Enterprise Risk Management of Artificial Intelligence for Healthcare

---

A
Dissertation
Presented to
the faculty of the School of Engineering and Applied Science
University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy in Systems Engineering

By

Negin Moghadasi

May 2024

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
ANCHOR	ANchored COunterfactual Explanations
BO	Bayesian Optimization
CEM	Counterfactual Explanation Method
CFX	Counterfactual eXplanations
CIU	Counterfactual Instance Uncertainty
CMM	Counterfactual Model Management
CMR	Cardiac Magnetic Resonance
CS	Cardiac Sarcoidosis
CTRL	Control
DALEX	Model Agnostic Exploration, Explanation, and Learning
DICE	Deep Interpretable Counterfactual Explanations
DOD	Devise Outer Diameter

EBM	Explainable Boosting Machines
EEM	Exemplar Explanation Method
ELI5	Explain Like I am 5
FACE	Feature Attribution by Contrastive Explanation
GAM	Generalized Additive Model
GB	Gradient Boosting
GLOCALX	Global-Local Explanations
GP	Gaussian Process
ICE Plot	Individual Conditional Expectation Plot
ID	Inner Diameter
KDE	Kernel Density Estimation
KNN	k-Nearest Neighbors
LIME	Local Interpretable Model-agnostic Explanations
LORE	Local Rule-Based Explanations
LR	Logistic Regression
LRP	Layer-wise Relevance Propagation
MAPLE	Model Agnostic Predictive Local Explanations
MCA	Multiple Criteria Analysis
MSFT	Microsoft
NAM	Numerically Approximated Explanation Method
NC	Non-Cardiac
NIST AI RMF	National Institute of Standards and Technology AI Risk Management Framework
OD	Outlier Diameter

PDP	Partial Dependence Plot
PROTODASH	Prototype-Based Dashboard
RF	Random Forest
ROC	Receiver Operating Characteristic
ROT	Rule of Thumb
RULEMATRIX	Rule-Based Explanation Matrix
SCALABLE-BRL	Scalable Bayesian Rule Lists
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
UCB	Upper Confidence Bound
WT	Width of Vessel
XAI	eXplainable Artificial Intelligence
XGB	eXtreme Gradient Boosting

ACKNOWLEDGEMENTS

The work presented in this dissertation benefited greatly from the advice, approval, criticism, and support of numerous individuals and organizations. Collaboration with both public and private partners interested in AI risks in healthcare was facilitated by the Commonwealth Center of Advanced Logistics Systems (CCALS) and the U.S. National Science Foundation Center for Hardware and Embedded Systems Security and Trust (NSF CHEST). The Heart and Diabetes Center North-Rhine Westphalia Hospital, Johns Hopkins University, Western University of Medical Sciences, and Johnson & Johnson MedTech provided medical resources. Also, I am grateful to the University of Virginia School of Engineering for awarding me the UVA School of Engineering Endowed Graduate Fellowship in recognition of academic excellence and scholarship, which greatly contributes to supporting my research endeavors.

Prof. James H. Lambert provided invaluable guidance and support throughout my PhD journey and assisted in gathering information and creating models for this dissertation. I extend my gratitude to Prof. Stephen Baek for his insightful advice over the past two years and his expertise in artificial intelligence and data science. Special thanks to Prof. Michael Porter for sharing his technical proficiency in statistical programming and data sciences. I am grateful to Prof. Rupa S. Valdez for her expert assessments on techniques and subjects concerning risk and uncertainty in healthcare and for guiding me with her comments on several of my publications. Also, I am grateful to Prof. Seokhyun Chung for his contribution to sharing his technical expertise in statistical modeling and optimization. Lastly, Dr. Misagh Piran offered valuable advice on medical leadership development and terminology.

Special recognition goes to my parents, Dr. Hesam Moghadasi and Dr. Haydeh Majlessipour, and my sister, Dr. Negar Moghadasi, for their unwavering support and medical advice. I am grateful for the unwavering academic and emotional support of my spouse, Dr. Misagh Piran, whose encouragement has been invaluable throughout this journey. Their support has been a constant source of motivation, enabling me to overcome challenges encountered during my research endeavors in graduate school.

ABSTRACT

Artificial intelligence (AI) is increasingly being adopted across technology domains, including healthcare, commerce, economy, energy, environment, trust and cybersecurity, transportation, etc. However, system owners, experts, regulators, developers, and other actors describe concerns regarding the risks associated with AI applications. This dissertation develops a framework for management of risk, cost, and schedule in AI applications in enterprise systems, focusing on healthcare technologies. The framework combines risk analysis and systems modeling with an understanding of recent AI healthcare applications. A risk register, which includes the *Purpose (Pi), Structure (Sig),* and *Function (Phi) characteristic* layers of a system, serves as the foundation of the framework. The proposed method identifies success criteria, research and development initiatives, and emergent conditions of AI healthcare systems within each layer. The outcomes have insights into the requirements and policies for healthcare organizations that are prioritizing initiatives and tracking potential disruptions. To demonstrate the framework, three cases of scenario-based disruption of priorities are described across three systems modeling layers: First, an analysis of hospital priorities is developed in the *Purpose (Pi)*/sector layer; this tracks the most disruptive system stressors. Second, an AI-assisted design optimization of a vascular anastomosis device is developed in the *Structure (Sig)*/device layer; this avoids costly physical experiments. Third, an analysis of AI-based diagnosis of cardiac sarcoidosis using multi-chamber wall motion is developed in the *Function (Phi)*/disease diagnosis layer; this avoids waste in programming examinations and procedures. Various eXplainable AI (XAI) techniques are then employed to interpret the outputs of the second and third cases. These techniques aid in improving communication between AI systems and non-technical users, enhancing understanding of AI outputs, reducing distrust in the AI results, and assisting in data evaluation. In addition, the framework is extended to quantify the dynamics of the system layers using resilience curves of order disruption. This scale-free quantification of resilience allows for the deployment of the framework across various application domains.

Chapter 1 | Introduction

This chapter describes the motivation, purpose and scope, and organization of the dissertation.

1.1. Motivation

AI has the potential to bring about a significant transformation in healthcare systems, impacting how medical professionals approach diagnosis, treatment, patient care, and medical device design, among other areas [1]. According to FDA Commissioner Scott Gottlieb, M.D., "Artificial intelligence and machine learning have the potential to fundamentally transform the delivery of health care. As technology and science advance, we can expect to see earlier disease detection, more accurate diagnosis, more targeted therapies, and significant improvements in personalized medicine" [2].

As with any technological advancement, AI carries inherent risks that require effective management. During the launch of the NIST AI RMF (National Institute of Standards and Technology Artificial Intelligence Risk Management Framework), published in 2023, Dr. Alondra Nelson, deputy director of the Office of Science and Technology Policy, acknowledges that "We know that artificial intelligence and other automated systems are shaping almost every part of our lives: The way we work, the way we learn, how we access healthcare, and how we find a good job." He added, "And yet, too often, the use of these technologies comes with serious risks. They can be used and abused to track our communities and to limit access to fundamental opportunities. Our Administration—like so many in industry, in Congress, and across the United States—is clear-eyed about these risks." Moreover, "for high-risk settings like diagnostic decision making, such over-reliance on advice can be dangerous." [3]. "AI can pose certain risks and a slew of unexpected risks" [4].

Dr. John Smith, a leading expert in AI and healthcare, said, "The NIST AI Risk Management Framework is a valuable tool for organizations in the healthcare industry that are developing and deploying AI systems. It provides a framework for understanding and managing the risks associated with AI, and it can help organizations to build trustworthy and responsible AI systems that can improve patient care."

With the above background, it is important to identify the NIST AI Risk Management Framework gaps and identify, assess, and mitigate the risks associated with AI in healthcare. As former FDA Commissioner Scott Gottlieb said, "We need to be proactive in modeling and managing these risks, so that we can ensure that AI is used for good, not for harm" [2] and Li said, "In the end, it

is about a human-centered approach". Thus, by actively addressing the risks in AI-based systems, healthcare organizations can fully leverage the benefits of AI while minimizing potential harm to patients, mitigating legal and financial risks, and safeguarding their reputation.

1.2. Purpose and Scope

Adopting an AI risk management framework is essential across diverse domains and systems. This framework fills the gaps in research within the Systems Engineering Body of Knowledge (SEBoK) related to AI risk management. Its implementation contributes to the field of systems engineering by providing a comprehensive approach to addressing AI-related risks in healthcare. This dissertation will be an intersection between risk analysis, systems modeling, and AI applications in healthcare. It will demonstrate enterprise risk management of AI in this domain using scenario-based preferences. Furthermore, it showcases the potential of trustworthy AI to enhance various engineering systems while effectively identifying associated risks.

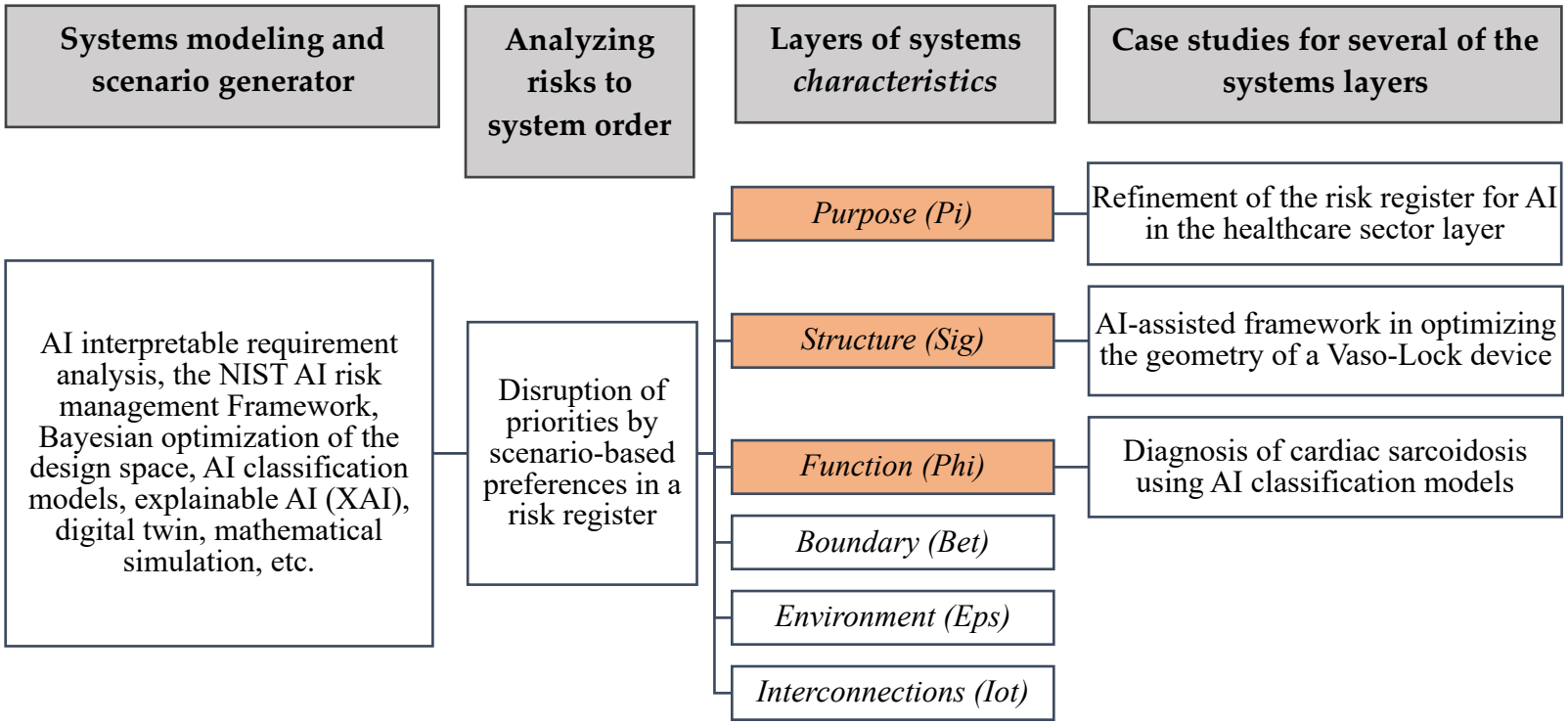

Figure 1. The dissertation theory and method conceptual diagram of systems modeling for enterprise risk management of AI in healthcare [5,6].

Figure 1 describes the dissertation theory and method conceptual diagram of systems modeling for enterprise risk management of AI in healthcare. It comprises four steps: 1. Systems modeling and scenario generator, 2. Analyzing risks to system order, 3. Systems *characteristics*, and 4. Case studies. Each step will be explained in detail in the following chapters.

1.3. Organization of the Dissertation

Table 1 is a roadmap for the structure of this dissertation. It starts with the introduction and literature review in Chapter 1 and Chapter 2. Chapter 3 introduces the theory and methods used in the dissertation. Chapter 4 describes the scenario-based disruption in priorities for the risk of AI in the healthcare *Purpose (Pi)* layer. Chapter 5 describes the scenario-based disruption of priorities in the risk of AI in the healthcare *Structure (Sig)*/device layer and describes methods to optimize and explain Vaso-Lock device design configurations using AI; then employ eXplainable AI (XAI) global and local interpreters to explain the output to non-technical users and Vaso-Lock designers. Chapter 6 describes the scenario-based disruption of priorities in the risk of AI in the healthcare *Function (Phi)* layer; this chapter describes methods to predict and explain the cardiac sarcoidosis diagnosis using machine learning classification models; then employs XAI global and local models to interpret the output for non-technical users. Chapter 7 describes two scenario-based disruptions of priorities in the risk of AI in the healthcare *Function (Phi)* layer considering two perspectives: First, from the perspective of the physicians and second, from the perspective of the patients. Chapter 8 describes the synthesis and comparison of cases, limitations,

and lessons learned. Chapter 9 is a discussion of research opportunities of evaluating system resilience by the degree of order disruption, and Chapter 10 concludes the dissertation with the dissertation contributions, summary conclusions, and future works.

Table 1. Organization of the dissertation on enterprise risk management of AI in healthcare.

Chapter 1	Introduction	• Overview of the motivation, purpose, and scope of the work, and organization of the dissertation
Chapter 2	Literature Review	• Identification of gaps and opportunities based on literature and state of the practice in the fields of risk analysis, systems engineering, and AI applications in healthcare
Chapter 3	Theory and Method	• Introducing multi-layered scenario-based disruption of priorities method
Chapter 4	Case 1: *Purpose (Pi)* Layer	• Scenario-based disruption of priorities method on healthcare *Purpose (Pi)* or sector layer
Chapter 5	Case 2: *Structure (Sig)* or Component Layer	• Scenario-based disruption of priorities method on *Structure (Sig)* or component layer • Optimization of Vaso-Lock geometry: Description of methods used for AI-assisted design optimization framework to determine the optimal geometry of Vaso-Lock, a vascular anastomosis device • XAI models to interpret the predictions
Chapter 6	Case 3: *Function (Phi)* Layer	• Scenario-based disruption of priorities method on *Function (Phi)* or disease diagnosis layer • Diagnosis of cardiac sarcoidosis: Description of a machine learning-based diagnosis of cardiac sarcoidosis using multi-chamber wall motion analysis • XAI models to interpret the predictions
Chapter 7	Case 3: *Function (Phi)* Layer – Perspectives Comparison	• Comparing Case 3 scenario-based disruption of priorities methods considering two groups of experts and actors involved: Physicians versus patients
Chapter 8	Synthesis and Comparison of Cases	• Highlight of the cases, limitations and lessons learned
Chapter 9	Discussion of Research Opportunities	• Evaluation system resilience by the degree of order disruption
Chapter 10	Conclusions	• Summary of contributions, conclusions, and publications

Chapter 2 | Literature Review

2.1. Introduction

This chapter describes the review of literature in six sub-sections: Theory and methods of enterprise risk management, examples of AI unintended harms, AI and its risks in healthcare, the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework (NIST AI RMF), opportunities to improve science and practice, and explainable artificial intelligent (XAI).

2.2. Theory and Methods of Enterprise Risk Management

There is urgency for systems modeling in terms of evolving priority orders of complex systems to complement existing systems models for *Purpose, Structure,* and *Function*. Priority orders involve assets, policies, investments, organizational units, locations, personnel, etc. Orders are disrupted by technologies,

environments, missions, obsolescence, regulations, behaviors, markets, human migrations, conflicts, etc. [7,8]. Systems engineering informs designs for unprecedented and unimagined disruptions. Risk, safety, security, trust, and resilience programs address scope, resources, and evaluation [9,10].

Risk is defined as the specifics of what can go wrong, the likelihood that an event will occur, and the consequences if it does. Complementary definitions of risk are as follows: Risk has also been defined as the influence of scenarios on priorities [9–13]. ISO 31000 defines risk as the effect of uncertainty on objectives. Systems are exposed to various risks, such as disruptive, uncertain, and unpredictable events over time that affect the performance of the system level [13–15]. Risk could also appear as a combination of threat, vulnerability, consequence, or hazard, exposure, and effect [16]. In other words, these three risk components are called triplets [13,17].

2.3. Examples of Artificial Intelligence (AI) Unintended Harms

This section describes various risks linked to the utilization of AI, including apprehensions regarding errors, biases, and unintended consequences, which have impeded its broad acceptance and led to a decline in trust. The adverse impacts of AI are not confined to individuals and entities but can also reverberate throughout society, impacting it at large [18].

AI algorithmic bias is a problem arising from the development and implementation of AI, which can have negative effects on effectiveness and fairness. Systemic bias, human bias, and statistical/computational bias are three main categories of AI bias. Systemic bias is a consequence of non-representative

samples, leading to errors. Algorithmic bias, such as under, over-fitting, and others, is a significant issue in machine learning, as it can be absorbed and perpetuated by AI systems, leading to systematically discriminatory outcomes.

Insufficient training data and the absence of data not only diminish potential advantages but also engender the potential for harm. The absence of proper representation can result in recommendations that are ill-suited for individuals not adequately accounted for in the dataset. The algorithm might struggle to differentiate between individuals for whom there is insufficient data, hampering its ability to comprehend and adapt to variations [18].

Measurement and misclassification errors in the dataset are another source of bias in observational studies. Differential misclassification can occur due to errors by practitioners, and implicit biases related to patient factors like sex, race, ethnicity, and practitioner-related factors may also impact the quality of care provided [18].

Although this section is not the primary focus of this dissertation, it is worth mentioning some examples of unintended harms caused by AI to illustrate the serious risks associated with the technology.

