explains the keywords and terminology used throughout the thesis. The third chapter reviews existing XAI methods and their relevance to clinical decision support, pinpointing areas where they fall short in practical application. It introduces an interactive visualization framework to bridge this gap and provide techniques that may be used by healthcare professionals in better comprehending AI models.

Chapter 4 introduces a visual analytic tool designed to explain ranking systems, with a case study focusing on patient ranking and prioritization in intensive care units (ICUs). Leveraging various XAI methods, this tool could help healthcare providers to grasp the inner workings of ranking models and prioritize patients based on critical factors.

Chapter 5 employs an XAI technique to identify areas where a machine learning model underperforms, offering users valuable insights to approach such situations with care. This is illustrated through case studies on the detection of septic shock in ICUs.

Chapter 6 introduces a visual analytics tool tailored to investigate potential biases within ranking systems, illustrated by a case study on ICU admissions. Chapter 7 gives a summary of the previous chapters and concludes the thesis, encapsulating the key findings and contributions. Overall, our research aims to strengthen the connection between AI and healthcare professionals, with the long-term goal of fostering transparency, trust, and fairness in AI-driven decision support.

Co-Authorship Statement

Chapter 1 represents my original work in explaining the motivation, identifying the problem, framing the dissertation, and providing an overview of the chapter structure. Chapter 2, which offers the necessary background to understand the technical foundations of the keywords and terminology in later chapters, is also my original work.

Chapter 3 represents my own research and writing, with supervisory guidance, editing, and direction provided by my advisors, Dr. Lizotte and Dr. Sedig. I was responsible for all research and writing.

Chapter 4 has been published at the 21st International Conference of Artificial Intelligence in Medicine (AIME) in Portoroz, Slovenia, as part of the XAI-Healthcare workshop. My supervisors, Drs. Lizotte and Sedig, are co-authors. Co-authors contributed to framing the problem, validating the experimental approach, and distinguishing the case study from the general algorithm.

Chapter 5 is developed in collaboration with Spassmed Inc. The co-authors were responsible for data analysis and developing the model. The conceptual idea was mine, and I was responsible for all research and writing.

Chapter 6 represents my own research and writing, with advisory input from my supervisors.

Lastly, Chapter 7 is my own writing, providing a summary and conclusion for the dissertation.

Acknowledgements

I would like to extend my gratitude to my supervisor, Dr. Dan Lizotte, whose continuous support and guidance have been invaluable throughout this thesis. I am deeply grateful for his unwavering dedication, patience, and mentorship. Dr. Lizotte's exceptional expertise, insightful feedback, and willingness to go above and beyond have been instrumental in shaping and refining this research. I am truly fortunate to have had the opportunity to work under his guidance.

I would also like to express my appreciation to my co-supervisor, Dr. Kamran Sedig, for his valuable feedback and assistance throughout this research journey. Dr. Sedig's expertise and insights have contributed significantly to my growth as a researcher, and I am grateful for his continuous support and mentorship.

I extend my appreciation to my family and friends for being my pillars of strength throughout this research journey. Your love and encouragement have been a constant reminder of the importance of the bonds we share. Your presence in my life is a gift I treasure beyond words.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

This dissertation is presented in the form of an integrated article, comprising a collection of individual materials that collectively contribute to a cohesive research program. These materials span across diverse and interdisciplinary domains, including machine learning, explainable artificial intelligence, clinical decision support systems, visual analytics, and human-computer interaction. Each work is presented in a self-contained form, maintaining its original structure with sections such as introduction, background, materials, results, and conclusion. This format allows for both independent and progressive reading of each material. in essence, the merging of these chapters constructs a unified storyline that progresses from the establishment of fundamental ideas, identifying gaps, proposing innovative frameworks, and culminating in a practical toolset that can provide comprehensible explanations, eventually forming a component of systems that foster trust, transparency, and fairness in the adoption of XAI within CDS systems. In the following sections, we provide a concise overview of the dissertation's general motivation and outline the subsequent chapters.

1.1 Motivation

In recent years, the integration of artificial intelligence (AI) methods into clinical decision support have shown immense potential for improving healthcare outcomes [1]. These AI models have the ability to assist healthcare professionals in critical decision-making processes, such as patient triage[2], diagnosis [3], and treatment recommendations[4]. However, the lack of transparency and interpretability in these AI models has hindered their widespread adoption in clinical settings [5]. The challenges and obstacles related to the integration of AI have spurred considerable research interest in the realm of interpretable and explainable AI, often referred to as XAI. This field has emerged to promote interpretability and tackle the issues surrounding accountability and trustworthiness [9, 8]. Various XAI approaches have been developed and employed to create and communicate explanations for the predictions made by ML models [8]. These XAI methods essentially serve as intermediaries for the models, generating explanations that, when consistent with domain knowledge, aim to aid users in gaining a deeper understanding of the models' underlying reasoning and in mitigating their opaque nature. Ultimately, the primary goal of these methods is to empower users to comprehend and have confidence in both the models and their outcomes. Consequently, XAI techniques hold substantial potential to enhance the application, effectiveness, and acceptance of AI within Clinical Decision Support (CDS) systems [8]. Existing XAI methods hold promise in enhancing accountability and transparency [9, 10]. Nevertheless, effectively communicating explanations generated by these methods to users through the system interface remains an obstacle, creating a gap between the tools and users in the adoption of AI within clinical practice. To address this challenge, this thesis aims to bridge the gap between AI and end-users by developing a framework that combines explainable AI (XAI) methods and visual analytics for effective clinical decision support.

The core motivation of this thesis stems from the pressing need to harness the power of AI while ensuring its responsible and accountable application in the clinical realm. Our research is driven by four main objectives, each playing a crucial role in advancing AI-powered healthcare systems:

1. Enhancing Clinical Decision Support with XAI: The primary motivation behind this research is the adoption of AI in clinical decision support systems, augmented by eXplainable AI. These AI-driven systems often make complex predictions and recommendations that impact patient care. However, the ”black-box” nature of many AI models can be a barrier to their widespread adoption in healthcare. By integrating XAI methods, we aim to bring transparency to these models, allowing healthcare professionals to understand the reasoning behind the AI's decisions. This transparency is a prerequisite not only for enhancing the accountability of AI systems but also for fostering trust among medical practitioners by helping them verify the validity of AI-driven recommendations.

2. Developing Visual Analytics for Transparent Ranking Systems: A significant challenge in healthcare is the prioritization of patient admissions, especially in critical settings like Intensive Care Units (ICUs). Here, the integration of AI for ranking patients based on urgency can significantly impact outcomes. However, simply presenting a ranked list of patients without context can be confusing and counterproductive. This is where the development of visual analytics tools utilizing XAI methods becomes pivotal. By creating interactive visualizations that explain the factors contributing to a patient's ranking, we provide a basis for empowering healthcare providers to make more informed decisions. This step is essential to ensure that AI-driven rankings are not just accepted blindly but are understood, questioned, and appropriately utilized to provide optimal care.

3. Promoting Trust through Cautious Decision-Making: Trust is the paramount importance of any healthcare system. Introducing AI, with all its complexities, requires healthcare professionals to have confidence in the technology's capabilities and limitations. By using XAI methods to highlight areas where healthcare practitioners should practice cautious decision-making, we emphasize the role of human expertise in conjunction with AI. To create an environment where this trust is possible, we employ XAI methods to delve into the model's inner workings and identify areas where users should practice cautious decision-making. By highlighting these aspects, we aim to empower medical practitioners to make informed decisions based on actionable insights derived from the AI system. Encouraging cautious decision-making can foster a collaborative relationship between AI and healthcare providers and contributes to better patient outcomes.

4. Addressing Bias and Fairness in Healthcare: Bias and discrimination are not just theoretical concerns; they have real-world consequences, particularly in healthcare. AI systems trained on biased data can perpetuate and exacerbate existing healthcare disparities. Developing a visual analytics tool that assesses bias within AI models and quantifies fairness metrics is of utmost importance. This tool can be used by healthcare organizations to proactively identify and rectify biases, thereby ensuring that AI systems provide equitable outcomes across different demographic groups. Addressing bias and fairness is a critical step towards creating an AI-powered healthcare ecosystem that upholds principles of justice and equality.

This thesis is driven by the convergence of advanced AI technologies, eXplainable AI methods, and healthcare needs. By enabling transparent clinical decision support, developing visual analytics for ranking systems, promoting cautious decision-making, and addressing bias and fairness, our research is a step on the road to enhancing patient care, empowering medical professionals, and establishing a strong foundation of trust in AI-powered healthcare systems. Each objective has been carefully crafted to address a unique facet of AI integration, with the goal of eventually contributing to a responsible transformation of healthcare practices.

Importance of this Work

This thesis presents a significant contribution to the field of clinical decision support by identifying the critical gap between the potential benefits of AI and the need for transparency and trust in healthcare settings and by offering methods can shrink this gap. By integrating XAI methods with visual analytics, the proposed framework can empower end-users to understand and effectively utilize AI models, ultimately enhancing their adoption and impact in clinical practice.

By providing healthcare professionals with transparent and interpretable AI models, this research aims to enhance trust, understanding, and the effective adoption of AI in clinical settings, ultimately improving patient care and outcomes. The exploration of discrimination in ranking systems further highlights the ethical considerations associated with AI implementation in healthcare, paving the way for more equitable and unbiased decision-making processes.

This thesis endeavors to bridge the gap in clinical decision support by developing a framework that combines XAI methods and visual analytics. Through a multidisciplinary approach, this research aims to provide healthcare professionals with transparent and interpretable AI models that are needed for establishing trust and facilitating the effective adoption of AI in clinical practice. The comprehensive visualization and interaction provided by the framework enable users to gain deeper insights into the decision-making processes, while addressing the critical issue of discrimination promotes fairness and equity in patient care. Ultimately, this work advances the field of clinical decision support, with the potential of enhancing patient outcomes and improving healthcare practices on a broader scale.

1.2 Structure of the Dissertation

This dissertation continues with the following chapters:

Chapter 2: Background

This chapter provides the foundational context for the subject matter that forms the backbone of this dissertation. Within this foundation, key terminologies utilized throughout the dissertation are distinctly defined. Given the interdisciplinary nature of this work, this chapter effectively establishes a common ground for readers across various disciplines, acquainting them with the fundamental concepts that form the basis of the dissertation.

Chapter 3: Rapid Review of XAI Methods and Their Application in Clinical Decision Support

This chapter provides a comprehensive and rapid review of XAI methods and their application in the context of clinical decision support. With the rapid advancement of AI techniques, it is crucial to assess the current state of XAI and its relevance to the healthcare domain. The chapter explores the existing landscape of XAI techniques, ranging from rule-based approaches to more complex machine learning interpretability methods.

In addition to discussing the potential benefits of XAI, the chapter also identifies challenges and limitations faced in its practical application within clinical settings. These limitations include the lack of interpretability in black-box models, difficulties in integrating XAI into complex medical workflows, and the need for effective communication between AI systems and end-users. By critically analyzing existing approaches, this chapter highlights the existing gap between the potential benefits of XAI and its practical application in clinical scenarios.

Furthermore, this chapter proposes a novel framework that aims to bridge this gap. The framework centers around the development of an interactive visualization tool that leverages XAI methods for clinical use. The tool is designed to enhance the interpretability and transparency of AI models, which is needed before end-users, such as healthcare professionals, can trust and effectively utilize AI-driven decision support systems. By providing a visual and intuitive interface, the framework empowers users to gain a deeper understanding of the AI models' decision-making processes and facilitates their incorporation into real-world healthcare workflows.

Chapter 4: Visual Analytic Tool for Explaining Ranking System

Building upon the proposed framework from Chapter 3, Chapter 4 focuses on the development of a visual analytic tool specifically tailored for ranking and triaging patients in the intensive care unit (ICU). This chapter addresses the challenge of understanding and explaining complex ranking systems employed in critical care settings. By integrating a variety of XAI methods, including counterfactual explanations and feature importance analysis, the tool can enable healthcare professionals to comprehend how the ranking model functions and how individual patients are prioritized.

The visual analytic tool not only allows users to explore the ranking outcomes but also provides interactive features for deeper investigation. It facilitates the examination of factors influencing the ranking, the impact of parameter adjustments, and the evaluation of alternate scenarios. Through this comprehensive visualization and interaction, healthcare professionals can gain insights into the underlying decision-making processes and better comprehend the reasoning behind the rankings. Such enhanced understanding fosters trust in the AI model and enables informed decision-making for optimal patient care.

Chapter 5: Exploring Error Analysis in Machine Learning Through Explainable AI

This chapter serves as an exploration into the analysis of errors within a machine learning model utilizing explainable AI. This endeavor holds significant importance in establishing user trust and in helping analysts iteratively improve their models. By analyzing misclassified instances, significant features contributing to suboptimal performance are identified. The analysis reveals regions where the classifier performs poorly, allowing the calculation of error rates within these regions. This understanding becomes extremely valuable for promoting cautious decision-making, especially in situations with critical consequences.

Chapter 6: Addressing Discrimination in Ranking Systems

Recognizing the potential for bias and discrimination in AI-driven ranking systems, Chapter 6 explores how to address these issues within the proposed framework by developing methodologies for investigating the impact of demographic factors, such as sex, race, and income on the decision-making process of AI models. It highlights ethical considerations associated with AI implementation in healthcare and emphasizes the need for fair and unbiased decision-making.

The chapter introduces a visual analytic tool that allows users to explore and evaluate the fairness and equity of a ranking system, identifying potential biases and disparities. By empowering healthcare professionals to actively assess and address discriminatory practices, the framework supports the promotion of equal treatment for all patients, regardless of their demographic characteristics.

Chapter 7: Summary and Conclusion

This chapter serves as the concluding chapter, offering concise summaries of the preceding chapters, highlighting their contributions, and outlining future research directions derived from the findings presented in this dissertation.

1.3 References

Bibliography

[1] Sandhu, S., Sendak, M., Ratliff, W., Knechtle, W., Fulkerson, W. & Balu, S. Accelerating health system innovation: principles and practices from the Duke Institute for Health Innovation. *Patterns*. **4** (2023)

[2] Hong, W., Haimovich, A. & Taylor, R. Predicting hospital admission at emergency department triage using machine learning. *PloS One*. **13**, e0201016 (2018)

[3] Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence In Medicine*. **23**, 89-109 (2001)

[4] Kohli, M., Kar, A., Bangalore, A. & Ap, P. Machine learning-based ABA treatment recommendation and personalization for autism spectrum disorder: an exploratory study. *Brain Informatics*. **9**, 1-25 (2022)

[5] Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing And Applications*. **32**, 18069-18083 (2020)

[6] Tjoa, E. & Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions On Neural Networks And Learning Systems*. **32**, 4793-4813 (2020)

[7] Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. & Others Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. **58** pp. 82-115 (2020)

[8] Dalvi-Esfahani, M., Mosharaf-Dehkordi, M., Leong, L., Ramayah, T. & Kanaan-Jebna, A. Exploring the drivers of XAI-enhanced clinical decision support systems adoption: Insights from a stimulus-organism-response perspective. *Technological Forecasting And Social Change*. **195** pp. 122768 (2023)

[9] Bunn, J. Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI). *Records Management Journal*. **30**, 143-153 (2020)

Chapter 2

Background

In order to ensure a comprehensive understanding for readers across various fields, this section offers concise explanations of key terminologies and concepts that will be used throughout this thesis. This background chapter acts as a guide, assisting readers from various fields navigate the complex relationships among CDS systems, AI, machine learning, XAI, and the fundamental principles of Human Computer Interaction.

2.1 Clinical Decision Support

Clinical decision support (CDS) systems are computer applications whose goal is to facilitate the decision-making process of clinicians [1]. CDS systems help clinicians utilize data and modeling techniques to improve the quality of decisions, which in turn can enhance healthcare delivery. These systems empower clinicians and patients by providing person-specific information [2]. Clinical Decision Support Systems are primarily employed at the point of care, where clinicians combine their expertise with suggestions from the CDS system [3]. Various CDSS have been developed and applied to a range of diseases and disorders, such as oncology [4], breast cancer [5], prediction of chronic kidney disease [6], Alzheimer's disease [7], diabetes care [8], and risk-level prediction of heart disease [9]. Beyond diagnosis, CDS systems serve many purposes, including treatment response prediction [10], treatment selection (personalization) [11, 12], prognosis [13], and prioritizing patient care based on risk [14].

CDS systems act as a valuable "second set of eyes" for clinicians, supplementing human expertise with embedded system knowledge. They contribute to reducing healthcare costs, enhancing patient safety, care quality, and healthcare effectiveness [19]. CDS systems improve patient safety by reducing medical errors and providing reminders regarding medications or medical events. Moreover, they are advantageous in low-resource settings where medical institutions, equipment, and qualified clinicians are limited.

There are several types of CDS systems [3]:

- **Knowledge-based systems**: These systems provide expert recommendations for treating specific medical conditions based on the latest research.
- **Alerts and reminders**: Systems offering real-time alerts or reminders when a patient's condition changes or requires immediate attention. They can also remind providers to order lab tests or prescribe medications, and provide information on adverse events.
- **Diagnostic support systems**: These systems assist providers in making diagnoses by using algorithms and other tools. They use patient records, lab tests, and other sources to narrow down potential diagnoses and offer recommendations.
- **Treatment support systems**: Based on recent evidence and guidelines, these systems aid providers in developing treatment plans.
- **Monitoring and surveillance systems**: These systems maintain a history of patient data and alert providers to changes or trends indicating further evaluation or treatment is needed.
- **Electronic health record (EHR) systems**: EHR systems store and manage patient information electronically, often including CDS functionality such as alerts, diagnostics, and treatment guidance.

2.2 Artificial Intelligence

Artificial intelligence (AI) refers to the development of computer systems that can perform tasks that normally require human intelligence, such as learning, problem-solving, and decision-making. There are several types of AI, including narrow or weak AI, which is designed to perform a specific task, such as playing a game or recognizing speech [16, 17]. General or strong AI is designed to perform potentially any intellectual task that a human can perform [18]. Super intelligent AI is a hypothetical type of AI that would surpass human intelligence in almost all areas [19].

AI systems can be trained using a variety of methods, including machine learning, in which the system is fed large amounts of data and uses algorithms to learn from it. Other approaches to AI include rule-based systems, in which the system is programmed with a set of rules to follow, and expert systems, in which the system is designed to mimic the decision-making processes of a human expert in a particular field. AI has the potential to revolutionize many fields, including healthcare, transportation, finance, and education. The use of artificial intelligence and machine learning algorithms in healthcare is increasing for the diagnosis and treatment of medical conditions [20]. The use of such tools can help increase the accuracy of clinical decisions and minimize the likelihood of clinical errors [21]. However, it also raises ethical and societal concerns and the need for responsible and transparent decision-making by AI systems.

2.2.1 Machine Learning

Machine learning (ML), a subfield of AI, discerns patterns within extensive medical data to forecast outcomes. ML involves crafting algorithms that can analyze data and predict or decide without explicit programming. It’s rooted in the concept that systems learn from data, discern patterns, and make decisions based on that data. ML models utilize large data volumes and statistical techniques to unveil data patterns and relationships, constructing models that predict or decide for new, unseen data. Several machine learning types exist:

- **Supervised Learning**: Train models using labeled data with known outputs, enabling predictions for new data based on patterns learned.

- **Unsupervised Learning**: Train models using unlabeled data to unveil data’s underlying structure or pattern.

- **Semi-Supervised Learning**: Train models with both labeled and unlabeled data.

- **Reinforcement Learning**: Train agents receiving rewards or penalties based on actions in a given environment.

Machine learning algorithms find application in diverse domains, including image and speech recognition, natural language processing, and outcome prediction in healthcare and finance. ML's success extends to various medical aspects such as disease prediction [22], medical imaging analysis [23], and clinical outcome forecasting like ICU admission [24].

A supervised ML model for prediction creates a mapping from inputs to outputs [3]:

$$f : X \rightarrow Y$$

Here, X represents input, Y is the output, and $f(.)$ signifies the mapping function (ML model). For instance, X could denote patient vital signs like heart rate, blood pressure, and respiration rate, while Y might indicate binary labels predicting patient cardiac arrest.

Despite their technical prowess, these systems encounter adoption challenges, their impact on healthcare remaining uncertain. A key factor is that AI/ML-based medical devices' effectiveness is significantly influenced by user behavior, often exposed to biases and algorithmic aversions [26]. These models heavily rely on input data and their internal decision logic, rendering them reliant on data quality and quantity [27]. Bias during training can affect ML models, leading to biased or erroneous predictions. For instance, a facial recognition system trained predominantly on a single ethnicity's images would introduce bias and compromise accuracy for diverse users. Trustworthiness and accountability are also concerns in clinical ML usage. The "black box" nature of predictive algorithms impacts trust, accountability, and adoption, which the forthcoming section on Explainable AI (XAI) aims to address.

2.3 Explainable AI (XAI)

The term Explainable AI (XAI) was first introduced by Van Lent et al. in 2004 [28]. XAI endeavors to elucidate the decision-making processes of machine learning models, offering transparency into how they arrive at specific conclusions. The aim of XAI is to render AI systems more interpretable and comprehensible for humans. XAI holds potential to enhance the transparency and accountability of AI systems, which is crucial in sectors like healthcare, finance, and criminal justice, where consequential decisions are made using AI, impacting individuals' lives.

In essence, XAI strives to make AI systems more understandable to humans, although a technical consensus on its definition remains elusive at present, demanding further clarity and consistency [5]. While the terms transparency, interpretability, and explainability are often used interchangeably [19], they encapsulate distinct concepts. Model interpretability refers to its ease of understanding, sometimes used synonymously with "explainability" [30]. "Transparency" can be understood as the holistic communication of information about the model's workings to users, encompassing training procedure documentation, assessment of data distribution, code release, and feature explanations [31]. "Explainability" offers insights into the rationale behind system decisions, sometimes linked to "understandability," defined as "tools enabling users to grasp model outcome reasoning" [32].

Our focus in this work is on explainability. "Black box" models that predict without explanation pose issues due to their lack of transparency and potential biases being concealed [14]. Similarly, biased data usage can pose challenges in precision medicine [34], as evident in biased prediction models stemming from medical datasets disregarding minorities. For instance, Framingham Heart Risk functions overestimated coronary heart disease risk among the German population [35], potentially attributed to lifestyle, dietary, and genetic differences between the German population and the Framingham study reference sample. This underscores the need for careful AI implementation.

Explainability's importance for AI systems is multifaceted [20]:

- **Trust**: AI systems lacking explanation may fail to gain user trust. XAI systems instill trust and confidence by furnishing explanations for their predictions.

- **Transparency**: AI can enhance transparency in decision-making, vital for accountability and fairness.

- **Debugging**: Absent an explanation, rectifying erroneous decisions by an AI system can be challenging. XAI systems offer this information, simplifying debugging and enhancing them.

The integration of XAI in medicine stems from medical professionals' desire to comprehend the rationale behind machine-generated decisions. As a result, there's a growing demand for AI approaches that are not only effective but also interpretable. This has implications not only for medical professionals but also for the public, governance, and policy [37]. Improved explanations can facilitate patient comprehension of the reasoning process, reducing over- and under-reliance, ultimately boosting confidence and leading to more effective decision-making [38].

2.4 Human Computer Interaction (HCI)

The development of Explainable AI (XAI) applications poses a challenge as the effectiveness of explanations hinges not solely on the model itself but also on the recipient's perception and understanding. Transparency does not ensure comprehension or prevent overwhelm. The quality of an explanation, measured by its appropriateness and usability, depends on factors such as the recipient's existing knowledge and goals. Hence, Human-Computer Interaction (HCI) research and User Experience (UX) design play an increasingly vital role in this domain.

HCI pertains to the interfaces between computers and users, encompassing the design and utilization of computer technology. HCI research aims to identify ways humans interact with computers and develop technology that enhances this interaction. The field emerged in the early 1980s, with the first known use reported by Carlisle in 1975 [40]. Combining cognitive science and human factors engineering within computer science, HCI draws from disciplines like computer science, behavioral sciences, design, and media studies [39]. HCI treats computer usage as a dialogue between users and machines, mirroring human-to-human interaction and informing theoretical considerations in the field [41]. In the subsequent sections, we explore HCI elements integral to explaining concepts, including visual representation, interaction, visual perception, and cognitive activity.

2.4.1 Cognitive Activity (Cognition)

As XAI-generated explanations must account for users' cognitive capacities and constraints, it's crucial to briefly discuss cognitive theories linked to human reasoning. Cognition denotes humans' ability to perceive, process, and comprehend information, all of which support reasoning. When dealing with intricate data, adaptation and accommodation are particularly valuable reasoning skills. Adaptation involves integrating newly perceived data into existing mental structures, facilitating effective handling of rapidly changing information [42]. Accommodation, on the other hand, permits categorizing incomplete information that doesn't align with existing knowledge structures, enabling sense-making of such data [43].

Various theories have sought to explain human cognition. Initially, cognition was likened to a computer with input, processing, and output components. However, this model was criticized as overly simplistic, incomplete, and assuming humans operate in isolation. Later, cognitive theories surfaced that emphasized the role of external artifacts and the environment. Concurrently, Vygotsky's activity theory gained prominence, valuing people's engagement and activities [44]. This theory subdivides activities into actions and tasks. While delving into activity theory implications is beyond this paper's scope, it's worth mentioning since it underpins the cognitive framework for analyzing human-computer interaction. These cognitive theories have influenced computational tools' design, with the subsequent section delving further into distributed cognition.

2.4.2 Distributed Cognition

The theory of distributed cognition posits that cognitive processes are distributed across individuals and extended over time, involving the coordination of both inner mental representations and external representations [45]. This perspective is pivotal for the development of computational tools as it challenges the notion that cognition solely occurs within the brain. The theory underscores the concept of a 'joint cognitive system,' encompassing both the user and the computational tool [62]. Consequently, user and tool collaborate in processing information rather than working independently. The various components of the tool interact with the user's cognitive processes to process information effectively. This concept holds particular relevance for visualization tools, where different components perform specific functions. According to this theory, the user's internal mental representations interact with the external representations presented on the tool's interface. Therefore, the success of cognitive tasks in the context of visualization tools hinges on diverse factors, including how visual representations of information align with each other, bridging the gap between the user's internal understanding and the tool's external representation [40].

2.4.3 Visual Perception

Visual perception involves the interpretation of information received from the light that reaches our eyes [48]. It is a dynamic process influenced by various factors such as attention, focus, and experience. Extensive research has explored the development of visual perception abilities and the interpretation of our surroundings. The processing of visual information by the brain is often delineated into two main stages. The initial stage encompasses parallel processing and the identification of five distinct features: orientation, texture, contour, color, and motion. In the subsequent stage, objects are identified and localized.

Numerous factors shape our perception of visual information, including expectations, motivations, past experiences, and cultural background [48]. Despite lingering questions in this realm, visual perception is widely acknowledged for its potency. Researchers highlight that humans possess enhanced input channels when utilizing visual abilities [50]. Additionally, visual recall is purportedly superior to verbal recall, and our brain exhibits a strong aptitude for recognizing visual patterns. Accordingly, appropriately designed computational tools that leverage our intrinsic capability to comprehend and process visual information offer increased efficiency in information processing.

2.4.4 Visual Representations

A visual representation entails presenting information, data, or ideas using visual elements like charts, diagrams, maps, graphs, or images. These representations are employed to elucidate complex or abstract concepts, enhance the effectiveness of information communication, or emphasize patterns or trends in data. Visual representations amalgamate diverse visual cues (e.g., lines, dots, shapes) into more intricate compositions (e.g., bar charts, scatter plots, heat maps) to encode information [51]. Various types of visual representations can be chosen based on the information type and presentation goals, including:

- Charts and graphs: Display numerical data, such as trends over time, group comparisons, or variable relationships. Examples comprise line graphs, bar charts, and scatter plots.
- Maps: Depict geographical information, such as place locations or distribution across an area.
- Diagrams: Illustrate relationships between components, parts, or processes. Examples include flowcharts, Venn diagrams, and organizational charts.
- Images: Convey information through pictures or photographs.

Visual representations aid understanding and retention of information by offering a visual means to grasp complex or abstract concepts. Utilized on digital interfaces, they expose stored information to users. Since cognition relies on the coordination of user's internal representations and tool's external representations [52], it is crucial to explore how external visual representations support user activities.

Visual representations harness the strength of the human visual perceptual system, designed to process information and recognize visual patterns. They enhance cognition by providing increased memory and processing resources, reducing information search, facilitating pattern detection, enabling perceptual inferences, and encoding information interactively.

The manner in which information is represented directly influences users' (clinicians') cognitive tasks. Cognitive science evidence reveals that different forms of representation impact cognitive activities. Subpar representation can impede users' task performance [52]. Clinicians thus necessitate appropriate tools for manipulating visual representations. The capacity to alter the form or content of visual representations can be achieved through interaction.

2.4.5 Interaction

Interaction can be defined as the user's actions on the tool interface, the ensuing reactions in the visual representation, and the user's perception of the resultant changes in the representation [55]. Interactions enable users to not only govern the form or content of visual representations but also the entire information dialogue. Interactions operate at different levels and facilitate dialogues between users and information. Users can manipulate information by performing actions on visual representations. Changes in representations, reflected on the interface, signal the interactive nature of the discourse. This interaction cycle completes when users perceive changes in visual representations. Thus, actions, reactions, and perceptions engender a two-way dialogue between users and information. Actions may be sequenced based on professional discretion, particularly valuable in fields like medicine where designers lack insight into users' data analysis strategies.

Static representations, though seemingly useful, lack manipulation capabilities, thereby placing a substantial information processing burden on users [56].

In contrast, interactive visual representations distribute information processing between the tool and the user. Interaction empowers users to bridge the gap between their internal (mental) and external (visual) representations, facilitating connections [62]. These interactive representations facilitate both convergent and divergent thinking, enabling users to tailor visual representations to their cognitive and perceptual requirements [53]. They offer several advantages. Interactive visual representations can unveil latent information [54], a particularly pivotal trait in healthcare due to the voluminous information. Clinical tasks involving analytical reasoning demand sequential analysis of information subsets. For example, doctors may initially focus on patient symptoms to narrow down potential diseases and subsequently assess demographic and geographical data. Apart from controlling displayed information subsets, interactive visual representations also grant users the authority to manipulate information presentation [54].

2.4.6 Visual Analytics

Visual analytics (VAs) is a multidisciplinary field that amalgamates analytics techniques with interactive visualization to facilitate insights from data [57]. VAs employs computational tools utilizing visual representations to enhance human cognition during data interaction [51]. Through a fusion of machine learning techniques, analytical processes, diverse visualizations, and various interaction mechanisms, VAs aids users in executing cognitive tasks [57]. These tasks encompass data-driven activities like decision-making, which involve analyzing, interpreting, comparing, and contrasting extensive data volumes. VAs empowers users to explore data at different levels of detail and abstraction, thus fostering enhanced comprehension. Its suitability for probing extensive datasets like electronic health records (EHRs) is attributed to its rapid data-to-visual mapping capability. By dynamically altering EHR data mapping, view, and scope through interaction, users can effectively fulfill their objectives.

VAs consists of two synergistic modules: analytics and interactive visualization [58]. The analytics module alleviates the user's cognitive burden in data-intensive tasks by merging machine learning with data processing techniques. Data mining algorithms and processing methods within the analytics module are tailored to specific domain requirements. Meanwhile, the interactive visualization module oversees the translation of data derived from the analytics module into visual components.