2.4. Artificial Intelligence (AI) and Its Risks in Healthcare

With scientific advances in the healthcare domain and medical field [19], the systems face more demand for enhancing user services. The surge in this demand encourages the utilization of technology such as AI. In one hand, AI has revolutionized healthcare by transforming state-of-the-art diagnoses [20], treatment, disease prevention, surgical devices, and more [21,22]. Thus,

healthcare systems became a promising application area for this technology. For instance, in Europe, AI in the healthcare market exceeded $1.15 billion in 2020 and is expected to grow more than 44.2% by 2027 [23]. AI in healthcare has the potential to significantly improve outcomes and reduce the costs and time of procedures [24]. However, applying AI in the healthcare domain has some limits and challenges. Managing risks caused by AI is as essential as managing risks caused by other technologies and disruptive events. AI holds significant promise; however, the present negative consequences of AI and the potential risks and harms associated with its ongoing advancement pose substantial threats. The dominance of large corporations in controlling technologies implies the need for caution when forecasting the potential advantages of AI [25]. AI algorithm apprehensions have emerged owing to errors, biases, and a lack of transparency; these concerns have led to a decline in trust among clinicians and patients. For instance, it poses the risk of further accentuating pre-existing biases against marginalized groups, leading to exacerbating inequities and unintended harms. Moreover, a lack of representation in the data can cause harm, for example, by advising courses of action that may not fit individuals not represented in the data or by failing to distinguish between individuals for whom there is insufficient data for the algorithm to understand variability [18]. The absence of well-established, reliable principles for properly utilizing AI and ML in healthcare has worsened the situation. The reliable application of AI in healthcare settings is a topic of discussion because there are no universally accepted guidelines for its implementation [26]. Using machine learning introduces diverse risks in “high-risk” AI systems such as healthcare [27]. Careful mitigation strategies are required to instill confidence in the system [28]. AI

should be valid and reliable, safe and fair [29], not biased [30,31], secure and resilient, explainable and interpretable, accountable and transparent [6,32–34]. However, transparency measures should depend on the area of AI application [35].

As mentioned, there are significant concerns regarding AI in healthcare, however, it is crucial to recognize the potential advantages that AI can offer to the field. Although there are difficulties, AI has the potential to transform healthcare by providing solutions that can improve patient care, streamline clinical processes, and enhance health results [6].

Moreover, important advantage of AI in the field of healthcare is its capacity to enhance and support clinical decision-making procedures. AI-driven diagnostic tools possess the capability to rapidly and accurately analyze extensive quantities of medical data, thereby assisting healthcare providers in making well-informed decisions pertaining to patient diagnosis and treatment strategies. This can result in the early identification of illnesses, the implementation of more tailored treatment strategies, and ultimately improved patient results [5,36].

Also, AI has the capacity to enhance operational efficiencies in healthcare systems. AI can enhance efficiency in administrative tasks, resource allocation, and scheduling by utilizing predictive analytics and optimization algorithms. This leads to cost reduction and better utilization of resources. This can enable healthcare professionals to allocate more time to direct patient care and decrease patient wait times [5].

Furthermore, the implementation of AI-powered technologies like remote patient monitoring, telemedicine, and wearable devices holds the promise of

expanding the availability of healthcare services, especially in areas that lack sufficient access or are geographically isolated. These technologies facilitate the ongoing monitoring of patient health in non-traditional healthcare environments, resulting in earlier intervention and improved management of chronic conditions.

Moreover, the utilization of AI in research and development for drug discovery and development shows potential for expediting the rate of advancement in the field of medicine. Researchers can accelerate the discovery of new treatments and therapies for different diseases by utilizing AI algorithms to analyze molecular structures, forecast drug interactions, and identify potential therapeutic targets [19,23].

Nevertheless, it is crucial to approach the incorporation of AI in healthcare with prudence and awareness of the potential hazards and difficulties mentioned previously. Ensuring the ethical and responsible development, deployment, and use of AI technologies in healthcare is crucial for fully realizing their benefits and maximizing their positive impact on patient care and public health. Effective collaboration among healthcare professionals, researchers, policymakers, and technologists is essential for successfully navigating the intricacies involved and utilizing the revolutionary potential of AI to improve healthcare delivery and outcomes.

2.5. National Institute of Standards and Technology Artificial Intelligence Risk Management Framework

The NIST AI RMF (National Institute of Standards and Technology Artificial Intelligence Risk Management Framework), published in 2023 [34], is a set of guidelines developed to help organizations manage risks associated with AI systems [34]. It covers technical, operational, and ethical considerations. The framework includes risk governance, assessment, mitigation, assurance, and communication. It emphasizes transparency, explainability [37], and accountability in AI systems. The framework assists organizations in identifying, assessing, and mitigating risks throughout the lifecycle of AI systems, promoting trust by evaluating trust and being responsive to reasons for mistrust and distrust, fairness, and reliability. In summary, the NIST AI Risk Management Framework addresses risks in designing, developing, using, and evaluating AI systems and products [34]. The framework describes the requirements that need to be addressed to have a trustworthy AI application. Figure 2 describes the NIST *principles* of trustworthy AI [28,34]. According to this figure, a trustworthy AI system should include all seven *principles*: 1. Accountable and transparent, 2.

Valid and reliable, 3. Safe, 4. Fair – With harmful bias managed, 5. Secure and resilient, 6. Explainable and interpretable, and 7. Privacy enhanced [34].

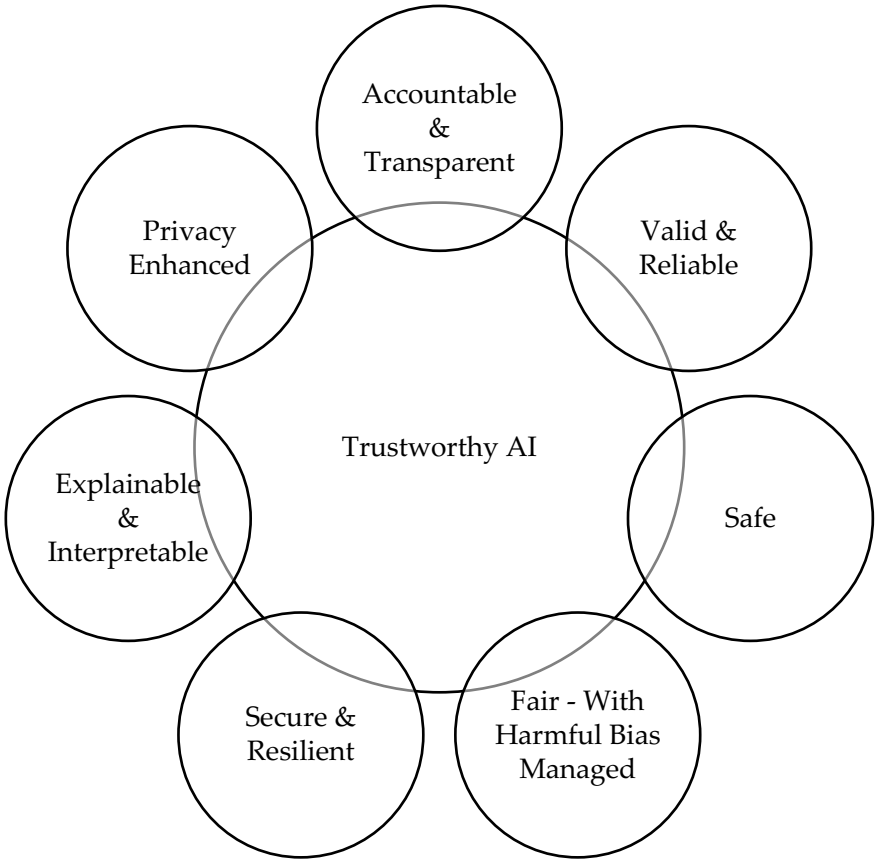

Figure 2. *Principles* of trustworthy AI adapted from the NIST AI Risk Management Framework [34].

2.6. Opportunities to Improve Science and Practice

The NIST AI risk management framework serves as a starting point for addressing the risks of AI. However, it is important to consider specific gaps and limitations. These include the need for more specificity in the framework, limited coverage of technical risks, the absence of clear guidance on adapting AI

applications, and challenges with integrating existing risk management frameworks specific to certain industries. Another area for improvement is the high-level guidance of the framework, which needs more detailed instructions for implementing specific risk management measures tailored to the unique AI applications of organizations. As a result, organizations may need help translating the framework into actionable steps. The NIST AI risk management framework is a dynamic framework that evolves. The NIST actively seeks feedback and input from various experts and actors to improve its effectiveness and address these limitations.

With the above background, there is an opportunity for mathematical explanations and systems analysis to assist the widespread adoption of the NIST framework by critical organizations.

2.7. Explainable Artificial Intelligence (XAI)

Concerns are growing over the risks associated with AI advancements [1]. It has been observed that AI has surpassed human performance in a wide range of intellectually challenging tasks. One common approach involves developing machine learning models and assessing the associated risks [38]. Any necessary adjustments are made to the model based on this evaluation. However, failing to consider and recognize the potential risks of the system during the initial stages of design could lead to disastrous consequences and unintended harm to individuals and society. Thus, it is crucial to take a proactive approach to anticipate and resolve potential problems well in advance to help maintain the ability to make corrections and prevent irreversible consequences [38–40].

Trustworthy AI and XAI ensure that AI predictions can be trusted and widely utilized. Due to this, research on XAI [41–44] has emerged. Transparency, fairness, bias avoidance, informativeness, causality, confidence, transferability, privacy and safety, and ease of use as essential factors in the design and development of XAI [39] should be considered. XAI also aims to provide insight into AI decision-making [45]. XAI is dedicated to creating prediction models that are easily comprehensible to humans. These models provide predictions and employ model-agnostic techniques to generate explanations for existing machine learning models [40,45–47].

Figure 3 describes the current taxonomy used to classify explanation methods. Two categories of explainable methods exist: 1. Explanation by design or intrinsic explainability, and 2. Black-box explanation or post-hoc explainability. One example of an intrinsic explanation is the decision tree model. The structure of the model can be easily understood due to its architecture. Some examples of post-hoc explainability include the support vector machine (SVM) model, extreme gradient boosting (XGBoost), and others. Figure 4 describes that as the performance of the models improves, their interpretability diminishes. Although the neural network (NN) model is highly accurate in its predictions, it is challenging to clearly explain how it arrives at its predictions [39,40].

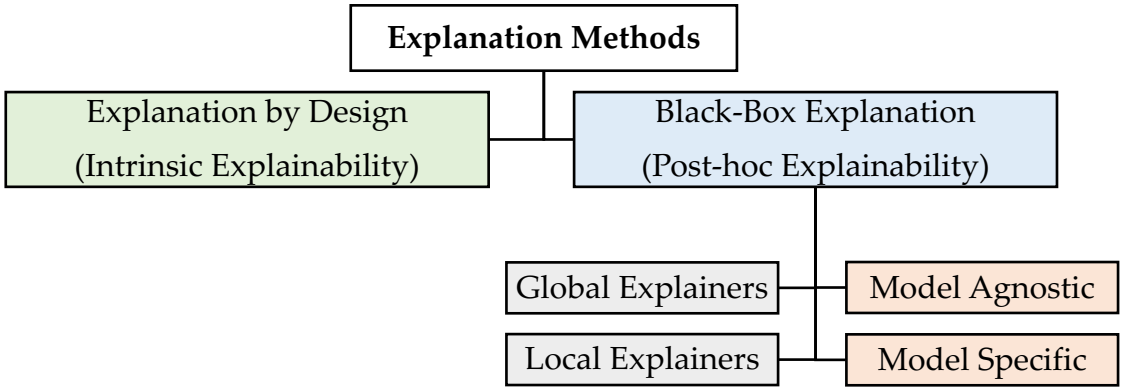

Figure 3. Classification of explanation methods; adapted from [39,40,48]

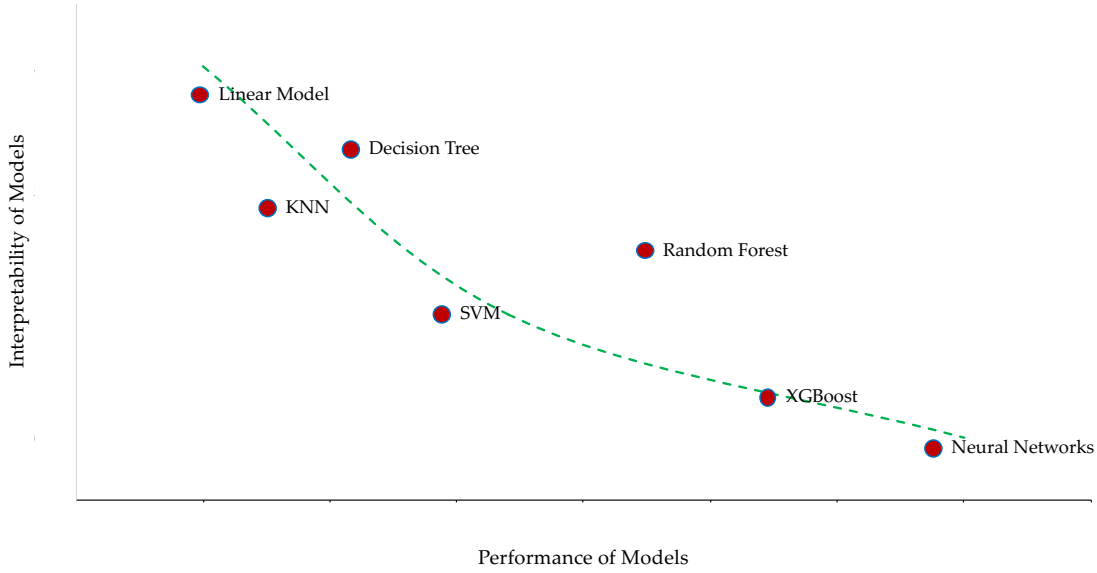

Figure 4. Scale-free interpretability and performance relations based in models type; as the performance of the models increase, the interpretability of the models decreases; adapted from [1,40,48].

Post-hoc explainability can be classified into various types, including local and global explainers, and model agnostic and model specific approaches. On the one hand, global explanation methods focus on elucidating the underlying

reasoning of a black-box model [38,46]. Thus, the explanation provided is a comprehensive and universal explanation that applies to any situation. On the other hand, local explainers focus on explaining the reasons behind a black-box model decision for a particular instance [38]. Explanation methods that are model-agnostic can be used to interpret any black-box model. However, model-specific explanation methods are specific to a particular type of black-box model [48].

Characteristics of other approaches for explaining decision-making in tabular data include feature importance, rule based, prototype, and counterfactual. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) [49,50] are two popular explainers that are considered feature importance types.

Table 2 is a summary of XAI models. In the following sections, several XAI techniques will be utilized and discussed in various cases, highlighting their importance in identifying and resolving model errors, addressing potential biases, addressing ethical concerns, and promoting trust and cooperation between individuals and their AI assistants. In this dissertation, some models, including feature importance and rule-based explainers, were implemented.

Table 2. List of black-box explainer models for tabular data; adapted from [39]

Type of the Explanations for Tabular Datasets	Name of the Models	Full Name of the Models	Sources
Feature Importance	LRP	Layer-wise Relevance Propagation	Bach et al. (2015)
	LIME	**Local Interpretable Model-Agnostic Explanations**	Ribeiro et al. (2016)
	SHAP	**SHapley Additive exPlanations**	Lundberg and Lee (2017)
	MAPLE	Model-Agnostic Projection and Local Explanation	Plumb et al. (2018)
	EBM	Explainable Boosting Machine	Nori et al. (2019)
	NAM	Neural Additive Model	Agarwal et al. (2021)
	CIU	Counterfactual Inference for Understanding	Anjomshoae et al. (2020)
	EEM	Empirical Explanatory Models	Chowdhury et al. (2022)
	DALEX	Descriptive mAchine Learning EXplanations	Lipovetsky (2022)
Rule Based	TREPAN	Decision Tree Analyzer	Shavlik (1995)
	MSFT	Microsoft (not a specific XAI model, but a company)	Chipman et al. (1998)
	CMM	Conceptual Memory Model	Domingos (1998)
	DECTEXT	Decomposable Attention Model for Natural Language Inference	Boz (2002)
	STA	Sequential Testing Approach	Zhou and Hooker (2016)
	SCALABLE-BRL	Scalable Bayesian Rule Lists	Yang et al. (2017)
	LORE	Local Rule-Based Explanations	Guidotti et al. (2019a)
	RULEMATRIX	Rule Matrix	Ming et al. (2019)

	ANCHOR	Anchors	Ribeiro et al. (2018)
	GLOCALX	Global-Local Explanations	Setzu et al. (2019)
	SKOPERULE	Scalable Rule-Based Models	Friedman and Popescu (2008)
Prototype	PS	Prototype Selection	Bien and Tibshirani (2011)
	MMD-CRITIC	Maximum Mean Discrepancy Critic	Kim et al. (2016)
	PROTODASH	Prototype Discovery for Anomaly Detection in SHapley space	Gurumoorthy et al. (2019)
	TSP	Time Series Pattern	Tan et al. (2020)
Counterfactual	CEM	Contrastive Explanation Method	Dhurandhar et al. (2018)
	CFX	**Counterfactual Explanations**	Albini et al. (2020)
	DICE	Diverse Counterfactual Explanations	Mothilal et al. (2020)
	C-CHAVE	Counterfactual-based CHAnge point Visual Explanation	Pawelczyk et al. (2020)
	FACE	Feature Attribution by Contrastive Explanation	Poyiadzi et al. (2020)
	ARES	Adaptive Rule-based Expert System	Ley et al. (2022)

2.8. Summary

This chapter has described the literature on risk management in the context of AI and its risks in healthcare. It discusses the importance of systems modeling and risk management in addressing complex systems. The NIST AI risk management framework provides guidelines for managing risks associated with

AI systems, emphasizing transparency, explainability, and accountability. The framework aims to identify, assess, and mitigate risks throughout the lifecycle of AI systems, promoting trust, fairness, and reliability. The framework is a starting point for addressing AI risks, but it has limitations such as lack of specificity, technical risks, and guidance on adapting AI applications. Then, the chapter describes the use of XAI in explaining prediction models that are easily comprehensible to humans. Two categories of explainable methods exist: Intrinsic and black-box explainability. XAI techniques are also implemented to identify and resolve model errors, address potential biases, and promote trust and cooperation between individuals and AI assistants.