2.5 References

Bibliography

[1] Moon, J. & Galea, M. Overview of clinical decision support systems in healthcare. *Improving Health Management Through Clinical Decision Support Systems*. pp. 1-27 (2016)

[2] Garg, A., Adhikari, N., McDonald, H., Rosas-Arellano, M., Devereaux, P., Beyene, J., Sam, J. & Haynes, R. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama*. **293**, 1223-1238 (2005)

[3] Sutton, R., Pincock, D., Baumgart, D., Sadowski, D., Fedorak, R. & Kroeker, K. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*. **3**, 17 (2020)

[4] Walsh, S., Jong, E., Timmeren, J., Ibrahim, A., Compter, I., Peerlings, J., Sanduleanu, S., Refaee, T., Keek, S., Larue, R. & Others Decision support systems in oncology. *JCO Clinical Cancer Informatics*. **3** pp. 1-9 (2019)

[5] Mazo, C., Kearns, C., Mooney, C. & Gallagher, W. Clinical decision support systems in breast cancer: a systematic review. *Cancers*. **12**, 369 (2020)

[6] Hamedan, F., Orooji, A., Sanadgol, H. & Sheikhtaheri, A. Clinical decision support system to predict chronic kidney disease: A fuzzy expert system approach. *International Journal Of Medical Informatics*. **138** pp. 104134 (2020)

[7] Sanchez, E., Toro, C., Carrasco, E., Bonachela, P., Parra, C., Bueno, G. & Guijarro, F. A knowledge-based clinical decision support system for the diagnosis of Alzheimer disease.

2011 IEEE 13th International Conference On E-Health Networking, Applications And Services. pp. 351-357 (2011)

[8] Sim, L., Ban, K., Tan, T., Sethi, S. & Loh, T. Development of a clinical decision support system for diabetes care: A pilot study. *PloS One*. **12**, e0173021 (2017)

[9] Anooj, P. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal Of King Saud University-Computer And Information Sciences*. **24**, 27-40 (2012)

[10] Kessler, R. The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current Opinion In Psychiatry*. **31**, 32-39 (2018)

[11] Mokhles, S., Nuyttens, J., Mol, M., Aerts, J., Maat, A., Birim, Ö., Bogers, A. & Takkenberg, J. Treatment selection of early stage non-small cell lung cancer: the role of the patient in clinical decision making. *BMC Cancer*. **18** pp. 1-10 (2018)

[12] Yoon, J., Davtyan, C. & Schaar, M. Discovery and clinical decision support for personalized healthcare. *IEEE Journal Of Biomedical And Health Informatics*. **21**, 1133-1145 (2016)

[13] Rinott, R., Carmeli, B., Kent, C., Landau, D., Maman, Y., Rubin, Y. & Slonim, N. Prognostic data-driven clinical decision support-formulation and implications. *User Centred Networked Health Care*. pp. 140-144 (2011)

[14] Topaz, M., Trifilio, M., Maloney, D., Bar-Bachar, O. & Bowles, K. Improving patient prioritization during hospital-homecare transition: A pilot study of a clinical decision support tool. *Research In Nursing & Health*. **41**, 440-447 (2018)

[15] Antoniadi, A., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. & Mooney, C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*. **11**, 5088 (2021)

[16] Perez-Liebana, D., Samothrakis, S., Togelius, J., Schaul, T. & Lucas, S. General video game ai: Competition, challenges and opportunities. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **30** (2016)

[17] Le, Q., Miralles-Pechuán, L., Kulkarni, S., Su, J. & Boydell, O. An overview of deep learning in industry. *Data Analytics And AI*. pp. 65-98 (2020)

[18] Ng, G. & Leung, W. Strong artificial intelligence and consciousness. *Journal Of Artificial Intelligence And Consciousness*. **7**, 63-72 (2020)

[19] Gill, K. Artificial super intelligence: beyond rhetoric. *AI & Society*. **31** pp. 137-143 (2016)

[20] Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthcare Journal*. **6**, 94 (2019)

[21] Miller, D. & Brown, E. Artificial intelligence in medical practice: the question to the answer?. *The American Journal Of Medicine*. **131**, 129-133 (2018)

[22] Chen, M., Hao, Y., Hwang, K., Wang, L. & Wang, L. Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*. **5** pp. 8869-8879 (2017)

[23] An, K., Kim, M., Teplansky, K., Green, J., Campbell, T., Yunusova, Y., Heitzman, D. & Wang, J. Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks.. *Interspeech*. pp. 1913-1917 (2018)

[24] Fernandes, M., Mendes, R., Vieira, S., Leite, F., Palos, C., Johnson, A., Finkelstein, S., Horng, S. & Celi, L. Predicting Intensive Care Unit admission among patients presenting to the emergency department using machine learning and natural language processing. *PloS One*. **15**, e0229331 (2020)

[25] Shamout, F., Zhu, T. & Clifton, D. Machine learning for clinical outcome prediction. *IEEE Reviews In Biomedical Engineering*. **14** pp. 116-126 (2020)

[26] Dietvorst, B., Simmons, J. & Massey, C. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*. **64**, 1155-1170 (2018)

[27] Obermeyer, Z. & Emanuel, E. Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal Of Medicine*. **375**, 1216 (2016)

[28] Van Lent, M., Fisher, W. & Mancuso, M. An explainable artificial intelligence system for small-unit tactical behavior. *Proceedings Of The National Conference On Artificial Intelligence*. pp. 900-907 (2004)

[29] Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M. & Kagal, L. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference On Data Science And Advanced Analytics (DSAA)*. pp. 80-89 (2018)

[30] Lipton, Z. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.. *Queue*. **16**, 31-57 (2018)

[31] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. & Eckersley, P. Explainable machine learning in deployment. *Proceedings Of The 2020 Conference On Fairness, Accountability, And Transparency*. pp. 648-657 (2020)

[32] Bhatt, U., Andrus, M., Weller, A. & Xiang, A. Machine learning explainability for external stakeholders. *ArXiv Preprint ArXiv:2007.05408*. (2020)

[33] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. & Pedreschi, D. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*. **51**, 1-42 (2018)

[34] Ferryman, K. & Pitcan, M. Fairness in precision medicine. (Data & Society Research Institute,2018)

[35] Hense, H., Schulte, H., Löwel, H., Assmann, G. & Keil, U. Framingham risk function overestimates risk of coronary heart disease in men and women from Germany—results from the MONICA Augsburg and the PROCAM cohorts. *European Heart Journal*. **24**, 937-945 (2003)

[36] Adadi, A. & Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. **6** pp. 52138-52160 (2018)

[37] Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery*. **9**, e1312 (2019)

[38] Bussone, A., Stumpf, S. & O'Sullivan, D. The role of explanations on trust and reliance in clinical decision support systems. *2015 International Conference On Healthcare Informatics*. pp. 160-169 (2015)

[39] Thieme, A., Belgrave, D. & Doherty, G. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions On Computer-Human Interaction (TOCHI)*. **27**, 1-53 (2020)

[40] Carlisle, J. Evaluating the impact of office automation on top management communication. *Proceedings Of The June 7-10, 1976, National Computer Conference And Exposition*. pp. 611-616 (1976)

[41] Diederich, S., Brendel, A., Morana, S. & Kolbe, L. On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal Of The Association For Information Systems*. **23**, 96-138 (2022)

[42] Richardson, K. Models of cognitive development. (Psychology Press,2019)

[43] Komatsu, L. Recent views of conceptual structure.. *Psychological Bulletin*. **112**, 500 (1992)

[44] Nardi, B. Concepts of cognition and consciousness: Four voices. *ACM SIGDOC Asterisk Journal Of Computer Documentation*. **22**, 31-48 (1998)

[45] Hollan, J., Hutchins, E. & Kirsh, D. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions On Computer-Human Interaction (TOCHI)*. **7**, 174-196 (2000)

[46] Sedig, K., Parsons, P., Dittmer, M. & Haworth, R. Human-centered interactivity of visualization tools: Micro-and macro-level considerations. *Handbook Of Human Centric Visualization*. pp. 717-743 (2014)

[47] Sedig, K. & Parsons, P. Interaction design for complex cognitive activities with visual representations: A pattern-based approach. *AIS Transactions On Human-Computer Interaction*. **5**, 84-133 (2013)

[48] Benyon, D. Designing user experience. (Pearson UK,2019)

[49] Man, D. & Vision, A. A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman And Company, San Francisco*. **1** (1982)

[50] Karray, F., Alemzadeh, M., Abou Saleh, J. & Arab, M. Human-computer interaction: Overview on state of the art. *International Journal On Smart Sensing And Intelligent Systems*. **1**, 137 (2008)

[51] Sedig, K., Parsons, P., Dittmer, M. & Ola, O. Beyond information access: Support for complex cognitive activities in public health informatics tools. *Online Journal Of Public Health Informatics*. **4** (2012)

[52] Zhang, J. External representations in complex information processing tasks. *Encyclopedia Of Library And Information Science*. **68** pp. 164-180 (2000)

[53] Thomas, J. & Cook, K. A visual analytics agenda. *IEEE Computer Graphics And Applications*. **26**, 10-13 (2006)

[54] Sedig, K. Interactive Mathematical Visualisations: Frameworks, Tools and Studies. *Trends In Interactive Visualization: State-of-the-Art Survey*. pp. 343-363 (2009)

[55] Calvary, G., Coutaz, J., Thevenin, D., Limbourg, Q., Bouillon, L. & Vanderdonckt, J. A unifying reference framework for multi-target user interfaces. *Interacting With Computers*. **15**, 289-308 (2003)

[56] Yi, J., Kang, Y., Stasko, J. & Jacko, J. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions On Visualization And Computer Graphics*. **13**, 1224-1231 (2007)

[57] Cook, K. & Thomas, J. Illuminating the path: The research and development agenda for visual analytics. (Pacific Northwest National Lab.(PNNL), Richland, WA (United States),2005)

[58] Sedig, K., Parsons, P. & Babanski, A. Towards a characterization of interactivity in visual analytics.. *J. Multim. Process. Technol.*. **3**, 12-28 (2012)

Chapter 3

Explainable AI for Clinical Decision Support: Literature Review, Key Gaps, and Research Synthesis

3.1 Introduction

Clinical decision support (CDS) systems are computer applications whose goal is to facilitate the decision-making process of clinicians [1]. CDS systems help clinicians utilize data and modeling techniques to improve the quality of decisions, which in turn can improve the delivery of healthcare. CDS systems have a variety of applications, ranging from information management to patient-specific recommendations [2]. In recent years, CDS developers have become interested in applying Artificial intelligence (AI) to make clinical outcome predictions [3]. Machine learning (ML), as a subfield of AI, can learn from past experiences (such as patient history) and recognize useful patterns in health data. ML models receive data features as inputs, and based on the underlying patterns, provide a prediction output. A review paper on ML models by Montani and Strianti [4] suggests that such models are integral to CDS.

Although applying ML models in CDS systems may ultimately improve clinical outcomes, the black-box nature of these models limits their utility. Predictions made by ML models are often characterized by lack of interpretability [5]. As such, Rudin [6] argues that these predictive models should be avoided in high-stakes decisions, particularly in the medical domain where the interpretability and trustworthiness of a model are as important as its accuracy. Clinicians or patients cannot trust a prediction if they cannot understand the logic behind it [7]. As a result, the limited interpretability of many ML models is a major barrier to clinical adoption.

The challenges and barriers to the adoption of ML models have sparked research interest in interpretable and explainable AI, commonly known as XAI. This field has been developed to facilitate interpretability and address barriers related to accountability and trustworthiness. Different XAI methods are developed and used to construct and communicate explanations of the predictions made by ML models [8]. These XAI methods behave as “translator” for the models; the explanations they produce, when consistent with domain knowledge, are intended to help users develop a deeper understanding of the logic of the models and to mitigate their black-box nature. Ultimately, when warranted, the objective of such methods is to enable users understand and trust the models and their outputs [8, 9]. As such, XAI methods can have great potential to improve the implementation, utility, and adoption of AI in CDS systems.

The existing XAI methods show promise in helping make systems more accountable and transparent [10]. However, communication of explanations to users through the system interface is still an obstacle to the adoption of ML in clinical practice, resulting in a gap between tools and users. To accelerate the integration of XAI, it is crucial to reduce this gap by carefully considering the clinicians’ needs and task objectives in the design of CDS systems. In this chapter, we emphasize the rule of users in the design of XAI methods in clinical applications. The main contributions of this chapter are:

- Synthesizing different XAI methodologies through providing a comprehensive background literature of the subject

- Providing an organized overview of XAI applications in CDS systems

- Using the above to identify shortcomings in the existing body of research that contribute to the widening gap between CDS systems and clinicians

- Creating a framework for developing XAI methods whose aim is to reduce this gap

To the best of our knowledge, no studies have reviewed the development and application of XAI methods in CDS systems, though there are related studies. For instance, Holzinger et al. [11] provide an overview of the topic of explainable AI systems in the medical domain dealing with legal and privacy aspects; and, Nunes and Jannach [12] review the taxonomy of explanations in decision support to answer questions such as what the characteristics of an explanation are, how explanations can be generated, and how explanations can be evaluated. However, these studies do not describe the development or application of XAI methods. The objective of this chapter is to review and discuss issues surrounding XAI in CDS systems to help identifying challenges that have not been adequately addressed. In other words, our goal in reviewing XAI in the context of CDS systems is to find out what hinders the uptake of XAI methods in real-world applications. Due to the challenges involved in applying such methods in practical settings, there are few studies in this field. By reviewing these studies, however, and investigating the existing challenges, the gaps between clinicians and CDS systems can be highlighted. In this chapter, we propose a strategy to help bridge the gap.

The remainder of the chapter is organized as follows. Section II provides a structured overview of XAI methods surveyed from the literature. Section III presents the criteria that we have used for surveying and selecting the relevant literature. Using our structured overview of XAI methods as a lens, Section IV briefly reviews the application of XAI in CDS systems. Section V identifies the factors that contribute to gaps between clinicians and CDS systems and impede the application of XAI in CDS systems. Using human-computer interaction and cognitive task analysis concepts, Section VI discusses strategies and proposes a framework for bridging the existing gap. Finally, Section VII concludes this chapter and suggests specific future research directions.

3.2 Background

In this section, we provide a comprehensive review of different XAI methods. This review is in the context of a unified framework that illustrates the rationale for different methods. The goal is to provide a foundation for a clear understanding of the key aspects of the existing XAI methods. In addition, we provide a brief summary of distributed cognition which we will elaborate upon its role in the adoption of the CDS system by clinicians in a further section.

3.2.1 XAI

There are two general categories of XAI methods: ante-hoc and post-hoc [20]. Ante-hoc (or explainable modeling) methods are specifically designed models to be transparent or "glass box" whose logic is understandable to end users. Examples of these methods are RuleFit, Additive Models, GAMS, Fuzzy, Decision Tree, and Linear Regression (references for each method). Post-hoc (or post-modeling) methods are used to explain existing black box models whose logic is not understandable to end users.

Since post-hoc methods can be applied to already trained models, they have wider applicability. Post-hoc methods may be categorized in terms of three dimensions: specificity of models, scope of explanation, and type of explanation. In terms of the specificity of models, XAI methods can be divided into two major categories: model-specific and model-agnostic. Model-specific methods try to explain what occurs within a specific machine learning model–i.e., they are limited to one particular machine learning model. For instance, Layer-wise Relevance Propagation (LRP) is a model-specific explanation method for neural networks (ref). Model-agnostic methods are not tied to a specific model and try to explain a prediction without referring to what happens inside the model. As model-agnostic methods deal with models built by different machine learning algorithms, they are popular.

In terms of scope of explanation, XAI methods may be categorized as: global and local. The two dimensions of specificity and scope are orthogonal–that is, XAI methods can be: model-specific and global, model-specific and local, model-agnostic and global, or model-agnostic and local. Global methods explain the logic of and reasoning behind a model for a whole dataset, whereas local methods explain the reason for a specific prediction made for a single data point.

In terms of type of explanation, XAI methods may be categorized as: simplification, influence, and example-based. To categorize these methods, we have adapted categorization used by Adadi and Berrada [20], and Guidotti et al. [14] (see Figure 1). These types either present information about the inner workings of a model and the logic behind its predictions or expose a model's input-output relationships. These types of explanation methods can be used

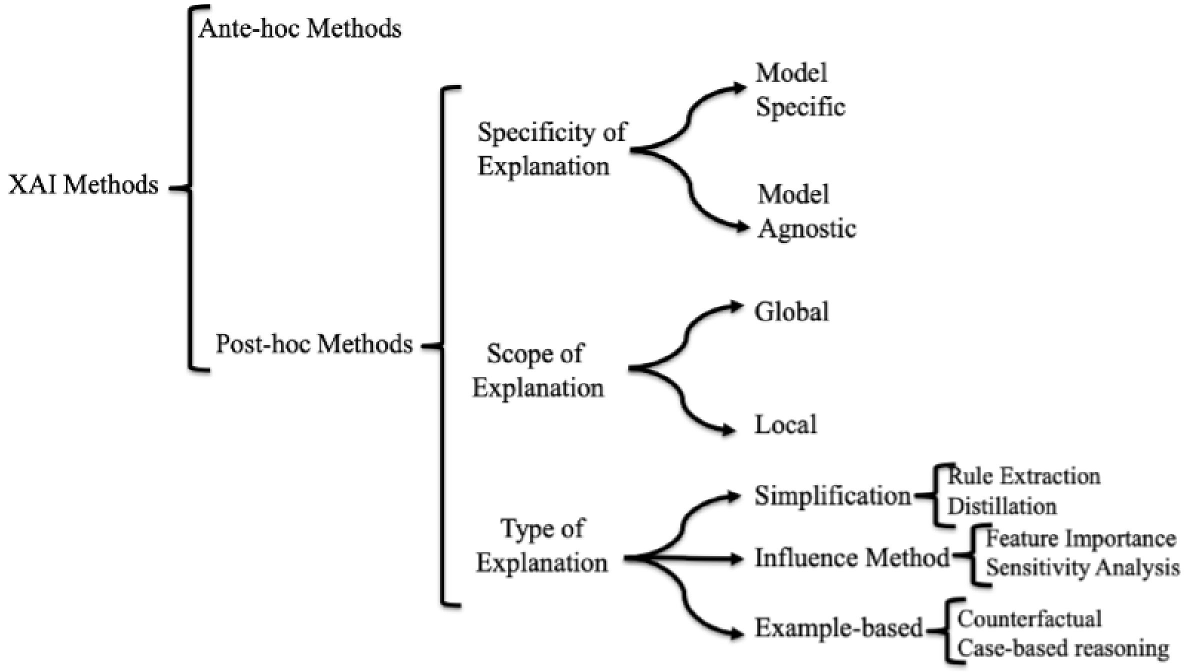


Figure 3.1: Division of Post-hoc Methods Based on Specificity, Scope, and Type of explanation.

individually or in combination.

1. Simplification: Strategies that fall into this category simplify the model so that it is understandable to the end-user, which provides insight into complex models. These strategies produce knowledge that approximates the process of making a prediction using the complex model and represent the digested information.

   - Knowledge Extraction: This strategy extracts and represents a simplified and symbolic description of knowledge acquired by a machine learning model during training. It has two subcategories:

     - Distillation: This strategy distills and compresses a machine model to make it simpler and therefore easier for humans to understand [15, 16, 17].
     - Rule Extraction: This strategy uses rule extraction to generate explanations [18, 19, 20, 21, 22]. A symbolic description of the knowledge learned by the model is extracted during training and these extracted rules approximate

the model's decision-making process. The interpretability of this strategy decreases as the number of features in the ML model increases, and therefore is most effective for models with fewer features.

- Surrogate Building: This strategy explains a complex model with a simple and interpretable model such as a linear model or decision tree. The interpretable model is trained on the predictions of the original complex model in order to explain it. Local Interpretable Model-agnostic Explanations (LIME) is one example of a surrogate method that explains individual predictions [23]. One of the drawbacks of this approach is that the surrogate model can be very close to the original complex model for one subset of the data, but very divergent for another subset; therefore, it may only provide useful explanations for local rather than global predictions [24].

2. Influence Assessment: This strategy determines and communicates the importance of each feature used in the model in terms of its influence on the value of a prediction. This strategy determines which feature values have led to this decision and allows decision-makers to compare the model decision to their own judgment; this is very important in case of any mismatch between these two. This strategy measures the importance of a feature by changing the input or internal component which can be further subdivided into three classes:

   - Sensitivity Analysis: This refers to how much the input can affect the output. This strategy is used to verify a model behavior and stability [25]
   - Feature Importance: it quantifies the importance and contribution of each feature to making a prediction [26]. This measurement is done with three different following strategies:
     - Perturbation/permuting-based: The effect of omitting or changing the value of an input feature is examined. Features are perturbed and their effects on the output are recorded. The feature that has the biggest effect is considered the most important [27]. The main drawback of this method is that perturbation of features can result in unrealistic data instances when two or more features are correlated. For example, two features of height and weight are correlated, and

changing just one of them may lead to an unrealistic data sample (height:170cm and weight:30kg). Using these instances to measure the importance generates misleading results. In other words, the characteristics of the model are measured with values we would never observe in reality.

  - SHAP: SHAP (SHapley Additive exPlanations) method [28, 6] uses a cooperative game theory approach to explain the machine learning model predictions with SHAP values which are calculated as a weighted average of features' marginal contribution. This strategy is most often used for a local explanation, but it also approximates a global solution using a means SHAP values metric.
  - Saliency Mask: This strategy visually highlighted what causes a certain outcome. It is generally used to explain deep neural networks treating images or text [30, 31].

- Neuron Contribution: This strategy is model-specific and used for neural networks. Two different strategies are used to calculate the importance or contribution of each neuron.

  - Propagation-based: This strategy determines the contribution of each input by back-propagating a quantity of interest through a neural network. Layer-wise Relevance Propagation (LRP), discussed earlier, is a popular method that uses this strategy [32, 33, 34].
  - Activation Maximization: This strategy is model specific and is applicable to neural networks. It looks for input patterns that maximize the activation of neurons. For this purpose, it inspects which neuron is activated with respect to a particular input [35].

3. Example Based: This strategy explains the model behavior using particular instances. Most example-based strategies are model agnostic. We identified three strategies:

   - Prototype/Criticism: This approach selects instances from the data which are intended to represent the overall data set. To avoid overgeneralization, "critics" or rare instances that are not represented by prototypes are identified [36].

- Counterfactual: In the counterfactual explanation method, the minimum change in input data for one specific feature that would have led to a different result is determined [37, 24].
- Adversarial: The adversarial method tries to reverse the prediction with small feature perturbations, and then shows the perturbations to the user. In other words, adversarial examples are counterfactual examples with the aim of causing a model to make an incorrect prediction [39]. These types of explanations help humans to understand the data distribution.
- Case-based reasoning:

Type of Explanation	Simplification	Knowledge extraction	Rule Extraction
			Distillation Methods
		Surrogate models	LIME
	Influence Method	Sensitivity Analysis	PDP, ICE
		Feature Importance	Perturbation/Permutation
			Shapley
			Saliency
		Neuron Contribution	LLC-backward propagation
			Activation Maximization
	Example based	Prototype/Criticism	
		Counterfactual	
		Adversarial	
		Case-based reasoning	

Table 3.1: Different type of explanation for Post-hoc methods.

3.2.2 Joint Cognitive Systems

A decision-making process in the presence of complicated conditions involves complex cognitive activities [40]. Cognitive activities, according to the theory of distributed cognition, are not only taking place in an individual's brain but also involve the external environment [41]. This distribution occurs over time and is the result of interactions among internal factors, such as analytical reasoning ability and background knowledge, and external resources, such as the computational power of computers and representations of information. In this way, external environments assist the cognitive system, become coupled with it, and can even be extended by it. When a clinician uses a CDS system to perform complex cognitive activities (decision making), they form a joint cognitive system. In this joint cognitive system, cognitive activities are the result of a coupling that takes place between the clinician's cognitive system and the external representation of the CDS system [42]. In this way, the clinician and the CDS system share information processing required to perform complex cognitive activities. In fact, how effectively a complex cognitive activity is performed is largely influenced by both the characteristics of the user [1] and the CDS systems, as well as the strength of the coupling between them. The coupling between a user and a tool can be weak or strong. With weak coupling, external aids do not actively contribute to the information processing [41].

3.3 Review Methodology

A literature search was conducted to collect research papers describing the development or application of XAI methodology in the context of CDS systems. A set of relevant keywords indicating research in XAI and CDS systems was developed and used to search several databases; these are shown in Table 1. All possible combinations of keywords in the two columns were used including explainable AI in clinical decision support. The search engines PubMed, IEEE, ACM, and Google Scholar were used, restricting to peer-reviewed papers, preprints, and gray literature publications between 2010 and 2020. Duplicates were removed and the abstracts of those remaining were reviewed to assess whether they address the topic of explainable AI in clinical decision support; papers not addressing this topic were excluded. For example, papers that only discussed decision support systems in general or that did not discuss applications in health were excluded. Furthermore, studies that were exclusively focused on public health, genomic, administrative data, and guidelines were excluded since our focus is on systems de-

[1] By user, we mean clinician and these two words are used interchangeably in this chapter.

ployed in clinical care. The reference lists of the retrieved papers were manually reviewed to find additional relevant papers that were not retrieved in the original search; these were subjected to the same screening procedure. Duplicates were removed and the abstracts of those remaining were reviewed to assess whether they address the topic of explainable AI in clinical decision support; papers not addressing this topic were excluded. For example, papers that only discussed decision support systems in general or that did not discuss applications in health were excluded. Furthermore, studies that were exclusively focused on public health, genomic, administrative data, and guidelines were excluded since our focus is on systems deployed in clinical care. The reference lists of the retrieved papers were manually reviewed to find additional relevant papers that were not retrieved in the original search; these were subjected to the same screening procedure.

Table 3.2: Keywords used in the search engine.

Keywords used in the search engine	
Explainable AI	Clinical Decision Support
Interpretable AI	CDS SYSTEMS
Interpretable Machine Learning	Clinical Decision Support Systems
Black box	Clinical Decision Support Tools
Interpretable algorithm	Clinical
Explainable algorithm	Medical

Having described the landscape of XAI methods, we next present a review of the literature that addresses the application of XAI methods in CDS systems.

3.4 Results

This section describes the literature that was retrieved as a result of our search. In total, we found 232 papers that included both keywords from both columns in table 3.2. Specifically, we identified papers that had a combination of keywords from the first column (for example, ”XAI” or ”Explainable AI”) and keywords from the second column (such as ”clinical” or ”Clinical

Decision Support"). The review of reference lists of obtained papers added 29 additional papers for a total of 261. After title screening and removing duplicate papers, 89 articles were excluded, leaving 172 for abstract screening. Following the abstract screening, a total of 14 papers met the final inclusion criteria (deployment of XAI in clinical setting) considered for full-text screening and in-depth analysis. One paper was not available in its full text, which left us with 13 papers. We note that two papers [43, 44] out of the total papers retrieved described a deployment of XAI in real-world CDS systems and the remaining papers are about the potential deployment of XAI in clinical or medical domain. Finally, of the 13 papers, four papers identified the intended users of the explanation, and none discussed the potential for use by patients.

In the following, we summarize the key articles that we found, considering their importance in providing useful explanations for clinicians that allow them to appropriately trust the decisions made by the machine learning models. We group the articles according to their XAI methodology based on the categories we identified in the background section. This allows us to examine which categories of XAI methodology have been investigated the most for use in CDS systems. Four out of the 13 papers used Ante-hoc methods, while the rest used Post-hoc methods.

- Ante-hoc: An interpretable predictive model called REverse Time AttentIoN model (RETAIN) was developed for risk prediction of heart failure. The model consists of a two-level neural attention network to identify influential past visits as well as relevant clinical variables in those visits. The contribution of variables for the diagnosis of heart failure for an individual person per visit was visualized using a diagram [45] and an interactive visualization [46]. Case-Based Reasoning (CBR) method was used to retrieve similar cases to the query (an individual patient) from the database. Using a visual interface, the query is compared to similar cases in quantitative and qualitative terms. The proposed approach was applied to a real dataset in breast cancer [44]. An incremental explanation of inference that can be applied to the Bayesian network was proposed to identify important evidence supporting or contradicting the prediction for an individual. A real clinical case study was used to illustrate the explanation [47].

- Post hoc: Different types of strategy have been employed for different applications, as

Table 3.3: Summary of Applications and Explanatory Techniques

Application	**Ante/post-hoc**	**ML model**	**Model specific/agnostic**	**Type of explanation (Strategy)**	**Local/global**	**User interface**	**Evaluation**
Risk prediction of heart failure [45]	Ante hoc	RETAIN	NA	NA	Local	Diagram	×
Risk prediction of heart failure [46]	Ante hoc	RETAIN	NA	NA	Local	Interactive visualization	×
Breast cancer management [44]	Ante hoc	KNN	NA	NA	Local	Interactive Rainbow boxes	✓
Predict Coagulopathy [47]	Ante hoc	Bayesian networks	NA	NA	Local	Text	✓
Antibiotic prescription [43]	Post hoc	Preference Learning	specific	Feature importance	Global	Rainbow boxes	✓
Stroke outcome prediction [48]	Post hoc	MLP	specific	Feature importance	Global	Graphical representation of features	×
Survival time [53]	Post hoc	Neural network	Specific	LRP	Local	Table of relevance score	×
Alzheimer disease detection [54]	Post hoc	Neural network	Specific	LRP	Local	Heatmap feature importance	×
Clinical gate classification [55]	Post hoc	Neural Network	Specific	LRP	Local	Score relevance on signal	×
Predictive therapy in breast cancer [49]	Post hoc	RNN	Specific	Knowledge extraction	Local	Distance matrix	×
Breast cancer detection [50]	Post hoc	Trained classifiers	Agnostic	Knowledge extraction	Local/Global	Interactive visualization	✓
Ventilator free days prediction [51]	Post hoc	Gradient boosting tree	Specific	Knowledge distillation	Global	Feature importance score	×
ICU Mortality [52]	Post hoc	Random Forest	Agnostic	LIME	Local	Feature importance score	×

illustrated in the following:

1. Influence based: Feature importance strategy was used to provide global explanations for preference learning [43] and MLP model [48] for antibiotic prescription and stroke outcome prediction, respectively. The preference model was visualized using a rainbow box integrated into a CDSS called AntibioHelp. In this study, an evaluation of the learning process was performed. The important features of stroke prediction were explained through graphical representation. The neuron contribution strategy was applied to the neural network to provide local explanations in three different clinical applications of survival time prediction, Alzheimer's disease detection, and clinical gate classification. For the explanation, a table of relevance scores was used for survival time, a heatmap that highlighted important Alzheimer's disease features, and a graph illustrating score relevance was used for the gate signal.
2. Knowledge extraction strategies were applied in predictive therapy in breast cancer [49], breast cancer detection [50], and ventilator-free days prediction [51].
3. Simplification strategy has been used for ICU mortality prediction and the feature importance score of each prediction was recognized [52].

Table 3 summarizes the XAI methods used in different clinical applications. Along with describing the characteristics of XAI methods, the table provides information regarding the ML model that was used, how the explanation was represented (interface), and whether the evaluation of interpretability was considered.

3.5 The CDS-Clinician Gap

The main purpose of XAI is to bridge the gap between clinicians and the CDS system. We discuss three factors that play an important role in the existing gap between clinicians and the CSD system. To create a better CDS-clinician partnership, these three factors must be considered carefully. These factors contribute to the degree of coupling between the clinician and the CDS system. We will elaborate on these factors next.

1. Clinicians' needs and abilities

   Based on our review, existing work on XAI in CDS systems does not choose XAI methodology based on user characteristics. This represents a missed opportunity, since different XAI methods may be used to tailor CDS systems to the needs and reasoning processes of different users. The main goal of XAI is to explain the model to users. Users will be able to trust the model only if they understand how its outputs are produced. However, studies show that XAI methods are not designed based on users' needs; rather they are based on developers' intuition of what a good explanation is [14]. In this regard, for producing an explanation, it is useful to consider how explanations have been defined in different fields, such as philosophy and psychology. These fields define an explanation as a conversation or interaction for the purpose of transferring knowledge, implying that the explainer must leverage the understanding of the person receiving the explanation to improve that understanding. Different people have different types of reasoning to understand and receive a piece of information differently. Their reasoning process and how they perceive information are different. Therefore, accommodating the user is a key determinant in designing a useful, effective, and practical XAI. This issue of not considering end users can result in models that end users do not wish to use; consequently, ignoring the users is a major flaw in model design. Designing an explanation model in isolation of the decision that a user is likely to make does not lead to effective support. Decision-making is information-intensive and in the presence of complex conditions involves complex human cognition. As discussed in the background section, according to distributed cognition theory, cognitive activities not only take place in an individual's brain but also involve the external environment. This collaboration occurs over time and is the result of interactions among internal factors, such as the analytical reasoning process and background knowledge, and external resources, such as the computational and representational power of computers. Therefore, to design an explanation, the reasoning processes of users must be considered.