Chapter 3 | Theory and Method

3.1. Introduction

This chapter describes the theory and method. It introduces the initiatives, success criteria, emergent conditions, and scenarios as the components of the analysis. The multi-layered scenario-based disruption of priorities aims to find the most and least disruptive scenarios and prioritize the highest ranked initiatives for AI users at each system *characteristic* layer. Scenarios reflect the most uncertainties in the system life cycle identified by expert and decision-maker points of view [9–11,51]. This section is divided into two subsections, as follows:

3.2. Layers of System Characteristics in Enterprise Risk Management of AI in Healthcare

In healthcare, a distinct void exists that calls for the convergence of risk analysis, systems modeling, and AI applications. Figure 1 and Figure 5 show that

in systems modeling, the defining *systems characteristic layers* of any system are *Purpose*[1] *(π, Pi), Structure*[2] *(σ, Sig), Function*[3] *(ϕ, Phi), Interconnections (ι, Iot), Environment (ε, Eps),* and *Boundary (β, Bet)* [6,52–55]. The Roman alphabets are employed to facilitate fluent reading and enhance annotations throughout this dissertation. Other studies may find additional layers for the AI risk management analysis.

- ***Purpose (Pi) layer:*** This layer focuses on the objectives and overall goal of the system and includes the strategic and operational objectives of the system. This layer includes domain experts and actors in healthcare such as health center board members and clinicians responsible for the operation of a clinic section/sector.

- ***Structure (Sig) layer:*** This layer includes the physical framework of the system which could resemble physical medical devices. These are device developers and designers involved in implementing AI in healthcare.

- ***Function (Phi) layer:*** This layer includes a specific operation or a task defined and performed by medical professionals, such as disease diagnosis. For instance, physicians specializing in radiology and cardiology contribute to the functional aspects of AI applications in healthcare.

---

[1] In some literatures, *Purpose (Pi)* is also referred as behavior [52–54].
[2] In some literatures, *Structure (Sig)* is also referred as elements or components [52–54].
[3] In some literatures, *Function (Phi)* is also referred as process or operations [52–54].

• *Interconnections (Iot) layer:* This layer describes the interactions and connectivity of medical components.

• *Environment (Eps) layer:* This layer includes any external factors or environments that could affect the medical system outside its boundary.

• *Boundary (Bet) layer:* This layer defines the limits of the medical system scope. This layer distinguishes the medical system from its external *Environment (Eps)*.

The scope of the dissertation is limited to the *Purpose (Pi), Structure (Sig),* and *Function (Phi)* layers.

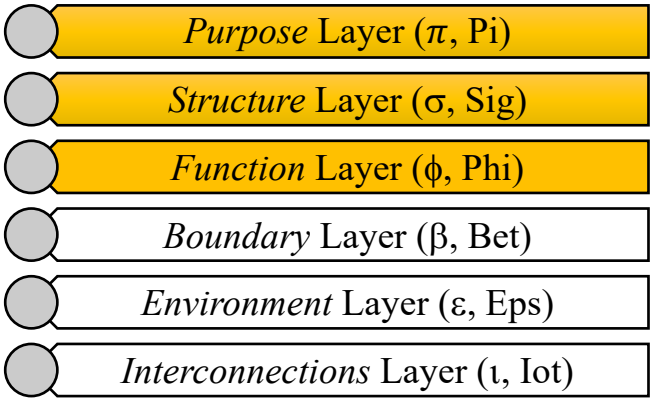

Figure 5. Six layers of system *characteristics* that can be used in enterprise risk management of AI in healthcare. Orange cells indicate the scope of this dissertation [6].

The following chapters will describe three scenario-based preference risk registers for deploying AI in complex engineering systems built on top of the NIST AI Risk Management Framework *principles*.

The risks of AI in the healthcare context should be considered for involved subjects, including AI developers, healthcare clinicians, and patients distributed

at three layers: Insider, internal, and external layers, respectively[4]. The scope of Chapters 4, 5, and 6 is limited to the internal trustworthiness that AI providers address to AI users listed below. AI users and perspectives for the three healthcare layers are:

- *Purpose (Pi)* layer: Domain experts and actors in healthcare, i.e., health center board members and clinicians[5] responsible for the operation of a clinic section/sector;
- *Structure (Sig)* layer: Device developers and designers[6];
- *Function (Phi)* layer: Physicians with specialties in radiology and cardiology[7].

3.3. Scenario-Based Disruption of Priorities

This section describes a scenario-based disruption of priorities that is a mathematical decision model for system priorities. This framework extends a model applied to the trust and security of electric vehicle-to-grid systems and

[4] These AI actors may include trade associations, standards-developing organizations, researchers, advocacy groups, environmental groups, civil society organizations, end users, and potentially impacted individuals and communities.

[5] We acknowledge that trust relationships between patients and insurance providers could significantly differ. While the users of AI systems can be diverse, including patients and insurance providers, the focus of this case study is limited to the domain experts and actors in healthcare, i.e., health center board members and clinicians—a collaboration with Binagostar Eye Surgical Hospital board members in Shiraz, Iran.

[6] A collaboration with Johns Hopkins University, the department of mechanical engineering, the Massachusetts Institute of Technology, the University of Virginia, the department of data science, and the department of systems engineering.

[7] A collaboration with HDZ-NRW hospital, the department of radiology, nuclear medicine and molecular imaging, the heart and diabetes center North-Rhine Westphalia, the Ruhr University of Bochum.

hardware supply chains as R&D priorities for the security of embedded hardware devices [10,11]. The purpose of the framework is to assist decision-makers with prioritizing resilience capabilities and identifying disruptive scenarios of trustworthy AI in healthcare systems. It also describes an elicitation of scenario-based preferences that aids in identifying system initiatives, success criteria, and emergent conditions.

Figure 6 describes a conceptual diagram of the risk assessment methodology for the risk of AI in healthcare. The dissertation identifies criteria, initiatives, emergent conditions, and scenarios. The next step is to assess criteria-initiative, criteria-scenario effects, and emergent conditions-scenarios. Finally, the highest ranked initiatives and the most and least disruptive scenarios were identified.

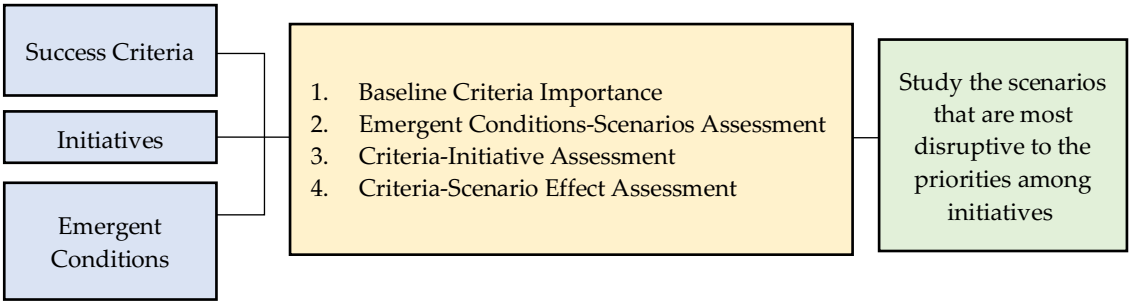

Figure 6. Conceptual diagram of risk assessment methodology for enterprise risk management of AI in healthcare, adapted from [6,10,11,18,40].

Success criteria, as the first element of the framework, are developed to measure the performance of investment initiatives based on the system objectives [8,10,11]. Success criteria are mainly derived from research on technological analyses, literature reviews, and expert opinions, which mainly describe the goals of the system. Their relationships with the initiatives are developed to

measure the potential impact of investing in specific initiatives [51]. Any changes in success criteria affect expectations of success and represent the expert values. The set of success criteria is {c.01, c.02, …, c.m}.

Since this framework builds on the NIST AI Risk Management Framework, the success criteria are the seven trustworthy AI system *principles* in Figure 2. The list of success criteria is comprised of *c.01 – Safe, c.02 – Secure & Resilient, c.03 – Explainable & Interpretable, c.04 – Privacy Enhanced, c.05 – Fair (With Harmful Bias Managed), c.06 – Accountable & Transparent, c.07 – Valid & Reliable.*

The baseline relevance of criteria is established by interviewing experts and actors to assess each criterion under normal conditions, with the relative emphasis being scored as *Low, Medium*, and *High*. Based on this determination, each success criteria baseline weight is assigned. Table 3 describes that numbers 0, 1, 2, and 4 are assigned to dashes (-), *Low, Medium*, and *High,* respectively. The values could change, considering various experts, actors, and contexts. A sensitivity analysis is provided in Chapter 7 to evaluate how the ranking of initiatives will change by increasing or decreasing the criteria scenario relative importance weights.

Table 3. Criteria-scenario relative importance weights

Criteria Scenario Relative Importance	Weights
High	4
Medium	2
Low	1
-	0

Initiatives, as the second element of the model, represent a set of decision-making alternatives in the form of technologies, policies, assets, projects, or other such investments [8,56]. The set of initiatives is {x.01, x.02, …, x.n}. Initiatives are developed through literature reviews and interviews with experts and actors to determine what components, actions, assets, organizational units, policies, locations, and allocations of resources constitute the system [6,9,51,57].

To assess how well each initiative addresses the success criteria for each *system characteristic* layer in the enterprise risk management of AI in healthcare, experts and actors are interviewed as part of the criteria-initiative (C-I) assessment. In criteria-initiative assessment, the decision-makers assess to what degree they agree on initiative x.i address criterion c.j. In the C-I assessment, *neutral* entries are represented by a dash (-), *somewhat agree* is represented by an unfilled circle (○), *agree* is represented by a half-filled circle (◐), and *strongly agree* is represented by a filled circle (●) in the matrix [9,10,57]. Table 4 describes the weights associated with each degree initiative x.i address criterion c.j.

Table 4. Criteria-initiative assessment weights

Criteria-Initiative Assessment	Degree	Weights
●	*strongly agree*	1
◐	*agree*	0.667
○	*somewhat agree*	0.334
–	*neutral*	0

The qualitative results of the project constraint matrix are converted into numerical weights [58,59] following a rank-sum weighting method [60] based on Equation 1:

$$w_j = \frac{m-rankj+1}{\sum_{j=1}^{m} \mathrm{m}-rankj+1} \Rightarrow \forall_j \in C \quad (1)$$

Where w_j is the weight of the j-th criterion, m is the total number of criteria, and $rank_j$ is the ordinal rank of the j-th criterion [6,61].

Emergent conditions, as the third element of the model, are events, trends, or other factors impacting the priorities of decision-makers in future strategic planning contexts. Emergent uncertainties significantly contribute to project failure and impact the ability of the system to meet success criteria [51]. The set of emergent conditions is {e.01, e.02, …, e.k}. In the framework, emergent conditions influence the relevance weights of individual prioritization criteria, increasing or decreasing.

Scenarios are comprised of one or more emergent conditions [6,56]. The set of scenarios is {s.01, s.02, …, s.p}.

The next step is to adjust the success criteria weights for each scenario. The effect of disruptive emergent conditions is operationalized by changing in the criteria weights. For each scenario, the user is asked to assess to what degree the relative importance of each criterion change given the scenario will occur [62]. Responses include *decreased (D)*, *decreased somewhat (DS)*, *no change*, *increased somewhat (IS),* and *increased (I)*. These changes are recorded in the *W* matrix. In Equation 2, α is a scaling constant equal to {8, 6, 1, 1/6, 1/8} for *increases, increases somewhat, no change, decreases somewhat, and decreases,* respectively. The scaling constant is intended to be consistent with the swing weighting rationale. The swing weight technique accommodates adjustments for the additional scenarios. The procedure for deriving weights for an additive value function using the

swing weight method is thoroughly documented in the MCDA literature, as evidenced by works such as those by Keeney and Raiffa [59], Keeney [63], Belton and Stewart [64] and Clemen and Reilly [65]. Karvetski and Lambert explain the rationale for swing weighting as follows: α serves as a value multiplier, adjusting the trade-off between exchanging a high-level of performance for a low level of performance in one criterion and an exchange of a low level of performance for a high level of performance in another criterion. [61,66]. The swing weight technique was adopted to derive the baseline criteria weights (W_j) and the adjusted weights for each scenario [66].

$$W_{jp} = \alpha * W_j \Rightarrow \forall_j \in C, \forall_p \in S \quad (2)$$

Table 5 describes the weights assigned to each criterion relevant across each scenario. These changes are recorded in the *W* matrix.

Table 5. Criteria-scenario importance change

Criteria-Scenario Importance Change	Weights
Increases	8
Increases Somewhat	6
-	1
Decreases Somewhat	0.1667
Decreases	0.125

The initiatives are prioritized with a linear additive value function, defined in Equation 3. vj(x.i), the partial value function of initiative x.i, and criterion c.j,

defined through the C-I assessment. V is a matrix that contains the relative importance scores for each initiative across each scenario.

$$V_p(x.i) = \sum_{j=1}^{m} w_{jp} v_j(x.i) \Rightarrow \forall_i \in X, \forall_p \in S, \forall_j \in C \quad (3)$$

Each initiative score is ranked for each scenario, providing a ranking from 1 to n. Equation 4 describes this process, where > indicates that an initiative has a higher ordinal ranking. That is if the score of an initiative x.i is higher than that of an initiative x.a, then it has a higher ordinal ranking. For instance, they may be ranked 1 and 2 [6,40,67].

$$IF \quad V_p(x.i) > V_p(x.a) \quad THEN \quad x.i > x.a \quad \Rightarrow \quad \forall_{i,a} \in X, \forall_p \in S \quad (4)$$

The next step is to calculate the disruptiveness score. Disruptiveness measures the degree to which priority orders change under a given scenario [67]. It is defined based on the sum of the squared differences in priority for each initiative compared to the baseline scenario [10,51,67]. Equation 5 describes the disruptiveness score for scenario s.p.

$$D_p = \sum_{i=1}^{n} \left(r_{i0} - r_{ip}\right)^2 \quad \Rightarrow \quad \forall_i \in X, \forall_p \in S \quad (5)$$

r_{ip} is the rank of initiative x.i under scenario s.p, and r_{i0} is the rank of initiative x.i under the baseline scenario. These scores are then normalized on a 0-100 scale [6,8,11,56,57].

3.4. Summary

The chapter has described the mathematical decision framework applied to the risks of AI in healthcare systems. Determining which assets are vulnerable to disruption will assist in prioritizing resilient strategies to identify the risks to the system using AI as an application to operate, design, develop, and diagnose. Chapters 4, 5, 6, and 7 will demonstrate the method for each *characteristic* of system layers.

Chapter 4 | Case 1 (*Purpose (Pi)* Layer)

4.1. Introduction

This chapter describes a risk register that acknowledges the *Purpose (Pi)* of the system. Initially, a mathematical decision framework is employed to assess the risks of AI in this layer. This layer focuses on the goals and objectives of the system, explicitly emphasizing the internal trustworthiness that involves AI providers catering to healthcare AI users across various roles. The users include those involved in hospital or healthcare institute/clinic operations and team members at all levels within the healthcare organization [6].

For demonstration, experts and actors from different medical specialties were engaged in the process and interviewed from the early stages of the study, from identifying the initiatives, emergent conditions, and scenarios to scoring and ranking assessments. The experts and actors for the *Purpose (Pi)* layer are the board members of Binagostar Eye Surgical Hospital. Three interviews were

conducted with board members of Binagostar Eye Surgical Hospital through the online platform [6].

4.2. Scenario-Based Disruption of Priorities (Purpose (Pi) Layer)

Table 6 describes seven success criteria that were identified for this layer. These seven success criteria are the *principles* of the NIST AI risk management framework that are shared as success criteria between all three *Purpose (Pi), Structure (Sig),* and *Function (Phi)* layers [6].

Table 6. Success criteria for the *Purpose (Pi), Structure (Sig),* and *Function (Phi)* layers in enterprise risk management of AI in healthcare. Success criteria are adapted from the NIST AI risk management framework [6,18,34,40,68].

Index	Criterion
c.01	Safe
c.02	Secure & Resilient
c.03	Explainable & Interpretable
c.04	Privacy Enhanced
c.05	Fair – With Harmful Bias Managed
c.06	Accountable & Transparent
c.07	Valid & Reliable
c.i	Others

Table 7 describes forty-three identified initiatives by interviewing the experts, actors, and literature reviews [6,9,34,35,69,70]. *x.Pi.11 - Identify Roles and Responsibilities of Humans Involved in the AI Lifecycle* is an example of an identified initiative [6].

Table 7. Initiatives for the *Purpose (Pi)* layer in enterprise risk management of AI in healthcare [6]. Abridged from various sources that are identified in the narrative [6].