2. Interface

   The interface that conveys explanations to users has received little attention. Only two papers out of 13 focus on the design of the interface component of CDS systems. Indeed, most of the literature focuses mainly on the XAI methods, not on designing interfaces that communicate machine learning explanations to users. The human-computer interface plays an essential role in transferring the external representation of information to the internal mental conceptualizations of humans [57]. However, there is a gap between how information is represented externally and how humans conceptualize this information. The goal of an interface is to decrease the gap between these two. The more human-centered the interface is, the smaller the gap. A human-centered explanation interface can enable users to understand and interact with CDS systems and ultimately gain trust in the system. However, the main concern of developers is the XAI method itself rather than communicating the explanation to the users. Communicating a complex process such as an explanation to users, in addition to having knowledge of machine learning, requires Human-Computer Interaction (HCI) expertise as well. The visualization component of a human-computer interface makes information perceptible to users by encoding information and giving a tangible form to it. The quality of the visualization impacts the quality of the perception; for example, multiple tables or a huge decision tree used for explanations are not helpful in quickly perceiving the required information. A representation that uses complex visualization can be confusing for users. Furthermore, representing different types of explanations on a single screen, while visually interesting, can be confusing and overwhelming if there is too much information for the user. Indeed, the majority of explanations provided by XAI methods are static. Many researchers have traditionally assumed that explanations are a single message to be conveyed. However, users need to interact with the interface and iteratively explore large amounts of complex information and glean insights, with the result that the information is more willingly and accurately received. In fact, interactions give the user the ability to modify and control the amount of represented information in order to accomplish a complex activity such as decision-making. Although interaction has a key role in conveying the information, how the interactions can be designed in a human-centered way that involves cognitive activity tasks has not been discussed in the literature.

3. Evaluation

   Evaluation of XAI methods in the context of CDS systems has received scant attention. Therefore, the degree to which the provided explanation is understandable to users is unknown. It is clear that there is a great need to evaluate the interpretability and explainability of XAI methods so that their usability and adaptability can be assured. Despite the increasing body of research on XAI methods, relatively few works have addressed evaluating these methods and assessing their relevance. Only 4 papers out of 12 conducted the evaluation in their studies. Furthermore, only one study we found has investigated whether explanation can help users to trust the system and use the model's prediction in their decision-making [46]. This implies that there is a great need for evaluation so that we be able to validate, compare, quantify, and evaluate different explanations [20].

3.6 Discussion and Framework Proposal

Clinical Decision Support Systems (CDS systems) hold significant promise in enhancing healthcare outcomes by assisting healthcare providers in making well-informed decisions. However, the adoption of Explainable Artificial Intelligence (XAI) in CDS systems faces several challenges, including accommodating the needs and capacities of end users, employing interactive visualization techniques, and appraising the effectiveness of XAI methods in delivering satisfactory explanations. As discussed earlier, historically, XAI methods within CDS systems have been developed without a specific focus on human-computer interactions. To bridge this gap, this proposal introduces a comprehensive framework that advocates for user involvement throughout the entire lifecycle of XAI algorithm development, interface design, and evaluation.

- Phase 1- User-oriented XAI method application:

  To ensure active user involvement in the developmental process of Explainable Artificial Intelligence (XAI), the early engagement of users becomes imperative, particularly in the selection of suitable XAI methods. The choice of an appropriate XAI method is pivotal, as it entails aligning the selected method's explanations with users' intended reasoning objectives. This requires following an approach that focuses on users' needs and how humans naturally think and reason.

As discussed in the background section, there are plenty of methods that provide explanations with different rationales. It is important to know which reasoning method is more suitable and acceptable for the potential users. To select appropriate explanations based on various XAI methods, we adopted Wang et al.'s [58] framework, which is based on human reasoning. This framework explains how to decide which XAI methods can satisfy users' reasoning goals. Based on this framework, investigating how users seek explanations and reason for a specific task articulates explanation types that satisfy users' reasoning goals. This investigation can be conducted by interviewing potential users and asking questions such as the following: "What kind of explanation would you expect from a colleague if they provide you with information that helps you make a decision?". After that, linking the connection between how users reason and how an XAI method supports reasoning methods enables us to select appropriate explanations and justify whether XAI could be useful.

Once we have gathered the necessary information, we can establish a link between how users naturally reason and how a specific XAI method supports those reasoning processes. This will enable us to select appropriate explanations that align with users' cognitive abilities and reasoning preferences. Justifying whether XAI could be useful for specific user groups will be a crucial aspect of this phase. The insights gained from this phase will inform the subsequent steps in the XAI development process, ensuring that the system is designed to meet the specific needs and preferences of end users.

The proposed guideline for user-involved XAI development can integrate with the following principles to enhance the design and implementation of the system within the healthcare domain:

1. **User-Centered Design Approach:** Engage end users, including healthcare professionals (e.g., doctors, nurses), patients, and other stakeholders, throughout the XAI development process. Employ techniques such as participatory design and focus groups to understand their needs, expectations, and cognitive abilities concerning explanations in the Clinical Decision Support System (CDSS). Healthcare providers

often have different levels of expertise and understanding of AI models. Some may require detailed technical insights, while others may prefer simplified explanations that emphasize clinical implications. By involving end users from the beginning, researchers can create an XAI system that aligns with the users' preferences and enhances their decision-making process.

2. **Customizable Explanation Level:** Offer different explanation granularity levels based on users' expertise and preferences. Provide simplified explanations for non-experts and more detailed insights for experienced medical professionals. For example, a CDSS aimed at radiologists might offer explanations in the form of heatmaps, highlighting regions of interest in medical images. Simultaneously, explanations for primary care physicians might be presented in plain language, focusing on the key factors influencing the model's prediction.

3. **User Profiling:** Develop user profiles based on their experience, domain knowledge, and preferred interaction styles to personalize the explanation delivery. Understanding the users' backgrounds, familiarity with AI, and specific needs enables tailoring the XAI system to their unique requirements. For instance, novice users might benefit from step-by-step explanations, while expert users could be presented with comprehensive visualizations for deeper exploration.

- Phase 2- User-oriented Interface Design:

  In order to achieve user-oriented interfaces, we have drawn from the rich body of research on Interaction Design, Context-awareness, Software Learnability in HCI as Xie and Gao [58] point out that these concepts are essential for designing XAI. For a context-aware system, users need to know which actions will be performed by systems and what they are going to do next, and so software learnability is strongly connected with the ease of use of a system. Recommendations for designing a visualization to represent an explanation for clinicians in order to follow this generated advice in their decision-making processes are provided below.

  In light of the factors contribute to the gap, to form a joint cognitive system between a system and a human, more focus should be put on human issues such as cognition, perception, reasoning, and insight, all of which are recognized as components of human visual perception [59]. Our first recommendation, based on Parsons and Sedig's study [42], is to consider the context and activity and how the CDS systems can be used during design. To have a contextual design, we need to answer the question "In what tasks and activities might the clinician engage?". Afterward, visual representation properties need to be discussed with users. These properties include appearance, complexity, dynamism, fidelity, fragmentation, interiority, scope, and type, all of which Parsons and Sedig [60] explain in detail.

Furthermore, how users interact with an explanation affects their receptiveness to that explanation. If the users' mental model is different from what the system shows, those users are likely to be confused about the decision-making process when interacting with such systems [61]. Indeed, to ensure that tools are effective, focusing on only visual perception is not adequate, and systematic thinking about interaction design is required. Interactivity is defined at two levels: the micro and macro [62]. The micro-level is concerned with individual interactions whereas the macro-level is concerned with how different interactions are combined to perform tasks. For the micro-level, there is a comprehensive interaction catalog [41] so one can refer to this list to perform a systematic design of interactivity in the visual representation. For interactivity at the macro level, designing strategies of interaction and helping users predict system behavior requires thinking broadly. We need to consider the number, diversity, and types of transactions that are available to the users, the relationships among them, and more importantly the users' needs. Asking questions such as "how does interactivity facilitate their decision-making process?", "which actions should be supported in CDS systems?" and "what properties should be adjustable for clinicians?" can be helpful in the design process.

By incorporating context-aware design, software learnability, and interactivity design, the user-oriented interface will foster a joint cognitive system between the XAI system and the human user. This phase will ensure that the interface aligns with users' needs, enhances their decision-making process, and improves their receptiveness to the explanations provided by the XAI system. The insights gained from this phase will inform the subsequent steps in the XAI development process, creating an effective and user-friendly system that enhances transparency and trust in AI-driven decision-making.

These guidelines are proposed to facilitate a user-centered interactive visualization:

1. **Interactivity:** Facilitate immediate exploration of the model's behavior and predictions using interactive visualization tools. Provide users with the capability to engage the Clinical Decision Support System (CDSS) through hypothetical scenarios, enabling a more profound grasp of the decision-making process.

   Interactive visualizations empower users to experiment with "what-if" scenarios, adjusting input variables and observing the corresponding responses of the model's predictions. This approach yields valuable insights into the model's behavior, bolstering trust by empowering users to authenticate the system's recommendations.

2. **Multi-modal Explanations:** Utilize various visual representations (e.g., heatmaps, line graphs, bar charts) to present different aspects of the model's reasoning and support a more comprehensive understanding.

   Different users may comprehend information better through different modalities. By offering multiple visualizations, researchers can cater to diverse user preferences and enhance the overall comprehensibility of the eXplainable AI (XAI) system.

3. **Interactive Explanations:** Allow users to interact with the visualizations to control the explanation output, adjust parameters, and explore alternative decision pathways.

   Enabling users to interact with the explanations fosters a sense of ownership and control over the decision-making process. They can manipulate visualization elements to gain deeper insights, leading to a more engaged and informed use of the CDSS.

- Phase 3- Evaluation:

In order to develop a user-centered explanation that is effective, it's crucial to have a clear methodology for evaluating its impact. This evaluation process is heavily dependent on the specific context in which the visualization environment is applied, as highlighted by

Hundhausen [63]. Understanding the exact task that requires support becomes a critical first step. Moreover, beyond comprehending the tasks and context, it is essential to define what is meant by “effectiveness.” As Hundhausen [63] suggests, effectiveness within a visualization environment can be operationalized by aligning with the objectives of that environment. The initial stage of evaluation involves translating effectiveness into measurable terms.

Despite the comprehensive guidelines outlined by Freitas et al. [64] for evaluating visualizations, they emphasize the persistent challenge of conducting user-centered usability assessments. Additionally, existing literature points out that the evaluation process primarily revolves around an intuitive sense of what constitutes a ’quality’ explanation [14], rather than focusing on how well users (clinicians) integrate these explanations into their decision-making. To tackle this, it’s crucial to establish well-defined evaluation criteria that take into account user needs and system characteristics, gaining consensus from both researchers and clinicians.

To put these principles into practical application, we suggest the following guidelines:

1. **User Satisfaction Surveys:** Engage in user satisfaction surveys and interviews to collect feedback from end users. The aim is to gauge the clarity, usefulness, and trustworthiness of the explanations provided by the XAI system. Recognizing the significance of user satisfaction in the success of XAI systems, it’s important to regularly gather input from healthcare professionals and patients. This iterative process aids in identifying areas for enhancement and aligning the XAI system with user expectations.

2. **Diagnostic Accuracy with Explanations:** Conduct a comparative assessment of the diagnostic accuracy of healthcare professionals utilizing the CDS System with and without explanations. The objective is to discern whether the inclusion of explanations through XAI positively impacts decision-making. By evaluating how the CDS system performs both with and without explanations, we gain insights into the tangible influence of XAI on clinical outcomes. This analysis also serves to verify that the integration of explanations does not compromise the accuracy of

the model.

3. **Comprehensibility Metrics:** Create quantifiable metrics to evaluate the comprehensibility of explanations, employing measures like the Flesch-Kincaid readability score [65] for explanations in text format. The purpose of these metrics is to assess whether the conveyed information can be easily understood by the intended users, irrespective of their level of expertise. Ensuring that explanations are comprehensible contributes to effective communication between the XAI system and its users.

Through the integration of these assessment methodologies, our intention is to develop an XAI system that resonates with the needs and expectations of healthcare professionals and other stakeholders. By capturing user satisfaction, evaluating diagnostic accuracy, and quantifying comprehensibility, we can achieve a more refined and user-friendly XAI system that makes a positive impact on decision-making processes. These guidelines, informed by research and practical insights, guide us toward tangible outcomes in the realm of Explainable AI.

By following this proposed framework, researchers can contribute to the development of more effective and user-friendly XAI methods for Clinical Decision Support Systems. The integration of end users' needs and abilities, interactive visualization, and robust evaluation processes will pave the way for enhanced transparency and trust in AI-driven clinical decision-making, ultimately leading to improved patient care and outcomes. The application of XAI in CDS systems will bridge the gap between the "black-box" nature of AI models and the need for interpretable, actionable explanations in the healthcare domain.

In addition, Uncertainty Quantification, Bias and Fairness Analysis, and Continuous Improvement and User Feedback are vital considerations in XAI design. By integrating these additional aspects into the framework, researchers and developers can create more responsible and effective XAI systems for CDS systems. Uncertainty Quantification involves offering metrics of uncertainty for the model's predictions, and notifying users when the CDS system might lack certainty or adequate data for confident decisions. This process assists in quantifying and conveying the confidence and reliability of AI-generated recommendations, enabling healthcare providers to make more informed decisions by relying on the level of certainty provided by the system. Bias and fairness analysis is essential to prevent XAI algorithms from perpetuating biases, thereby fostering impartial healthcare outcomes across diverse patient populations. It involves assessing both the data and models utilized within the CDS system to pinpoint potential biases that could disproportionately impact different patient groups. This process focuses on mitigating and rectifying biases to ensure equitable treatment and unbiased healthcare recommendations. Furthermore, integrating continuous improvement mechanisms and User Feedback loops empowers ongoing refinement of the XAI system, making it adaptable to changing medical landscapes and user needs. This iterative process not only enhances system performance but also fosters transparency and accountability in the decision-making process. Addressing ethical concerns, ensuring model reliability, and enabling continuous improvement through user feedback will contribute to building a transparent, trustworthy, and impactful CDS system with Explainable AI, ultimately leading to better healthcare outcomes for patients and healthcare providers.

3.7 Conclusion

Approaches that leverage XAI to provide an explanation in the context of clinical decision support systems are promising and gaining attention although they are in the early stages of their maturity. Our overview of the existing XAI methods based on the literature analysis of 13 papers identified the factors contributing to the gap regarding the design of XAI systems in the clinical context. We proposed a framework to bridge the gap between CDS systems and users and to represent how XAI methods should be fitted into a cognitive activity such as decision-making. We also argued that more interdisciplinary research teams, including clinicians, physicians, cognitive scientists, and computer scientists, are needed to advance XAI so that it can be used in CDS systems by producing user-centered explanations.

There are several opportunities for future work in this area. For example, in response to the provided explanations, clinicians could provide feedback on the quality of any recommended decisions, which may help to detect bias in the model and improve its performance. There is also significant potential for XAI-based CDS systems to support shared decision-making by enabling patients to more easily engage in the process of making decisions that impact their health. Designing tailored XAI interfaces for different categories of users of the same tool (e.g., patients, clinicians, researchers) will further increase the importance of user-centered design as described in our framework.[1]

3.8 References

Bibliography

[1] Berner, E. & La Lande, T. Overview of clinical decision support systems. *Clinical Decision Support Systems: Theory And Practice*. pp. 1-17 (2016)

[2] Kubben, P., Dumontier, M. & Dekker, A. Fundamentals of clinical data science. (Springer Nature,2019)

[3] Shamout, F., Zhu, T. & Clifton, D. Machine learning for clinical outcome prediction. *IEEE Reviews In Biomedical Engineering*. **14** pp. 116-126 (2020)

[4] Montani, S. & Striani, M. Artificial intelligence in clinical decision support: a focused literature survey. *Yearbook Of Medical Informatics*. **28**, 120-127 (2019)

[5] Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M. & Kagal, L. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference On Data Science And Advanced Analytics (DSAA)*. pp. 80-89 (2018)

[6] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. **1**, 206-215 (2019)

[7] Tonekaboni, S., Joshi, S., McCradden, M. & Goldenberg, A. What clinicians want: contextualizing explainable machine learning for clinical end use. *Machine Learning For Healthcare Conference*. pp. 359-380 (2019)

[8] Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. & Others Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. **58** pp. 82-115 (2020)

[9] Tjoa, E. & Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions On Neural Networks And Learning Systems*. **32**, 4793-4813 (2020)

[10] Bunn, J. Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI). *Records Management Journal*. **30**, 143-153 (2020)

[11] Holzinger, A., Biemann, C., Pattichis, C. & Kell, D. What do we need to build explainable AI systems for the medical domain?. *ArXiv Preprint ArXiv:1712.09923*. (2017)

[12] Nunes, I. & Jannach, D. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling And User-Adapted Interaction*. **27** pp. 393-444 (2017)

[13] Adadi, A. & Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. **6** pp. 52138-52160 (2018)

[14] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. & Pedreschi, D. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*. **51**, 1-42 (2018)

[15] Tan, S., Caruana, R., Hooker, G. & Lou, Y. Distill-and-compare: Auditing black-box models using transparent model distillation. *Proceedings Of The 2018 AAAI/ACM Conference On AI, Ethics, And Society*. pp. 303-310 (2018)

[16] Xu, K., Park, D., Yi, C. & Sutton, C. Interpreting deep classifier by visual distillation of dark knowledge. *ArXiv Preprint ArXiv:1803.04042*. (2018)

[17] Che, Z., Purushotham, S., Khemani, R. & Liu, Y. Distilling knowledge from deep networks with applications to healthcare domain. *ArXiv Preprint ArXiv:1512.03542*. (2015)

[18] Tickle, A., Andrews, R., Golea, M. & Diederich, J. The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions On Neural Networks*. **9**, 1057-1068 (1998)

[19] Su, C. & Chen, Y. Rule extraction algorithm from support vector machines and its application to credit screening. *Soft Computing*. **16** pp. 645-658 (2012)

[20] De Fortuny, E. & Martens, D. Active learning-based pedagogical rule extraction. *IEEE Transactions On Neural Networks And Learning Systems*. **26**, 2664-2677 (2015)

[21] Bologna, G. & Hayashi, Y. A rule extraction study from svm on sentiment analysis. *Big Data And Cognitive Computing*. **2**, 6 (2018)

[22] Hailesilassie, T. Rule extraction algorithm for deep neural networks: A review. *ArXiv Preprint ArXiv:1610.05267*. (2016)

[23] Ribeiro, M., Singh, S. & Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*. pp. 1135-1144 (2016)

[24] Magesh, P., Myloth, R. & Tom, R. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Computers In Biology And Medicine*. **126** pp. 104041 (2020)

[25] Zhang, Y. & Wallace, B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *ArXiv Preprint ArXiv:1510.03820*. (2015)

[26] Hooker, S., Erhan, D., Kindermans, P. & Kim, B. A benchmark for interpretability methods in deep neural networks. *Advances In Neural Information Processing Systems*. **32** (2019)

[27] Hara, S., Ikeno, K., Soma, T. & Maehara, T. Maximally invariant data perturbation as explanation. *ArXiv Preprint ArXiv:1806.07004*. (2018)

[28] Lundberg, S. & Lee, S. A unified approach to interpreting model predictions. *Advances In Neural Information Processing Systems*. **30** (2017)

[29] Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge And Information Systems*. **41** pp. 647-665 (2014)

[30] Fong, R. & Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 3429-3437 (2017)

[31] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. & Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. *International Conference On Machine Learning*. pp. 2048-2057 (2015)

[32] Montavon, G., Samek, W. & Müller, K. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*. **73** pp. 1-15 (2018)

[33] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. & Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*. **10**, e0130140 (2015)

[34] Samek, W., Montavon, G., Binder, A., Lapuschkin, S. & Müller, K. Interpreting the predictions of complex ml models by layer-wise relevance propagation. *ArXiv Preprint ArXiv:1611.08191*. (2016)

[35] Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T. & Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances In Neural Information Processing Systems*. **29** (2016)

[36] Bien, J. & Tibshirani, R. Prototype selection for interpretable classification. (2011)

[37] Sharma, S., Henderson, J. & Ghosh, J. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *ArXiv Preprint ArXiv:1905.07857*. (2019)

[38] Mothilal, R., Sharma, A. & Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings Of The 2020 Conference On Fairness, Accountability, And Transparency*. pp. 607-617 (2020)

[39] Yuan, X., He, P., Zhu, Q. & Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions On Neural Networks And Learning Systems*. **30**, 2805-2824 (2019)

[40] Sedig, K. & Parsons, P. Interaction design for complex cognitive activities with visual representations: A pattern-based approach. *AIS Transactions On Human-Computer Interaction*. **5**, 84-133 (2013)

[41] Sedig, K., Naimi, A. & Haggerty, N. Aligning information technologies with evidence-based health-care activities: A design and evaluation framework. *Human Technology*. **13** (2017)

[42] Parsons, P. & Sedig, K. Distribution of information processing while performing complex cognitive activities with visualization tools. *Handbook Of Human Centric Visualization*. pp. 693-715 (2013)

[43] Lamy, J., Sedki, K. & Tsopra, R. Explainable decision support through the learning and visualization of preferences from a formal ontology of antibiotic treatments. *Journal Of Biomedical Informatics*. **104** pp. 103407 (2020)

[44] Lamy, J., Sekar, B., Guezennec, G., Bouaud, J. & Séroussi, B. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence In Medicine*. **94** pp. 42-53 (2019)

[45] Choi, E., Bahadori, M., Sun, J., Kulas, J., Schuetz, A. & Stewart, W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances In Neural Information Processing Systems*. **29** (2016)

[46] Kwon, B., Choi, M., Kim, J., Choi, E., Kim, Y., Kwon, S., Sun, J. & Choo, J. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions On Visualization And Computer Graphics*. **25**, 299-309 (2018)

[47] Kyrimi, E., Mossadegh, S., Tai, N. & Marsh, W. An incremental explanation of inference in Bayesian networks for increasing model trustworthiness and supporting clinical decision making. *Artificial Intelligence In Medicine*. **103** pp. 101812 (2020)

[48] Zihni, E., Madai, V., Livne, M., Galinovic, I., Khalil, A., Fiebach, J. & Frey, D. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *Plos One*. **15**, e0231166 (2020)

[49] Yang, Y., Fasching, P. & Tresp, V. Predictive modeling of therapy decisions in metastatic breast cancer with recurrent neural network encoder and multinomial hierarchical regression decoder. *2017 IEEE International Conference On Healthcare Informatics (ICHI)*. pp. 46-55 (2017)

[50] Ming, Y., Qu, H. & Bertini, E. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE Transactions On Visualization And Computer Graphics*. **25**, 342-352 (2018)

[51] Che, Z., Purushotham, S., Khemani, R. & Liu, Y. Interpretable deep models for ICU outcome prediction. *AMIA Annual Symposium Proceedings*. **2016** pp. 371 (2016)

[52] Katuwal, G. & Chen, R. Machine learning model interpretability for precision medicine. *ArXiv Preprint ArXiv:1610.09045*. (2016)

[53] Yang, Y., Tresp, V., Wunderle, M. & Fasching, P. Explaining therapy predictions with layer-wise relevance propagation in neural networks. *2018 IEEE International Conference On Healthcare Informatics (ICHI)*. pp. 152-162 (2018)

[54] Böhle, M., Eitel, F., Weygandt, M. & Ritter, K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers In Aging Neuroscience*. **11** pp. 194 (2019)

[55] Slijepcevic, D., Horst, F., Lapuschkin, S., Raberger, A., Zeppelzauer, M., Samek, W., Breiteneder, C., Schöllhorn, W. & Horsak, B. On the explanation of machine learning predictions in clinical gait analysis. *ArXiv Preprint ArXiv:1912.07737*. (2020)

[56] Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*. **267** pp. 1-38 (2019)

[57] Benbasat, I., Bruckman, A., Carey, J., Djamasbi, S., Farooq, U., Gefen, D., Germonprez, M., Hassanein, K., Head, M., Hess, T. & Others Transactions on Human-Computer Interaction. *AIS Transactions On Human-Computer Interaction*. **3**, 1-25 (2011)

[58] Wang, D., Yang, Q., Abdul, A. & Lim, B. Designing theory-driven user-centric explainable AI. *Proceedings Of The 2019 CHI Conference On Human Factors In Computing Systems*. pp. 1-15 (2019)

[59] Tory, M. User studies in visualization: A reflection on methods. *Handbook Of Human Centric Visualization*. pp. 411-426 (2013)

[60] Parsons, P. & Sedig, K. Adjustable properties of visual representations: Improving the quality of human-information interaction. *Journal Of The Association For Information Science And Technology*. **65**, 455-482 (2014)

[61] Xie, Y., Gao, G. & Chen, X. Outlining the design space of explainable intelligent systems for medical diagnosis. *ArXiv Preprint ArXiv:1902.06019*. (2019)

[62] Sedig, K., Parsons, P., Dittmer, M. & Haworth, R. Human-centered interactivity of visualization tools: Micro-and macro-level considerations. *Handbook Of Human Centric Visualization*. pp. 717-743 (2014)

[63] Hundhausen, C. Evaluating visualization environments: Cognitive, social, and cultural perspectives. *Handbook Of Human Centric Visualization*. pp. 115-145 (2013)

[64] Freitas, C., Pimenta, M. & Scapin, D. User-centered evaluation of information visualization techniques: Making the HCI-InfoVis connection explicit. *Handbook Of Human Centric Visualization*. pp. 315-336 (2014)

[65] Kincaid, J., Fishburne Jr, R., Rogers, R. & Chissom, B. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. (Institute for Simulation,1975)

Chapter 4

Unlocking the Power of Explainability in Ranking Systems: A Visual Analytics Approach with XAI Techniques

This chapter, initially published at the 21st International Conference of Artificial Intelligence in Medicine (AIME) in Portoroz, Slovenia, as part of the XAI-Healthcare workshop on June 14th, 2023, has been expanded and revised to integrate with the overall structure of this dissertation. The modifications made maintain the integrity and coherence of the original paper within the context of the broader research presented here.

4.1 Introduction

Ranking systems have become prevalent in various domains, including e-commerce [1], social media [2], and human resources [3, 4]. These systems are often complex, with many input variables and black box algorithms, making it challenging for users to understand how they generate rankings. the lack of transparency and accountability in such systems presents a significant challenge, which may lead to user mistrust and reluctance in adopting them. Thus, there has been a growing demand to provide users with a deeper understanding of these black box models, giving rise to the development of Explainable Artificial Intelligence (XAI) methods.

XAI methods are designed to enhance the transparency and interpretability of artificial intelligence systems for human comprehension. These methods aim to alleviate the difficulties in understanding how the ranking system generates its output by providing users with insights into its workings. By doing so, XAI methods can improve transparency and user trust, and can lead to greater accountability and effective utilization of the system [5].

In this chapter, we present a visual analytics tool that uses XAI methods and feature importance visualization to explain ranking systems. Our tool is intended to provide users with a better understanding of how these systems generate rankings by identifying the impact of each feature on the output. We also introduce a customized counterfactual explanation method that considers a dynamic threshold for ranking items instead of the traditional approach of identifying the minimum changes required to change class prediction. To demonstrate the effectiveness of our tool, we illustrate its use in a healthcare scenario of triage patients and ranking them for admission to the ICU based on their severity. We argue that our tool can provide clinicians with a transparent and understandable system for decision-making, which is a prerequisite for improving their confidence in the system and enabling them to make informed decisions about patient care.

We believe our tool can be applied to other ranking systems in various domains, providing users with a transparent and understandable system for ranking-based decision support. The rest of the chapter is structured as follows: Section 2 reviews related work in explainable AI and ranking systems. Section 3 describes our methodology. Section 4 presents the experimental results of our tool applied to the healthcare scenario. Section 5 discusses the strengths and limitations of our approach, and Section 6 concludes the chapter.

4.2 Background

4.2.1 SHAP

SHAP (SHapley Additive exPlanations) is an XAI method that explains how machine learning models make predictions. It assesses the importance of each feature in the input data by determining how much each feature contributes to the model's output.

The basis of SHAP values originated from cooperative game theory ideas developed in 1953 by Lloyd Shapley [7]. These values quantify the contribution (payout) of each player in a game. In the context of machine learning, SHAP values measure the contribution of each feature to the prediction made by the model. They are computed by examining all possible combinations of features and calculating the amount each feature contributes to the prediction in comparison to its contribution in the absence of other features. The SHAP approach considers the marginal contribution of each feature in the prediction, which is the difference between the prediction made by a model that has the feature included and the prediction made by a model that does not have the feature included[6]. By assuming that the prediction is a weighted sum of feature values, SHAP is able to identify the weight of each feature based on its contribution to the prediction, thus allowing it to capture feature interactions.

The Shapley value is defined through a value function val. It represents the contribution of a specific feature value to the overall prediction, considering all possible combinations of feature values [8]. This concept can be mathematically expressed using Equation 4.1:

$$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{x_j\}) - val(S)) \tag{4.1}$$

Here, S represents a subset of the features used in the model, ϕ_j signifies the contribution (Shapley value) of the j-th feature to the prediction, x is the vector of feature values for the instance being explained, and p is the number of features. The term $val(S)$ represents the prediction for the feature values in set S, with other features marginalized (not included) in S.

This prediction can be mathematically expressed as:

$$val(S) = \int \hat{f}(x_1, \ldots, x_p) dP_{x \notin S} - E_X(\hat{f}(X)) \tag{4.2}$$

In Equation $\hat{f}(x_1, \ldots, x_p)$ represents the expected model prediction when the features in $\mathcal{S}$ are fixed, and the rest ($x \notin \mathcal{S}$) are randomly sampled from their marginal distributions, while $E_X(\hat{f}(X))$ denotes the expected prediction according to the model for all possible feature value combinations.

A notable strength of the Shapley value is its ability to fairly assess the contributions of various features, backed by a mathematically proven theory. However, its drawback lies in the algorithm's complexity, which escalates due to the exponential growth in feature combinations. As a result, when applied in real-world scenarios, practical constraints necessitate approximating the Shapley value by considering only a subset of combinations [9].

4.2.2 Counterfactual Explanation

Counterfactual explanation in Explainable AI (XAI) refers to a technique used to provide interpretable and understandable explanations for the predictions or decisions made by a machine learning (ML) model. The term "counterfactual" comes from the idea of creating a scenario that is contrary to the actual outcome [10]. In other words, a counterfactual explanation explains what changes to the input data would have resulted in a different prediction or outcome from the ML model [22, 23]. The counterfactual explanation provides insight into how the input features influenced the ML model's prediction. It helps users understand why a certain prediction was made and what specific factors played a crucial role. For example, it can explain why a certain medical diagnosis was made and what changes in the patient's symptoms or medical history would have led to a different diagnosis.

Given a classifier, denoted as f, that generates a decision $y = f(x)$ for a given instance x, a counterfactual explanation involves generating an instance x'. In this new instance, the classifier's decision for x', denoted as $f(x')$, is different from the original decision y. This implies that $f(x') \neq y$, while ensuring that the dissimilarity between the original instance x and the counterfactual instance x' is kept to a minimum [10]. The classifier f is often an opaque model, such as a black-box machine learning model like a neural network or an ensemble.

Counterfactual explanations, also referred to as counterfactuals, belong to the category of example-based explanations. Other forms of example-based explanations include prototypes, criticisms, and influential instances, as discussed in Chapter 3. However, in these approaches, other instances x' share the same class label as the instance x, indicating $f(x) = f(x')$, whereas in counterfactual explanations, $f(x) \neq f(x')$. On the other hand, the difference between x and a counterfactual x' reveals precisely what needed to be distinct in x to yield an alternative outcome.