Index	Initiative
x.Pi.01	Identify At-Risk Components
x.Pi.02	Understanding ML Tools to Uncover Subtle Patterns in Data
x.Pi.03	Record-Keeping, Reserving, and Storing
x.Pi.04	Data Governance and Management
x.Pi.05	Data Traceability of the Process
x.Pi.06	Clear and Plain Language
x.Pi.07	Concise, Transparent, Easily Accessible Form and Process
x.Pi.08	Human-AI Collaboration and Consulting
x.Pi.09	Accurate, Appropriate, Clear and Accessible Information
x.Pi.10	Providing of Information/Documents
x.Pi.11	Identify Roles and Responsibilities of Humans Involved in the AI Lifecycle
x.Pi.12	Ensuring Safety and Quality of AI in Healthcare
x.Pi.13	Making Informed Decisions (Which in the Case of Patients Also Leads to the Realization of Individual Rights)
x.Pi.14	Guaranteeing Quality and Safety
x.Pi.15	Continuous Collecting, Generating, and Verification of Data, Information, and Knowledge
x.Pi.16	Ex-Ante and Ex-Post Control Over the AI Outcomes
x.Pi.17	Outcome Assessment Through Explanations, and Keeping the Records of AI Development and Testing
x.Pi.18	Interpretation for an Erroneous Prediction to Understand the Cause of the Error
x.Pi.19	Comprehension of AI-Based Devices and their Decisions
x.Pi.20	Clinicians Shall be Provided with the Information that Enables them to Choose in What Situations to Apply AI Tools, How to Use them, and How to Verify the Results that an AI System Suggests
x.Pi.21	The Safety Risks Specific to AI Systems Covered by the Requirements of the AI Regulation, and the Sectorial Legislation
x.Pi.22	Tries to Minimize Risks to the Maximum Possible Extent
x.Pi.23	Clinicians to be Convinced that Specific AI System Outcome is Safe
x.Pi.24	Clinicians to be Convinced that AI Device is Generally Useful, Safe, and Efficient
x.Pi.25	Clinicians to have All the Necessary Information, Documents, and Explanations Provided Directly (Through Interaction and Cooperation) or Indirectly (Through the Process of AI Device Verification and Authorization) by AI Developers
x.Pi.26	Ability to Perform the Functions of the AI Internal Transparency
x.Pi.27	Ensure the Safety and Quality of AI Medical Devices before and after they are Placed on the Market
x.Pi.28	Identify Residual Risks, Contra-Indications, and Any Undesirable Side Effects, Including Information to be Conveyed to the Patient in this Regard
x.Pi.29	Provide Specifications the User Requires to Use the Device Appropriately, e.g., If the Device Has a Measuring Function, the Degree of Accuracy Claimed for it

x.Pi.30	Provide Any Requirements for Special Facilities, or Special Training, or Particular Qualifications of the Device User and/or Other Persons
x.Pi.31	Some Adaptations or Interpretations to the Nature of AI Technologies, their Opacity, and Self-Learning Shall be Made by Policy Makers for Ensuring Transparency at the Internal Level
x.Pi.32	Data Governance and Management Practices Shall be Developed by AI Providers
x.Pi.33	To Inform the Users on What Kind of Data was Used to Train and Validate an AI System and How System Parameters Might Change Depending on the Input Data
x.Pi.34	The Required Information About the Device Risks, Side Effects, and Limitations Shall Cover the Risk of Algorithmic Changes, AI Providers Predictions of that Changes, As Well As the Explainability Limitations When it is Applicable (Black-Box Models)
x.Pi.35	AI Providers Have to Also Inform Users on Why and How the Benefits of the Use of this System Overweigh its Risks (Level of Accuracy, Comparison with Other Technologies, or Practices Available on the Market)
x.Pi.36	AI Providers Shall Develop Possible Technical Measures to Implement Automatically Generated Explanations into their AI Systems ('Explanations by Design')
x.Pi.37	Application-Grounded Evaluation of Interpretability Involves Conducting Human Experiments Within a Real Application
x.Pi.38	The Quality of Explanations Developed by the AI Provider to be Supplied Together with the Device in Question Shall be Assessed by Healthcare Professionals Specializing in the Area of the Device Use
x.Pi.39	Some Sort of Independent Bodies Representing Healthcare Professionals for their Participation in the AI-Device Evaluation Can be Created
x.Pi.40	Accepting Some Degree of Opacity If the Benefits of the Use of AI Application Overweigh this Opacity Risk
x.Pi.41	Providing Quality Records as A Part of the Quality Management System
x.Pi.42	The Requirement to Use State-Of-The-Art Explainability Techniques Shall be Part of the Conformity Assessment Process
x.Pi.43	Enables AI Providers to Justify the Acceptance of Some Level of Opacity (Because the Technologies that Solve it Do Not Exist)
x.Pi.i	Others

Table 8 describes twenty-five emergent conditions identified through literature reviews [6,35,69] and interviews with the experts and actors regarding the risk of AI in the healthcare *Purpose (Pi)*/sector layer. An emergent condition for this layer could be *e.Pi.19 - One-Size-Fits-All Requirements AI Model Challenges* that may disrupt the system [6].

Table 8. Emergent conditions were used to create sets of scenarios for the *Purpose (Pi)* layer in enterprise risk management of AI in healthcare [6]. Abridged from various sources that are identified in the narrative [6].

Index	Emergent Condition
e.Pi.01	Lack of Algorithmic Transparency
e.Pi.02	Low Quality and Relevance of the Inputs and thus the Relevant Procedures to Verify it
e.Pi.03	Concerning the Trade-Off Between AI Accuracy and Explainability
e.Pi.04	Impossible Reaching Zero Risks in AI Area
e.Pi.05	Lack of Full Predictability (Due to Constant Self-Learning) of Some AI Applications
e.Pi.06	Concerns About Information Provision
e.Pi.07	Some AI Models are Opaque even for their Creators, and this Issue Brings us to the Last Level of Transparency – Insider Transparency
e.Pi.08	Some AI Models are Inherently Opaque and Cannot be Fully Explained
e.Pi.09	The Availability for Explanations and the Quality of the Data Used in the AI Development and Training Process
e.Pi.10	Different Automated Explanations Techniques Available for the Model in Question
e.Pi.11	Legislative Requirements (Already Existing or to be Adopted in the Future)
e.Pi.12	Banning and Limitations of AI Technologies Usage in High-Risk Areas Such as Healthcare
e.Pi.13	The Level of Algorithmic Opacity that Cannot be Technically Solved at the Moment
e.Pi.14	Limitations in Always Accurately Predict the Outcomes of Medical Treatment
e.Pi.15	Limitations in Explaining the Health Conditions of the Specific Person (Diagnose Him/Her) Or Explain Why His/her Treatment Did Not Help
e.Pi.16	Shortage of AI in Cognitive Empathy
e.Pi.17	Hard to Track and Measuring Emergent Risks by Organizations
e.Pi.18	Security Concerns Related to the Confidentiality, Integrity, and Availability of the System and its Training and Output Data
e.Pi.19	One-Size-Fits-All Requirements AI Model Challenges
e.Pi.20	Unexpected Changes in the Environment or Use
e.Pi.21	Data Poisoning
e.Pi.22	Privacy Intrusions
e.Pi.23	Lack of Access to the Ground Truth in the Dataset
e.Pi.24	Intentional or Unintentional Changes During Training
e.Pi.25	Cyber Attacks
e.Pi.i	Others

Table 9 describes ten scenarios that were identified by grouping one or more emergent conditions. For instance, *e.Pi.02 - Low Quality and Relevance of the Inputs and thus the Relevant Procedures to Verify it* and *e.Pi.14 - Limitations in Always*

Accurately Predict the Outcomes of Medical Treatment are grouped under *s09. Uncontrollable Environment* [6].

Table 9. Emergent conditions are grouped for the *Purpose (Pi)* layer in enterprise risk management of AI in healthcare which describes which emergent conditions fit in each scenario [6]. Abridged from various sources that are identified in the narrative [6].

	s.01 - Funding Decrease	s.02 - Government Regulation and Policy Changes	s.03 - Privacy Attacks	s.04 - Cyber Security Threats	s.05 - Changes in AI RMF	s.06 - Non-Interpretable AI and Lack of Human-AI Communications	s.07 - Global Economic and Societal Crisis	s.08 - Human Errors in Design, Develop, Measurement and Implementation	s.09 - Uncontrollable Environment	s.10 - Expensive Design Process
e.Pi.01		✓			✓	✓		✓		
e.Pi.02	✓							✓	✓	
e.Pi.03	✓									
e.Pi.04						✓		✓		✓
e.Pi.05						✓		✓		
e.Pi.06		✓		✓						
e.Pi.07						✓				
e.Pi.08						✓				
e.Pi.09										
e.Pi.10	✓					✓		✓		
e.Pi.11	✓	✓			✓		✓			
e.Pi.12	✓	✓			✓		✓			
e.Pi.13						✓				
e.Pi.14					✓			✓	✓	
e.Pi.15						✓			✓	
e.Pi.16			✓			✓				
e.Pi.17	✓		✓			✓		✓	✓	✓
e.Pi.18	✓		✓				✓		✓	
e.Pi.19						✓			✓	
e.Pi.20	✓	✓			✓		✓		✓	
e.Pi.21								✓		
e.Pi.22			✓	✓						
e.Pi.23	✓							✓	✓	✓
e.Pi.24								✓	✓	
e.Pi.25				✓						

All criteria in Table 10 were categorized as having *High* relevance among the other criteria.

Table 10. Baseline relevance for the *Purpose (Pi)* layer in enterprise risk management of AI in healthcare [6].

The criterion c.xx has	s.00 - Baseline	relevance among the other criteria
c.01 - *Safe* has	*high*	relevance
c.02 - *Secure & Resilient* has	*high*	relevance
c.03 - *Explainable & Interpretable* has	*high*	relevance
c.04 - *Privacy Enhanced* has	*high*	relevance
c.05 - *Fair - With Harmful Bias Managed* has	*high*	relevance
c.06 - *Accountable & Transparent* has	*high*	relevance
c.07 - *Valid & Reliable* has	*high*	relevance

Table 11 describes the criteria-initiative assessment of how well each initiative addresses the success criteria for this system layer. For instance*, x.Pi.01 - Identify At-Risk Components* addresses the success criteria *c.01 – Safe* by using *Strongly agree* as a filled circle (●) and *agree* for c.02 – Secure & Resilient by a half-filled circle (◐) [6].

Table 11. The criteria-initiative assessment describes how well each initiative addresses the success criteria for the *Purpose (Pi)* layer in enterprise risk management of AI in healthcare. *Strongly agree* is represented by a filled circle (●), *agree* is represented by a half-filled circle (◐), *somewhat agree* is represented by an unfilled circle (○), and *neutral* is represented by a dash (━) [6].

	c.01	*c.02*	*c.03*	*c.04*	*c.05*	*c.06*	*c.07*
x.Pi.01	●	◐	○	○	◐	○	○
x.Pi.02	━	━	●	◐	○	◐	◐
x.Pi.03	━	━	○	●	━	◐	━
x.Pi.04	○	○	○	●	━	◐	━
x.Pi.05	●	○	○	◐	━	●	━
x.Pi.06	━	━	●	━	━	●	━
x.Pi.07	━	━	●	◐	━	●	━
x.Pi.08	━	━	●	━	━	●	◐
x.Pi.09	○	○	●	●	━	◐	━
x.Pi.10	○	○	○	●	━	◐	━
x.Pi.11	○	○	◐	◐	◐	◐	◐
x.Pi.12	●	●	━	━	○	◐	●
x.Pi.13	━	━	●	○	━	●	○
x.Pi.14	●	●	○	━	○	◐	●
x.Pi.15	○	◐	●	●	○	◐	○
x.Pi.16	◐	◐	◐	━	◐	◐	●
x.Pi.17	◐	◐	◐	━	◐	◐	●
x.Pi.18	◐	◐	●	━	━	●	●
x.Pi.19	●	●	●	○	●	●	●
x.Pi.20	●	●	●	◐	●	●	●
x.Pi.21	◐	◐	◐	◐	◐	◐	○
x.Pi.22	●	●	◐	◐	━	◐	◐
x.Pi.23	●	●	●	●	●	●	●
x.Pi.24	●	●	●	●	●	●	●
x.Pi.25	●	●	●	●	●	●	●
x.Pi.26	○	◐	●	●	●	●	●
x.Pi.27	●	●	●	●	●	●	●
x.Pi.28	○	○	●	○	○	●	○
x.Pi.29	●	●	●	●	●	●	●
x.Pi.30	◐	◐	●	◐	◐	●	◐
x.Pi.31	◐	◐	●	◐	◐	●	◐
x.Pi.32	○	○	○	●	━	◐	━
x.Pi.33	●	●	●	●	●	●	●
x.Pi.34	●	●	●	●	●	●	●
x.Pi.35	●	●	●	●	●	●	●
x.Pi.36	◐	◐	●	◐	◐	●	●
x.Pi.37	━	━	●	━	◐	●	◐

x.Pi.38	◐	◐	●	○	◐	●	◐
x.Pi.39	▬	▬	●	○	●	◐	○
x.Pi.40	◐	◐	●	○	○	●	◐
x.Pi.41	●	●	○	▬	○	◐	●
x.Pi.42	●	●	●	◐	◐	●	●
x.Pi.43	◐	◐	●	◐	◐	●	●

Table 12 describes the criteria-scenario relevance assessment for the *Purpose (Pi)* layer which shows how well each scenario fits the success criterion. For example, scenario *s.02 - Government Regulation and Policy, somewhat increases (SI)* criterion *c.04 – Privacy* [6].

Table 12. The criteria-scenario relevance describes how well each scenario fits the success criterion for the *Purpose (Pi)* layer in enterprise risk management of AI in healthcare. *Decrease Somewhat* = DS, *Decrease* = D, *Somewhat Increase* = SI, *Increase* = I [6].

	s.01	*s.02*	*s.03*	*s.04*	*s.05*	*s.06*	*s.07*	*s.08*	*s.09*	*s.10*
c.01	DS	SI	-	D	SI	DS	DS	DS	DS	DS
c.02	-	SI	-	D	SI	DS	DS	DS	DS	DS
c.03	-	SI	-	-	SI	DS	DS	DS	D	-
c.04	-	SI	D	D	SI	-	DS	-	-	-
c.05	DS	SI	-	-	SI	DS	DS	DS	-	-
c.06	DS	SI	-	-	SI	DS	DS	-	DS	DS
c.07	DS	SI	-	-	SI	DS	DS	DS	DS	DS

Ultimately, in Table 13, initiative-scenarios were ranked. The table describes the ranking of each initiative under each scenario for the *Purpose (Pi)* layer in the enterprise risk management of AI in healthcare and the results scores in the *R* matrix. This information is used to create the first artifact of the mathematical framework.

Table 13. Initiative-scenario ranking chart. This table describes the ranking of each initiative under each scenario for the *Purpose (Pi)* layer in enterprise risk management of AI in healthcare. The green filled cells show a higher ranking and the red and orange filled cells indicate a lower ranking.

	s.00 - Baseline	s.01 - Funding Decrease	s.02 - Government Regulation and Policy Changes	s.03 - Privacy Attacks	s.04 - Cyber Security Threats	s.05 - Changes in AI RMF	s.06 - Non-Interpretable AI and Lack of Human-AI Communications	s.07 - Global Economic and Societal Crisis	s.08 - Human Errors in Design, Develop, Measurement and Implementation	s.09 - Uncontrollable Environment	s.10 - Expensive Design Process
x.Pi.01	29	35	29	28	37	29	31	29	42	22	31
x.Pi.02	31	26	31	34	26	31	27	31	31	23	22
x.Pi.03	42	40	42	43	43	42	25	42	30	32	37
x.Pi.04	36	27	36	40	40	36	21	36	24	27	32
x.Pi.05	34	36	31	33	39	31	28	31	19	34	38
x.Pi.06	43	43	43	39	36	43	43	43	43	43	42
x.Pi.07	40	34	40	38	35	40	30	40	27	38	30
x.Pi.08	40	42	40	35	27	40	42	40	40	42	41
x.Pi.09	31	20	31	36	38	31	17	31	19	24	23
x.Pi.10	36	27	36	40	40	36	21	36	24	27	32
x.Pi.11	27	24	27	29	24	27	24	27	28	19	21
x.Pi.12	28	39	28	25	34	28	40	28	41	39	43
x.Pi.13	39	38	39	37	32	39	34	39	32	40	35
x.Pi.14	22	32	24	21	30	24	35	24	36	36	39
x.Pi.15	21	13	20	27	28	20	12	20	17	18	17
x.Pi.16	22	30	22	19	20	22	35	22	36	30	28
x.Pi.17	22	30	22	19	20	22	35	22	36	30	28
x.Pi.18	26	23	26	23	22	26	39	26	33	41	36
x.Pi.19	10	12	10	10	10	10	18	10	16	11	12
x.Pi.20	9	9	9	9	9	9	10	9	10	10	10
x.Pi.21	20	22	20	26	28	20	20	20	23	17	20
x.Pi.22	19	18	19	24	33	19	19	19	22	25	27
x.Pi.23	1	1	1	1	1	1	1	1	1	1	1
x.Pi.24	1	1	1	1	1	1	1	1	1	1	1
x.Pi.25	1	1	1	1	1	1	1	1	1	1	1
x.Pi.26	12	11	12	12	11	12	9	12	9	9	9
x.Pi.27	1	1	1	1	1	1	1	1	1	1	1
x.Pi.28	30	25	30	30	25	30	32	30	29	33	25

x.Pi.29	1	1	1	1	1	1	1	1	1	1	1
x.Pi.30	15	16	15	15	15	15	15	15	14	15	15
x.Pi.31	15	16	15	15	15	15	15	15	14	15	15
x.Pi.32	36	27	36	40	40	36	21	36	24	27	32
x.Pi.33	1	1	1	1	1	1	1	1	1	1	1
x.Pi.34	1	1	1	1	1	1	1	1	1	1	1
x.Pi.35	1	1	1	1	1	1	1	1	1	1	1
x.Pi.36	13	14	13	13	13	13	13	13	12	13	13
x.Pi.37	35	41	35	31	18	35	41	35	35	35	26
x.Pi.38	17	19	17	17	17	17	26	17	18	20	18
x.Pi.39	31	37	31	32	23	31	33	31	34	21	19
x.Pi.40	18	21	18	18	19	18	29	18	21	26	24
x.Pi.41	22	32	24	21	30	24	35	24	36	36	39
x.Pi.42	10	10	10	11	12	10	11	10	11	12	11
x.Pi.43	13	14	13	13	13	13	13	13	12	13	13

Based on the tables provided above, Figure 7 is generated, which describes how each scenario is given a disruptiveness score, where the higher the score, the more disruptive the scenario will be to a system. This figure describes that *s.06 – Non-Interpretable AI and Lack of Human-AI Communications* is the most disruptive scenario for trustworthy AI in the healthcare *Purpose (Pi)* layer [6]. In situations where the consequences of the actions of the system could be severe, such as when human life or liberty is at risk, AI developers and deployers should take proactive measures to adjust their transparency and accountability practices proportionally. Other disruptive scenarios to the system are *s.08 - Human Errors in Design, Develop, Measurement, and Implementation, s.09 - Uncontrollable Environment,* and *s.10 - Expensive Design Process*. These results will inform the development of subsequent resilience measures and could be expanded as the system model is improved. While scenarios *s.03 – Privacy Attacks* and *s.05 – Cyber Security Threats* hold significant roles within healthcare centers, they may not be

regarded as the most disruptive scenarios in this case. This is because these scenarios primarily focus on internal rather than external transparency, which involves patients and other external parties in the assessment process.