For example, consider a bank customer x applying for a loan. The bank's AI system, utilizing a black-box machine learning model f, declines the loan request. A counterfactual explanation might unveil that a hypothetical customer x' would have been approved for the loan. This hypothetical x' customer is almost identical to the original applicant x, except for a yearly income of $15,000$ instead of $12,000$, and no other outstanding debts with the bank. In this scenario, the hypothetical x' customer serves as a counterfactual example, and the counterfactual explanation $\delta_{x,x'}$ includes the increased income of $15,000$ and the absence of other bank debts. These minimal adjustments would have resulted in a different decision.

Presented in Figure 4.1 is a basic diagram illustrating a decision boundary between two classes: positive and negative. The aim is to determine a counterfactual solution for a given instance x. The conventional approach adopted by multiple algorithms involves making minor adjustments to instance x until it crosses the decision boundary. The outcome of this iterative process referred to as the counterfactual of x encapsulates instances $cf1$ and $cf2$ have shown in Figure 4.1. $cf1$ and $cf2$ are obtained by shifting x in two different directions. Since both counterfactuals are on the positive side of the decision boundary, they qualify as valid counterfactual explanations for x.

Diverse algorithms designed for generating counterfactual explanations differ in two key ways. The first concerns the strategy for selecting an optimal counterfactual candidate from potential options. The second involves the level of access required to the underlying predictive model for generating counterfactual explanations.

In 2017, Wachter et al.[27] introduced the concept of counterfactual explanations. They framed counterfactual explanations as an optimization problem. The objective, as stated in Equation 4.3, is to minimize the distance between the original data point (x) and a counterfactual point (x'), while ensuring that the classifier's output for the counterfactual matches a desired label ($y' \in Y$).

$$\underset{x'}{\arg\min}\, d(x, x') \quad \text{s.t.} \quad f(x') = y' \tag{4.3}$$

To make the objective differentiable and unconstrained, it is reformulated into two terms, as shown in Equation 4.4. The first term encourages the counterfactual (classifier output) to align with the desired class, while the second term enforces proximity between the counterfactual and the original data point. A distance metric (denoted as d) measures the separation between

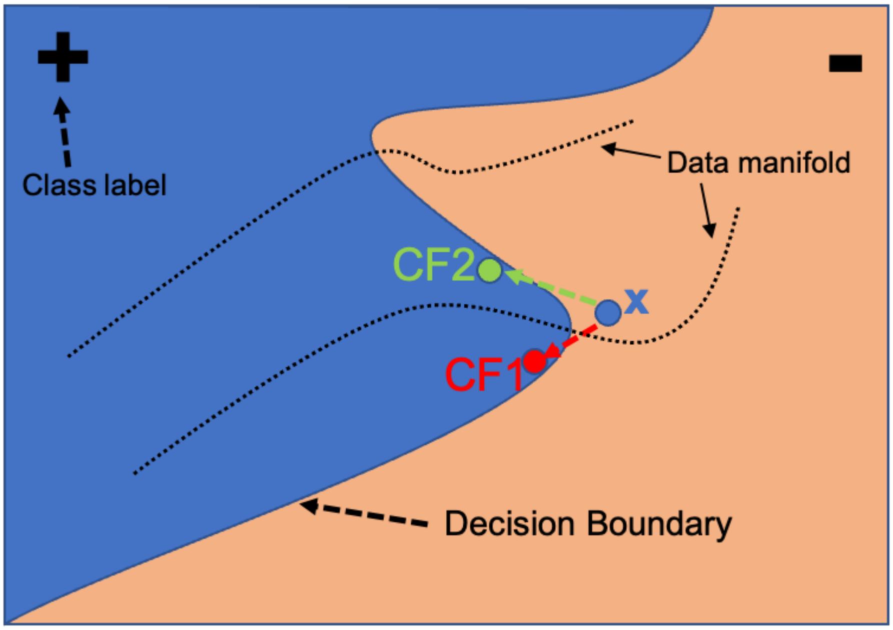


Figure 4.1: Exploring Counterfactual Paths for Blue Data Point denoted as x: The data point is initially classified as negative. The figure represents two potential paths (red and green) that cross the decision boundary, serving as valid counterfactuals. While the red path is the shortest, the green path closely adheres to the training data manifold, despite being longer.

two data points, x and x', which could be metrics like L1/L2 distance, quadratic distance, or distance functions based on feature cumulative distribution functions (CDF) [12]. Thus, the definition underscores the importance of keeping counterfactual changes relatively small compared to the starting point. However, this approach employs the normalized Manhattan distance as the distance metric and relies on gradient information from the predictive model, making it unsuitable for black box models [11].

$$\arg\min_{x'} (f(x') - y')^2 + d(x, x') \tag{4.4}$$

Furthermore, some of the counterfactuals generated may suggest infeasible changes, such as altering an applicant's race. To address this issue, Ustun et al.[13] proposed a modified objective function that considers the cost incurred in changing from x to x_0 and incorporates feasibility constraints. Incorporating the idea of a set of actionable features (A), the optimization problem is adapted accordingly. The loss function is updated to minimize the difference between the classifier output (counterfactual) and the desired label, subject to the constraint that changes are made only within the actionable feature set (Equation 4.5).

$$\arg\min_{x' \in A} (f(x') - y')^2 + d(x, x') \tag{4.5}$$

Achieving the counterfactual might involve finding a compromise between the number of modified features and the magnitude of change. Ideally, a counterfactual should modify as few features as possible. Given that finding shorter explanations is more comprehensible [14], sparsity becomes a significant consideration. The loss function is enhanced with a penalty term [15] that encourages sparsity in the difference between the modified and original data points ($g(x' - x)$), possibly using norms like L0/L1 (Equation 4.6).

$$\arg\min_{x_0 \in A} (f(x_0) - y_0)^2 + d(x, x_0) + g(x' - x) \tag{4.6}$$

Moreover, a counterfactual should not propose feature combinations that are vastly dissimilar from observed data. Realism is vital, implying that counterfactuals should adhere to the training data's characteristics and correlations among features. The idea involves the concept of data manifold closeness, where the goal is to generate counterfactuals that remain close to the original data distribution. This intuition ensures that counterfactuals generated are valid data points and not outliers far from the data distribution. Various methods exist for quantifying this adherence. The loss function is extended to include a penalty term promoting adherence to the data manifold defined by the training set ($l(x'; X)$) (Equation 4.7). In Figure 4.1, the dashed lines represent the data manifold. For the blue data point denoted as x, the green path adhering to the manifold is favored over the unrealistic red path, due to the added manifold loss term although the red path is shorter.

$$\arg\min_{x' \in A} (f(x') - y')^2 + d(x, x') + g(x' - x) + l(x'; X) \tag{4.7}$$

Another crucial consideration in generating counterfactual explanations is accounting for feature interactions. Features in a dataset are interconnected, so altering one feature can impact others due to causal relationships. For instance, obtaining a new educational degree might necessitate increasing the individual's age. To ensure realism and actionability, counterfactuals should maintain known causal relationships. By incorporating the causal graph's information, their approach provides counterfactuals that are more realistic and feasible, considering feature interdependencies. While access to the full causal graph may be impractical, partial knowledge or user inputs about feasibility constraints can be utilized effectively.

Different ways of generating counterfactual examples have been studied [24, 25, 26]. The most popular approaches use an optimization algorithm that was initially proposed by Wachter et al. [27]. Russell et al. used the idea of generative models to synthesize new input instances that are close to the original data but output models with different results [28]. Dandl et al. use a gradient-based optimization algorithm to find counterfactual examples [29]. Another approach to generating counterfactual examples is greedy search. In this approach, feature values are modified iteratively until the model prediction changes. Yang et al. used this approach to find counterfactual examples [30]. Prior works have focused on generating counterfactual examples using different approaches with the aim of explaining the minimum changes required to alter class predictions. In our approach, we modify the traditional greedy algorithm to address the specific needs of ranking problems. To achieve this, we generate counterfactual examples that take into account the relationship between items in a ranked list, providing a better understanding of the changes required to reach a desired rank.

Overall, counterfactual explanations hold great promise in enhancing transparency and interpretability in predictive models, ultimately leading to fairer and more accurate decision-making across a wide range of domains and applications. By addressing the challenges and expanding their application to ranking systems, we can unlock their full potential in providing actionable insights and empowering end users with recourse for improved outcomes.

4.3 Related Work

Recently, the lack of interpretability of existing ranking techniques has received attention, leading to the development of XAI methods and visualization frameworks to aid in the interpretation and analysis of ranking models. For instance, Srvis proposed by Di Weng et al. integrates spatial contexts with rankings through scalable visualizations and flexible spatial filtering and comparative analysis to support decision-making for large-scale spatial alternatives like selecting store locations [16].

RankViz, on the other hand, supports the analysis and interpretation of learning-to-rank (LtR) models by providing visualizations that give information on important data features and enable comparison of element positions to aid in understanding and creation of rankings [17]. LineUp is a bar chart-based technique that ranks items based on multiple attributes with different scales and semantics, allowing interactive combination and refinement of parameters to explore changes in rankings and enable comparison of multiple rankings on the same set of items [18]. Meanwhile, uRank is a tool that provides views summarizing the contents of a recommendation set and interactive methods to convey users' interests through a recommendation ranking, enabling users to understand, refine, and reorganize documents as information needs evolve [19].

Anahideh, et al. propose a hierarchical ranking explanation framework that uses a proper neighborhood construction approach to capture local explanations for competitive rankings, exploring various explanation techniques to identify the local contribution of ranking indicators based on an instance's position in the ranking and the size of the neighborhood [20]. Finally, Zhuang, et al. introduce the use of generalized additive models (GAMs) for ranking tasks, instantiated using neural networks, demonstrating that their approach outperforms traditional GAMs while maintaining similar interpretability, offering promise for the development of intrinsically interpretable ranking models [21].

4.4 Methodology

Our visual analytics tool combines two key strategies for explaining rankings: XAI-derived explanations and interactive visualizations. The XAI strategies we use are counterfactual explanations and feature importance, and we have proposed a customized algorithm to contextualize counterfactual explanations for use in ranking systems. For feature importance, we use the Shapley value method. The interactive visualization strategies we use contain two primary sub-visualizations, namely the ranking list and the what-if panel. These visual representations serve as a bridge between the explanations generated by the XAI module and the users' comprehension, allowing them to explore the ranking system and gain a deeper understanding of how it operates.

4.4.1 XAI

Feature Importance

To determine the importance of features, we employed SHAP values, a widely recognized interpretability technique, to gain insights into the individual contributions of features to the ranking outcomes. These values offer a comprehensive view of how each feature affects the final ranking. By employing Shap values, we were able to discern the relative importance of various factors in determining the ranking. Through analyzing Shap values, we determined that the most crucial features, in descending order of importance, are respiratory rate, heart rate, body temperature, blood pressure, and polymerase chain reaction (PCR) results.

We selected this approach as it offers an understandable interpretation of the model's prediction, and its ability to capture interactions between features. Furthermore, SHAP is a model-agnostic method. It can be used with various types of machine learning models, including neural networks, tree-based models, and linear models, making them a flexible tool for analyzing feature importance in a broad range of machine learning applications.

Counterfactual Explanations

A counterfactual explanation describes the smallest change to the feature values of an example that results in the model making a meaningful change in output. This is typically defined as a change in the predicted class. While counterfactual explanations have proven effective in classification tasks, one of the challenges lies in extending their application to ranking systems. Ranking systems often involve complex decision-making processes, and using counterfactual explanations in this context requires careful consideration. In this regard, we have proposed an approach that applies counterfactual explanations to a greedy algorithm for finding a counterfactual example in the context of ranking. However, our proposed approach is not limited to this specific algorithm and can be adapted to other counterfactual methods, such as genetic algorithms or other optimization-based approaches discussed earlier in this chapter.

Our proposed method uses a greedy algorithm to find counterfactual explanations that are applicable to ranked model outputs, which are a collection of outputs for a set of test data that have been sorted according to the output of a probabilistic classification model. The goal of our algorithm is to determine the minimum changes required for a data instance to achieve a different rank. For this purpose, we use a greedy approach given in Algorithm 1. The algorithm takes as input: *Mlmodel* - the machine learning model that produced the ranking; *dataInstance* - a specific data instance; and *RChange* - a desired rank change.

The algorithm first finds the most important features $\mathcal{F}^*$ sorted in descending order for the given *dataInstance* according to Shapley values. Then the algorithm initializes the set $\mathcal{F}$ with f^* (the first important feature in $\mathcal{F}^*$) and generates $InteractionList$ which is a list of features that have the most interaction with this feature. These are the features whose values have the biggest impact on the relationship between the f^* feature and the outcome; they are available in standard implementations of XGBoost and can be computed for other models as well. To generate counterfactual examples, the approach varies the feature values in the subset $Fsub$, which is obtained by adding one feature at a time from the feature list $\mathcal{F}$). The approach changes the values of these features for *dataInstance* along a specified range using a grid search, and observing the corresponding model outputs while holding all other features constant. It computes the rank assigned by the model for each new potential counterfactual example and compares it to the rank of the original data instance.

Given that the proposed approach aims to achieve the desired ranking with the minimum number of changes possible, if a single important feature modification is sufficient to meet the target ranking, then modifying other features can be avoided. However, if changing the value of one of the features in $Fsub$ alone is not able to change the rank of *dataInstance*, the algorithm considers changing multiple important features together to see if a change in rank can be produced. After each modification to the feature(s), the algorithm replaces the feature value with the modified value and evaluates the $MLmodel$'s output. The premise is that by simultaneously changing important features together with their most strongly interacting features, the algorithm has the best chance of being able to identify a counterfactual example whose rank is at least *RChange* away from the rank of *dataInstance*. The algorithm iteratively adds features from the list of most important features and their corresponding interaction list and performs a grid search after each is added. This process is repeated until a change in the feature values results in a counterfactual example whose rank is different from that of *dataInstance* by the desired amount (Rchange), at which point the algorithm will stop and return the new counterfactual example.

The objective of our approach is to identify a counterfactual example that involves the minimum possible number of feature modifications. To achieve this, we begin with the most important feature and add its interacting feature one at a time. If this process fails to yield a suitable counterfactual example, it will add another important feature (f^*) from $\mathcal{F}^*$ along with its interacting features to $\mathcal{F}$ and repeat the algorithm.

4.4.2 Interactive Visualization

Our interactive visualization is designed to provide users with a transparent and understandable system for decision support. Two primary sub-visualizations are employed, namely the ranking list and the what-if panel. These visualizations map the explanations produced in the XAI module into visual representations, allowing users to explore the ranking system and understand how it generates a ranking.

Interactive Ranking list

An interactive ranking list was developed based on the concept of semantic zoom to allow for a detailed exploration of each item locally and a comparison between items globally. Semantic zoom is a technique that provides users with distinct representations of data as they zoom in or out, with the aim of enhancing the overall understanding of the underlying semantic structure [31]. When an item in the ranking list is expanded, a treemap of its attributes is displayed, presenting the details of each item with different categories, if applicable. A treemap is a visualization technique that organizes hierarchical data into a set of nested rectangles. Each rectangle's size is proportional to a quantitative variable, and additional information about specific categories or variables can be conveyed through colors [32]. In our visualization the size of the boxes in the treemap represents the importance of each category and its contribution to the overall ranking, enabling users to identify the primary feature category responsible for the

Algorithm 1 Generating Counterfactual Explanations for Ranking Models Using a Greedy Algorithm

```
procedure CounterfactualRanking(Mlmodel, dataInstance, RChange)
    $InstanceRank \leftarrow$ Mlmodel$($dataInstance$)$
    if RChange is not provided then
        RChange $\leftarrow 1$
    $\mathcal{F}^* \leftarrow$ FindListOfMostImportantFeatures$($Mlmodel, dataInstance$)$
    $\mathcal{F} \leftarrow EmptyList$
    $Fsub \leftarrow EmptyList$
    for $f^*$ in $\mathcal{F}^*$ do
        $InteractionList \leftarrow$ GetFeaturesWithHighestInteractionWith$(f^*,$ Mlmodel, dataInstance$)$
        $\mathcal{F} \leftarrow append(\mathcal{F}, f^*, InteractionList)$
        for $f$ in $\mathcal{F}$ do
            $Fsub \leftarrow append(Fsub, f)$
            $\Delta_f \leftarrow$ {MinObserved($f$),0,MaxObserved($f$)}
            for $\delta$ in $\prod_{f \in Fsub} \Delta_f$ do    ▷ $\delta$ is a change to every feature (there are $3^{|Fsub|}$). This exhaustive search can be replaced with a heuristic search.
                $newInput \leftarrow$ ReplaceSelectedFeatureValues$($dataInstance, $\delta)$
                $newRank \leftarrow$ Mlmodel$(newInput)$
                if $newRank \geq InstanceRank -$ RChange or $newRank \leq InstanceRank +$ RChange then
                    return $newInput$
    return "No feasible changes achieve the desired ranking."
```

rank. Different colors are used for different feature categories. Additionally, each box is further divided into subcategories that specify the precise feature and its importance for the item's ranking through a prediction model. By using the treemap and its interactive features, such as zooming in and out, users can explore the categories, their features and values, and their contributions to each item in the ranking list. Semantic zoom facilitates the exploration of information at different levels of granularity. As an illustration, users can begin with an overview of the ranking list, subsequently zoom in to examine the top-ranked items more closely, and then zoom even further to inspect the particular attributes that contributed to their respective rankings. This feature enables users to navigate and comprehend large volumes of information without feeling overwhelmed.

What-if panel

We developed a *what-if* that allows for the exploration of counterfactual scenarios and the necessary adjustments to feature values to alter an item's ranking. This panel contains histograms of the top 5 global feature importance values, offering users a comprehensive overview of the most critical features across the entire dataset. Users can select an item from the ranking list and set their preferred rank for that item. Our customized counterfactual explanation algorithm then identifies the minimum changes required to achieve the desired rank, which the system presents to the user. The system displays the value of the most important feature of the selected item with a blue line on the histogram, along with an indicator of its rank. This feature enables users to compare the feature values of different items with various rankings. Additionally, there are red and black dotted lines on the histogram, which represent the feature value required for the item to rank one rank above or below its current rank, respectively. In some cases, a feature may lack either a black or red line, indicating that changing that feature alone cannot alter the item's rank, and modifying more features is required. Furthermore, our system includes a box for the desired rank, which users can use to explore the required changes to the feature values for a specific item to reach the desired rank. Our system considers both lower and higher desired ranks. Users can also use radio buttons to specify which features they want to modify. This panel provides users with an interactive tool to better comprehend how the ranking system works. By displaying the top global features and allowing users to modify the feature values and observe their impact on the ranking, our system facilitates a more nuanced and in-depth understanding of the ranking system, which can aid users in making more informed decisions.

The rest of the chapter will describe a potnential application of our tool to a healthcare scenario and discuss the strengths and limitations of our approach.

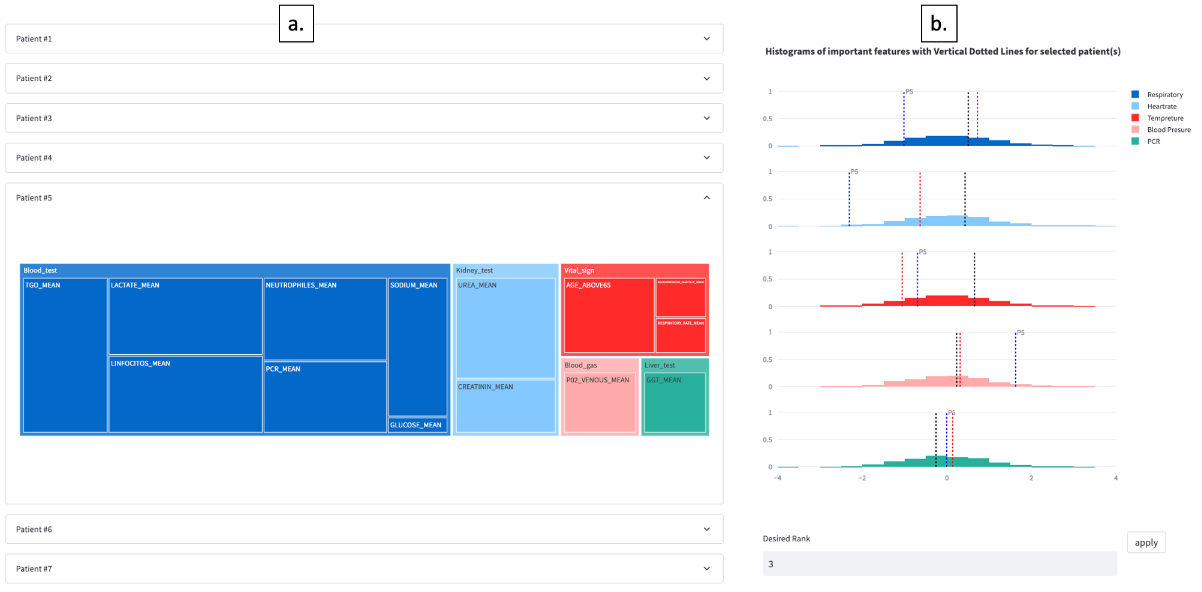


Figure 4.2: an overview of the entire visual analytics tool including interactive ranking list and what-if panel, a) Interactive List of ranked patients and Treemap Visualization of Patient ranked 1 (P1) Features. b) What-if Panel with Histograms of 5 Top Important Features.

4.5 Case Study: Explaining Triaging Patients to be Admitted to ICU

In order to evaluate the effectiveness of our proposed method, we conducted a case study focusing on ranking patients for admission to the ICU. For this purpose, we used a dataset from Sírio Libanês Hospital, which contains patient demographic information, previous disease groupings, blood results, vital signs, and blood gases [33]. The dataset includes labels indicating whether a patient was admitted to the ICU or not. We used this label to train an XGBoost model to predict patient admission probability. We then used these prediction probabilities to rank a test set of patients for triage purposes. After ranking the patients in the test set, our proposed method was

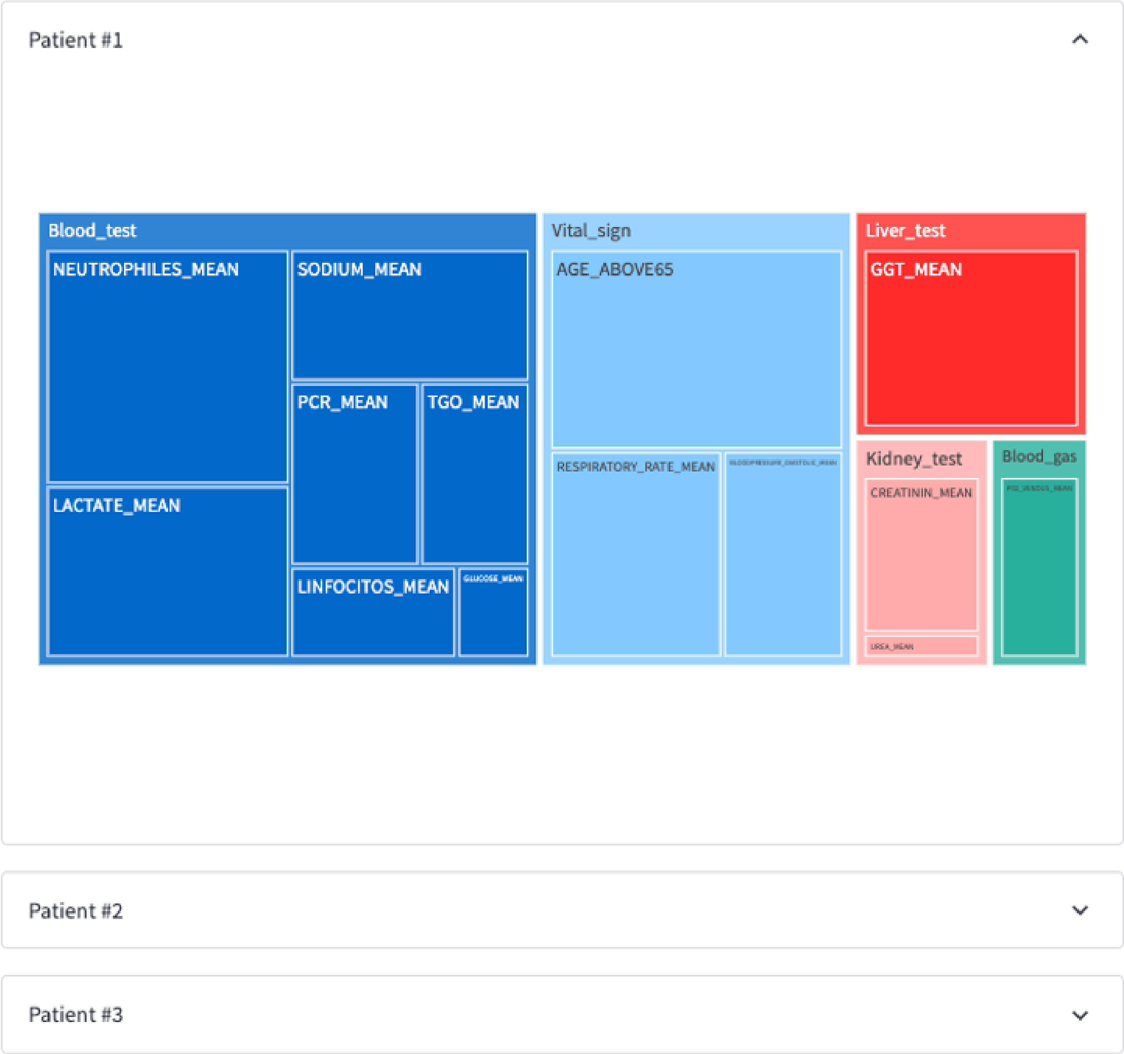


Figure 4.3: Interactive List of ranked patient and Treemap Visualization of Patient ranked 1 (P1) Features

applied to give a better understanding of how patients were ranked.

Figure 4.2 provides an overview of the entire visual analytics tool including the interactive ranking list on the left and the what-if panel on the right. The interactive list of ranked patients, as demonstrated in Figure 4.2.a, can be expanded to reveal a treemap showing the importance of the features used in the ranking. The treemap is designed to enable zooming in and out to explore the ranking importance of each category while hovering over each box generates a tooltip displaying the feature value. The size of each box corresponds to the importance of

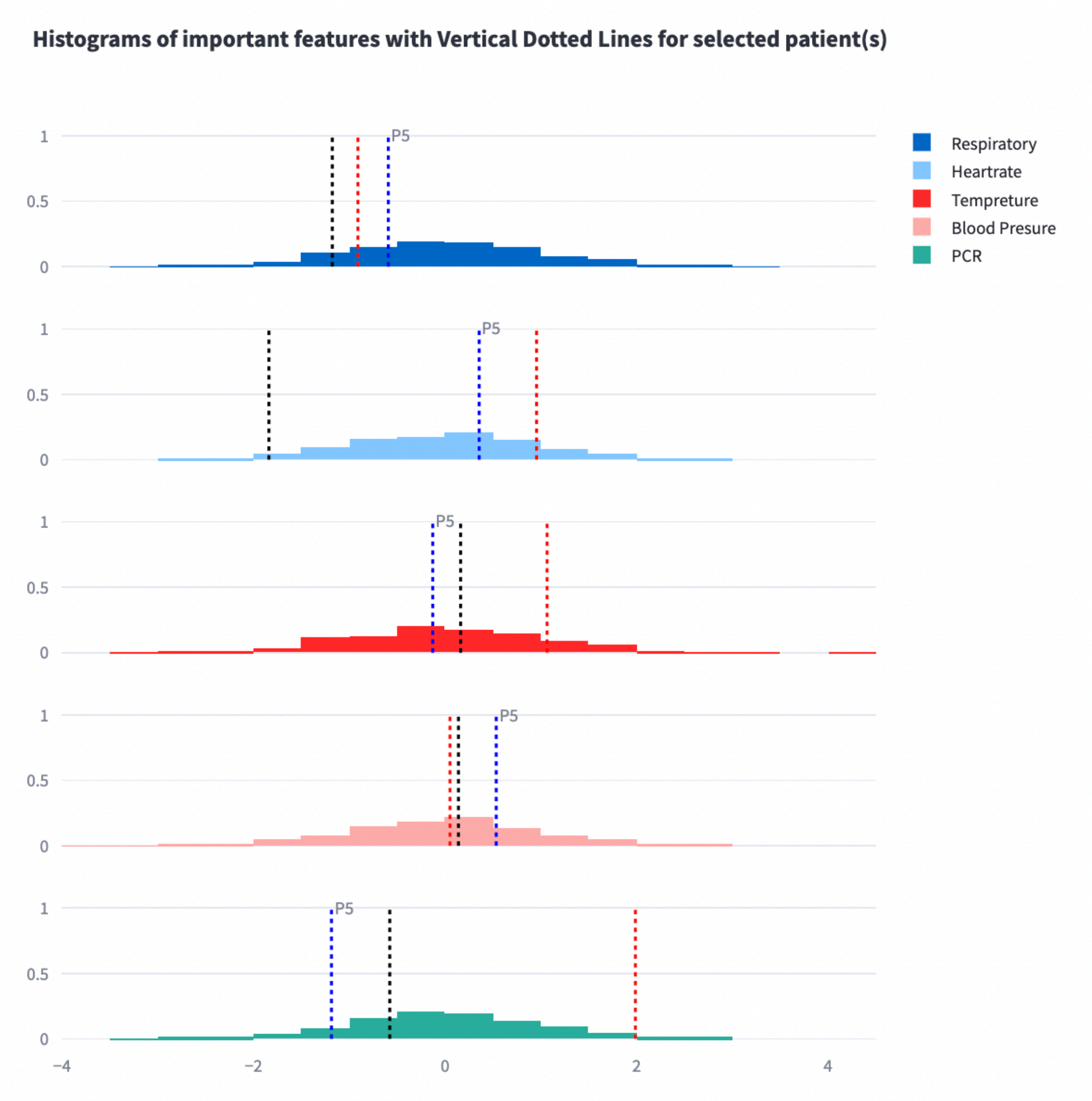


Figure 4.4: What-if Panel with Histograms of 5 Top Important Features.

the feature. For instance, in Figure 4.3 the blood test has the most impact on the ranking, and among the kidney tests, Creatinine is the dominant feature, as reflected in Figure 4.3 where the box representing Creatinine is much larger than other boxes in this category. The boxes for PCR and TGO in the blood test category have comparable sizes displayed indicating that they have a similar impact on the ranking. The tool's semantic zoom feature allows users to expand the attributes of two or more patients to compare their feature importance. Users can also compare the feature values of selected patients on the histogram.

The treemap offers insight into the local important feature for each patient, providing users with a more granular understanding of the local feature importance influencing the patient's ranking. In contrast, the histogram showcases the top five globally important features, giving users an understanding of the feature's importance across the whole population. By comparing feature importance at both the local and global levels, users can develop a better understanding of the ranking system and the features that contribute to it.

In the what-if panel depicted in Figure 4.2.b the blue dotted line denotes the selected patient (P5) feature value, while the black and red dotted lines indicate the amount of P5 feature that would result in a higher or lower ranking, respectively. The absence of a red or a black dotted line for each feature implies that modifying only this feature of the patient is insufficient to achieve a higher ranking, and modifying more than one feature (including this feature) is necessary. Our visualization tool allows users to set a desired ranking for a selected patient (in the text box) and presents the minimum changes required for the patient to attain that ranking, using the updated black and red lines. To illustrate, suppose a desired rank of 3 is entered for a patient currently ranked 5th. The updated red line in Figure 4.4 represents the minimum changes required in features, namely Temperature, PCR, and HeartRate, to achieve a higher rank of 2 compared to the previous rank of 5. Conversely, modifying these features as indicated by the black lines results in a lower rank of 8. This example highlights how the visualization tool allows users to explore the impact of changing specific features on a patient's rank, providing valuable insights into the ranking system.

Overall, our proposed method can be utilized to explain the ICU patient ranking system, enabling users to gain insights into how patients were ranked and explore potential changes that could impact the ranking.