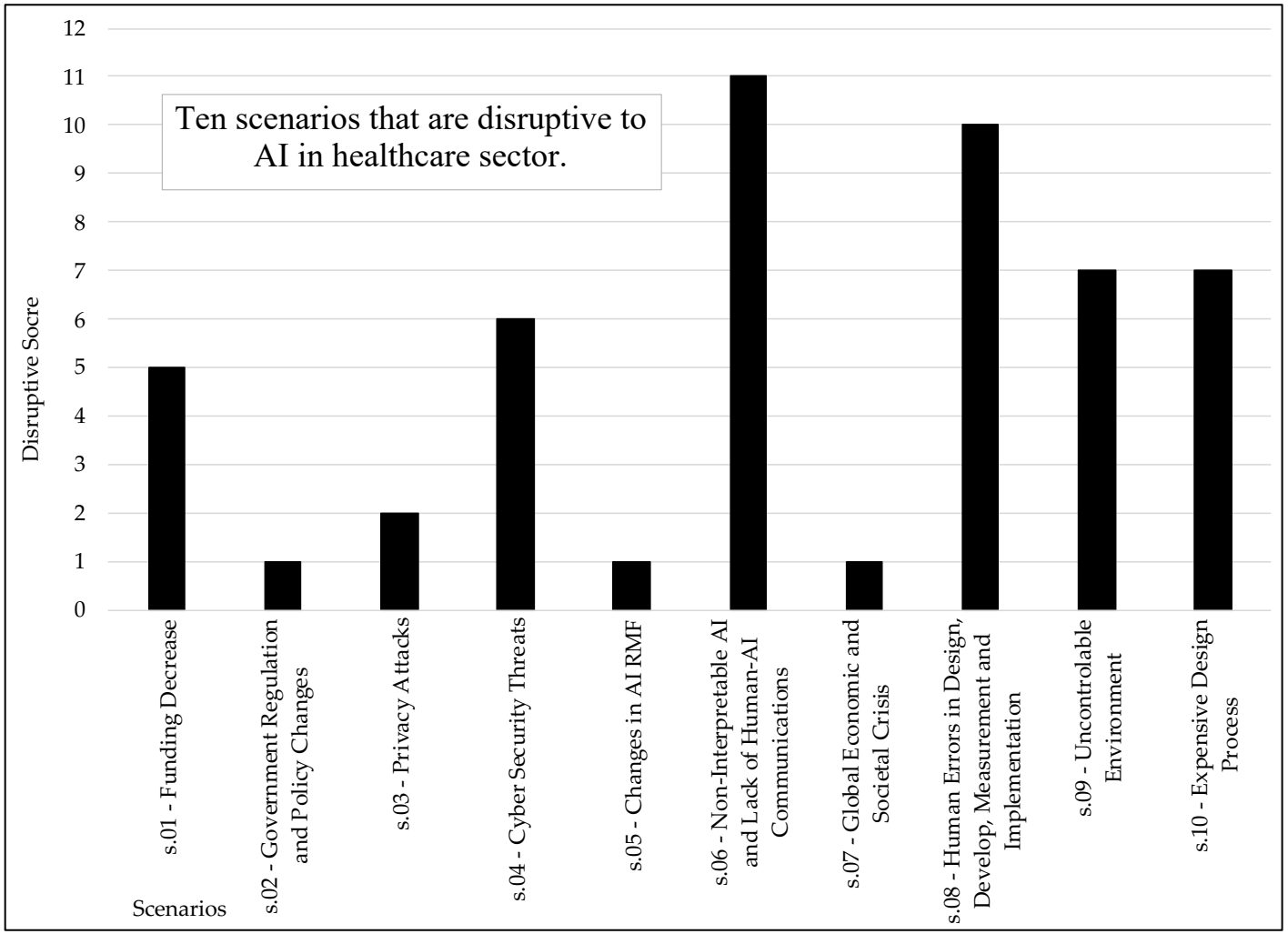

Figure 7. Disruptive score of scenarios is based on the sum of squared differences in the priority of initiatives relative to the baseline scenario for the *Purpose (Phi)* layer in enterprise risk management of AI in healthcare. These are scenarios where they caused low levels of trust in AI [6].

Figure 8 describes the variation in the prioritization of initiatives across scenarios. The blue dotted and red stripes highlight the possible upper and lower ranges of each initiative from the baseline rank of each initiative. In contrast, the black bar shows the baseline ranking of each initiative. The bar indicates the ranking range for each initiative, subject to disruptions by scenarios. More specifically, the red bar shows how far an initiative may fall in rank under various scenarios, and the blue bar shows how high an initiative may rise under various scenarios [6,9]. In other words, an initiative with a baseline ranking centered on a wide bar is sensitive to disruptions and does not consistently rank under disruptive scenarios. Suppose the baseline is positioned to the left of the bar with a long red segment. In that case, it indicates that the priority of the initiative has decreased because of one or more disruptions. When the initiative is represented by a long blue bar on the right side, it indicates that the initiative is likely to gain importance due to disruptions. Table 14 describes the highest ranked initiatives for the *Purpose (Pi)* layer [6].

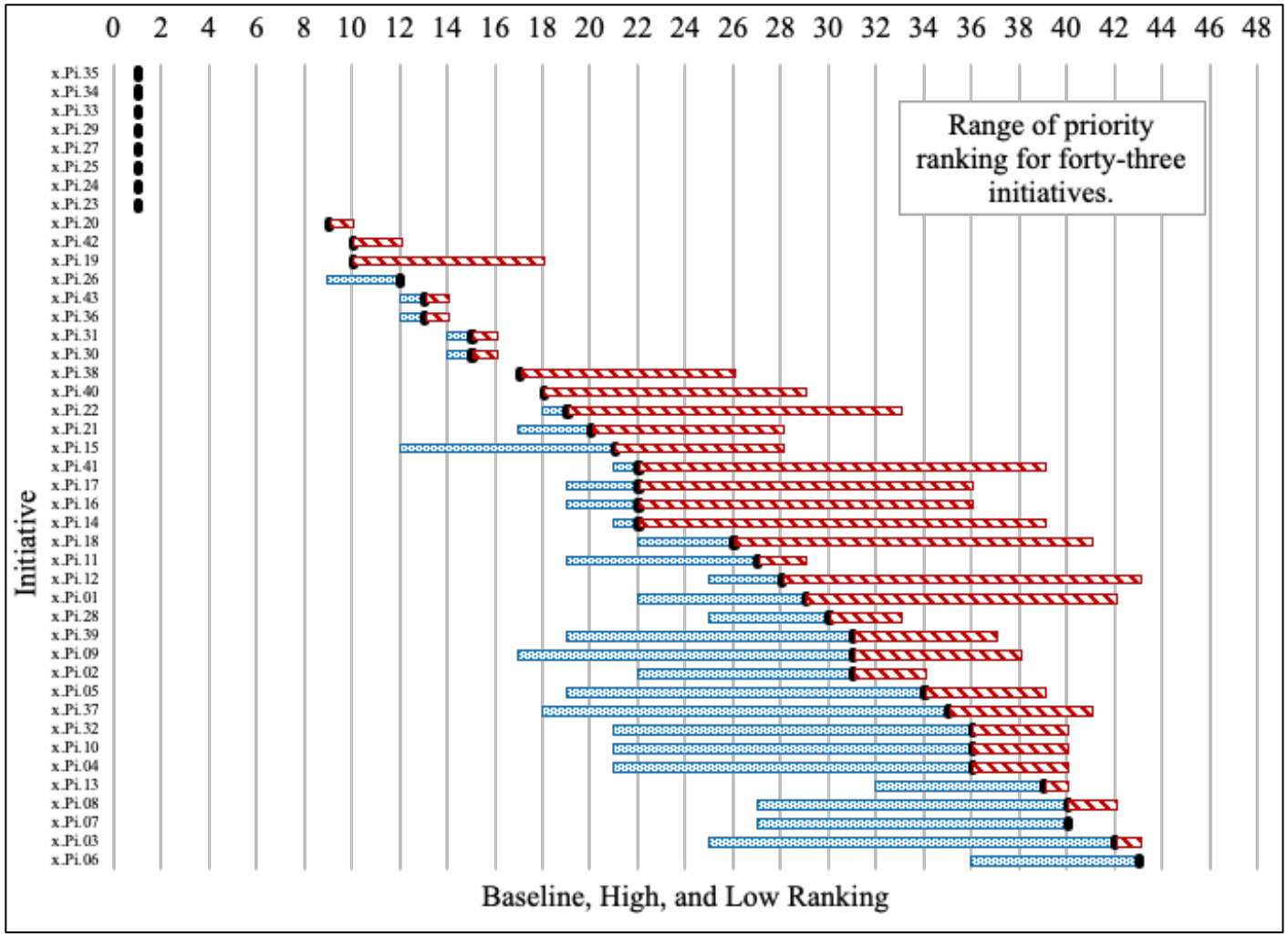

Figure 8. Distributions of initiatives influence rankings based on which emergent conditions that could arise more often or do not occur for the *Purpose (Phi)* layer in enterprise risk management of AI in healthcare; blue means promotion in ranking and red means demotion in ranking [6].

Table 14. The highest ranked initiatives of the *Purpose (Pi)* layer in enterprise risk management of AI in healthcare [6].

Index	Most Important Initiative
Purpose (Pi)	***x.Pi.35 - AI Providers Have to Also Inform Users on Why and How the Benefits of the Use of this System Overweigh its Risks (Level of Accuracy, Comparison with Other Technologies, or Practices Available on The Market)***
	x.Pi.34 - The Required Information About the Risks, Side Effects, and Limitations Shall Cover the Risk of Algorithmic Changes, AI Providers Predictions of that Changes, As Well As the Explainability Limitations When it is Applicable (Black-Box Models)
	x.Pi.33 - To Inform the Users on What Kind of Data Was Used to Train and Validate an AI System and How System Parameters Might Change Depending on the Input Data
	x.Pi.29 - Provide Specifications the User Requires to Use the Device Appropriately, e.g., If the Device has a Measuring Function, the Degree of Accuracy Claimed for it
	x.Pi.27 - Ensure the Safety and Quality of AI Medical Devices before and after They Are Placed on the Market
	x.Pi.25 - Clinicians to Have All the Necessary Information, Documents, and Explanations Provided Directly (Through Interaction and Cooperation) or Indirectly (Through the Process of AI Device Verification and Authorization) by AI Developers
	x.Pi.24 - Clinicians to be convinced that AI Device is Generally Useful, Safe, and Efficient
	x.Pi.23 - Clinicians to be Convinced that Specific AI System Outcome is Safe

4.3. Summary

This chapter has described a risk register for AI in healthcare, focusing on the *Purpose (Pi)* layer and employing a mathematical decision framework to gauge AI trustworthiness. It identifies forty-three initiatives, twenty-five emergent conditions, and ten scenarios impacting AI trustworthiness, emphasizing transparency, and accountability. Evaluations based on success criteria reveal varying initiative priorities across scenarios, underlining measures such as informing users about system benefits, ensuring the safety and quality of AI outcomes, and providing comprehensive AI risk information. Experts, actors, and managers will determine which resilience capabilities should be funded and applied to the system using a scenario-based preference framework.

The next step is applying the mathematical decision framework to the risk of AI in the *Structure (Sig)* or component layer.

Chapter 5 | Case 2 (*Structure (Sig)* Layer)

5.1. Introduction

This chapter describes a case study in the *Structure (Sig)* layer. This layer includes the physical framework of the system, which could resemble physical medical devices, particularly in this case, the design optimization of the surgical suturing anastomosis.

Every year, 310 million major surgeries are performed in the United States [71]. Vascular anastomosis is a common technique used in complex surgeries, with various handheld suturing techniques employed [72]. Figure 9 describes various handheld suturing techniques, such as end-to-end, side-to-end, and side-to-side. This chapter focuses on the end-to-end technique. End-to-end suturing is the most common technique [73]; however, it can lead to potential complications such as thrombosis, fistula formation, vessel occlusion,

hemorrhage, wound infection, aneurysm formation, nerve damage, and chronic pain [49–56]. The handheld technique is also time-consuming, subject to human error, and can cause vessel damage or blood flow turbulence [72]. To mitigate these risks, careful attention and execution are required.

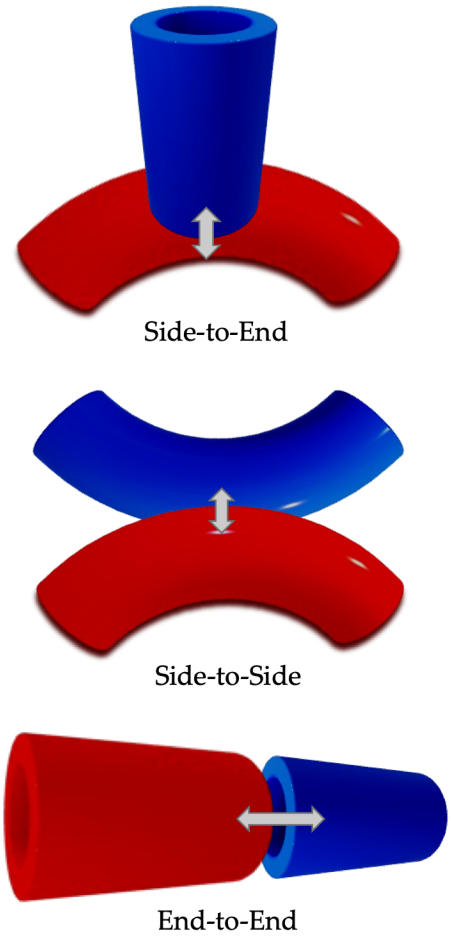

Figure 9. Vascular anastomosis types, adapted from [82].

Sutureless vessel anastomosis is considered another alternative in operations such as magnetic compression anastomosis (MCA); however, hardships in the installation of the device, potential hazards to performers, and more, limit its

popularity [72]. Having reliable devices to improve the prognosis and patient health outcomes of surgical procedures is crucial. The outcome of the surgeries could heavily depend on the devices used for the patients. Thus, Kang et al. proposed a 3D printing prototype, Vaso-Lock, that replaces handheld suturing by holding together free vascular ends [83]. Vaso-Lock is a bioinspired, additively manufactured device that aims to simplify vascular anastomoses during surgical procedures compared to the traditional method of blood vessel hand-sewing. Figure 10 describes the design concept of the device that resembles the structure of the rose plant in which prickles are arranged around a cylindrical configuration. Such arranged prickle structures, fabricated with biodegradable materials, allow the device to be firmly attached to the inner side of blood vessels and help connect two vessels.

The traditional design method will require an ensemble of 'cut-and-try' experiments. The experiment required identifying the device configurations, 3D modeling the device, and printing the device, the process of which may take several hours or even days.

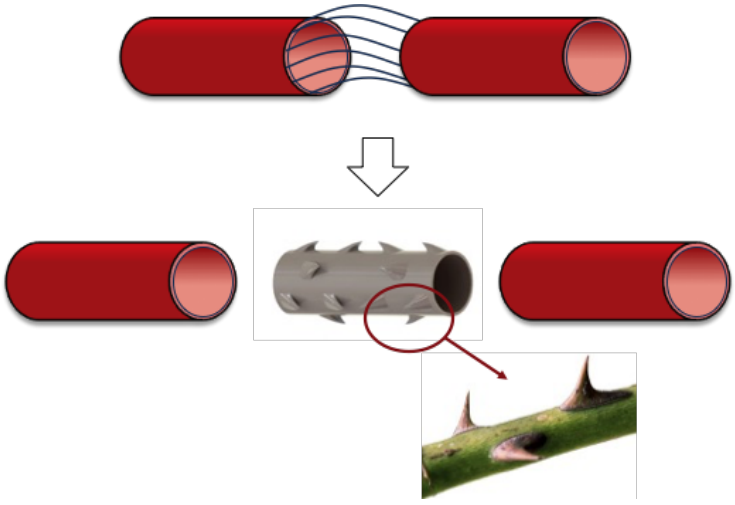

Figure 10. Bioinspired structure for blood vessel anastomosis.

A next step is to develop a scenario-based disruption of priorities for designing Vaso-Lock.

5.2. Scenario-Based Disruption of Priorities (Structure (Sig) Layer)

Scenario-based analysis is developed to identify the highest ranked initiatives and most and least disruptive scenarios in the risks of AI in device design (*Structure (Sig)* layer). This layer defines the physical structure of the healthcare system being modeled, which includes device and component design. The list of success criteria for Vaso-Lock is similar to the NIST AI Risk Management Framework seven *principles* that were listed in Figure 2.

The experts and actors for the *Structure (Sig)* layer are research scientists and device designers from the mechanical engineering department at Johns Hopkins University and the Western University of Health Sciences College of Dental Medicine. Bi-weekly meetings were held with experts and actors from Johns Hopkins University from July 2022. Seven interviews were carried out with dentists at the Western University of Health Sciences College of Dental Medicine. All interviews were conducted using an online platform.

Seven success criteria (Table 6), forty-seven initiatives (Table 15) [6,34], fifty emergent conditions (Table 16) [34], ten scenarios as the *Purpose (Pi)* layer (Table 17), baseline relevance (Table 18), criteria-initiative assessment (Table 19), criteria-scenario relevance (Table 20), and initiative-scenario ranking chart (Table 21) were identified and developed as follows for risk management of AI for Vaso-Lock design and development [9,34,83–85].

Table 15. Initiatives for the *Structure (Sig)* layer in enterprise risk management of AI in healthcare. Abridged from various sources that are identified in the narrative [6].

Index	Initiative
x.Sig.01	Identify At-Risk Components
x.Sig.02	Understanding ML Tools to Uncover Subtle Patterns in Data
x.Sig.03	Maintaining the Provenance of Training Data
x.Sig.04	Safety/Verifiability of Automated Analyses
x.Sig.05	Supporting Attribution of the AI System Decisions to Subsets of Training Data
x.Sig.06	Correctly Labeling the Data
x.Sig.07	Training Data to Follow Application Intellectual Property Rights Laws
x.Sig.08	Find the Maximum value of the Max Force of the Device
x.Sig.09	Maintain Organizational Practices Like Implement Risk Management to Reduce Harm Reduction and More Accountable Systems
x.Sig.10	Prioritization Policies and Resources Based on Assesses Risk Levels
x.Sig.11	Safety of Personally Identifiable Information
x.Sig.12	Appropriate Accountability Mechanism, Roles and Responsibilities, Culture, and Incentive Structures for Risk Management to be Effective
x.Sig.13	Identify the Right AI RMF in Different Context Based on Capabilities, Resources and Organization Size
x.Sig.14	Identify AI Actors with Diversity in Experience, Expertise, Background, Demographically and Disciplinary
x.Sig.15	Assist in Providing Context and Understanding Potential and Actual Impacts
x.Sig.16	Identify a Source of Formal or Quasi-Formal Norms and Guidance for AI Risk Management
x.Sig.17	Designate Boundaries for AI Operation (Technical, Societal, Legal, and Ethical)
x.Sig.18	Promote Discussion of the Tradeoffs Needed to Balance Societal Values and Priorities Related to Civil Liberties and Rights, Equity, the Environment and the Planet, and the Economy
x.Sig.19	Articulate and Document the System Concept and Objectives, Underlying Assumptions, and Context in Light of Legal and Regulatory Requirements and Ethical Considerations
x.Sig.20	Gather, Validate, and Clean Data and Document the Metadata and Characteristics of the Dataset, in Light of Objectives, Legal and Ethical Considerations
x.Sig.21	Pilot, Check Compatibility with Legacy Systems, Verify Regulatory Compliance, Manage Organizational Change, and Evaluate User Experience
x.Sig.22	Operate the AI System and Continuously Assess its Recommendations and Impacts
x.Sig.23	Balancing and Tradeoff Each of Trustworthy AI Systems Characteristics Based on the AI System Context of Use
x.Sig.24	Reduce the Number of Experiments to be Cost and Time Effective by Optimizing the Configurations
x.Sig.25	Ability of an Item to Perform as Required without Failure
x.Sig.26	Confirmation, Through the Provision of Objective Evidence that the Requirements for a Specific Intended Use Have been Fulfilled

x.Sig.27	Closeness of Results of Observations, Computations, or Estimates to the True Values or the Values Accepted as Being True
x.Sig.28	Human-AI Teaming
x.Sig.29	Demonstrate External Validity or Generalizable Beyond the Training Conditions
x.Sig.30	Ability of a System to Maintain its Level of Performance Under a Variety of Circumstances
x.Sig.31	Minimizing Potential Harms to People if it is Operating in an Unexpected Setting
x.Sig.32	Responsible Design, Development, and Deployment Practices
x.Sig.33	Clear Information to Deployers on Responsible Use of the System
x.Sig.34	Responsible Decision-Making by Deployers and End Users
x.Sig.35	Explanations and Documentation of Risks Based on Empirical Evidence of Incidents
x.Sig.36	Ability to Shut Down, Modify, or Have Human Intervention into Systems that Deviate from Intended or Expected Functionality
x.Sig.37	Resilient to Withstand Unexpected Adverse Events or Unexpected Changes in the Environment or Use
x.Sig.38	Maintain the Functions and Structure in the Face of Internal and External Change and Degrade Safely and Gracefully When this is Necessary
x.Sig.39	Managing Risks from Lack of Explainability by Describing How AI Systems Functions Considering Users' Role, Knowledge, and Skill Level
x.Sig.40	Communicating a Description of Why an AI System Made a Particular Prediction or Recommendation
x.Sig.41	Securing Individual Privacy, Anonymity, and Confidentiality
x.Sig.42	De-Identification and Aggregation for Certain Model Outputs
x.Sig.43	Strengthened Engagement with Interested Parties and Relevant AI Actors
x.Sig.44	AI Systems May Require More Frequent Maintenance and Triggers for Conducting Corrective Maintenance Due to Data, Model, or Concept Drift
x.Sig.45	Human Roles and Responsibilities in Decision Making and Overseeing AI Systems Need to be Clearly Defined and Differentiated
x.Sig.46	Explain and Identify Most Important Features Using AI Models
x.Sig.47	Incorporates Processes to Assess Potential Impacts
x.Sig.i	Others

Table 16. Emergent conditions are used to create sets of scenarios for the *Structure (Sig)* layer in enterprise risk management of AI in healthcare. Abridged from various sources that are identified in the narrative [6].

Index	Emergent Condition
e.Sig.01	Systematic Biases in Clinical Data Collection
e.Sig.02	Improperly Labeling the Data in Surgery-Specific Patient Registries
e.Sig.03	Misidentification of Variables Used in Surgery-Specific Patient Registries
e.Sig.04	Test Different Types of Transparency Tools in Cooperation with AI Deployers
e.Sig.05	AI to be Susceptible to Unrealistic Expectations from Media
e.Sig.06	Limitation in Types and Accuracy of Available Data
e.Sig.07	Expensive Data Collection
e.Sig.08	Time Consuming Data Collection
e.Sig.09	Policy and Regulation Changes
e.Sig.10	Difficult and Complex AI Algorithms Interpretability
e.Sig.11	Lack of AI Determination of Casual Relationships in Data at Clinical Implementation Level
e.Sig.12	Inability of AI in Providing an Automated Clinical Interpretation of its Analysis
e.Sig.13	Non-Intuitive Hidden Layers in DL
e.Sig.14	Abuse or Misuse of the Model or Data
e.Sig.15	Challenges with Training Data to be Subject to Copyright
e.Sig.16	Complicate Risk Measurement by Third Party Software, Hardware, and Data
e.Sig.17	Hard to Track and Measuring Emergent Risks by Organizations
e.Sig.18	Lack of Consensus on Robust and Verifiable Measurement Methods for AI Trustworthiness
e.Sig.19	Misidentification of Different Risk Perspective in Early or Late Stages of AI Lifecycle
e.Sig.20	Difference Between Controlled Environment vs. Uncontrollable and Real-World Settings
e.Sig.21	Inscrutable Nature of AI Systems in Risk Measurements
e.Sig.22	Hard to Find Human Baseline for AI Systems Intended to Replace Human Activity
e.Sig.23	Risk Tolerance Influence by Legal or Regulatory Requirements Changes
e.Sig.24	Unrealistic Expectations About Risk to Misallocate Resources
e.Sig.25	Residual Risk or Risk Remaining after Risk Treatment Directly Impacts End Users
e.Sig.26	Privacy Concerns Related to the Use of Underlying Data to Train AI Systems
e.Sig.27	Energy and Environmental Implications Associated with Resource-Heavy Computing Demands
e.Sig.28	Security Concerns Related to the Confidentiality, Integrity, and Availability of the System and its Training and Output Data
e.Sig.29	General Security of the Underlying Software and Hardware for AI Systems
e.Sig.30	One-Size-Fits-All Requirements AI Model Challenges
e.Sig.31	Neglecting the Trustworthy AI Characteristics
e.Sig.32	Difficult Decisions in Tradeoff and Balancing Trustworthy AI Characteristics by Organizations

e.Sig.33	Subject Matter Experts and Actors Can Assist in the Evaluation of TEVV Findings and Work with Product and Deployment Teams to Align TEVV Parameters to Requirements and Deployment Conditions
e.Sig.34	Different Perception of the Trustworthy AI Characteristics Between AI Designer than the Deployer
e.Sig.35	Potential Risk of Serious Injury or Death Call
e.Sig.36	Unexpected Changes in the Environment or Use
e.Sig.37	Data Poisoning
e.Sig.38	Negative Risk Stem from a Lack of Ability to Make Sense of, or Contextualize, System Output Appropriately
e.Sig.39	AI Allowing Inference to Identify Individuals or Previously Private Information About Individuals
e.Sig.40	Privacy Intrusions
e.Sig.41	Data Sparsity
e.Sig.42	Fairness Perceptions Difference Among Cultures and Applications
e.Sig.43	Computational and Statistical Biases Stem from Systematic Errors Due to Non-Representative Samples
e.Sig.44	Human-Cognitive Biases Relates to How the Experts and Actors Perceives AI System Information to Make a Decision
e.Sig.45	Lack of Access to the Ground Truth in the Dataset
e.Sig.46	Intentional or Unintentional Changes During Training
e.Sig.47	Increased Opacity and Concerns About Reproducibility
e.Sig.48	Computational Costs for Developing AI Systems and their Impact on the Environment and Planet
e.Sig.49	Inability to Predict or Detect the Side Effects of AI-Based Systems Beyond Statistical Measures
e.Sig.50	Presenting AI System Information to Humans is Complex
e.Sig.i	Others

Table 17. Emergent conditions grouping for the *Structure (Sig)* layer in enterprise risk management of AI in healthcare describes which emergent conditions fit in each scenario. Abridged from various sources that are identified in the narrative [6].

	s.01 - Funding Decrease	s.02 - Government Regulation and Policy Changes	s.03 - Privacy Attacks	s.04 - Cyber Security Threats	s.05 - Changes in AI RMF	s.06 - Non-Interpretable AI and Lack of Human-AI	s.07 - Global Economic and Societal Crisis	s.08 - Human Errors in Design, Develop, Measurement and	s.09 - Uncontrollable Environment	s.10 - Expensive Design Process
e.Sig.01		✓								
e.Sig.02								✓		
e.Sig.03								✓		
e.Sig.04						✓				
e.Sig.05						✓				
e.Sig.06										✓
e.Sig.07	✓									✓
e.Sig.08										✓
e.Sig.09		✓			✓					
e.Sig.10						✓				
e.Sig.11						✓				
e.Sig.12						✓				
e.Sig.13						✓				
e.Sig.14						✓		✓		
e.Sig.15		✓	✓			✓				
e.Sig.16						✓				✓
e.Sig.17						✓			✓	
e.Sig.18		✓			✓	✓		✓	✓	✓
e.Sig.19					✓	✓		✓		
e.Sig.20									✓	
e.Sig.21						✓				
e.Sig.22						✓				
e.Sig.23		✓								
e.Sig.24						✓		✓		
e.Sig.25						✓		✓	✓	
e.Sig.26			✓							
e.Sig.27		✓					✓			✓
e.Sig.28			✓	✓						
e.Sig.29				✓						
e.Sig.30						✓		✓	✓	✓

e.Sig.31						✓		✓		
e.Sig.32	✓				✓	✓				✓
e.Sig.33						✓				
e.Sig.34						✓				
e.Sig.35								✓	✓	
e.Sig.36									✓	
e.Sig.37			✓	✓				✓		
e.Sig.38						✓				
e.Sig.39			✓							
e.Sig.40			✓							
e.Sig.41								✓	✓	✓
e.Sig.42		✓			✓	✓				
e.Sig.43	✓					✓		✓	✓	✓
e.Sig.44	✓					✓				
e.Sig.45	✓							✓		✓
e.Sig.46					✓	✓		✓		
e.Sig.47						✓		✓		
e.Sig.48	✓						✓			✓
e.Sig.49						✓	✓		✓	
e.Sig.50						✓				

Table 18. Baseline relevance for the *Structure (Sig)* layer in enterprise risk management of AI in healthcare [6].

The criterion c.xx has	s.00 - Baseline	relevance among the other criteria
c.01 - *Safe* has	*medium*	relevance
c.02 - *Secure & Resilient* has	*medium*	relevance
c.03 - *Explainable & Interpretable* has	*high*	relevance
c.04 - *Privacy Enhanced* has	*low*	relevance
c.05 - *Fair - With Harmful Bias Managed* has	*low*	relevance
c.06 - *Accountable & Transparent* has	*high*	relevance
c.07 - *Valid & Reliable* has	*high*	relevance

Table 19. The criteria-initiative assessment describes how well each initiative addresses the success criteria for the *Structure (Sig)* layer in enterprise risk management of AI in healthcare. *Strongly agree* is represented by a filled circle (●), *agree* is represented by a half-filled circle (◐), *somewhat agree* is represented by an unfilled circle (○), and *neutral* is represented by a dash (−) [6].

	c.01	*c.02*	*c.03*	*c.04*	*c.05*	*c.06*	*c.07*
x.Sig.01	●	◐	−	−	−	−	○
x.Sig.02	−	−	○	−	−	−	●
x.Sig.03	−	−	○	−	−	○	●
x.Sig.04	−	−	●	−	−	◐	●
x.Sig.05	−	−	●	−	−	◐	−
x.Sig.06	○	−	−	−	●	●	●
x.Sig.07	○	○	◐	−	−	○	◐
x.Sig.08	●	●	◐	−	−	●	●
x.Sig.09	○	◐	●	−	○	●	○
x.Sig.10	−	−	◐	○	○	◐	◐
x.Sig.11	−	−	−	●	−	◐	−
x.Sig.12	○	○	◐	−	−	○	◐
x.Sig.13	○	○	○	○	○	○	○
x.Sig.14	−	−	−	−	◐	−	−
x.Sig.15	−	−	●	−	−	−	−
x.Sig.16	○	○	○	−	−	○	○
x.Sig.17	○	○	◐	−	−	○	◐
x.Sig.18	−	−	○	−	○	−	−
x.Sig.19	●	○	●	−	−	◐	◐
x.Sig.20	●	○	●	−	−	●	◐
x.Sig.21	○	−	◐	−	−	−	−
x.Sig.22	●	●	●	−	−	●	●
x.Sig.23	○	○	●	−	−	●	○
x.Sig.24	●	●	●	−	○	●	●
x.Sig.25	●	●	●	−	−	●	●
x.Sig.26	●	●	●	−	−	●	●
x.Sig.27	−	−	◐	−	−	●	●
x.Sig.28	◐	◐	●	−	−	◐	◐
x.Sig.29	◐	◐	●	−	−	◐	◐
x.Sig.30	●	●	●	−	−	●	●
x.Sig.31	●	●	●	−	−	●	●
x.Sig.32	●	●	●	−	−	●	●
x.Sig.33	●	●	●	−	−	●	●
x.Sig.34	◐	○	●	−	−	◐	◐
x.Sig.35	●	●	●	−	−	●	◐
x.Sig.36	●	●	●	−	−	●	−
x.Sig.37	●	●	○	−	−	●	−

x.Sig.38	◐	◐	○	−	−	○	◐
x.Sig.39	●	●	●	−	−	●	●
x.Sig.40	●	●	●	−	◐	●	●
x.Sig.41	−	−	−	○	−	−	−
x.Sig.42	−	−	●	○	−	−	−
x.Sig.43	◐	◐	●	−	−	◐	◐
x.Sig.44	◐	●	●	−	●	●	●
x.Sig.45	◐	◐	●	−	−	◐	◐
x.Sig.46	◐	◐	●	−	−	◐	◐
x.Sig.47	◐	◐	●	−	−	●	◐

Table 20. The criteria-scenario relevance describes how well each scenario fits the success criterion for the *Structure (Sig)* layer in enterprise risk management of AI in healthcare. *Decrease Somewhat* = DS, *Decrease* = D, *Somewhat Increase* = SI, *Increase* = I [6].

	s.01	*s.02*	*s.03*	*s.04*	*s.05*	*s.06*	*s.07*	*s.08*	*s.09*	*s.10*
c.01	D	SI	-	-	SI	D	DS	D	D	D
c.02	D	SI	-	-	SI	D	DS	DS	DS	D
c.03	DS	SI	-	-	I	D	DS	D	D	D
c.04	-	SI	D	DS	-	-	-	-	-	-
c.05	DS	-	-	-	SI	-	DS	DS	-	-
c.06	D	SI	-	-	I	D	DS	D	D	D
c.07	D	SI	-	-	I	D	DS	D	D	D

Table 21. Initiative-scenario ranking chart. This table describes the ranking of each initiative under each scenario for the *Structure (Sig)* layer in enterprise risk management of AI in healthcare. The green filled cells show a higher ranking and the red and orange filled cells indicate a lower ranking.

	s.00 - Baseline	s.01 - Funding Decrease	s.02 - Government Regulation and Policy Changes	s.03 - Privacy Attacks	s.04 - Cyber Security Threats	s.05 - Changes in AI RMF	s.06 - Non-Interpretable AI and Lack of Human-AI Communications	s.07 - Global Economic and Societal Crisis	s.08 - Human Errors in Design, Develop, Measurement and Implementation	s.09 - Uncontrollable Environment	s.10 - Expensive Design Process
x.Sig.01	40	43	40	40	40	42	43	42	42	43	43
x.Sig.02	39	41	39	39	39	37	41	41	41	41	41
x.Sig.03	35	39	35	35	35	35	37	37	37	37	37
x.Sig.04	26	27	26	26	26	25	29	26	27	29	29
x.Sig.05	36	38	36	36	36	36	39	38	39	39	39
x.Sig.06	28	30	29	28	28	28	4	29	31	4	4
x.Sig.07	31	33	31	31	31	31	33	32	34	34	33
x.Sig.08	12	13	12	13	13	13	13	12	12	13	13
x.Sig.09	23	23	23	23	23	22	15	23	22	15	15
x.Sig.10	30	25	30	30	30	30	16	28	25	18	16
x.Sig.11	43	29	43	44	44	44	29	31	27	29	29
x.Sig.12	31	33	31	31	31	31	33	32	34	34	33
x.Sig.13	37	32	37	37	37	38	27	36	32	27	27
x.Sig.14	46	47	47	46	46	46	41	47	47	41	41
x.Sig.15	42	42	42	42	42	41	45	43	43	45	45
x.Sig.16	38	40	38	38	38	39	40	40	40	40	40
x.Sig.17	31	33	31	31	31	31	33	32	34	34	33
x.Sig.18	45	46	45	45	45	45	44	46	46	44	44
x.Sig.19	21	16	16	16	16	16	19	16	23	25	19
x.Sig.20	14	14	14	14	14	15	17	14	15	17	17
x.Sig.21	44	44	44	43	43	43	46	44	44	46	46
x.Sig.22	4	4	3	4	4	4	5	4	4	5	5
x.Sig.23	25	26	25	25	25	27	28	25	26	28	28
x.Sig.24	2	3	2	2	2	2	3	2	3	3	3
x.Sig.25	4	4	3	4	4	4	5	4	4	5	5
x.Sig.26	4	4	3	4	4	4	5	4	4	5	5

x.Sig.27	26	28	26	26	26	25	29	26	27	29	29
x.Sig.28	16	16	16	16	16	16	19	16	17	20	19
x.Sig.29	16	16	16	16	16	16	19	16	17	20	19
x.Sig.30	4	4	3	4	4	4	5	4	4	5	5
x.Sig.31	4	4	3	4	4	4	5	4	4	5	5
x.Sig.32	4	4	3	4	4	4	5	4	4	5	5
x.Sig.33	4	4	3	4	4	4	5	4	4	5	5
x.Sig.34	24	24	24	24	24	24	26	24	24	26	26
x.Sig.35	12	12	12	12	12	12	13	12	12	13	13
x.Sig.36	22	22	22	22	22	23	25	22	16	19	25
x.Sig.37	29	31	28	29	29	29	32	30	30	32	32
x.Sig.38	31	36	31	31	31	34	33	32	33	33	33
x.Sig.39	4	4	3	4	4	4	5	4	4	5	5
x.Sig.40	1	1	1	1	1	1	2	1	1	2	2
x.Sig.41	47	45	46	47	47	47	47	45	45	47	47
x.Sig.42	41	37	41	41	41	40	37	39	37	37	37
x.Sig.43	16	16	16	16	16	16	19	16	17	20	19
x.Sig.44	2	2	11	2	3	2	1	2	2	1	1
x.Sig.45	16	16	16	16	16	16	19	16	17	20	19
x.Sig.46	16	16	16	16	16	16	19	16	17	20	19
x.Sig.47	14	14	14	14	14	14	17	14	14	16	17

Figure 11 describes that *s.06 – Non-Interpretable AI and Lack of Human-AI Communications* is one of the most disruptive scenarios. This indicates the importance of explainable AI in medical device and implant design. Explainability can help answer the question of "how" the machine made a decision. Also, *s.09 – Uncontrollable Environment,* and *s.10 – Expensive Design Process* are other most disruptive scenarios [6].