4.6 Discussion

The use of conventional statistical methods such as accuracy, precision, and sensitivity is often not sufficient to provide users with a clear understanding of why a particular item has been ranked in a certain way. This is particularly true for ranking systems that use machine learning models, which can be difficult to interpret. An alternative approach that is applicable to a variety of ranking systems and empowers users to explore and understand the results of the ranking system is needed.

The proposed visual analytic tool is model agnostic and provides users with the ability to identify the contributing factors that determine the ranking of a specific item. By presenting the important feature contributors of prediction, users can understand what factors play a crucial role in determining the ranking of a particular item. The interactive visualization feature of the system allows users to gain a global understanding of how the system ranks items.

The Treemap helps users to understand what feature is important for an item to be ranked as it was. The expanding and collapsing attribute of the ranking list is another useful feature that allows users to drill down into more information as needed. This feature can be helpful when users need to see the distribution of important features and how a particular item ranks in comparison to others. Users can also see the changes needed to be ranked differently, providing them with actionable insights to improve the ranking of a specific item.

The what-if panel shows what features are essential for the whole population. This feature enables users to compare the ranking of a specific item to the overall ranking system and gain insights into the underlying factors that determine the ranking. Additionally, users can compare two or more items, enabling them to identify similarities and differences between them and understand the changes required for them to be ranked differently. Furthermore, users can investigate what changes are needed for an item to be ranked differently.

To follow the framework for building user-centered explanations we addressed the diverse requirements and preferences of clinicians when generating explanations. In other words, this tool is designed to create explanations that are tailored to the specific needs and expectations of clinicians as users. Clinicians often seek concise and clinically relevant insights. Accordingly, our tool visually highlights the key features influencing the rankings, which may allow clinicians to grasp the essential information.

Interactivity is a pivotal component of our visual analytic tool. By allowing users to manipulate parameters and criteria in real-time, our system empowers them to explore the underlying factors influencing the ranking outcomes. This interactive approach can foster a deeper understanding of the AI's decision-making process and enables users to observe the impact of various inputs on the final rankings. Understanding that different user groups have varying levels of technical expertise and preferences, our visual analytic tool integrates multi-modal explanations. This approach offers a range of explanation methods combined with different visualization components and interactive elements, accommodating the diverse learning and comprehension styles of clinicians.

By providing users with a better understanding of the underlying factors that determine the ranking, the system can help users make more informed decisions and improve the overall quality of the ranking system. Furthermore, our visual analytic tool employs interactivity and multi-modal explanations to enhance the interpretability and usability of AI-driven ranking systems.

It is noted that our research endeavours faced significant challenges due to the widespread impact of the COVID-19 pandemic. The original plan involved conducting formal user studies to generate both qualitative and quantitative metrics, shedding light on the information-seeking process and the role of visual interface design in navigating extensive document sets.

Before the pandemic, we outlined study procedures and questionnaires (Appendix A). However, the severity of the pandemic led to the closure of educational institutions, the suspension of ethics review boards, and the enforcement of physical distancing measures. These circumstances made it impossible to proceed with our research as initially intended.

In response, we adapted our research directions, opting for an evolutionary design perspective. This involved in-depth analyses of relevant published research and their formal user studies, integrated with informal accounts gathered during formative assessment periods.

Despite the challenges, we believe we made the best effort possible to maximize research value within the constraints imposed by the pandemic. Our pivot in research approach reflects a pragmatic response to unforeseen circumstances while staying true to the essence of our original research objectives.

4.7 Conclusion

In this chapter, we proposed a visual analytic tool that combines XAI methods and interactive visualization to explain ranking systems by enabling users to investigate how changing the feature values of an item can impact the ranking. The use of our proposed method was illustrated using a case study on ICU patient triage. The case study demonstrated how our proposed tool could be used to provide users with a better understanding of how ranking systems work, which could ultimately improve decision-making processes. The counterfactual explanation method allowed users to explore how changes to individual patient features could have resulted in a different ranking, while feature importance provided insights into the importance of different features in the ranking system. Additionally, the interactive visualization allowed users to explore and experiment with different scenarios. Overall, our proposed tool has the potential to be applied in various domains, such as healthcare, finance, and education, to improve transparency and trust in ranking systems. Future work will investigate the integration of XAI methods, such as fairness metrics and algorithmic auditing, with interactive visualizations to detect and mitigate bias in the ranking system.

4.8 References

Bibliography

[1] Sivapalan, S., Sadeghian, A., Rahnama, H. & Madni, A. Recommender systems in e-commerce. *2014 World Automation Congress (WAC)*. pp. 179-184 (2014)

[2] Rappaz, J. Dynamic Personalized Ranking. (EPFL,2022)

[3] Faliagka, E., Iliadis, L., Karydis, I., Rigou, M., Sioutas, S., Tsakalidis, A. & Tzimas, G. On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed CV. *Artificial Intelligence Review*. **42** pp. 515-528 (2014)

[4] Yu, P., Lam, K. & Lo, S. Factor analysis for ranked data with application to a job selection attitude survey. *Journal Of The Royal Statistical Society: Series A (Statistics In Society)*. **168**, 583-597 (2005)

[5] Schoonderwoerd, T., Jorritsma, W., Neerincx, M. & Van Den Bosch, K. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal Of Human-Computer Studies*. **154** pp. 102684 (2021)

[6] Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge And Information Systems*. **41** pp. 647-665 (2014)

[7] Shapley, L. & Others A value for n-person games. (Princeton University Press Princeton,1953)

[8] Molnar, C. Interpretable machine learning, Local Model-agnostic methods: Shapley values. (Lulu.com,2020)

[9] Fatima, S., Wooldridge, M. & Jennings, N. A linear approximation method for the Shapley value. *Artificial Intelligence*. **172**, 1673-1699 (2008)

[10] Guidotti, R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining And Knowledge Discovery*. pp. 1-55 (2022)

[11] Singla, S., Eslami, M., Pollack, B., Wallace, S. & Batmanghelich, K. Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis*. **84** pp. 102721 (2023)

[12] Spangher, A., Ustun, B. & Liu, Y. Actionable recourse in linear classification. *Proceedings Of The 5th Workshop On Fairness, Accountability And Transparency In Machine Learning*. (2018)

[13] Ustun, B., Spangher, A. & Liu, Y. Actionable recourse in linear classification. *Proceedings Of The Conference On Fairness, Accountability, And Transparency*. pp. 10-19 (2019)

[14] Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*. **267** pp. 1-38 (2019)

[15] Sahil, V., Dickerson, J. & Hines, K. Counterfactual explanations for machine learning: A review. (2010)

[16] Weng, D., Chen, R., Deng, Z., Wu, F., Chen, J. & Wu, Y. Srvis: Towards better spatial integration in ranking visualization. *IEEE Transactions On Visualization And Computer Graphics*. **25**, 459-469 (2018)

[17] Pereira, M. & Paulovich, F. RankViz: A visualization framework to assist interpretation of Learning to Rank algorithms. *Computers & Graphics*, **93**, 25-38 (2020).

[18] Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H. & Streit, M. Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions On Visualization And Computer Graphics*. **19**, 2277-2286 (2013)

[19] Di Sciascio, C., Sabol, V. & Veas, E. Rank as you go: User-driven exploration of search results. *Proceedings Of The 21st International Conference On Intelligent User Interfaces*. pp. 118-129 (2016)

[20] Anahideh, H. & Mohabbati-Kalejahi, N. Local explanations of global rankings: insights for competitive rankings. *IEEE Access*. **10** pp. 30676-30693 (2022)

[21] Zhuang, H., Wang, X., Bendersky, M., Grushetsky, A., Wu, Y., Mitrichev, P., Sterling, E., Bell, N., Ravina, W. & Qian, H. Interpretable ranking with generalized additive models. *Proceedings Of The 14th ACM International Conference On Web Search And Data Mining*. pp. 499-507 (2021)

[22] Karimi, A., Barthe, G., Balle, B. & Valera, I. Model-agnostic counterfactual explanations for consequential decisions. *International Conference On Artificial Intelligence And Statistics*. pp. 895-905 (2020)

[23] Hashemi, M. & Fathi, A. Permuteattack: Counterfactual explanation of machine learning credit scorecards. *ArXiv Preprint ArXiv:2008.10138*. (2020)

[24] Mothilal, R., Sharma, A. & Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings Of The 2020 Conference On Fairness, Accountability, And Transparency*. pp. 607-617 (2020)

[25] Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., and Shah, C. (2020). Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

[26] Maragno, D., Röber, T. & Birbil, I. Counterfactual Explanations Using Optimization With Constraint Learning. *ArXiv Preprint ArXiv:2209.10997*. (2022)

[27] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, **31**, 841.

[28] Russell, C., Kusner, M., Loftus, J. & Silva, R. When worlds collide: integrating different counterfactual assumptions in fairness. *Advances In Neural Information Processing Systems*. **30** (2017)

[29] Dandl, S., Molnar, C., Binder, M. & Bischl, B. Multi-objective counterfactual explanations. *Parallel Problem Solving From Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part I*. pp. 448-469 (2020)

[30] Yang, W., Li, J., Xiong, C. & Hoi, S. MACE: An Efficient Model-Agnostic Framework for Counterfactual Explanation. *ArXiv Preprint ArXiv:2205.15540*. (2022)

[31] Dunsmuir, D. Selective Semantic Zoom of a Document Collection. *Available At,(Oct. 30, 2009)*. pp. 1-9 (2009)

[32] Johnson, B. TreeViz: treemap visualization of hierarchically structured information. *Proceedings Of The SIGCHI Conference On Human Factors In Computing Systems*. pp. 369-370 (1992)

[33] Sírio-Libanês Hospital. COVID-19 - Clinical Data to assess diagnosis. Available at: `https://www.kaggle.com/datasets/Sírio-Libanes/covid19`. Accessed on: 2023-01-20.

Chapter 5

Investigating Poor Performance Regions of Black Boxes: LIME-based Exploration in Sepsis Detection

This chapter represents the final outcome of a project conducted in collaboration with a company, where the main focus was on dealing with the implementation of XAI (eXplainable Artificial Intelligence) methods. The content presented in this chapter was originally showcased as a demonstration during The 1st World Conference on eXplainable Artificial Intelligence (xAI 2023). For this work, I conceptualized and implemented a comprehensive approach utilizing LIME to analyze all misclassified samples, enabling the identification of regions where the classifier exhibited suboptimal performance. This valuable insight was achieved through collaboration with the co-authors who provided their model as input. Additionally, I took the lead in drafting and refining the manuscript, which was revised by my co-authors.

As part of the thesis integration process, the chapter has undergone refinements and enhancements to better align with the overall theme and structure of the thesis. The project's integration within the thesis showcases a real-world application of XAI in an applied setting, emphasizing the significance of explainability in artificial intelligence and its potential impact on decision-making processes in various domains.

5.1 Introduction

Machine learning has exhibited impressive achievements in diverse fields, including healthcare [1]. The complexity of these models, however, creates challenges for their adoption in healthcare [19]. To address this issue, eXplainable AI (XAI) has been introduced, enabling machine learning models to provide explanations for their predictions. Model explainability is essential for gaining a deeper understanding of a model's decision-making process [20]. In critical domains such as sepsis detection [18] in the ICU, where incorrect predictions can result in fatal consequences, the reliability of these models is of utmost significance. This chapter aims to tackle a specific aspect of the interpretability challenge associated with these models, specifically the identification and explanation of scenarios in which black box predictive models fail or exhibit unexpected performance. Examining instances in which machine learning models exhibit deviations from their usual performance holds significant importance. These insights empower decision-makers to exercise caution in deploying models in situations where their predictions are prone to errors, thereby mitigating potential adverse consequences. Previous research endeavors have primarily centered on assessing the overall performance of these models through the adoption of evaluation metrics and methodologies aimed at gauging their reliability [11, 12]. W. Duivesteijn et al. [13] present an evaluation method that assesses the performance of a classifier, highlighting subspaces where the classifier excels or struggles in classification tasks, however, the method's applicability is limited to binary datasets and lacks model agnosticism. L. Torgo et al. [14] propose approaches that aim to offer interpretable descriptions of expected performance; however, the proposed visualization may not be well-suited when dealing with a high number of features. This chapter provides an analysis by focusing on the identification of specific regions where the models exhibit significant deviations from their usual performance. The identification of these regions empowers healthcare practitioners to make informed decisions by exercising caution when relying on the model. Additionally, these findings offer valuable insights that can guide the development of potential strategies aimed at improving and refining the model's overall performance [15, 16]. To achieve this, we propose an analytical approach that combines visual techniques to identify regions in the input space where a model's performance significantly diverges from its average performance. This visualization empowers users to grasp how various values of a particular predictor impact the model's performance.

5.2 Motivation

Sepsis, a life-threatening condition resulting from an uncontrolled response to infection, leads to inflammation, organ dysfunction, and potential fatality [17]. Timely diagnosis and intervention are vital for enhancing patient outcomes. Machine learning algorithms have demonstrated promise in this domain [18]; however, their black-box nature creates challenges for their adoption in healthcare, where life-or-death decisions are at stake [19]. To address this issue, eXplainable AI (XAI) has been introduced, enabling machine learning models to provide explanations for their predictions. Model explainability is essential for gaining a deeper understanding of a model's decision-making process [20], especially in critical applications like sepsis prediction, where false negatives can have fatal consequences.

While examining individual misclassified samples and determining the primary contributing features are essential, our focus lies in identifying patterns that reveal the features responsible for misleading the classifier. In sepsis diagnosis, misclassified samples are crucial due to their impact on patient care and resource allocation. False negatives can delay critical treatments, increasing the risk of complications and death, while false positives lead to unnecessary treatments and strain healthcare resources.

5.3 Background

5.3.1 Sepsis

Sepsis is a serious and potentially life-threatening medical condition that arises when the body's response to an infection becomes uncontrolled and triggers widespread inflammation. This inflammation can lead to organ dysfunction and, in severe cases, septic shock, which is a condition where blood pressure drops to dangerously low levels, potentially causing multiple organ failure [17].

Sepsis and septic shock can occur in response to various types of infections, such as bacterial, viral, fungal, or even parasitic infections. Common sources of infection include urinary tract infections, pneumonia, abdominal infections, and skin infections. The body's immune system response, intended to fight off the infection, can sometimes go into overdrive, causing harm to the body's own tissues and organs [5].

Early detection and intervention are crucial in managing sepsis and preventing it from progressing to septic shock [6]. Some of the main factors and features that medical professionals consider when detecting sepsis include Fever or Hypothermia, increased heart rate, and rapid breathing [7].

Recent advances in technology have led to a growing interest in using machine learning methods to identify sepsis and predict septic shock at an early stage [8]. Machine learning involves using computers to find patterns and gather information from large sets of data. This approach has proven useful in spotting subtle signs and connections that might not be easily noticed using traditional analysis methods. The goal is to identify the initial signs that indicate the beginning of sepsis, which allows for quick action and improved care for patients.

In this context, a variety of algorithms have been used to study different types of clinical information, including vital signs, lab test results, and medical notes. By carefully examining these complex sets of data, machine learning algorithms aim to uncover hidden patterns that suggest the potential development of sepsis. These algorithms play a role in predicting the occurrence of sepsis, providing healthcare providers with valuable insights to take proactive medical steps.

However, the complexity of the machine learning models might hinder their integration into real-world clinical settings. Moreover, building trust in the accuracy of these algorithms is crucial for practical use in healthcare. Ensuring this trust is vital to guarantee their dependability, especially in healthcare where decisions have serious consequences.

5.3.2 LIME

Local Interpretable Model-Agnostic Explanations (LIME) is a model-agnostic technique used in explainable artificial intelligence (XAI). It provides interpretable explanations for predictions made by complex machine learning models, even those considered black-box models. The main motivation behind LIME is to address the challenge of understanding complex models that lack easily interpretable relationships between input features and predictions [9].

LIME assigns importance weights to features to indicate their contribution to individual predictions. It achieves this by creating a simplified, interpretable surrogate model that approximates the black-box model's predictions locally around a specific data instance. This is accomplished through an optimization problem. Consider a scenario where we have a black-box model trained on a dataset with two features, like fever and body aches, to predict flu. The decision boundary of this complex model might be highly non-linear, making it challenging to provide straightforward explanations for its predictions. LIME's core idea revolves around zooming into the local area around a specific data instance and generating a simple explanation that is both valid and meaningful for that local region. By focusing on the local neighborhood, LIME can avoid trying to summarize the entire decision boundary while still offering relevant explanations.

To generate an explanation for a specific data instance, LIME perturbs the instance and creates a new dataset that represents a local neighborhood. It then obtains predictions from the black-box model for this perturbed dataset. A surrogate model, which is simple and interpretable, is trained on the perturbed dataset to approximate the black-box model's behavior in the local region. This surrogate model serves as a transparent proxy for the black-box model within the vicinity of the data instance of interest.

LIME commences by perturbing input features of a target instance, generating subtle variations through controlled modifications, while maintaining the instance's original label. This perturbation introduces diversity into the feature space, capturing the intricacies of the instance's surroundings.

Next, these perturbed instances traverse through the original complex model, yielding predictions that reveal the model's behavior under distinct feature conditions. This step effectively provides insight into how the model responds to perturbations, offering a window into its decision-making process.

LIME orchestrates the construction of a dataset by amalgamating altered instances with their corresponding predictions from the original model. Additionally, LIME computes "weights" to quantify the closeness between each modified variant and the initial instance. These weights encapsulate proximity, enriching our comprehension of local relationships.

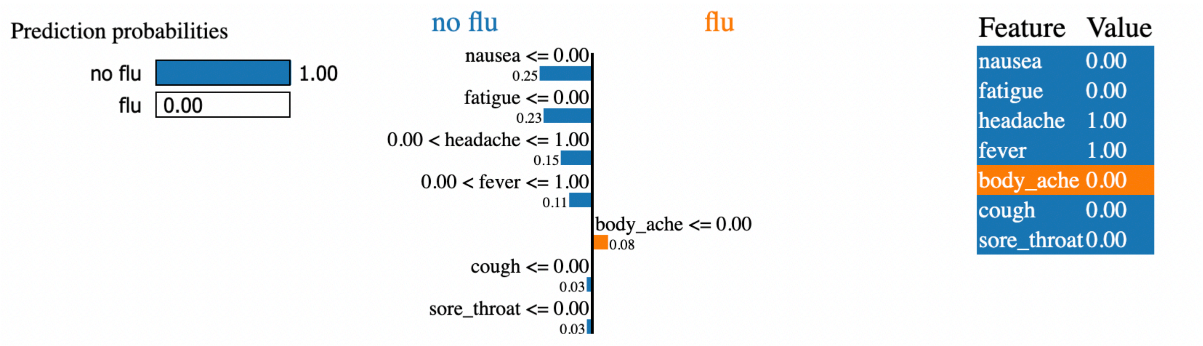
LIME introduces a local interpretable model, often a simpler linear regression, to mimic the intricate behavior of the complex model in the immediate vicinity of the target instance. By training this interpretable model on the set of modified instances and their predictions, LIME aims to distill the intricate dynamics of the model into a more accessible form.

The crux of LIME resides in the coefficients stemming from the local interpretable model. These coefficients function as a magnifying lens, exposing the individual impacts of features on specific predictions. Through their interpretation, practitioners glean insights into how adjustments in input features resonate through the prediction mechanism, offering a nuanced grasp of cause-and-effect relationships. In essence, LIME's seamless process unveils the decision-making of complex models, furnishing a coherent framework for interpretation.

The outcome produced by LIME is a set of explanations that outline the individual influence of each feature on predicting a specific data point as shown in Figure 5.1. This image presents a LIME-generated explanation for a particular patient, illustrating a classification outcome of ”no flu.” The explanation highlights features that align with and diverge from this classification, along with their corresponding weights. As depicted in the figure, nevertheless, the absence of body aches suggests a ”flu” diagnosis, but the cumulative weight of other features strongly leans toward the ”no flu” classification, resulting in an overall verdict of ”no flu”. This explanation not only offers a localized understanding of predictions but also helps pinpoint which features have the most significant impact on the predictions.

Creating an explanation with LIME involves approximating the behavior of the underlying model within a specific area using a simpler, interpretable substitute. These interpretable models could be linear models with strong regularization, decision trees, and more. They are trained using slightly altered versions of the original data point, focusing on generating a reliable local estimate. This ”dataset” is generated by introducing variations like adding noise to continuous features, removing certain words, or obscuring parts of an image. By concentrating on approximating the complex model only within a local range near the data point, LIME significantly simplifies the task.

LIME aims for local faithfulness, ensuring its explanations are accurate within the local



Feature	Value
nausea	0.00
fatigue	0.00
headache	1.00
fever	1.00
body_ache	0.00
cough	0.00
sore_throat	0.00

Figure 5.1: An example of LIME applied to a classification problem at a specific data point. Blue indicates features that, if their values were increased starting from the given data point, would induce the classifier to have more confidence in a “no flu” classification. On the other hand, the value of the orange feature would lead to more confidence in a ”flu” classification.

neighborhood of the explained data instance. While global faithfulness may not hold across the entire dataset, LIME’s focus on local interpretability offers valuable insights into individual predictions and the model’s behavior at a local level.

Mathematical Formulation

LIME formulates an optimization problem to find the best-fitting surrogate model within the local neighborhood. Given a complex black-box model f and a simple model $g \in \mathcal{G}$, where $\mathcal{G}$ is a family of interpretable models, LIME seeks to minimize the discrepancy between f and g in the local region around the data instance x_i. The optimization function includes a loss term that measures the difference between the predictions of f and g for the perturbed data points in the local neighborhood. Additionally, LIME incorporates a regularization term (denoted by ω) to encourage sparsity in the weights of the interpretable model g. This regularization helps simplify the explanation by considering only a few significant features. The trained local surrogate model now serves as an explanation for the prediction made by the black box model for the instance of interest. The interpretable model can be easily analyzed to understand the factors influencing the prediction.

Mathematically, the local surrogate models with interpretability constraints can be formalized through the following optimization problem [21]:

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{5.1}$$

Where:

- explanation(x) is the explanation model for the instance x.
- f is the model to be explained
- G is the family of possible explanations (e.g., all possible linear regression models).
- $L(f, g, \pi_x)$ is the loss function that measures how close the explanation is to the prediction of the original black box model f while considering the proximity of the perturbed samples to the instance x.
- $\Omega(g)$ is the model complexity term that ensures the explanation model is kept simple and interpretable.

The first term of the loss function aims to find the optimal approximation of f by g in the vicinity of x. The second term enforces simplicity upon g. In essence, we seek g to mirror f within this local zone while maintaining minimal complexity—an intricate balance. These interpretable models often take the form of sparse linear models, focusing on key features to simplify explanations [22].

We generate new data points around x, slightly perturbing the features. These perturbed points are then passed through the complex model f to generate predictions, forming a new dataset. Utilizing this dataset, we train a classifier for g, minimizing the differences between predictions made by g and f. We introduce a proximity measure, π, reflecting the closeness of each point to x and impacting its contribution weight, similar to a heatmap where closer points carry a stronger influence.

Lastly, we ensure the simplicity of g by incorporating a complexity measure, Ω. For sparse linear models, this translates to driving numerous weights toward zero, thereby focusing the explanation on pivotal variables. By solving this optimization problem, we attain our local surrogate model g, illuminating the behavior of the complex model within that specific neighborhood.

Solving this optimization problem empowers LIME to determine the feature importance weights (Φ) for the specific instance x. This calculation is based on the similarity between the predictions of the complex model for the specific instacne and the prediction for the original instance. A widely used metric for this similarity is the cosine similarity.

Let $f(x)$ be the prediction of the complex model for the instance x, and $f(x')$ be the prediction for a perturbed instance x'. The cosine similarity ϕ_i for each feature i is calculated using the following formula:

$$\phi_i = \frac{\sum_{j=1}^{N}(f(x)_j \cdot f(x')_j)}{\sqrt{\sum_{j=1}^{N}(f(x)_j)^2} \cdot \sqrt{\sum_{j=1}^{N}(f(x')_j)^2}}$$

Here, N represents the number of classes if the problem is a classification problem. For regression problems, N would be 1.

The significance of ϕ_i lies in its representation of how much the prediction for feature i in the original instance aligns with the predictions for the perturbed instances.

To ensure that the weights collectively represent the relative impact on the prediction, the calculated weights ϕ_i are normalized to sum to 1.

In practical implementations of LIME, such as in R and Python, linear regression is often chosen as the interpretable surrogate model. The user needs to decide the number of features, K, to include in the interpretable model. A smaller K results in a more interpretable model, while a larger K can potentially improve the fidelity of the model to the black box predictions.

To select the features for the interpretable model, one can use methods like Lasso with a regularization parameter λ, which gradually removes features with zero weights, or employ strategies like forward or backward selection to determine the optimal set of features with K features in the model.

Advantages of LIME

LIME offers multiple benefits when contrasted with other methods within the domain of eXplainable Artificial Intelligence (XAI). A significant merit lies in its model-agnostic characteristic. Irrespective of the intricacy or underlying algorithm of a machine learning model, LIME can elucidate decision-making processes. This adaptability renders LIME a versatile tool within the realm of XAI.

Another advantage of LIME pertains to its capacity for generating localized explanations. By developing a custom local model that mimics the actions of the complex model, LIME generates personalized explanations for individual cases. This proves particularly advantageous when explanations need to be customized for individual users or specific contexts.

Limitations of LIME

While the fundamental concept behind LIME may seem straightforward, there exist a few potential drawbacks that warrant consideration.

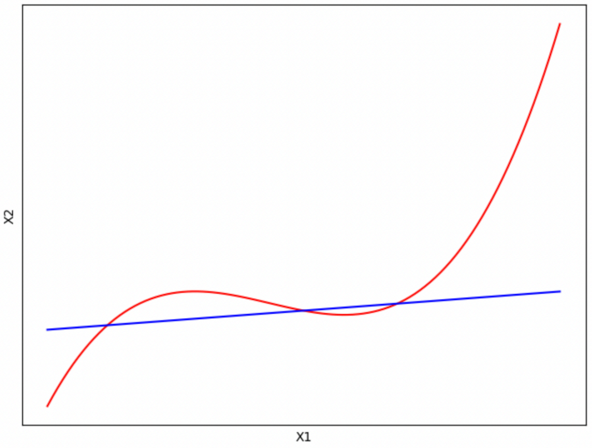


Figure 5.2: Inadequacy of Linear Approximation to Portray Local Behavior for Two Features and Its Inability to Capture the Model's Highly Non-Linear Behavior.

In the current implementation, LIME exclusively employs linear models to approximate local behaviors. This approach holds true to a certain extent when focusing on a very narrow region surrounding the data point. However, as this scope expands, it becomes conceivable that a linear model might lack the potency required to elucidate the intricacies of the original model's behavior. Non-linearity tends to emerge within localized regions, particularly in datasets demanding intricate, less interpretable models. The inability to apply LIME in such scenarios constitutes a noteworthy drawback.

Furthermore, the modifications needed to yield accurate explanations often exhibit specificity tied to the particular use case. The authors provide an illustrative instance in their paper:

consider a model that predicts a retro aesthetic for sepia-toned images—a prediction not easily explained by the presence or absence of superpixels.

Frequently, basic perturbations do not suffice. Ideally, the perturbations should mirror the variability observed in the dataset. However, manually directing these perturbations might not be a prudent approach, as it could potentially introduce bias into the model's explanations. This underscores a potential concern associated with steering perturbations through manual intervention.

Another limitation of LIME emerges from its susceptibility to perturbations. While LIME operates by perturbing instances to create a dataset of akin examples, even minor modifications to an instance can lead to notably distinct explanations. As such, LIME's explanations may not consistently withstand alterations in input data, thereby introducing an element of fragility to its explanatory outcomes.

5.4 Methodology

In this study, we adopted a modified visualization approach inspired by L. Torgo et al. [14] to identify the regions where a black-box model exhibits poor performance. L. Torgo et al. utilized the confusion matrix (CM) and cross-validation, and employed error distribution plots for each individual feature to demonstrate areas of inadequate model performance. However, we recognized the challenge of visual clutter arising from a large number of features. To address this limitation, we focused our analysis on identifying recurrent conditions associated with misclassifications, rather than visualizing misclassifications for each individual feature.

We applied LIME to misclassified data samples, allowing us to pinpoint the specific features responsible for incorrect predictions made by the classifier. This process was performed for each misclassified sample, enabling us to accumulate the features with high importance over the samples.

LIME calculates the feature importance scores based on the surrogate model g. These scores indicate the contribution of each feature to the prediction made by the black-box model for the specific data instance x_i. In the case of a linear surrogate model, the feature importance corresponds to the weights assigned to each feature. Positive weights indicate that an increase in the feature value leads to higher predictions, while negative weights suggest the opposite. For instance, a positive weight for the "fever" feature in a flu prediction model implies that a higher fever increases the likelihood of predicting flu.

Subsequently, we conducted an analysis involving the intersection of these features, with particular attention to those that consistently emerged as contributing elements to instances of misclassification. To achieve this, we compiled a list of features associated with each misclassified sample. From these lists, we identified features that were frequently shared and selected the top 10 features, along with their corresponding conditions, which were common across the misclassified samples. In particular, we focused on features that not only aligned positively with the misclassification but also carried considerable weight, emphasizing their dual importance as key indicators. This allowed us to discern regions or intervals in which the classifier demonstrated poor performance and was prone to misclassification.

Finally, we calculated the error rates within these regions by examining how often instances with these specific features were correctly classified versus misclassified. This analysis provided quantitative insights into the areas where the black-box model exhibited suboptimal performance and contributed to a deeper understanding of its limitations and potential areas for improvement.

In the following, we delve into the step-by-step explanation of the algorithm:

Given a set of misclassified samples, a black-box model for classification, and a threshold value, the algorithm proceeds to unravel the factors contributing to these misclassifications.

To begin, the algorithm initializes a dictionary named `misclassified_features` to capture the significant features driving misclassifications. It processes each misclassified sample, constructing a surrogate model using `train_surrogate_model()` and quantifying feature importance scores through `calculate_feature_importance()`. Subsequently, features with weights exceeding the predefined threshold are identified as "important features" and stored in the `misclassified_features` dictionary.

Building on this, the algorithm evaluates the frequency of these important features across the misclassified samples. It compiles a dictionary, `feature_frequency`, which counts how often each feature emerges. The dictionary is then sorted in descending order of feature frequency. The top ten features, termed `top_10_features`, are selected from this sorted list.