Figure 12 describes the variation in the prioritization of initiatives across scenarios.

Table 22 describes the highest ranked initiatives.

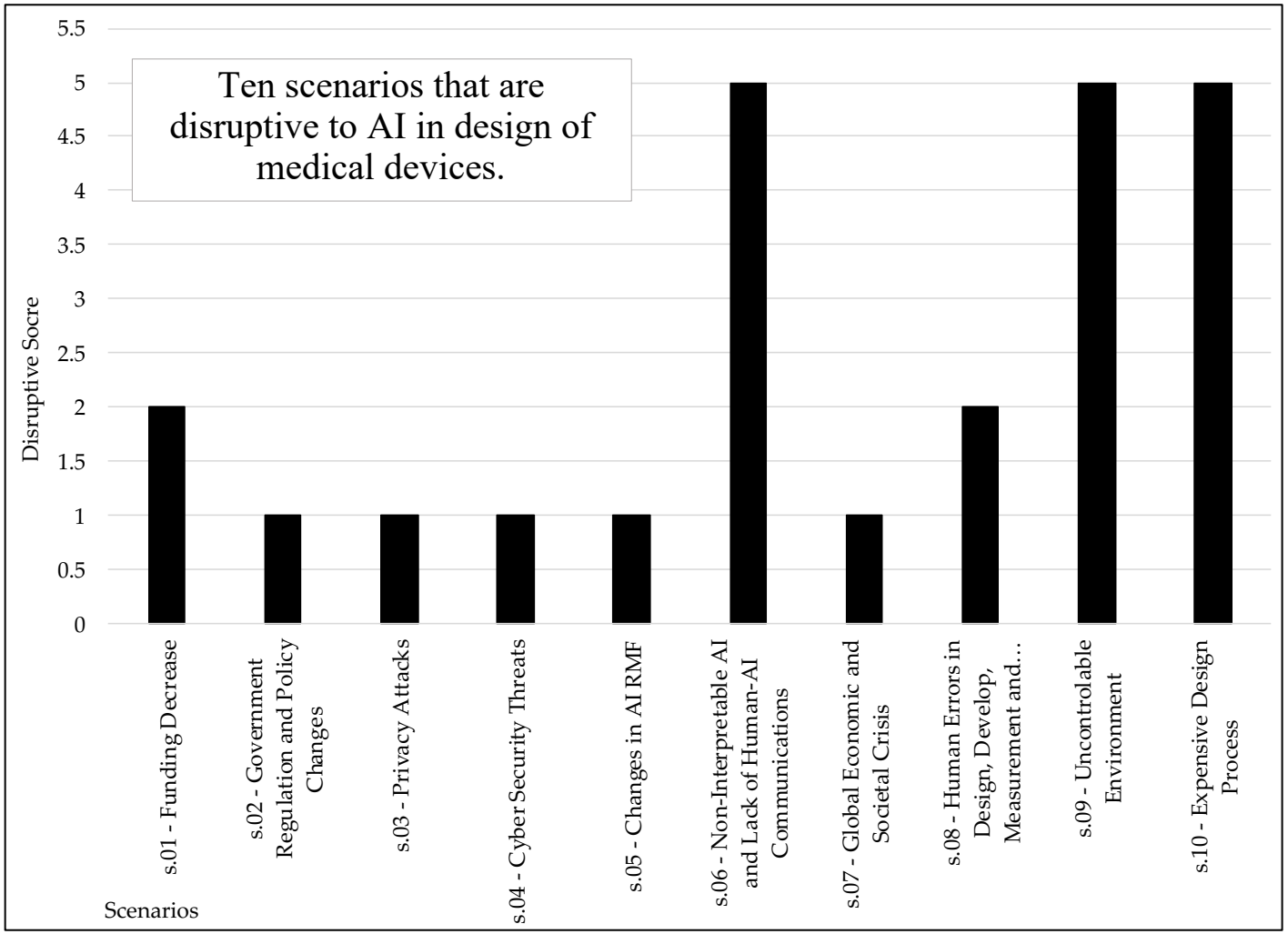

Figure 11. Disruptive score of scenarios is based on the sum of squared differences in the priority of initiatives, relative to the baseline scenario for the *Structure (Sig)* layer in enterprise risk management of AI in healthcare. These are scenarios where they caused low levels of trust in AI [6].

Table 22. The highest ranked initiatives of the *Structure (Sig)* layer in enterprise risk management of AI in healthcare.

Index	Most Important Initiative
Structure (Sig)	***x.Sig.40 - Communicating a Description of Why an AI System Made a Particular Prediction or Recommendation***
	x.Sig.44 - AI Systems May Require More Frequent Maintenance and Triggers for Conducting Corrective Maintenance Due to Data, Model, or Concept Drift
	x.Sig.24 - Reduce the Number of Experiments to be Cost and Time Effective by Optimizing the Configurations
	x.Sig.39 - Managing Risks from Lack of Explainability by Describing How AI Systems Functions Considering Users' Role, Knowledge, and Skill Level
	x.Sig.33 - Clear Information to Deployers on Responsible Use of the System
	x.Sig.32 - Responsible Design, Development, and Deployment Practices
	x.Sig.31 - Minimizing Potential Harms to People if it is Operating in an Unexpected Setting
	x.Sig.30 - Ability of a System to Maintain its Level of Performance Under a Variety of Circumstances
	x.Sig.26 - Confirmation, Through the Provision of Objective Evidence that the Requirements for a Specific Intended Use Have Been Fulfilled
	x.Sig.25 - Ability of an Item to Perform as Required without Failure
	x.Sig.22 - Operate the AI System and Continuously Assessing its Recommendations and Impacts

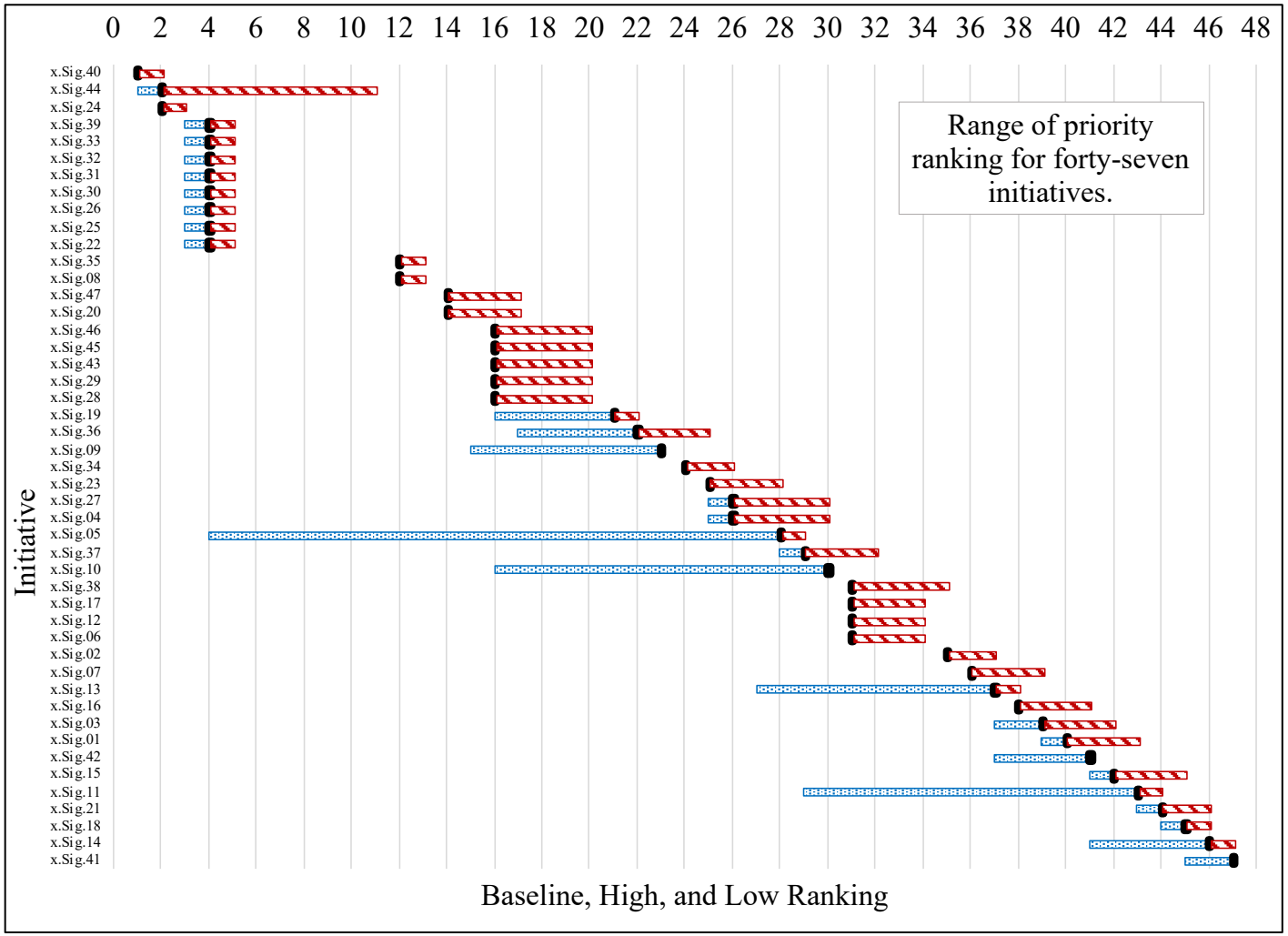

Figure 12. Distributions of initiatives influence rankings based on which emergent conditions that could arise more often or do not occur for the *Structure (Sig)* layer in enterprise risk management of AI in healthcare; blue means promotion in ranking and red means demotion in ranking.

Figure 12 describes that one of the highest ranked initiatives is *x.Sig.24 - Reduce the Number of Experiments to be Cost and Time Effective by Optimizing the Configurations*, which is one of the main objectives of the experts and actors in this case. This initiative requires experiments to determine the optimal geometry design of the device without the cut-and-try method. Also, designing the device has challenges, such as the experiments to test the vessel holding capability of Vaso-Lock are expensive and laborious, leading to a lack of training data. In addition, the operating condition of Vaso-Lock is highly random in a stochastic environment, as human vessel dimensions are not always similar and deterministic. This finding aligns with scenario *s.09 – Uncontrollable Environment*, which is one of the most disruptive scenarios to the system. Stochasticity can arise from various sources, including natural variations in blood vessel size and shape, unpredictable blood pressure and flow changes, and more. Uncontrollable features are identified as the outer diameter, inner diameter, and width of the vessel wall dimensions.

5.3. AI-Assisted Framework in Optimizing the Geometry of Vaso-Lock[8]

This section describes statistical analysis of Vaso-Lock and suggests new sample configurations with the maximum force a vessel could hold, and XAI analysis [39] to improve communication between AI systems and non-technical

[8] This case is a collaboration between the University of Virginia School of Data Science and Johns Hopkins University. The PIs are Prof. Stephen Baek and Prof. Sung H. Kang. The collaborators are Dr. Phong Nguyen, Negin Moghadasi, and Kate Beatrice Concannon.

users, enhance understanding of AI outputs, mitigate distrust in the AI outputs, and facilitate data evaluation.

5.3.1. Data Exploratory

This section explores the data provided by the experiments that were conducted. A total of 187 observations were generated from the experimental design. Observations were drawn from each cut-and-try experiment. Multiple configurations were measured for each experiment, which included the outer diameter of the vessel (OD), the inner diameter of the vessel (ID), the width of the vessel (WT), the outer diameter of the device (DOD), and the length of the vessel (Length). The uncertain environment of ID could be formulated as ID = $f(\mathrm{OD}, \mathrm{WT})$. All configurations were measured in millimeters (mm). Ultimately, a value for maximum force for each experiment was measured. The maximum force value is measured in Newtons (N) and defined as the maximum force with which the device could hold two vessels together without tear, breakage or leakage.

There are some constraints associated with the optimization process, such as:

1. The operating condition is random and has a wide range of distribution.
2. The wide range of distribution of operating conditions causes a variation in the device maximum force reaction. The variation range is vast, which results in the device not being robust.
3. The experiment is costly and time-consuming, so the team intends to find the optimal device configuration with the minimum number of experiments.

The vessel configurations are defined in various ranges, resulting in balanced bins. The range defined for the number of rows is {1, 2, 3, 4}, the number of prickles is {2, 4, 6, 8, 10, 12, 14}, and the number of DODs is {3.5, 4.0, 4.5, 5.0, 5.5, 6.0}. Considering various defined configurations, there are 168 combinations of configurations. Figure 13 describes the distributions for vessel OD, WT, and length.

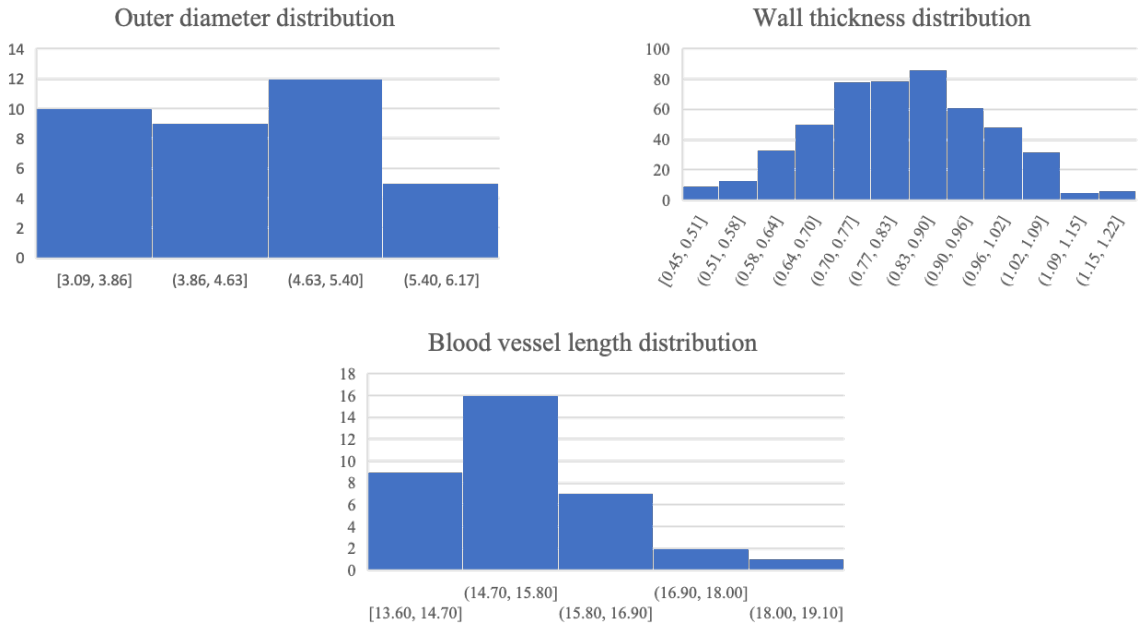

Figure 13. Blood vessel dimension-wide distribution that shows how broad the environment distribution is.

Figure 14 describes the reaction force considering various device configurations and vessel diameters. There are three layers per sub-figure. The top layer represents the upper bound, the bottom layer shows the lower bound, and the middle layer shows mean values.

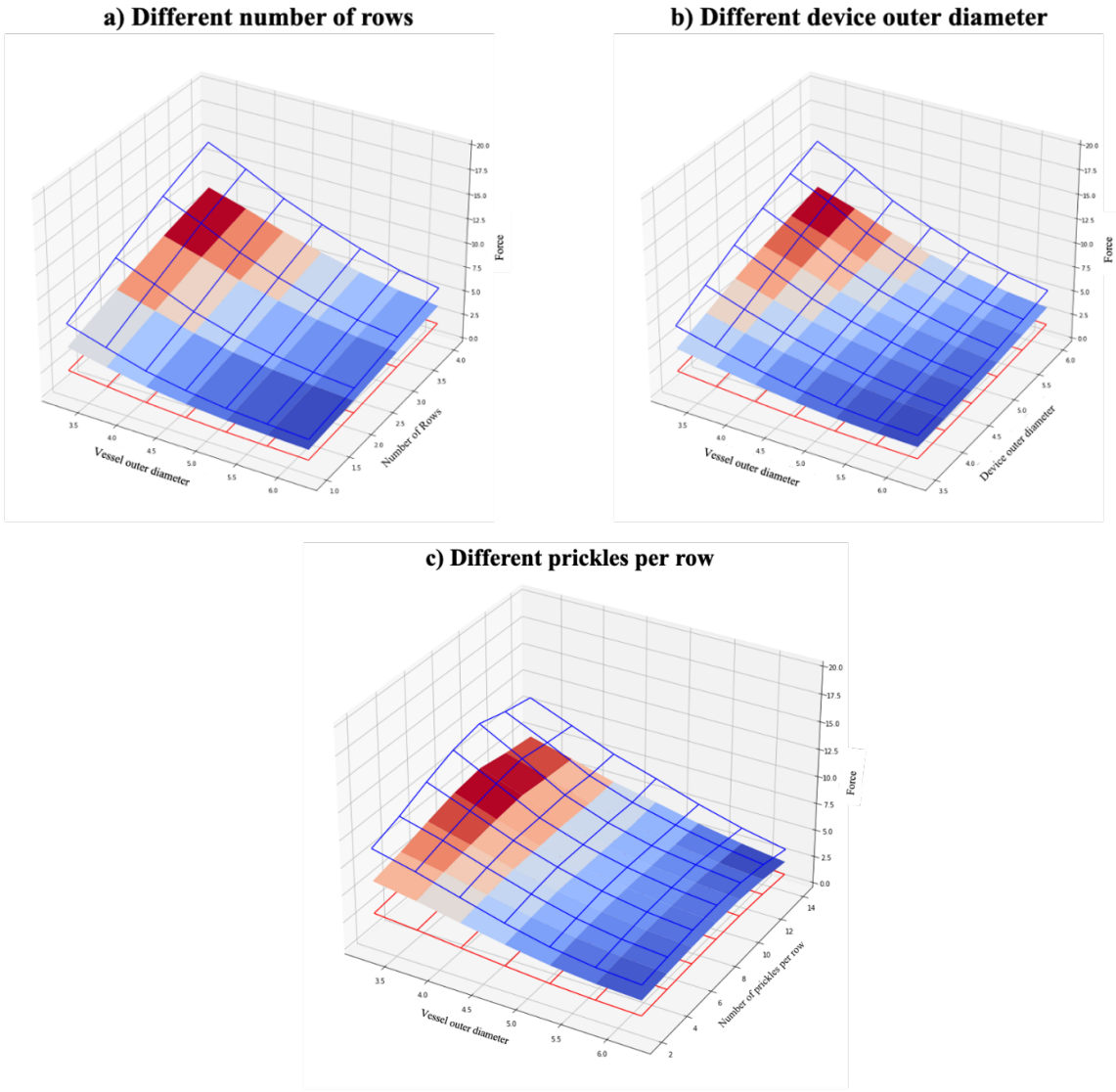

Figure 14. Reaction force with different device configurations and vessel diameters. Figure a) describes the reaction force considering various outer diameters and various numbers of rows. Figure b) describes the reaction force considering various outer diameters and devices.