Algorithm 2 Algorithm for Analyzing Misclassifications using LIME

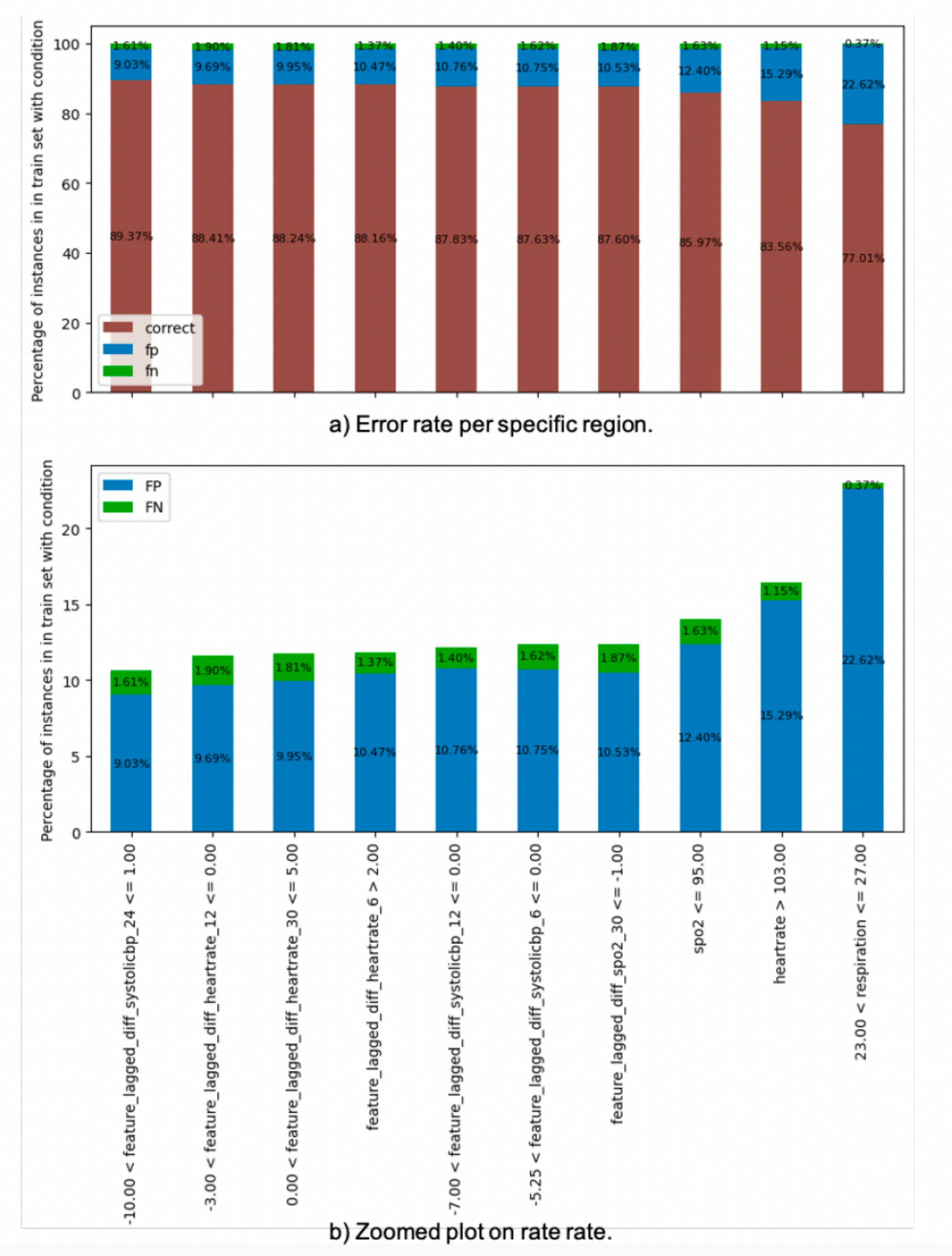
```
Input: misclassified_samples, black_box_model, threshold
Initialize dictionary misclassified_features
for all misclassified_sample in misclassified_samples do
    surrogate_model = train_surrogate_model(misclassified_sample)
    feature_importance_scores = calculate_feature_importance(surrogate_model, misclassified_sample)
    important_features = $\{f \mid \text{weight}(f) > \text{threshold}\}$
    Store important_features in misclassified_features[misclassified_sample]
Initialize dictionary feature_frequency
for each features in misclassified_features do
    for each feature in features do
        if feature exists in feature_frequency then
            Increment feature_frequency[feature] by 1
        else
            Initialize feature_frequency[feature] to 1
Sort feature_frequency by frequency in descending order
Get top 10 features as top_10_features from feature_frequency
Initialize an empty list samples_with_top_features
for each training sample training_sample do
    for each feature, condition in top_10_features do
        if satisfies_condition(training_sample, feature, condition) then
            Append training_sample to samples_with_top_features
            break
Initialize correctly_classified = 0, misclassified = 0
for all sample in samples_with_top_features do
    if black_box_model.predict(sample) == instance.true_label then
        correctly_classified += 1
    else
        misclassified += 1
error_rate = misclassified / (correctly_classified + misclassified)
Output: top_10_features, error_rate
```

Next, the algorithm identifies samples in the training set that align with the `top_10_features`. For each training sample, it evaluates whether any feature-condition pair from `top_10_features` is satisfied. If so, the sample is added to the list `samples_with_top_features`.

The algorithm's subsequent phase delves into performance assessment. It segregates samples from `samples_with_top_features` into correctly classified and misclassified categories. By comparing the model's predictions with the true labels, it computes the error rate—a ratio of misclassified samples to the total of correctly classified and misclassified samples.

5.5 Result

In this study, we employed the publicly available eICU dataset [24] to develop a predictive model for sepsis. The dataset comprises the vital signs of thousands of patients sampled at various rates. The vital signs considered in our experiments included systolic blood pressure, diastolic blood pressure, heart rate, respiration rate, oxygen saturation (SpO2), and gender of the patients. To standardize the sampling rates, all vital signs were resampled at a frequency of 5 minutes. To build a classifier for this time-series data, we employed LightGBM [25]. The time-series data was transformed into a format compatible with the LightGBM classifier [13] by calculating rolling statistical properties such as mean, standard deviation, and lag values from previous timestamps. The model's parameters were optimized using Python library Optuna [26]. The model achieved a recall score of 0.9308 and 0.8125 on in-sample and out-of-sample splits. To gain a deeper understanding of the model's performance and identify areas where it exhibits suboptimal results, we applied the proposed method. To visualize and communicate the regions of poor model performance, we present Figure 5.3. Figure 5.3.a demonstrated the error distribution over specific regions where the classifier exhibited suboptimal performance. Additionally, Figure 5.3.b provides a magnified view of the error distribution, offering a clearer resolution and facilitating a more detailed examination of the error rates within these identified regions. As indicated in the graph depicted in Figure 5.3, there is an important observation related to respiration levels between 23 and 27. During this range, the model's accuracy notably drops, resulting in a high error rate of 22.9%, which is significantly higher than the average error rate. Interestingly, a significant portion—about 22.62%—of this error can be attributed to



a) Error rate per specific region.

b) Zoomed plot on rate rate.

Figure 5.3: an overview of the entire visual analytics tool including interactive ranking list and what-if panel, a) Interactive List of ranked patients and Treemap Visualization of Patient ranked 1 (P1) Features. b) What-if Panel with Histograms of 5 Top Important Features.

cases where the model incorrectly identifies negatives. In other words, the model tends to classify instances as negative when they should be positive within this respiration range. Another noteworthy observation pertains to instances where the heart rate surpasses 103, as illustrated in Figure 5.3. Notably, the error rate in such cases is comparatively elevated, measuring 16.44%. This graph substantiates the notion that when a sample presents with any of these conditions depicted in the graph, exercising caution is imperative when utilizing the model's predictions. Further investigation is recommended to ascertain the model's accurate classification in such circumstances.

These figures serve as visual aids, aiding in the comprehension and interpretation of the model's performance shortcomings. In order to gain insights into the causes of misclassifications, we conducted a detailed analysis to determine which feature regions were most influential in contributing to these errors. Employing the LIME technique, we extracted the most significant features that consistently played a role in misclassification instances. By identifying and examining these recurring features, we revealed specific regions where the classifier exhibited poor performance. Figure 5.3 visually illustrates the feature regions that meet this criterion, highlighting the factors associated with the model's suboptimal predictions.

5.6 Discussion

In this study, we utilized LIME to identify regions where a black-box model exhibits poor performance. This approach allows us to investigate the error distribution across misclassification regions in both training and test data. The proposed method is model agnostic and can be utilized for any classifier. By analyzing the model's fit to the training data, we gain insights into its performance and identify areas where it inadequately represents the underlying patterns in the feature space. This assessment helps us understand the model's limitations in capturing the complexities of the training dataset. When evaluating the model's generalization error on test data, we pinpoint specific regions within the feature space that contribute to erroneous predictions for unseen data. In our case study, this was when respiration had the value between 23 and 27 and also where heartrate was greater than 103. This knowledge is crucial for important decision-making situations, such as sepsis, where being aware of regions requiring caution is essential when relying on the classifier's predictions. By conducting this analysis, we obtain a comprehensive understanding of the model's limitations and areas of poor performance. This knowledge empowers healthcare professionals and decision-makers to make informed judgments, taking into account the regions in the feature space where the classifier's predictions may be less reliable.

5.7 Conclusion

Our study contributes to the understanding of machine learning models' performance by introducing a modified visualization approach that identifies regions of poor performance. By leveraging LIME for the rule extraction method, we effectively pinpointed specific features responsible for misclassifications, allowing us to identify recurrent conditions associated with the classifier's suboptimal performance. The application of this methodology to the eICU dataset demonstrated its effectiveness in capturing regions where the classifier exhibits poor performance. These findings enhance interpretability and provide insights for decision-makers, enabling them to make informed choices regarding the deployment of machine learning models in critical domains such as sepsis detection. In light of the study's insights, our future work aims to enhance the model's performance by making specific modifications to the model architecture, feature engineering, and training strategies.

5.8 References

Bibliography

[1] Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *New England Journal Of Medicine*. **380**, 1347-1358 (2019)

[2] Antoniadi, A., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. & Mooney, C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*. **11**, 5088 (2021)

[3] Adadi, A. & Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. **6** pp. 52138-52160 (2018)

[4] Hotchkiss, R., Moldawer, L., Opal, S., Reinhart, K., Turnbull, I. & Vincent, J. Sepsis and septic shock. *Nature Reviews Disease Primers*. **2**, 1-21 (2016)

[5] Sagy, M., Al-Qaqaa, Y. & Kim, P. Definitions and pathophysiology of sepsis. *Current Problems In Pediatric And Adolescent Health Care*. **43**, 260-263 (2013)

[6] Yealy, D., Huang, D., Delaney, A., Knight, M., Randolph, A., Daniels, R. & Nutbeam, T. Recognizing and managing sepsis: what needs to be done?. *BMC Medicine*. **13** pp. 1-10 (2015)

[7] Sullivan, B. & Fairchild, K. Predictive monitoring for sepsis and necrotizing enterocolitis to prevent shock. *Seminars In Fetal And Neonatal Medicine*. **20**, 255-261 (2015)

[8] Fagerström, J., Bång, M., Wilhelms, D. & Chew, M. LiSep LSTM: a machine learning algorithm for early detection of septic shock. *Scientific Reports*. **9**, 15132 (2019)

[9] Khedkar, S., Subramanian, V., Shinde, G. & Gandhi, P. Explainable AI in healthcare. *Healthcare (April 8, 2019). 2nd International Conference On Advances In Science & Technology (ICAST)*. (2019)

[10] Fleuren, L., Klausch, T., Zwager, C., Schoonmade, L., Guo, T., Roggeveen, L., Swart, E., Girbes, A., Thoral, P., Ercole, A. & Others Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*. **46** pp. 383-400 (2020)

[11] Cerqueira, V., Torgo, L. & Mozetič, I. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*. **109** pp. 1997-2028 (2020)

[12] Flach, P. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **33**, 9808-9814 (2019)

[13] Duivesteijn, W. & Thaele, J. Understanding where your classifier does (not) work–the SCaPE model class for EMM. *2014 IEEE International Conference On Data Mining*. pp. 809-814 (2014)

[14] Torgo, L., Azevedo, P. & Areosa, I. Beyond Average Performance–exploring regions of deviating performance for black box classification models. *ArXiv Preprint ArXiv:2109.08216*. (2021)

[15] Roshan, K. & Zafar, A. Utilizing XAI technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (SHAP). *ArXiv Preprint ArXiv:2112.08442*. (2021)

[16] Fryer, D., Strümke, I. & Nguyen, H. Shapley values for feature selection: The good, the bad, and the axioms. *Ieee Access*. **9** pp. 144352-144360 (2021)

[17] Hotchkiss, R., Moldawer, L., Opal, S., Reinhart, K., Turnbull, I. & Vincent, J. Sepsis and septic shock. *Nature Reviews Disease Primers*. **2**, 1-21 (2016)

[18] Fleuren, L., Klausch, T., Zwager, C., Schoonmade, L., Guo, T., Roggeveen, L., Swart, E., Girbes, A., Thoral, P., Ercole, A. & Others Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Medicine*. **46** pp. 383-400 (2020)

[19] Antoniadi, A., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. & Mooney, C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*. **11**, 5088 (2021)

[20] Adadi, A. & Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. **6** pp. 52138-52160 (2018)

[21] Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. (2022), https://christophm.github.io/interpretable-ml-book

[22] Slack, D., Hilgard, S., Jia, E., Singh, S. & Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *Proceedings Of The AAAI/ACM Conference On AI, Ethics, And Society*. pp. 180-186 (2020)

[23] Ribeiro, M., Singh, S. & Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*. pp. 1135-1144 (2016)

[24] Pollard, T., Johnson, A., Raffa, J., Celi, L., Mark, R. & Badawi, O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*. **5**, 1-13 (2018)

[25] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. Lightgbm: A highly efficient gradient boosting decision tree. *Advances In Neural Information Processing Systems*. **30** (2017)

[26] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. *Proceedings Of The 25th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. pp. 2623-2631 (2019)

Chapter 6

Unveiling Bias and Discrimination in Ranking through Visualization

6.1 Introduction

With the increasing influence of advanced machine learning in both our online and offline experiences, there is a growing concern that automated decision models might incorporate biased practices [1]. In today's digital age, ranking algorithms play a crucial role in shaping our online experiences, influencing the information we consume, and guiding our decision-making processes. However, these algorithms are not immune to bias and discrimination, raising concerns about their fairness, transparency, and the potential harm they can perpetuate in various domains. Recognizing and understanding the presence of bias and discrimination in ranking systems has become an essential research area with profound societal implications.

The field of algorithmic fairness research aims to ensure unbiased practices, particularly with regard to sensitive attributes like race, gender, and age, which legal regulations prohibit from influencing certain decision outcomes. Research to date [2, 3, 35] has primarily focused on classification tasks, where predictive models determine binary outcomes. Several fairness criteria have been suggested for such tasks, and their relative advantages and trade-offs have been investigated [5]. Generally, it has been established that not all criteria can be fulfilled simultaneously [6]. The choice of appropriate fairness criteria heavily relies on the specific problem domain.

More recently, attention to fairness in machine learning has expanded to encompass fair ranking, which holds significant importance in tasks like information retrieval (IR) that underlie socio-technical systems. Fairness researchers have noted the necessity for a comprehensive analysis of equitable fairness ranking metrics [7], although this subject has not received substantial attention thus far.

Rankings now play a crucial role in many domains such as information retrieval, university admission, and more [8, 9, 10]. The fairness of ranked results can be influenced by various factors, including historical biases, the misrepresentation of groups in training data [11], biases embedded in tools used for data interpretation, such as those for image [14] and text analysis [15], as well as implicit biases inherent in user interaction patterns [16].

Fairness in ranking holds profound significance due to its impact on equitable opportunities and the prevention of discrimination. In various domains, ranging from information retrieval to critical sectors like healthcare and education, ranking systems are employed to make decisions that directly influence individuals' lives. Ensuring fairness in these systems is paramount to prevent biases from perpetuating and exacerbating existing societal inequalities. When rankings favor certain groups based on sensitive attributes such as race, gender, or socioeconomic background, it can result in unjust outcomes and hinder equal access to opportunities. Fair ranking mitigates these concerns by employing objective criteria that focus solely on relevant attributes, thus safeguarding against discrimination and fostering a more just and inclusive environment. Moreover, fair ranking contributes to the establishment of trust in automated decision-making processes, bolstering transparency and accountability in systems that impact diverse populations.

Ranking systems serve a multitude of purposes, each tailored to meet the distinct needs of different users. As a consequence of this adaptability, fairness interpretations become linked to the specific scenario at hand. This dynamic landscape necessitates the availability of a range of fairness metrics, each tailored to address different contextual demands. It is of paramount importance to thoroughly grasp these proposed metrics, empowering experts to select the metric that aligns best with the nuances of their specific application.

This chapter aims to explore the subject of bias and discrimination in ranking systems, emphasizing the importance of visualization as a powerful tool for uncovering and comprehending these underlying issues. By visualizing the patterns and disparities present in ranking algorithms, we can shed light on the mechanisms that contribute to bias and discrimination, providing valuable insights into the operation and impact of these systems.

The importance of understanding the presence of bias and discrimination in ranking algorithms cannot be overstated. Firstly, it is a prerequisite for providing equal opportunities and treatment for all individuals, irrespective of their background, race, gender, or other protected attributes. By identifying and mitigating biases, we can strive towards creating fairer and more inclusive ranking systems that promote diversity and avoid reinforcing existing societal inequalities.

Secondly, acknowledging the biases present in ranking algorithms is essential for transparency and accountability. As these algorithms increasingly shape our lives, it is imperative that we comprehend their inner workings and the potential consequences they have on individuals and communities. Visualization serves as a powerful tool in this endeavor, as it provides a tangible representation of the biases and discrimination embedded within these systems, making them more comprehensible and accessible to both researchers and the general public.

Finally, by exploring the work that has been done in this field, we can build upon existing knowledge and develop strategies to mitigate bias and discrimination in ranking algorithms effectively. Previous research efforts have highlighted various techniques and methodologies for understanding and visualizing bias, laying the foundation for further exploration and improvement.

In summary, this chapter aims to explore the intricate relationship between bias, discrimination, and ranking algorithms. By employing visualization techniques, we seek to reveal the hidden biases within these systems and advance our understanding of their impact on individuals and society. By building upon previous research, we hope to contribute to the ongoing efforts of creating fairer, more transparent, and accountable ranking algorithms in the digital landscape.

This chapter makes a contribution to the field by developing a novel visual analytics tool specifically designed to assess bias in ranking systems. The tool provides users with diverse options to explore and interact with the interface, empowering them to uncover hidden biases and obtain relevant information effectively. This study incorporates interdisciplinary research from fields such as Epidemiology, fairness, XAI-Healthcare, interactive visualization, HCI, and causal inference, reinforcing the comprehensive nature of our work.

6.2 Background

6.2.1 Bias

Bias or discrimination in ranking systems refers to the presence of unfair or unjust treatment based on specific *sensitive attributes* when generating rankings or recommendations [17]. A sensitive attribute refers to a characteristic or personal attribute that is considered protected or sensitive due to its potential for discrimination or bias. These attributes are typically intrinsic to individuals and include factors such as sex, race, age, income, disability status, sexual orientation, or religion. Sensitive attributes are protected under various anti-discrimination laws and regulations to ensure fair treatment and equal opportunities for individuals regardless of these attributes [18]. When ranking systems exhibit bias or discrimination, it means that the system's outcomes or recommendations are influenced by the sensitive attributes of individuals, leading to unequal treatment or disadvantages based on these attributes.

For instance, in a job search ranking system, bias may occur if the system favors candidates of a particular race or gender, resulting in individuals from other racial or gender groups receiving lower rankings or fewer opportunities.

Bias or discrimination in ranking systems can arise due to various factors, including biased training data, algorithmic design choices, or systemic societal biases [19] that reflect historical inequalities. These biases can arise from various factors, including:

Sampling Bias: Bias can occur if the training data used to develop the ranking system is not representative of the diverse population it aims to serve. For example, if the training data predominantly consists of individuals from certain demographic groups, the system may inadvertently favor those groups, leading to biased outcomes [20].

Stereotyping Bias: Ranking systems may inadvertently perpetuate stereotypes by associating certain attributes with specific outcomes. For instance, if a job ranking system favors candidates from prestigious universities, it may inadvertently reinforce existing biases in hiring practices, disadvantaging candidates from underrepresented backgrounds who may have equally relevant skills and experiences.

Proxy Bias: Proxy variables, such as zip codes or educational attainment, may be used as substitutes for sensitive attributes like race or income [21]. However, these proxies can introduce bias if they are correlated with the protected attributes and result in unfair rankings or recommendations. For example, if a loan ranking system uses zip code as a proxy for income, it may disadvantage individuals residing in historically marginalized neighborhoods.

Feedback Loop Bias: Ranking systems that incorporate user feedback may be susceptible to feedback loop biases. If the feedback itself is influenced by existing biases or discriminatory practices, the system may perpetuate and amplify those biases over time. This can create a cycle of bias, where certain groups consistently receive lower rankings or recommendations due to historical inequalities [22].

Contextual Bias: Ranking systems may fail to account for the contextual factors that affect individuals' experiences and qualifications [23]. For instance, a college ranking system that does not consider the socioeconomic background of students may unfairly favor institutions that predominantly enroll students from privileged backgrounds, overlooking the achievements of students who have overcome significant barriers.

It is crucial to identify and address these biases to ensure fair and equitable outcomes in ranking systems, promoting equal opportunities and reducing unjust disparities based on sensitive attributes.

6.3 Fairness Metrics

The proposed notions of fairness in ranking primarily focus on achieving fairness among groups [12, 31, 32]. This approach aims to secure equal treatment or results for groups of individuals based on attributes such as race, gender, or age that are protected by law. Most of these efforts employ statistical measures of parity. Statistical parity, one of the simplest definitions of fairness, entails ensuring that each group receives a just share of favorable outcomes. This concept is especially valuable when there exists a necessity for diversity as a means to attain equitable distribution for groups that have previously faced historical discrimination.

However, it has been recognized that imposing statistical parity might come at the expense of predictive accuracy and potentially compromise fairness for individuals [33]. Due to such concerns, adaptations of fairness definitions used in classification have also been suggested for ranking scenarios. The idea of individual fairness in classification proposes that similar individuals should be treated similarly. This principle has been extended to the ranking by Biega et al. [34]. Equalized Odds criteria, introduced in the context of classification by Hardt et al. [35], aim to ensure that an object's probability of receiving a specific label from the classifier remains independent of its group membership, given the true class label. Equalized Odds require that the false positive and true positive error rates are comparable across all groups. Lastly, fairness definitions grounded in causality [36, 37] try to understand the connections between data attributes and predicted outcomes, offering an alternative approach to assessing fairness beyond evaluation metrics.

Fairness definitions tailored for classification tasks revolve around the fact that individuals under evaluation will receive either positive or negative outcomes, corresponding to positive and negative classes. However, in the context of ranking tasks, determining a preferred outcome is more nuanced. Here, the position in the ranking dictates the outcome for the ranked items, granting advantages to those placed in higher positions. Yet, in rankings, the position is relative, influenced by multiple factors such as the quality of other items on the list and the significance of specific ranks (referred to as position bias [38]). Proposed fairness metrics for ranking aim to address this complexity by gauging group advantage within a ranking, employing established approaches in information retrieval (IR): including top-k analysis [39], pairwise inversions [40], and cumulative discounted metrics [11]. In this research, we will extend this IR approach to our ranking system, utilizing it to assess bias and subsequently create a visualization that enables users to interact with and discover the information they require.

6.4 Formulating the Fair Ranking Problem

The concept of ranking can carry different interpretations in various contexts, and models designed for ranking can be trained using diverse types of ground truth data. Rank predictions can be generated from training data that employ binary labels or discrete labels with ordered categories. Traditional regression assigns ranks based on continuous scoring functions. Learning-to-rank approaches also encompass pairwise and listwise models [12]. To ensure broad applicability, we adopt a model-agnostic approach targeting general rankings. Our assumption is that an order is established for a set of candidates x_i from the set X. This order defines a ranking of X, represented as a permutation $\pi = [x_1 \prec x_2 \prec \ldots \prec x_n]$ encompassing all candidates. Within this, π signifies a complete ordering relation on X, where $x_i \prec x_j$ implies that x_i is more favorably positioned than x_j in the ranking π. The position of a candidate x_i within the ranking π is denoted as $\pi(x_i)$. Following the convention, lower numerical positions are deemed more favorable, with $\pi(x_i) = 1$ representing the highest rank position.

In a unique context relevant to fairness analysis, each ranked candidate also possesses associated protected attributes, such as race, gender, or age. These attributes divide the dataset into distinct or overlapping groups, denoted as $\{G_1, \ldots, G_m | \cup_{i=1}^{m} |G_i| = |X|\}$. In many cases, a specific group corresponds to a minority or disadvantaged based on their sensitive attribute.

Definition 6.1. Given a group error metric $L_{G_i}(\pi, \hat{\pi})$, a Fairness Criterion (FC) is an evaluation rule that designates a ranking $\hat{\pi}$ as fair concerning a true ranking π if: $L_{G_i}(\pi, \hat{\pi}) = L_{G_j}(\pi, \hat{\pi})$ for all G_i and G_j where $i \neq j$. Fairness is assessed by comparing the errors for each group, determining if they are similar or within a predefined threshold, denoted as ϵ. The greater the disparity in errors among the groups, the more unfair the ranking is considered to be. Hence, our evaluation hinges on selecting a suitable group error function L tailored for rankings.

6.4.1 Defining Groups

While much of the research on algorithmic fairness centers around the scenario of two binary groups, the real-world situation involves candidates with complex, intersecting identities that encompass more than one protected group. In certain cases, a single candidate item might relate to multiple individuals. Often, practical scenarios may lack access to sensitive data for analysis [13]. We explore the potential expansion of our methodologies to scenarios with multiple overlapping groups as relevant. To simplify matters, we will focus on analyzing fairness within two distinct groups for the remainder of this study.

6.5 Related Work

In this section, we explore a range of innovative visual analytics systems that address fairness and discrimination concerns within machine learning, decision-making, and ranking contexts. FAIRVIS [24] is introduced as a mixed-initiative visual analytics system that incorporates a novel subgroup discovery technique to assess the fairness of machine learning models. This tool allows users to apply domain knowledge to generate and explore known subgroups, as well as investigate suggested and similar subgroups. FAIRVIS's coordinated views provide both a high-level overview of subgroup performance and detailed investigation capabilities. Through its interactive visualization, FAIRVIS aids in discovering biases in real-world datasets used for income prediction and recidivism. The system's primary focus is on helping data scientists and the general public comprehend and build more equitable algorithmic systems. FairSight [25] proposes a visual analytic system designed to achieve various notions of fairness in ranking decisions. The tool identifies actions required to enhance fairness in decision-making processes, including understanding, measuring, diagnosing, and mitigating biases. Through a case study and user study, FairSight's visual analytic and diagnostic modules are demonstrated to effectively aid in understanding the fairness-aware decision pipeline and achieving fairer outcomes.

DiscriLens [26] offers an interactive visualization tool to comprehensively analyze discrimination in machine learning. It employs causal modeling and classification rules mining to identify potentially discriminatory itemsets. The tool's combination of extended Euler diagrams and matrix-based visualization introduces a novel set visualization method. Through this approach, DiscriLens facilitates the exploration and interpretation of discriminatory itemsets, as confirmed by a user study. The tool proves informative in understanding and reducing algorithmic discrimination. RMExplorer [27] is an interactive visualization system developed for risk model assessment. Users can define patient subgroups based on various characteristics, explore risk model performance and fairness within these subgroups, and understand feature contributions to risk scores. A case study involving atrial fibrillation risk models demonstrates the tool's utility. RMExplorer empowers researchers to assess risk model performance and biases within specific subpopulations, contributing to a better understanding of model behavior.

FairFuse [28] introduces a visualization system to generate, analyze, and audit fair consensus rankings. The tool employs parallel-coordinates style rank visualizations to encode group fairness measures. Users can generate and explore fair consensus rankings through interactions. FairFuse supports decision-makers in ranking scenarios where fairness is a concern, contributing to a more balanced decision-making process. FairRankVis [29] presents a visual analytics framework for exploring multi-class bias in graph mining algorithms. The tool supports group and individual fairness comparisons, enabling developers to assess algorithmic debiasing impacts. The framework showcases two usage scenarios investigating algorithmic fairness. Fairlearn [30] includes an interactive visualization dashboard and unfairness mitigation algorithms. It aids in navigating trade-offs between fairness and model performance. Fairlearn acknowledges that complete debiasing is challenging due to complex sources of unfairness, emphasizing the importance of mitigating fairness-related harms.

Introducing a novel viewpoint, we propose a visual analytics tool engineered to evaluate bias and fairness within ranking systems. This tool's standout feature is its model-agnostic nature, allowing it to integrate with diverse ranking methodologies, making it adaptable to a wide array of scenarios. While we specifically demonstrate its effectiveness in the context of ICU patient prioritization, its adaptability extends beyond, fitting various applications. By harnessing intersectional sensitive attributes, coupled with fairness metrics and interactive visualization, we introduce a comprehensive framework. This empowers users to conduct in-depth examinations, enabling them to thoroughly investigate and assess potential biases.

6.6 Methodology

In this section, we provide a detailed description of the methods employed to assess and visualize bias in the ranking system. Our approach utilizes several techniques aimed at gaining insights into the presence and impact of bias based on sensitive attributes. The following methods are utilized:

Rank Equality

The fairness assessment in classification, known as the Equalized Odds criteria, focuses on the rate at which different groups are incorrectly assigned to preferred or non-preferred categories. In the context of evaluating rankings, the determination of preference is not binary but is instead based on the relative positions within the ranking. The top position in a ranking can be likened to the positive class, symbolizing a preferred outcome.

When a model overestimates the position of an object, it inaccurately assigns it a more favorable ranking than its true position. This can be compared to a false positive error made by a classifier. Similarly, underestimating an object's position penalizes it incorrectly, similar to a false negative. Following this principle, we quantify the Rank Equality error for a group (G_i) by counting the number of instances where pairs of items in the predicted ranking incorrectly favor that group over another group (G_j). This metric, described in Definition 7.1, captures the frequency of objects from group G_i being erroneously overestimated compared to those from group G_j.

The Rank Equality error is normalized by the total number of mixed pairs, ensuring that the error value falls within the range of [0, 1]. This normalization process provides an understandable measure of preference and addresses any disparities in group sizes.

Definition 7.1: Rank Equality Error Given a true ranking (ρ) and a predicted ranking (ρ') of items (x_i) belonging to mutually exclusive groups (G_1 and G_2), where ($\Phi_{D_{i \prec j}}(X)$) represents the count of discordant pairs favoring group G_1 over G_2 in the predicted ranking, and ($\Phi_{i,j}(X)$) is a count related to the number of pairs within the ranking, the Rank Equality error for group G_1 is computed as follows:

$$Req_{G_1}(\rho, \rho') = \frac{\Phi_{D_{i \prec j}}(X)}{\Phi_{i,j}(X)}.$$

The concept of Rank Equality emphasizes that no group should experience unfair advantages or penalties compared to other groups.

Rank Parity

We utilize the concept of pair inversion to formulate a metric that aligns with the fairness criteria of statistical parity, as observed in prior research on equitable ranking [39, 31]. In this context, the objective is to ensure an equitable representation of individuals from each group within items that receive favorable ranking positions. Our proposal involves quantifying pairs in which one group is favored over the other in the acquired ranking, without considering their positions in the original ranking. To make this measure interpretable, we once again standardize it by the total count of mixed pairs within the acquired ranking.

Definition 7.3: Rank Parity Error Given a predicted ranking (ρ') of items ($x_i \in X$) belonging to two distinct and mutually exclusive groups (G_i and G_j), where ($\Phi_{i \prec j}(X)$) denotes the number of mixed pairs of objects favoring group G_i over G_j in the predicted ranking, and ($\Phi_{i,j}(X)$) signifies the total count of pairs within the ranking, the Rank Parity error for group G_i is computed as follows:

$$R_{\mathrm{eq}G_i}(\rho, \rho') = \frac{\Phi_{i \prec j}(X)}{\Phi_{i,j}(X)}.$$

The central aim of Rank Parity is to ensure fairness in the distribution of ranking privileges across various groups. This is achieved by assessing pairs where one group is favored over the other within the predicted ranking, regardless of their original positions. To compute the Rank Parity errors for groups G_i and G_j, we evaluate the count of mixed pairs favoring G_i over G_j. This calculation emphasizes equity and guards against any potential ranking imbalances.

6.6.1 Interactive Visualizaiton

Our comprehensive visual analytic tool comprises five principal components that users can seamlessly navigate through: Home, Compare, What-If, Intersectional Analysis, and Group Fairness. Within each of these sections, users are presented with a range of valuable insights, including ranking outcomes, statistical summaries of sensitive attributes, and fairness metrics. By actively engaging with the system's interactive features, users can delve into detailed investigations aimed at assessing bias and fairness within the system's operations. The subsequent sections will provide detailed explanations of each component, shedding light on their distinct purposes and roles within the broader analytical framework. This will enable a comprehensive understanding of how each component functions and contributes to the overall analytical structure.

Home component

In the “Home” component, users are presented with a vertically listed display of the top 10 ranked items, each represented as a separate row. These rows encompass a comprehensive depiction of the associated attributes, including four sensitive factors: race, age percentile, sex, and income interval. Concurrently, statistical insights into these sensitive attributes are visually encapsulated through pie charts and stacked bar plots.

A pie chart visually represents categorical data, illustrating the proportionate distribution of different categories within the given attributes. It particularly serves to depict gender and race attributes, revealing the male-to-female ratio and the distribution of racial categories within the population.

On the other hand, a stacked bar chart presents a graphical representation where each bar is segmented into different sub-components, reflecting the composition of two variables in relation to each other. In this context, it specifically elucidates the interrelation between sex and income interval, displaying how the two attributes intersect within the population.

These visual aids—pie charts and stacked bar plots—employed in the “Home” component serve as illuminating tools, providing insight into gender and race distributions as well as the correlation between sex and income, with the goal of enhancing the users’ grasp of the population’s attributes and dynamics.

Compare Component

Within the “Compare” component, users are presented with two distinct sets of ranking results for comparison. The initial list incorporates sensitive attributes during both the training and ranking processes, while the subsequent list omits these attributes from the training phase.

The first ranking list provides a direct representation of ranks from 1 to 10. Meanwhile, the second list, which exhibits the top 10 rankings, employs a different approach. In this list, the numerical indicators denote the corresponding ranks in the alternative list. For instance, an entry marked as “P8” signifies that, when sensitive attributes were excluded from the training, the item would have achieved the second rank, whereas, in the context where these attributes were incorporated, it occupied the second position.

In summary, the ”Compare” component furnishes users with the means to contrast ranking outcomes under differing conditions, offering insights into the impact of sensitive attributes on the system’s performance. the comparison can unveil the impact of sensitive attributes on the ranking order. When an entry is assigned a different rank between the two lists, it signals the potential influence of sensitive attributes on the system’s decisions. Such discrepancies can indicate instances, where items favored due to their sensitive attributes in one context, might be ranked differently when these attributes are not considered.

Using the indicators from the first ranking result to the second group provides a quantitative representation of how much the inclusion or exclusion of sensitive attributes alters the ranking position. This can highlight cases where certain attributes exert significant influence, potentially leading to biases favoring or disadvantaging specific attributes.

Furthermore, the comparison can also bring to light cases where the system is consistently or significantly biased across multiple items. For instance, if a particular sensitive attribute consistently leads to higher or lower rankings across the board, it suggests a potential systemic bias that requires scrutiny.

The side-by-side comparison of these two ranking lists equips users with the means to detect potential biases arising from sensitive attributes. By assessing discrepancies and understanding the degree of influence these attributes wield on the rankings, users can uncover patterns of potential bias and gain valuable insights into the fairness and equity of the system’s outcomes.

What-if Component

Within the ”What-If” component, users can modify the sensitive attribute associated with each item and subsequently observe the potential repercussions on its ranking. This dynamic feature not only permits users to explore hypothetical scenarios but also serves as a powerful tool for investigating the sensitivity of rankings to variations in sensitive attributes.

Through these ”what-if” experiments, users can uncover valuable insights into the extent to which specific attributes wield influence over an item’s position in the ranking. These investigations allow users to gauge whether certain attributes hold disproportionate sway over rankings, potentially leading to preferential or adverse treatment. Moreover, this facet provides an avenue to discern whether alterations in attributes lead to significant shifts in ranking positions, thereby illuminating the degree of sensitivity in the ranking process.

Furthermore, the ”What-If” component offers users the opportunity to identify attributes that might carry a higher propensity to introduce disparities or biases within the ranking outcomes. By interactively modifying attributes and observing corresponding changes in rankings, users can pinpoint attributes that may warrant closer examination for potential fairness-related concerns.

Intersectional Analysis Component

We use a heatmap visualization to represent intersectional analysis. A heatmap is a graphical representation commonly employed in data visualization to portray the distribution and relationships among two-dimensional data points through the use of colors. It constitutes a grid-like structure, where each cell corresponds to a specific combination of two variables. The colors within these cells are indicative of the magnitude or intensity of a third variable, typically represented by a numerical value.

Heatmaps visualize information by assigning colors to cells in a grid based on the values of the third variable. Darker or brighter colors are used to represent higher or lower values, respectively, creating a visual gradient that allows for easy interpretation. This technique allows patterns, trends, and anomalies within the data to become evident. This analysis involves the intersection of two sensitive attributes, allowing for an in-depth examination of potential disparities within minor groups. By employing heatmap visualization, it becomes possible to discern whether these minor groups are being treated similarly or if differences exist, thus facilitating investigations into group fairness. The colors within the heatmap cells serve as a clear visual cue to highlight variations in the analyzed data, offering a comprehensive and accessible means of understanding intricate relationships and patterns. This technique is particularly valuable for unraveling complex interactions and disparities within multidimensional datasets. We enhanced the functionality of this heatmap by introducing interactivity, enabling users to select their preferred sensitive attributes. This interactive feature empowers users to generate a heatmap specific to the intersection of their chosen attributes. This innovation serves as a powerful tool for discerning patterns and conducting diverse intersectional analyses. By offering the flexibility to investigate various attribute combinations, users are equipped to uncover nuanced insights and delve into the complexities of the dataset, which can contribute to a more comprehensive understanding of underlying relationships and disparities.

Group Fairness Component

In the section 6.6 and 6.6, we established the definitions of Equality of Ranking Position and Rank Parity as our fairness metrics. Building upon this foundation, we have chosen to utilize a bubble chart to visually depict these fairness metrics. A bubble chart serves as a graphical tool to represent data that involves three distinct variables. It operates within a two-dimensional grid, where one variable is positioned along the horizontal axis, another along the vertical axis, and the third is denoted by the size of circles, commonly referred to as "bubbles," at the intersections of this grid.

Through this visual representation, the values of each variable are conveyed by the position of their corresponding data points on the axes, while the magnitude of the third variable is symbolized by the size of the bubbles. This approach allows for the simultaneous examination of multiple dimensions of the data, thereby facilitating the identification of patterns, relationships, and discrepancies.

In our bubble chart, the size of each bubble corresponds to the fairness values, and various sensitive attribute groups are differentiated using distinct colors. This visual technique effectively presents fairness trends, offering a clear understanding of these values across different attribute groups. By combining bubble size, color differentiation, and interactivity, we enhance the clarity and depth of the information presented.

Furthermore, our bubble chart is designed to be interactive, empowering users to dynamically explore the fairness metric alongside different attributes. This interactive feature enables users to delve into potential instances of discrimination by observing how the fairness metric changes across various sensitive attributes. By facilitating this investigative process, the interactive bubble chart can be used to recognize disparities and can light on how different attributes influence fairness outcomes.

6.7 Case Study: Prioritizing Patients to ICU: Assessing Bias

In this case study, we aim to prioritize patients for admission to the Intensive Care Unit (ICU) using the SIRIO dataset [41] and the XGBoost machine learning algorithm. Additionally, we employ the methodology described previously to assess and visualize any potential bias in the prioritization process based on sensitive attributes. By leveraging advanced techniques, such as partial dependence plots, LIME, counterfactual explanations, fairness metrics, and sensitivity analysis, we gain insights into the presence and impact of bias in the ICU admission ranking system.

6.7.1 Data Preparation

We start by acquiring the SIRIO dataset, which comprises anonymized medical records of patients. The dataset consists of 1925 observations and 231 columns. Each observation is associated with a unique patient identifier, PATIENT_VISIT_IDENTIFIER. It is important to note that there may be multiple entries for the same PATIENT_VISIT_IDENTIFIER, representing different stages of the patient since admission.

The dataset includes a wide range of patient information, including demographic data, previously grouped diseases, blood results, vital signs, and blood gases. Additionally, there are 42 features that have been expanded to include the mean, maximum, minimum, difference, and relative difference.

Upon examining the dataset, we find that there are 385 unique patients, and each patient has five entries in the original dataset. These five entries correspond to the different windows during which the patients were monitored. However, it is crucial to exclude data from the windows when the patients were already transferred to the intensive care unit (ICU). This exclusion is recommended because the target event may have occurred before the results were obtained, as advised by the source of the dataset.

To analyze the relationship between patient admission to the ICU and their historical data, we need to restructure the dataset. Our objective is to have each entry in the dataset provide information about patients admitted to the ICU in the current window, as well as their data from the two immediate previous windows. Consequently, entries that contain data when the target variable is present (i.e., the patient was already in the ICU) should be excluded from the final dataset.

To address missing values, we utilize the values from the neighboring windows of the same patient to fill in the gaps. This approach helps maintain the temporal relationship and contextual relevance of the data, enabling us to obtain a more comprehensive understanding of the target variable and its association with the patient's historical information.

By restructuring the dataset and incorporating the necessary adjustments, we can create a refined dataset that allows us to analyze the relationship between patient characteristics, historical data, and ICU admission. This data preparation process ensures that the dataset is compatible with the subsequent steps, such as training the XGBoost model and assessing bias using the methodology described previously.

6.7.2 Model Training

After preprocessing the dataset, we proceed to train a predictive model using the XGBoost algorithm. XGBoost is a powerful gradient boosting algorithm widely recognized for its capability to handle complex patterns and generate accurate predictions. It has gained popularity in various domains due to its superior performance and ability to capture intricate relationships within the data.

In this case study, we compare the performance of different machine learning algorithms, including Support Vector Machines (SVM), Neural Networks, and XGBoost. We evaluate these algorithms based on their ability to predict the urgency or severity of a patient's condition, which serves as a proxy for ICU prioritization. XGBoost outperforms the other algorithms in terms of accuracy.

The XGBoost model is trained on the preprocessed dataset, utilizing a gradient-boosting framework that combines the outputs of multiple weak learners (decision trees) to form a robust and accurate ensemble model. XGBoost employs a combination of boosting and regularization techniques to iteratively improve the model's performance by minimizing prediction errors and preventing overfitting.

During the training process, the XGBoost model learns complex patterns and relationships from the dataset, capturing important features and their interactions. It is trained to predict the urgency or severity of a patient's condition based on the available input features, which encompass demographic information, medical history, vital signs, and other relevant factors. The model outputs predictive probabilities, which represent the estimated likelihood of a patient having a more severe condition.

These predictive probabilities serve as the basis for ranking patients in the ICU prioritization process. A higher probability indicates a higher estimated severity of the patient's condition, suggesting a greater need for immediate admission to the ICU. By utilizing these probabilities as a ranking criterion, we can effectively prioritize patients based on their estimated urgency, ensuring that those with more severe conditions receive the necessary care sooner.

Through extensive experimentation and performance evaluation, XGBoost demonstrates superior predictive capabilities compared to SVM, Neural Networks, and other algorithms considered in this study. Its ability to handle complex patterns, optimize model performance, and generate accurate predictions, including predictive probabilities, makes it well-suited for the task of ICU prioritization based on the severity or urgency of a patient's condition.

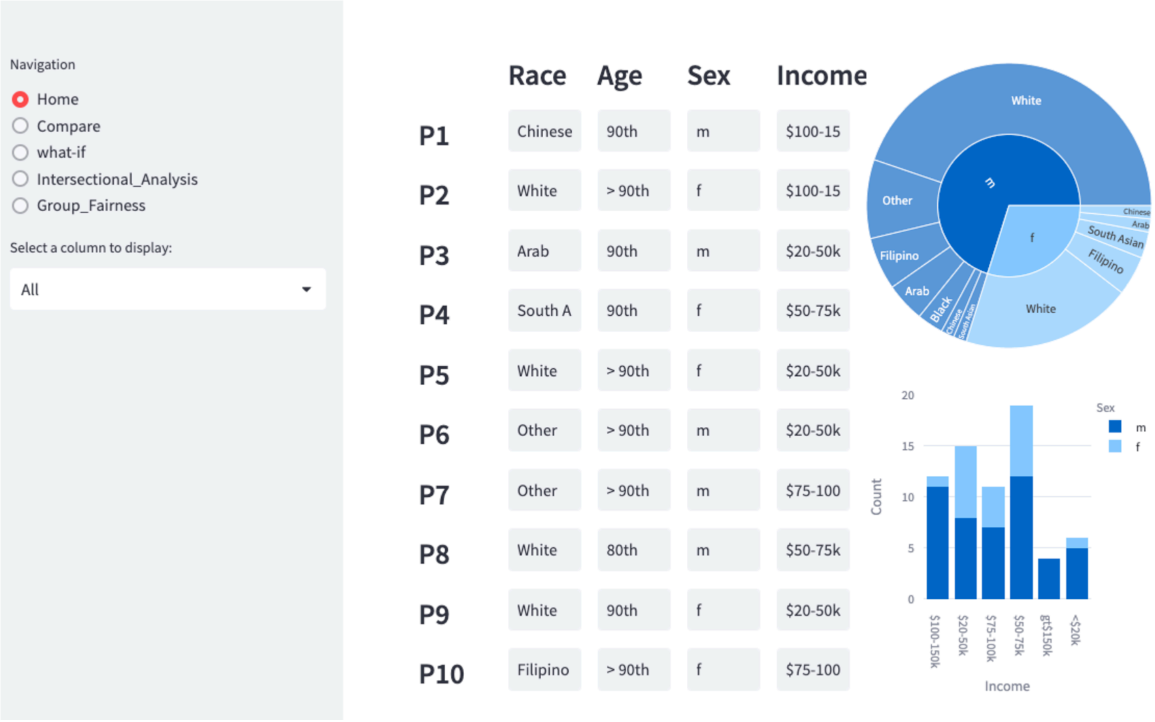
This example shows how utilizing XGBoost as our predictive model and leveraging the associated predictive probabilities as the ranking criterion, we could make informed decisions regarding ICU prioritization. This could help ensure that patients with higher estimated probabilities, indicating more severe conditions, are admitted promptly and receive the critical care they require.

6.7.3 Results: Ranking and Visualization of ICU Admissions

We applied our ranking algorithm to patients being considered for admission to the Intensive Care Unit (ICU) and subsequently employed a comprehensive visual analytic tool to interpret the outcomes. The tool consists of five distinct components: Home, Compare, What-If, Intersectional Analysis, and Group Fairness visualization. Each component offers unique insights into the ranking results, with the goal of providing a comprehensive understanding of the fairness and bias implications within the system.

Home Component

The Home component presents an intuitive overview of the top 10 ranked items. Figure 6.1 illustrates this component, with each row representing an item and displaying the sensitive attributes of race, age percentile, sex, and income interval. The associated statistics are depicted using pie charts and stacked bar plots, illustrating both gender and race distributions and the correlation between sex and income.

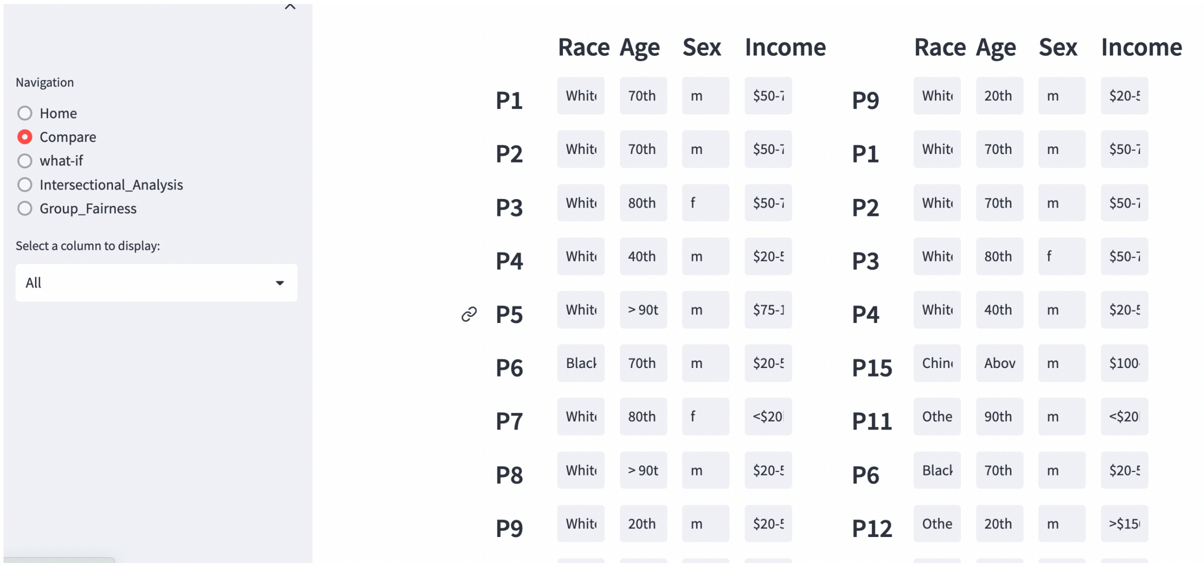


	Race	Age	Sex	Income
P1	Chinese	90th	m	$100-15
P2	White	> 90th	f	$100-15
P3	Arab	90th	m	$20-50k
P4	South A	90th	f	$50-75k
P5	White	> 90th	f	$20-50k
P6	Other	> 90th	m	$20-50k
P7	Other	> 90th	m	$75-100
P8	White	80th	m	$50-75k
P9	White	90th	f	$20-50k
P10	Filipino	> 90th	f	$75-100

Figure 6.1: Home Component Visualization

Compare Visualization

In the Compare component (Figure 6.2), two ranking lists are aligned for comparison. The first list incorporates sensitive attributes in both training and ranking processes, while the second list excludes these attributes during training. This comparison allows users to evaluate the influence of sensitive attributes on ranking outcomes and discern potential biases introduced by their inclusion.



Navigation

○ Home
● Compare
○ what-if
○ Intersectional_Analysis
○ Group_Fairness

Select a column to display:

All

	Race	Age	Sex	Income		Race	Age	Sex	Income
P1	Whit	70th	m	$50-7	P9	Whit	20th	m	$20-5
P2	Whit	70th	m	$50-7	P1	Whit	70th	m	$50-7
P3	Whit	80th	f	$50-7	P2	Whit	70th	m	$50-7
P4	Whit	40th	m	$20-5	P3	Whit	80th	f	$50-7
P5	Whit	> 90t	m	$75-1	P4	Whit	40th	m	$20-5
P6	Blac	70th	m	$20-5	P15	Chin	Abov	m	$100
P7	Whit	80th	f	<$20	P11	Othe	90th	m	<$20
P8	Whit	> 90t	m	$20-5	P6	Blac	70th	m	$20-5
P9	Whit	20th	m	$20-5	P12	Othe	20th	m	>$15

Figure 6.2: Side-by-Side Comparison of Ranking Lists: The first list integrates sensitive attributes throughout both the training and ranking stages, whereas the second list omits these attributes during training. Exploring the Impact of Attribute Inclusion on Rankings.

What-If Visualization

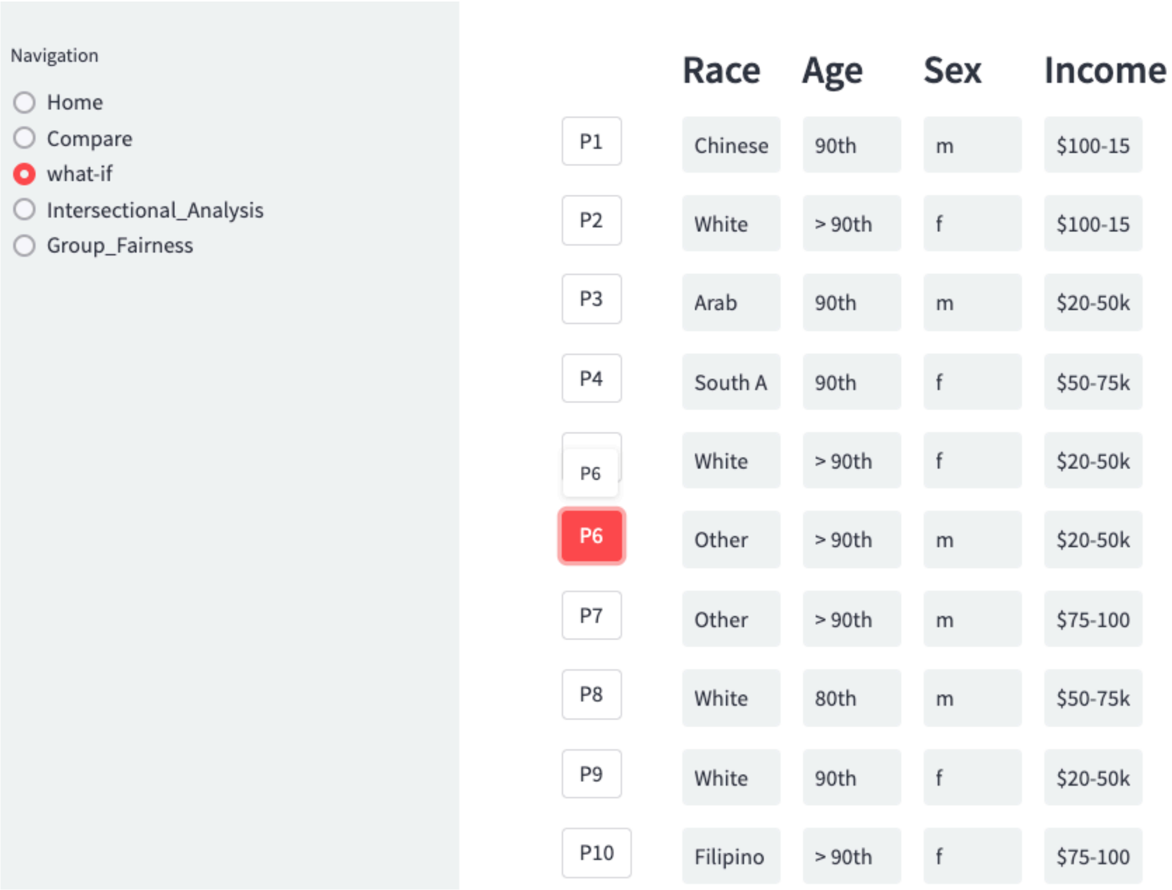
Figure 6.3 showcases the What-If component, where users can interactively modify sensitive attributes and observe resultant changes in item rankings. This functionality provides insights into the influence of attributes on rankings and enables users to identify attributes that significantly impact an item's position, thus shedding light on potential sources of bias.

Intersectional Analysis Visualization

The Intersectional Analysis component (Figure 6.4) delves into the crossroads of two sensitive attributes. By visualizing the intersections, users can discern patterns and disparities that might be obscured by singular attribute analysis, thereby enabling deeper exploration of group fairness and potential biases.

Group Fairness Visualization

Figure 6.7 represents the Group Fairness component, where users can examine group-specific fairness metrics. This visualization offers a comprehensive view of how different sensitive



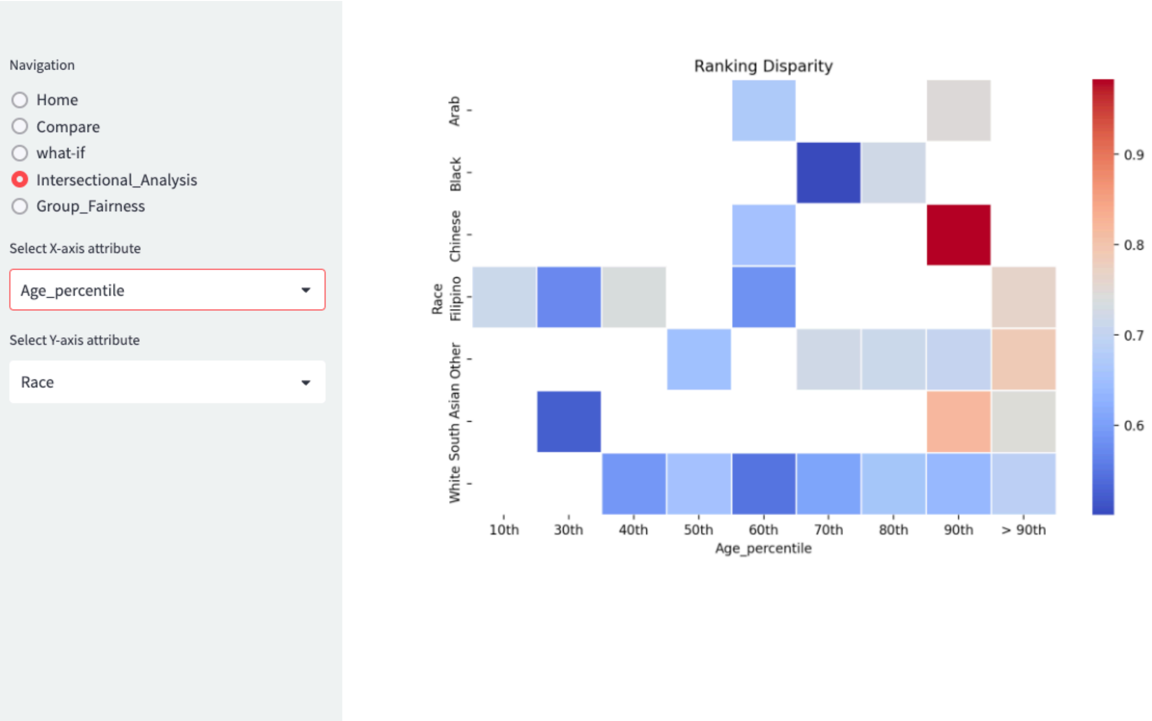
	Race	Age	Sex	Income
P1	Chinese	90th	m	$100-15
P2	White	> 90th	f	$100-15
P3	Arab	90th	m	$20-50k
P4	South A	90th	f	$50-75k
P6	White	> 90th	f	$20-50k
P6	Other	> 90th	m	$20-50k
P7	Other	> 90th	m	$75-100
P8	White	80th	m	$50-75k
P9	White	90th	f	$20-50k
P10	Filipino	> 90th	f	$75-100

Figure 6.3: What-If Analysis Component.

attributes impact fairness outcomes, aiding users in identifying groups that may experience disparate treatment.

6.8 Discussion

While the visual analytics tool may suggest that biases exist within the ranking system, it is essential to conduct causal analysis to understand the underlying factors contributing to the observed disparities. The tool's ability to filter sensitive attributes and explore fairness metrics provides a starting point for investigation, but additional statistical techniques and careful examination of potential confounding variables are necessary to establish causality. For instance, if the tool reveals that male applicants have a significantly higher acceptance rate, it does not automatically imply that discrimination or biases are present. It is crucial to consider other rel-



Navigation

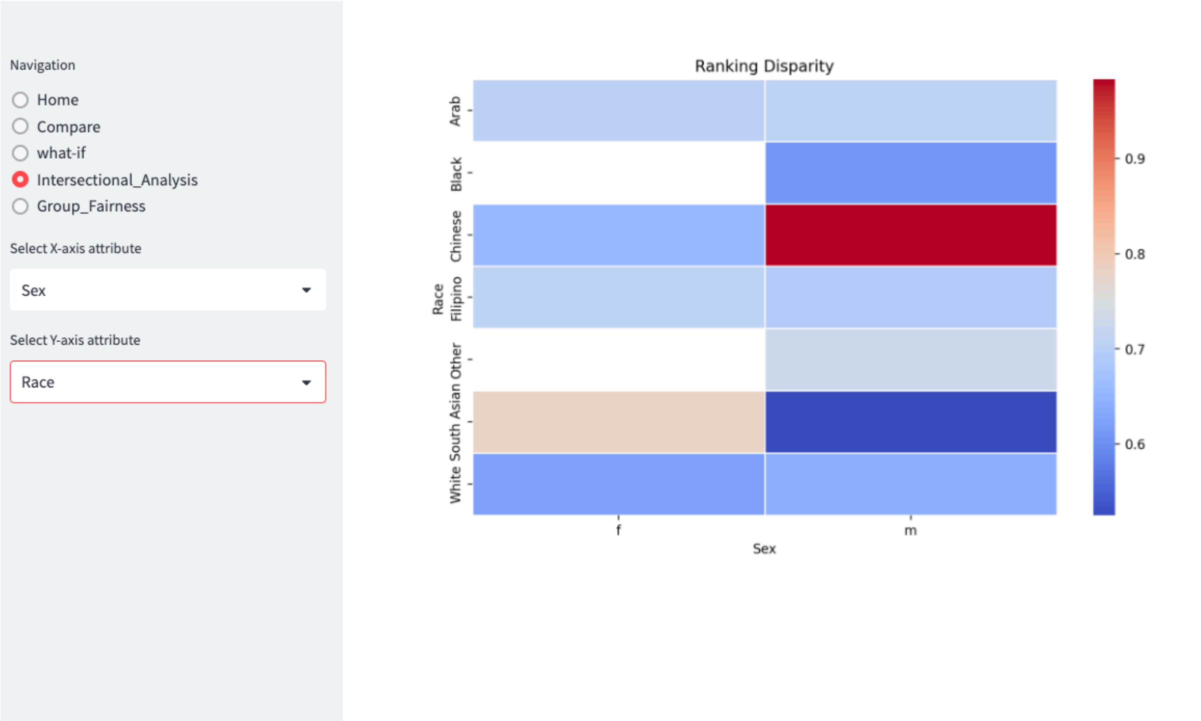
○ Home
○ Compare
○ what-if
● Intersectional_Analysis
○ Group_Fairness

Select X-axis attribute: Age_percentile

Select Y-axis attribute: Race

Figure 6.4: Heatmap Illustrating Intersectional Analysis of Age and Race.

evant variables that could explain this disparity, such as the applicants' scores or educational backgrounds. To conduct a causal analysis, researchers can utilize statistical techniques like regression analysis or propensity score matching. By controlling for potential confounders such as scores or educational qualifications, researchers can isolate the effect of sensitive attributes (in this case, gender) on the rankings. If, after adjusting for relevant factors, a significant disparity still persists, it suggests the presence of biases in the ranking system. Additionally, researchers can explore additional dimensions within the visual analytics tool, such as socioeconomic status or prior educational opportunities, to gain a more comprehensive understanding of the factors influencing the rankings. This multi-dimensional analysis helps disentangle the complex relationships between various attributes and outcomes, facilitating more accurate causal inference. It is crucial to approach causal analysis with caution and consider the limitations of the data and the potential for alternative explanations. Causal inference requires careful design, rigorous statistical methodologies, and an understanding of the specific context under investigation. In summary, while the visual analytics tool can suggest the presence of biases within the ranking system, it is necessary to conduct a thorough causal analysis to understand



Navigation

○ Home
○ Compare
○ what-if
● Intersectional_Analysis
○ Group_Fairness

Select X-axis attribute

Sex

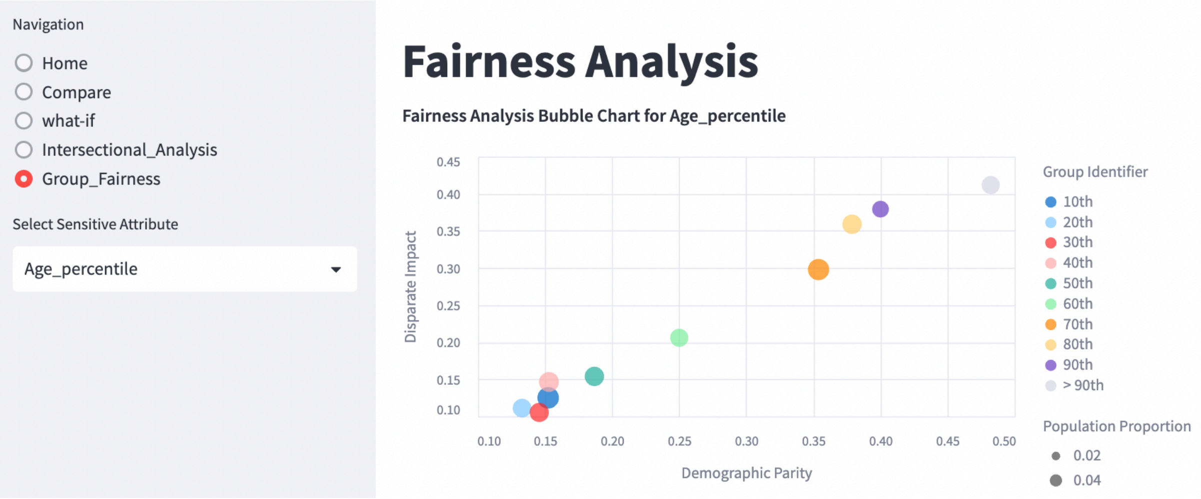
Select Y-axis attribute

Race

Figure 6.5: Heatmap Illustrating Intersectional Analysis of Sex and Race.

the underlying factors driving the observed disparities. By employing statistical techniques, controlling for confounding variables, and considering multiple dimensions, researchers can discern whether the observed differences in acceptance rates are indeed the result of biases or a consequence of other factors such as scores or educational backgrounds. By combining expertise from these diverse domains, our research brings a unique perspective to the assessment of bias in ranking systems. We emphasize the importance of incorporating clinical epidemiological perspectives, ensuring that healthcare-related rankings are examined thoroughly, leading to insights that can address healthcare disparities and improve patient outcomes. Moreover, our visual analytics tool integrates fairness metrics and techniques from XAI-Healthcare to provide interpretable and transparent visualizations that can enhance the fairness and accountability of ranking algorithms.

We underscore the importance of multi-modal explanations, interactive features, and catering to the distinct needs of a specific user group. Our tool places a strong emphasis on multi-modal explanations, employing a blend of visual components and interactive elements. This multifaceted approach ensures that individuals with varying technical backgrounds can grasp


Navigation

○ Home
○ Compare
○ what-if
○ Intersectional_Analysis
● Group_Fairness

Select Sensitive Attribute

Age_percentile

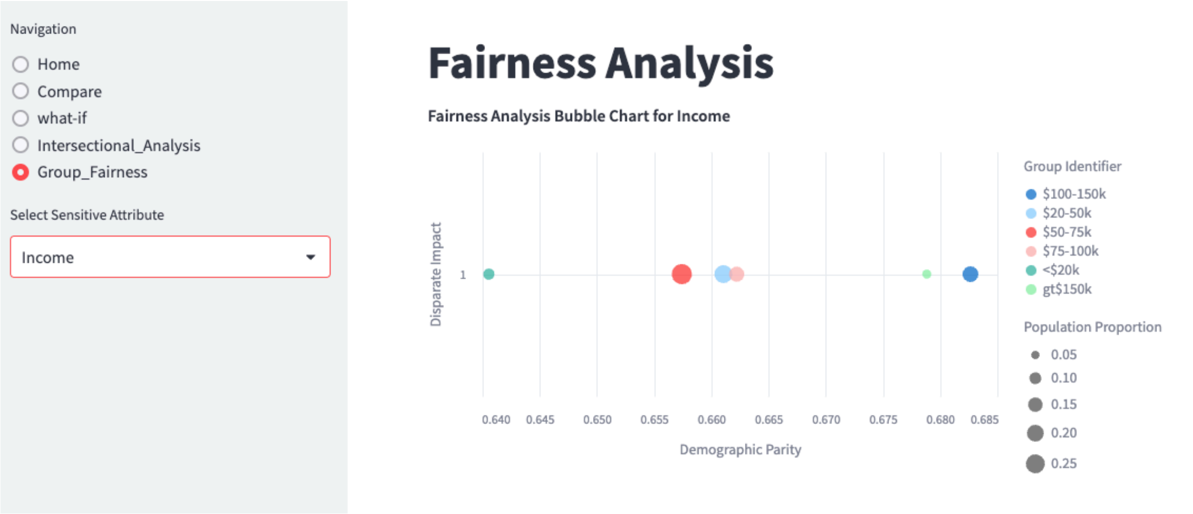
Figure 6.6: Group Fairness Assessment based on Age Percentile.

the intricacies of bias in these systems.

Interactivity is a core principle embedded in our tool's design. By offering a range of interactive options, our tool immerses users in the data, allowing them to actively explore and uncover insights. This engagement can not only aid in the identification of potential bias but can also empower users to play an active role in addressing bias within ranking systems.

Furthermore, we have tailored our tool to cater to the specific requirements of policymakers as the target user group. Recognizing their unique needs, the tool furnishes explanations that directly align with policy concerns. The language and insights provided are tailored to integrate into policy discussions and to enable policymakers to make informed decisions that have far-reaching implications.

In summary, our work is an endeavour that integrates multiple disciplines and methodologies. The development of our visual analytics tool, combined with our exploration of epidemiology, fairness, XAI-Healthcare, interactive visualization, and HCI, emphasizes the significance of our research in tackling bias in ranking systems. This contribution has the potential to drive positive change, promoting fairness, transparency, and accountability in the digital landscape while fostering improved decision-making and user experiences.


Navigation

○ Home
○ Compare
○ what-if
○ Intersectional_Analysis
● Group_Fairness

Select Sensitive Attribute: Income

Fairness Analysis

Fairness Analysis Bubble Chart for Income

Figure 6.7: Group Fairness Assessment based on Income Categories.

6.8.1 Conclusion

In conclusion, this chapter presents an investigation into bias and discrimination in ranking systems, with a specific focus on the prioritization of patients for admission to the ICU. By applying advanced techniques and fairness metrics to the SIRIO dataset and utilizing the XGBoost algorithm, we can gain valuable insights into the presence and impact of bias based on sensitive attributes.

The case study presented in this chapter demonstrates the application of our methodology for assessing and visualizing bias in ICU prioritization. Through the use of advanced techniques such as partial dependence plots, LIME, counterfactual explanations, fairness metrics, and sensitivity analysis, we are able to gain a deeper understanding of the biases inherent in the ranking system.

The findings from this study contribute to the ongoing efforts to improve fairness and mitigate bias in healthcare decision-making processes. By leveraging visualization techniques and incorporating fairness metrics, we gain valuable insights into the presence and impact of bias, which in turn can inform strategies for promoting equitable access to critical care resources.

This research emphasizes the importance of addressing bias and discrimination in ranking systems to ensure fairness, transparency, and accountability in healthcare settings. By identifying and mitigating biases, we can strive towards creating ranking systems that provide equal opportunities and treatment for all individuals, irrespective of their sensitive attributes.

Furthermore, the insights gained from this study have implications beyond healthcare, as they contribute to the broader field of ranking system design and implementation. By considering diverse perspectives and incorporating interdisciplinary research, such as Epidemiology, fairness, XAI-Healthcare, interactive visualization, HCI, and causal inference, we provide a comprehensive assessment of bias in ranking systems.

In conclusion, this chapter underscores the significance of visual analytics tools and methodologies in assessing and visualizing bias in ranking systems. The application of these techniques in the context of ICU prioritization using the SIRIO dataset and XGBoost algorithm provides valuable insights into the presence and impact of bias. The findings contribute to ongoing efforts to promote fairness, mitigate bias, and ensure equitable access when deploying clinical decision support.

Future Work

Based on the insights gained from the bias assessment, we will develop strategies to mitigate bias and enhance fairness in the ICU admission ranking system. This may involve adjusting the weighting or treatment of sensitive attributes, incorporating additional fairness constraints into the model, or refining the decision-making process to minimize disparities based on sensitive attributes.

6.9 References

Bibliography

[1] Barocas, S. & Selbst, A. Big data's disparate impact. *California Law Review*. pp. 671-732 (2016)

[2] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*. **5**, 153-163 (2017)

[3] Dwork, C., Immorlica, N., Kalai, A. & Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. *Conference On Fairness, Accountability And Transparency*. pp. 119-133 (2018)

[4] Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *Advances In Neural Information Processing Systems*. **29** (2016)

[5] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. Algorithmic decision making and the cost of fairness. *Proceedings Of The 23rd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining*. pp. 797-806 (2017)

[6] Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *ArXiv Preprint ArXiv:1609.05807*. (2016)

[7] Olteanu, A., Garcia-Gathright, J., Rijke, M., Ekstrand, M., Roegiest, A., Lipani, A., Beutel, A., Olteanu, A., Lucic, A., Stoica, A. & Others FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval. *ACM SIGIR Forum*. **53**, 20-43 (2021)

[8] Waters, A. & Miikkulainen, R. Grade: Machine learning support for graduate admissions. *Ai Magazine*. **35**, 64-64 (2014)

[9] Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M. & Wilson, C. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. *Proceedings Of The 2017 ACM Conference On Computer Supported Cooperative Work And Social Computing*. pp. 1914-1933 (2017)

[10] Geyik, S., Ambler, S. & Kenthapadi, K. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. *Proceedings Of The 25th Acm Sigkdd International Conference On Knowledge Discovery & Data Mining*. pp. 2221-2231 (2019)

[11] Singh, A. & Joachims, T. Fairness of exposure in rankings. *Proceedings Of The 24th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. pp. 2219-2228 (2018)

[12] Kuhlman, C., VanValkenburg, M. & Rundensteiner, E. Fare: Diagnostics for fair ranking using pairwise error metrics. *The World Wide Web Conference*. pp. 2936-2942 (2019)

[13] Kuhlman, C., Gerych, W. & Rundensteiner, E. Measuring group advantage: A comparative study of fair ranking metrics. *Proceedings Of The 2021 AAAI/ACM Conference On AI, Ethics, And Society*. pp. 674-682 (2021)

[14] Buolamwini, J. & Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference On Fairness, Accountability And Transparency*. pp. 77-91 (2018)

[15] Bolukbasi, T., Chang, K., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances In Neural Information Processing Systems*. **29** (2016)

[16] Celis, L., Mehrotra, A. & Vishnoi, N. Interventions for ranking in the presence of implicit bias. *Proceedings Of The 2020 Conference On Fairness, Accountability, And Transparency*. pp. 369-380 (2020)

[17] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*. **54**, 1-35 (2021)

[18] Žliobaitė, I. & Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence And Law*. **24** pp. 183-201 (2016)

[19] Hovy, D. & Prabhumoye, S. Five sources of bias in natural language processing. *Language And Linguistics Compass*. **15**, e12432 (2021)

[20] Gianfrancesco, M., Tamang, S., Yazdany, J. & Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*. **178**, 1544-1547 (2018)

[21] Alexander, L. What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies. *University Of Pennsylvania Law Review*. **141**, 149-219 (1992)

[22] Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B. & Burke, R. Feedback loop and bias amplification in recommender systems. *Proceedings Of The 29th ACM International Conference On Information & Knowledge Management*. pp. 2145-2148 (2020)

[23] Price, W. & Nicholson, I. Medical AI and contextual bias. *Harv. JL & Tech.*. **33** pp. 65 (2019)

[24] Cabrera, Á., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J. & Chau, D. FairVis: Visual analytics for discovering intersectional bias in machine learning. *2019 IEEE Conference On Visual Analytics Science And Technology (VAST)*. pp. 46-56 (2019)

[25] Ahn, Y. & Lin, Y. Fairsight: Visual analytics for fairness in decision making. *IEEE Transactions On Visualization And Computer Graphics*. **26**, 1086-1095 (2019)

[26] Wang, Q., Xu, Z., Chen, Z., Wang, Y., Liu, S. & Qu, H. Visual analysis of discrimination in machine learning. *IEEE Transactions On Visualization And Computer Graphics*. **27**, 1470-1480 (2020)

[27] Kwon, B., Kartoun, U., Khurshid, S., Yurochkin, M., Maity, S., Brockman, D., Khera, A., Ellinor, P., Lubitz, S. & Ng, K. RMExplorer: A visual analytics approach to explore the

performance and the fairness of disease risk models on population subgroups. *2022 IEEE Visualization And Visual Analytics (VIS)*. pp. 50-54 (2022)

[28] Shrestha, H., Cachel, K., Alkhathlan, M., Rundensteiner, E. & Harrison, L. FairFuse: Interactive Visual Support for Fair Consensus Ranking. *2022 IEEE Visualization And Visual Analytics (VIS)*. pp. 65-69 (2022)

[29] Xie, T., Ma, Y., Kang, J., Tong, H. & Maciejewski, R. Fairrankvis: A visual analytics framework for exploring algorithmic fairness in graph mining models. *IEEE Transactions On Visualization And Computer Graphics*. **28**, 368-377 (2021)

[30] Bird, S., Dudik, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H. & Walker, K. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*. (2020)

[31] Yang, K. & Stoyanovich, J. Measuring fairness in ranked outputs. *Proceedings Of The 29th International Conference On Scientific And Statistical Database Management*. pp. 1-6 (2017)

[32] Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M. & Baeza-Yates, R. Fa* ir: A fair top-k ranking algorithm. *Proceedings Of The 2017 ACM On Conference On Information And Knowledge Management*. pp. 1569-1578 (2017)

[33] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. *Proceedings Of The 3rd Innovations In Theoretical Computer Science Conference*. pp. 214-226 (2012)

[34] Biega, A., Gummadi, K. & Weikum, G. Equity of attention: Amortizing individual fairness in rankings. *The 41st International Acm Sigir Conference On Research & Development In Information Retrieval*. pp. 405-414 (2018)

[35] Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *Advances In Neural Information Processing Systems*. **29** (2016)

[36] Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M. & Diaz, F. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. *Proceedings Of The 27th Acm International Conference On Information And Knowledge Management*. pp. 2243-2251 (2018)

[37] Wu, Y., Zhang, L. & Wu, X. On discrimination discovery and removal in ranked data using causal graph. *Proceedings Of The 24th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. pp. 2536-2544 (2018)

[38] Joachims, T., Granka, L., Pan, B., Hembrooke, H. & Gay, G. Accurately interpreting clickthrough data as implicit feedback. *Acm Sigir Forum*. **51**, 4-11 (2017)

[39] Celis, L., Straszak, D. & Vishnoi, N. Ranking with fairness constraints. *ArXiv Preprint ArXiv:1704.06840*. (2017)

[40] Narasimhan, H., Cotter, A., Gupta, M. & Wang, S. Pairwise fairness for ranking and regression. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **34**, 5248-5255 (2020)

[41] Sírio-Libanês Hospital. COVID-19 - Clinical Data to assess diagnosis. Available at: `https://www.kaggle.com/datasets/Sírio-Libanes/covid19`. Accessed on: 2023-01-20.

Chapter 7

Conclusion

In this chapter, we present concise summaries of the five integrated article chapters, which together constitute the core of this dissertation. These summaries offer a comprehensive snapshot of the key insights and findings derived from our research in Explainable AI and its application in Clinical Decision Support Systems.

Moreover, we highlight the overall contributions of this dissertation to the field. Our work endeavors to bridge the gap between AI and healthcare by fostering transparency and interpretability in decision-making processes. By elucidating the significance of involving end-users, emphasizing interactive visualization, and addressing bias in ranking systems, we contribute to the development of trustworthy and accountable AI-driven CDS systems.

As we conclude this dissertation, we also acknowledge that the journey of research is never truly complete. Thus, we provide reflections on potential areas for future investigation that extend beyond the confines of this study. These thoughts on future research aim to inspire further advancements in the realm of Explainable AI, HCI, cognitive science, and their intersection with clinical applications.

This chapter encapsulates the essence of our research, its wider implications, and an invitation to the ongoing quest for more transparent, ethical, and effective AI systems in the healthcare domain.

7.1 Chapter Summaries

The following provides high-level summary of the four integrated article chapters.

In chapter 3, we have explored the realm of Explainable Artificial Intelligence (XAI) and its application in Clinical Decision Support Systems (CDS). We began by categorizing different approaches to XAI, laying the foundation for understanding the significance of transparency and interpretability in complex AI systems. With this understanding, we conducted a quick review of the current state of XAI in CDS systems, which revealed a crucial gap in the design process: the neglect of end-users, visualization, and evaluation aspects. Our investigation led us to propose an interdisciplinary framework that embraces Human-Computer Interaction (HCI), cognitive science, and human psychology to design and develop Explainable CDS systems that are not only reliable and accurate but also adaptable in real-world clinical settings.

In chapter 4, our research focused on developing an Explainable Interactive Visualization tool capable of explaining complex ranking systems. This tool takes into account the varying needs of users and provides multiple visualizations to cater to their unique requirements. Through its application in triaging patients at ICU admission based on their health history and conditions, we demonstrated how such tools could support medical practitioners in making informed decisions. By recognizing clinicians as potential end users of our system, we have tailored explanations to cater to their specific needs. Throughout the design process, the emphasis was on providing clinically actionable insights through interpretable explanations.

Moreover, we have incorporated various visualization components employing different eXplainable Artificial Intelligence (XAI) methods. This approach serves to accommodate multimodal explanations, ensuring that individuals with varying needs and levels of technological familiarity can effectively engage with the system and attain the explanations they seek.

In chapter 5, we employed XAI methods to identify regions in machine learning models where poor decisions were made. This granular explanation enables users to exercise caution and discern when to trust the model's predictions, specifically in situations like septic shock detection. By shedding light on the decision-making process, we take a necessary step toward instilling trust and confidence in the machine learning model's capabilities, and hence toward allowing healthcare professionals to integrate AI technologies more effectively into their clinical practice.

This tool was designed with policymakers as the intended end users. Adhering to the framework for creating user-centered explanations, we took into account the varying preferences and expectations of policymakers during the explanation generation process. Additionally, by incorporating interactive features, users can actively delve into the system to access the information they seek. Moreover, our approach involved offering explanations through various components, ensuring that individuals with differing levels of technological familiarity and diverse ways of reasoning can extract valuable insights from the tool.

In chapter 6, our research delved into the critical aspect of fairness and bias in ranking systems. We developed a visual analytics tool that employed XAI methods and fairness metrics to assess how the model generates rankings and how fair the system is in its predictions. By addressing bias in the CDS system, we take a key step toward ensuring equitable treatment of patients but also toward enhancing the overall trustworthiness and accountability of the system.

In conclusion, this thesis presents a comprehensive exploration of Explainable AI in the context of Clinical Decision Support Systems. By categorizing XAI approaches, addressing the end-user aspect, and developing novel visualization and evaluation techniques, we have taken significant steps toward building trust and accountability in CDS systems. The interdisciplinary approach that incorporates HCI, cognitive science, and human psychology strengthens the practicality and adoption of these systems in real-world medical settings.

The importance of this work lies in its potential to revolutionize the healthcare landscape, fostering a symbiotic relationship between medical professionals and AI systems. As we continue to integrate AI technologies into healthcare, ensuring that these systems are explainable and transparent becomes paramount to fostering trust, encouraging adoption, and ultimately improving patient outcomes. Our contributions towards creating trustworthy, accountable, and explainable CDS systems pave the way for accountable, more effective, and ethical AI-driven healthcare practices.

7.2 Contributions

The general contribution of this research is multifaceted and holds significance in several key areas:

1. Advancement of Explainable AI in Healthcare: This research contributes to the growing field of Explainable AI, specifically within the healthcare domain. By categorizing and reviewing various approaches to XAI, this work provides a foundation for developing transparent and interpretable AI models that can be applied in complex clinical decision-making scenarios.

2. Bridging the Gap between AI and Healthcare: The integration of AI technologies in healthcare has the potential to revolutionize patient care and outcomes. However, the lack of transparency and interpretability in AI models poses challenges for their acceptance and adoption in critical medical contexts. This research aims to bridge this gap by proposing an interdisciplinary framework that leverages insights from HCI, cognitive science, and human psychology to design Explainable Clinical Decision Support Systems that are both accurate and user-friendly.

3. End-User Involvement: By emphasizing the importance of involving end-users in the development of AI systems, this research ensures that the resulting tools cater to the needs and expectations of medical practitioners. Such involvement fosters user trust, confidence, and acceptance, promoting the successful integration of AI technologies in real-world clinical settings.

4. Explainable Interactive Visualization: The development of an Explainable Interactive Visualization tool represents a novel contribution to the field. This tool enables clinicians to comprehend complex ranking systems and AI predictions effectively. By providing multiple visualizations tailored to various user needs, this research can support medical professionals in aking informed and timely decisions in critical situations, such as patient triaging at ICU admission.

5. Trust and Confidence in AI: Addressing the regions where machine learning models make poor decisions and providing granular explanations is critical for instilling trust and confi-

dence in AI technologies. By employing XAI methods to explain machine learning model outputs, this research can be used to equip medical practitioners with the knowledge to exercise caution and make informed decisions when relying on AI systems, leading to safer and more responsible AI adoption.

6. Assessing Bias and Fairness: The development of a visual analytics tool to assess bias in ranking systems represents a vital contribution towards building equitable AI systems. By applying fairness metrics and XAI techniques, this research enables the evaluation of AI model rankings in terms of fairness, which is needed to ensure that the CDS system provides equitable treatment to all patients, irrespective of their demographics or backgrounds.

Overall, the general contribution of this research lies in its holistic approach to creating trustworthy, accountable, and explainable Clinical Decision Support Systems. By integrating insights from multiple disciplines, providing interactive visualization tools, and addressing bias and fairness, this research sets the stage for more ethical and responsible AI applications in healthcare, thereby fostering improved patient care and outcomes.

7.3 Future work

It is important to acknowledge that, due to the limitations imposed by the COVID-19 pandemic, we were unable to conduct formal studies to evaluate the performance and effectiveness of the proposed systems and their interaction mechanisms. Despite this, we made significant efforts to enhance the system's interpretability by offering multiple explanations with diverse interactive visualization components to cater to the varying needs of different users.

Formal studies in the future will be essential in assessing the efficiency and usability of these systems for both expert and non-expert users. Although the systems presented in this research were developed and tested using healthcare databases, conducting studies with different datasets and settings will provide valuable insights into the efficacy and generalizability of these systems.

By conducting such studies, we can gain a better understanding of how these Explainable AI systems perform in practical scenarios and how effectively they aid healthcare professionals in decision-making. Additionally, comparative studies with various datasets will help us identify potential strengths and limitations, enabling us to refine and improve the systems further for broader applicability and impact.

As discussed in this study, our endeavors to ensure a fair and unbiased ICU admission ranking system are ongoing. In the pursuit of this goal, several avenues for future work emerge, each geared toward mitigating bias and enhancing fairness within the system.

One promising direction involves the development of advanced strategies aimed at mitigating bias associated with sensitive attributes. To this end, we intend to explore methods for adjusting the weighting or treatment of these attributes during the ranking process. By judiciously calibrating the influence of sensitive attributes, we aspire to attain a more equitable distribution of rankings across the diverse population of patients.

Additionally, the incorporation of supplementary fairness constraints into the model represents another avenue for future research. These constraints could be designed to explicitly address disparities stemming from sensitive attributes, fostering a ranking system that is not only accurate but also sensitive to issues of fairness and equitable treatment.

Furthermore, we anticipate delving into the refinement of the decision-making process itself. This involves designing algorithms that proactively minimize disparities based on sensitive attributes while adhering to critical medical considerations. By fine-tuning the decision-making framework, we aim to minimize the impact of sensitive attributes on rankings, thereby cultivating a more just and impartial system.

Appendix A

Evaluation Methodology

In evaluating the effectiveness and user-centric nature of our visual analytics tool implemented in the thesis, we propose an evaluation methodology encompassing both quantitative and qualitative components. This approach aims to provide a nuanced understanding of user interactions and expert opinions, ensuring a thorough assessment of the tool's performance. The proposed methodology comprises the following key elements:

1. Quantitative Component: Usage Metrics

Data Collection: Utilizing advanced usage analytics tools, we intend to collect quantitative data on user interactions with the visual analytics tool. The metrics to be captured include:

- The total number of users accessing the tool within a specified period.
- Frequency of tool usage per user, categorized into daily, weekly, and monthly intervals.
- Patterns of interaction with the ranking list, including the frequency of zooming in and out.
- Exploration patterns within the treemap, measured by the number of clicks or time spent on each feature or category.
- Usage patterns within the what-if panel, specifically focusing on the frequency of adjustments to feature values and exploration of counterfactual scenarios.

Analysis: The collected quantitative metrics will be subjected to analysis to identify usage patterns and trends. Key analytical tasks include:

- Exploring correlations between the frequency of tool usage and the depth of exploration within the ranking list and what-if panel.
- Identifying popular features or categories within the treemap to discern which aspects users find most engaging.
- Examining whether specific times or user segments exhibit distinct usage patterns, providing insights into potential user preferences or needs.

2. Qualitative Component: Expert Evaluation or Feedback

Data Collection: To complement the quantitative analysis, we will engage domain experts, including data scientists and visualization experts, to gather qualitative data. This involves:

- Conducting structured interviews or surveys to solicit expert opinions on the effectiveness of the visualizations.
- Seeking feedback on the clarity of the treemap and its ability to convey information meaningfully.
- Assessing the perceived usefulness of the what-if panel in exploring counterfactual scenarios and adjusting feature values.

Analysis: Qualitative data will be analyzed to identify themes and insights from expert feedback. This includes:

- Identifying consensus or divergence among experts regarding the strengths and weaknesses of the visual analytics tool.
- Recognizing specific aspects of the treemap design that experts find particularly effective or challenging.
- Assessing whether the what-if panel meets the expectations of experts in providing a practical tool for scenario exploration.

3. Cognitive Interviewing: Participants' Think-Aloud Protocol

During the user testing phase, incorporate cognitive interviewing where participants are asked to "think aloud" as they interact with the visual analytics tool. This involves verbalizing their thoughts, reactions, and decision-making processes in real-time.

Data Collection:

- Participants will be instructed to express their thoughts and explain their actions while using the tool.
- Encourage participants to vocalize any confusion, insights, or difficulties they may encounter during the interaction.

Analysis:

- Analyze the recorded think-aloud sessions to gain insights into participants' cognitive processes and understanding of the tool.
- Identify common patterns in the explanations provided by participants and areas where confusion or misunderstandings arise.
- Use the think-aloud data to supplement the quantitative and expert feedback, providing a deeper understanding of users' cognitive experiences.

4. Integration: Synthesis and Interpretation

Synthesize Findings: We will synthesize the quantitative and qualitative findings and think-aloud insights to compare usage patterns with expert opinions. Key tasks include:

- Identifying instances where high tool usage aligns with positive expert feedback, indicating that users find the tool valuable and engaging.
- Identifying areas where expert opinions shed light on specific challenges or opportunities that may not be evident in the quantitative data alone.
- Look for discrepancies or areas where think-aloud sessions provide additional context to user behavior.

Interpretation: The integrated findings will be interpreted to draw conclusions about the user-centered nature of the visual analytics tool. This involves:

- Assessing whether user interactions align with the intended goals of detailed exploration and global comparison as described in the tool's description.
- Determining whether expert opinions provide valuable insights into areas for improvement or adjustment in the tool's design and functionality.
- Assess whether think-aloud sessions reveal nuances in user understanding that may not be evident in other evaluation components.

5. User Feedback Questionnaire

This questionnaire is a sample designed to assess the usability of visual analytics tools. It includes select questions derived from established standard usability questionnaires.

1. Pre-Task Questionnaire:

Participant Information:

- Occupation:
  - Doctor
  - Nurse
  - Data Scientist
  - Other (please specify)
- Experience:
  - How many years of experience do you have in your current role?
  - Have you used similar visual analytics tools before? (Yes/No)
- Confidence:
  - How confident do you feel about using visual analytics tools for decision-making? (Scale: 1-5, 1. Not Confident, 2. Less Confident, 3. Moderately Confident, 4. Quite Confident, 5. Very Confident.)

2. Task-Specific Questionnaire:

Task 1: Rank Patients Based on Visualizations:

Effectiveness:

- How accurate do you think the rankings were based on the visualizations provided? (Scale: 1-5, 1. Not Accurate, 2. Less Accurate, 3. Moderately Accurate, 4. Quite Accurate, 5. Very Accurate.)

- Were you able to easily identify the most influential features impacting the rankings? (Yes/No)

- To what extent did the visualizations contribute to your understanding of the decision-making process for patient triage? (Scale: 1-5, 1. Minimal Contribution, 2. Minor Contribution, 3. Moderate Contribution, 4. Significant Contribution, 5. Very Significant Contribution.)

- Do you believe the visualizations provided a transparent and understandable representation of how the patient ranking model works? (Yes/No)

Usability:

- How intuitive did you find the visual analytics tool for completing the ranking task? (Scale: 1-5, 1. Not intuitive, 2. Less intuitive, 3. Moderately intuitive, 4. Quite intuitive, 5. Very intuitive.)

- Did you encounter any challenges or difficulties while using the tool? (Open-ended)

Satisfaction:

- Overall, how satisfied are you with your experience using the visual analytics tool? (Scale: 1-5, 1. Very Dissatisfied, 2. Dissatisfied, 3. Neutral, 4. Satisfied, 5. Very Satisfied.)

- What aspects of the tool did you find most helpful or effective? (Open-ended)

3. Post-Task Survey:

Overall Experience:

Usability:

- Rate the overall usability of the visual analytics tool for understanding the ranking system. (Scale: 1-5, 1. Poor, 2. Below Average, 3. Average, 4. Good, 5. Excellent.)

- Were you able to efficiently navigate the visual analytics tool to understand the ranked patient information? (Yes/No)

- How intuitive did you find the tool for interpreting the visual representations of the ranked patients? (Scale: 1-5, 1. Not intuitive, 2. Less intuitive, 3. Moderately intuitive, 4. Quite intuitive, 5. Very intuitive.)

- Were you able to easily identify the features or factors influencing the ranking of patients? (Yes/No)

Learnability:

- How quickly were you able to grasp how to use the visual analytics tool to explore the ranked patient information? (Scale: 1-5, 1. Not Quickly, 2. Slowly, 3. Average Speed, 4. Quickly, 5. Very Quickly.)

Recommendation:

- On a scale from 1 to 5, how likely are you to recommend this visual analytics tool for explaining the ranking system to your colleagues?

- What specific features or aspects of the visual analytics tool do you believe contribute most to its effectiveness in explaining the ranking system? (Open-ended)

- Are there any improvements or additional features you would suggest to enhance the tool's capability in explaining the patient ranking system? (Open-ended)

Additional Considerations:

User Feedback: If available, we will consider incorporating direct user feedback, such as comments or suggestions, into the analysis for a more comprehensive understanding of user perspectives.

Iterative Improvement: The integrated findings will inform iterative improvements to the tool, ensuring that both quantitative and qualitative insights contribute to its ongoing development.

Curriculum Vitae

Education

Doctor of Philosophy, Computer Science

Dissertation: Enhancing Trust and Transparency in Clinical Decision Support: Integrating XAI Methods and Visual Analytics

Supervisors: Dr. Daniel Lizotte, Dr. Kamran Sedig

Western University, London, Ontario, Canada

Master of Science, Health Information Technology: 2017

Dissertation: Design and Implementation of a Portable Fetal Health Monitoring System

Supervisors: Dr. Hamidreza Memarzadeh

University of Tehran, Tehran, Iran

Visiting Student, Computer and Electrical Engineering: 2016-2017

Supervisors: Dr. Jamal Deen

McMaster University, Hamilton, Ontario, Canada

Publications

1. Salimiparsa, M., Sedig, K., & Lizotte, D. (2023). Unlocking the Power of Explainability in Ranking Systems: A Visual Analytics Approach with XAI Techniques. (Presented at

AIME Conference and in the Process of being Published)

2. Salimiparsa, M, et al. ”Investigating Poor Performance Regions of Black Boxes: LIME-based Exploration in Sepsis Detection.” *arXiv preprint arXiv:2306.12507* (2023).

3. (Prepared for Submission to Publisher) Salimiparsa, M., Sedig, K., & Lizotte, D. (2023). Unveiling Bias and Discrimination in Ranking through a Visual Analytics Tool.

Conferences and Posters

1. Salimiparsa, M. (2023). Counterfactual Explanations for Rankings. Proceedings of the Canadian Conference on Artificial Intelligence. `https://doi.org/10.21428/594757db.15b61c8c`

2. Salimiparsa, M, Daniel J. Lizotte, and Kamran Sedig. ”A User-Centered Design of Explainable AI for Clinical Decision Support.” In *Canadian Conference on AI*, 2021.

Conferences and Presentations

1. UWORCS (2023): Transparency in Ranking: An Analytic Visualization Tool - Awarded 2nd Best Presenter

2. UWORCS (2022): Triaging Patients to the ICU: A Visual Analytics Tool for Admission Decision

3. Canadian Celebration of Women in Computing Conference (CAN-CWiC) (2022): Bridging the Gap: Enhancing AI Adoption in Clinical Decision Support Systems through XAI

4. UWORCS (2021): Explainable AI in Clinical Decision Support