The next step is assessing the correlations between the features of the data. Figure 15 depicts the correlation among the features. The data demonstrates a strong positive connection between OD and ID, with a correlation coefficient of 0.91. This indicates that when ID grows, there is a linear increase in OD.

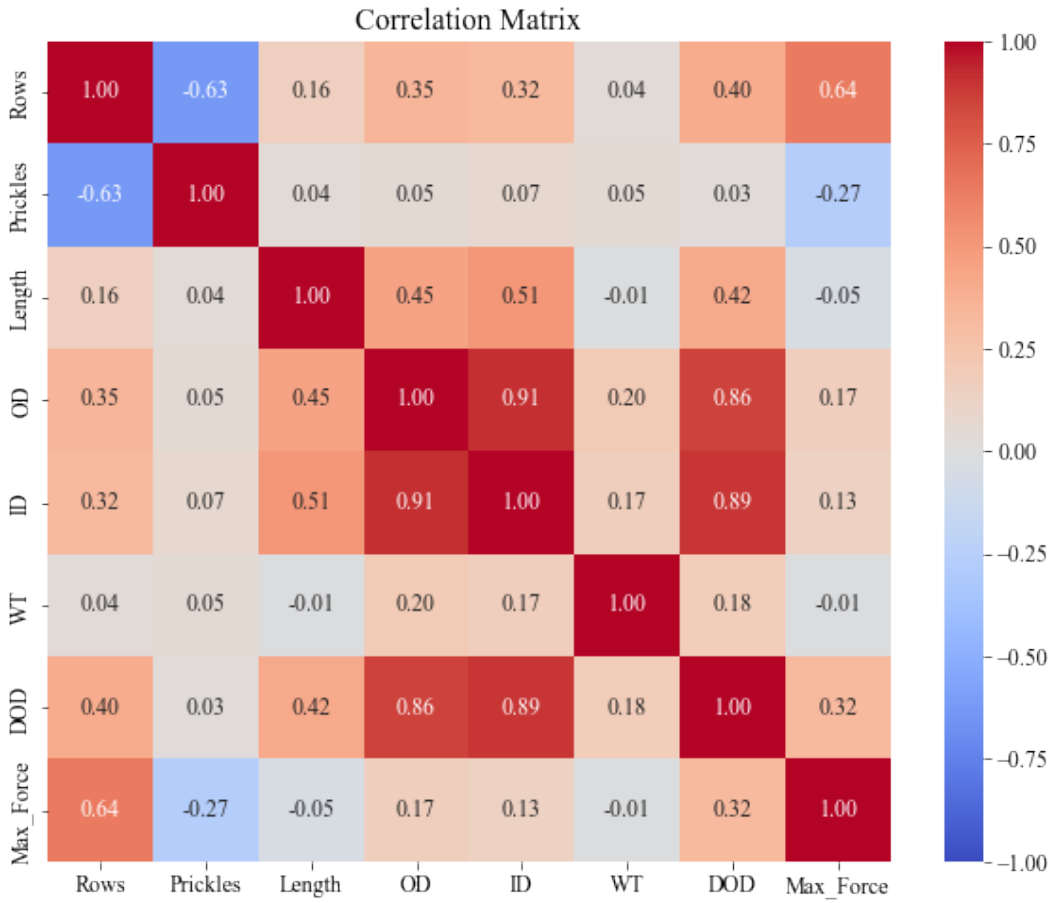

Figure 15. The correlation matrix of the variables in the dataset shows the relationships between various variables.

5.3.1.1. FEATURE IMPORTANCE

A next step is determining the feature importance by splitting the dataset, which consisted of 187 observations, into a test set comprising 25% of the data and a train set comprising 75% of the data. Subsequently, Figure 16 describes that the number of rows is the most important feature among others using the random forest classifier to train the train set. The results show that the number of rows is an important factor in designing the device configuration.

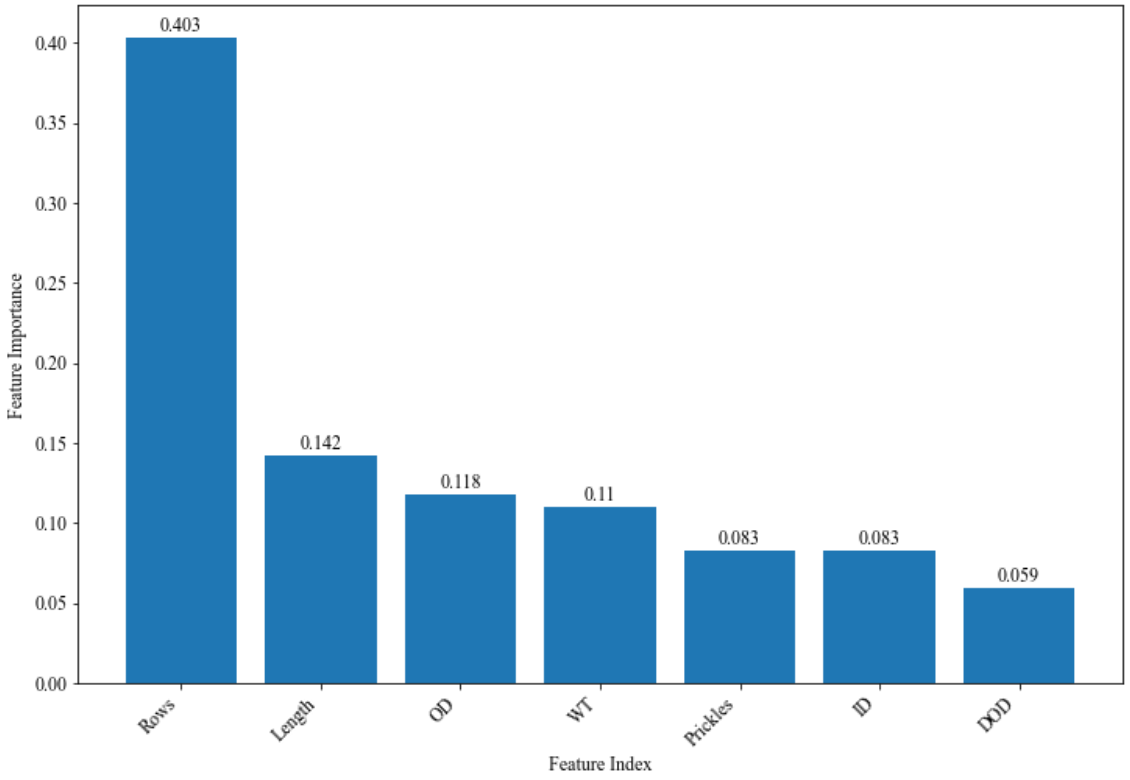

Figure 16. Feature Importance.

5.3.1.2. KERNEL DENSITY ESTIMATION (KDE)

Most approaches in the literature have focused on modeling the probability distribution of the environment as a simple Gaussian distribution. This occurred because the Gaussian distribution has been extended to be applicable to numerous random variables. The analytical form provides a computational speed advantage to the optimization process. However, in some cases, environmental variables do not follow the Gaussian distribution, and therefore, attempting to make this assumption is incorrect. In this case, knowledge about

the distribution of the environmental variables was acquired by directly analyzing the data using the kernel density estimation (KDE) method [86–89].

The KDE is a statistical method employed to estimate the probability density function of a random variable using a collection of observed data points [88,89]. Equation 6 is the KDE formulation:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad (6)$$

The formulas utilize the variable h to denote the bandwidth, while n represents the number of data points. These rules offer bandwidth approximations derived from the characteristics of the dataset. As the value of h increases, the probability density function (PDF) becomes wider, resulting in a smoother curve. A decrease in the value of h corresponds to a narrower curve of the probability density function (PDF) and an increase in the level of detail in the data distribution. A broader curve indicates a wider distribution, whereas a narrower curve implies a more focused distribution.

In this section, the KDE technique was employed to assess the probability distribution of the features in the data, utilizing the Gaussian distribution. This evaluation is illustrated in Figure 17.

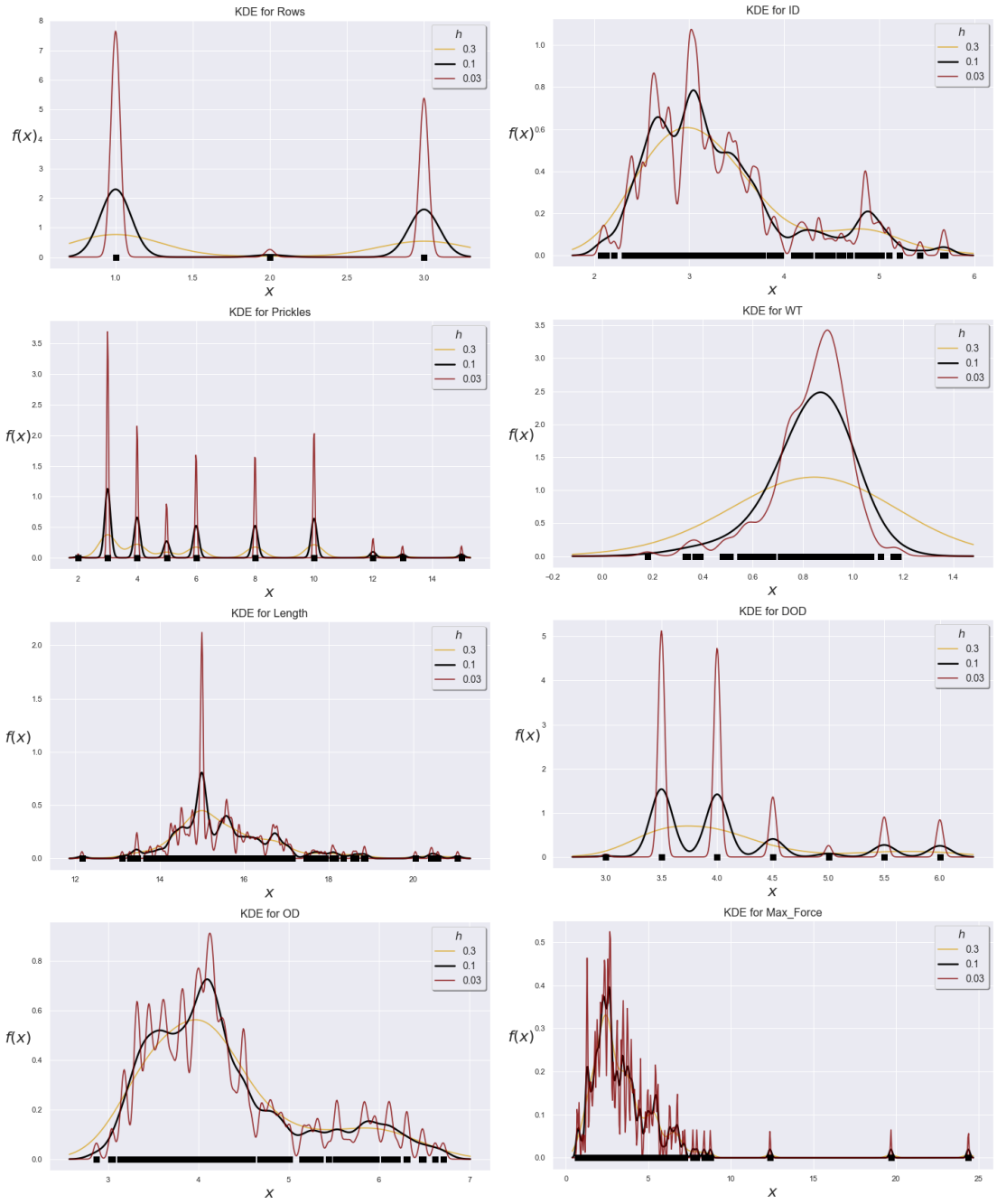

Figure 17. Gaussian KDE is used for features such as rows, prickles, length, OD, ID, WT, DOD, and Max_Force. A broader curve indicates a wider distribution, whereas a narrower curve implies a more focused distribution.

However, there exist alternative distributions, such as the cosine distribution [90], which may yield more precise outcomes. Prior to proceeding, it is essential to determine the optimal bandwidth. Practically, there are multiple techniques available for choosing the bandwidth, such as rule-of-thumb methods like Scott's rule or Silverman's rule, cross-validation techniques, and optimization algorithms. In this section, the optimal value of h was determined using Silverman's rule. The KDE-Silverman Rule of Thumb (ROT) [90] is not the most optimal method for selecting bandwidth. However, it is commonly used as a quick and reasonably accurate estimator or as an initial estimator in multistage bandwidth selection processes. Equation 7 shows the formula for Silverman's rule:

$$h = \left(\frac{4 * standard\ deviation^5}{3n}\right)^{\frac{1}{5}} \tag{7}$$

The cosine kernel is applied using Silverman's bandwidth [89,91] of 0.26921718387928867. To achieve the desired level of smoothing, it may be necessary to modify the bandwidth based on the complexity of the data. Figure 18 describes the distributions of each feature.

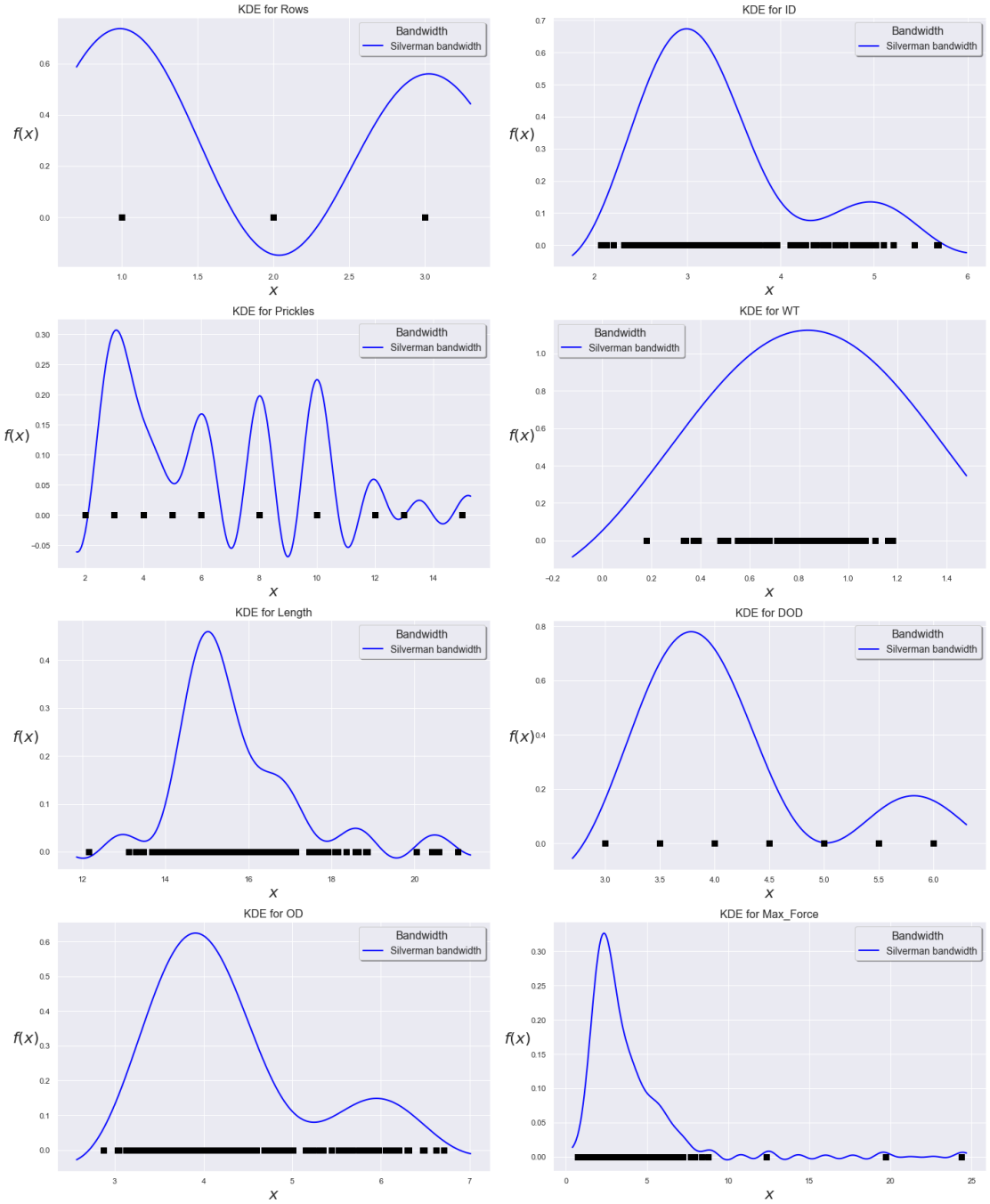

Figure 18. Cosine KDE is used for features such as rows, prickles, length, OD, ID, WT, DOD, and Max_Force. A broader curve indicates a wider distribution, whereas a narrower curve implies a more focused distribution.

The KDE shape of the curve is an indication of the underlying data distribution. Peaks in the curve signify areas with a high density, whereas valleys indicate areas with a low density. The modes in the KDE plot are associated with the peaks observed in the data distribution. Every mode corresponds to a unique group or cluster within the dataset.

In the following sections, the ID KDE plot was utilized as clustering cut-off points. Since Gaussian KDE gives more information on the peaks of the features than cosine KDE, it will be utilized for x_e to be drawn from this learning distribution.

5.3.2. Clustering

This section describes clustering the observations that are spread out across the average number of rows, the number of prickles per row, and the outer diameter of the device.

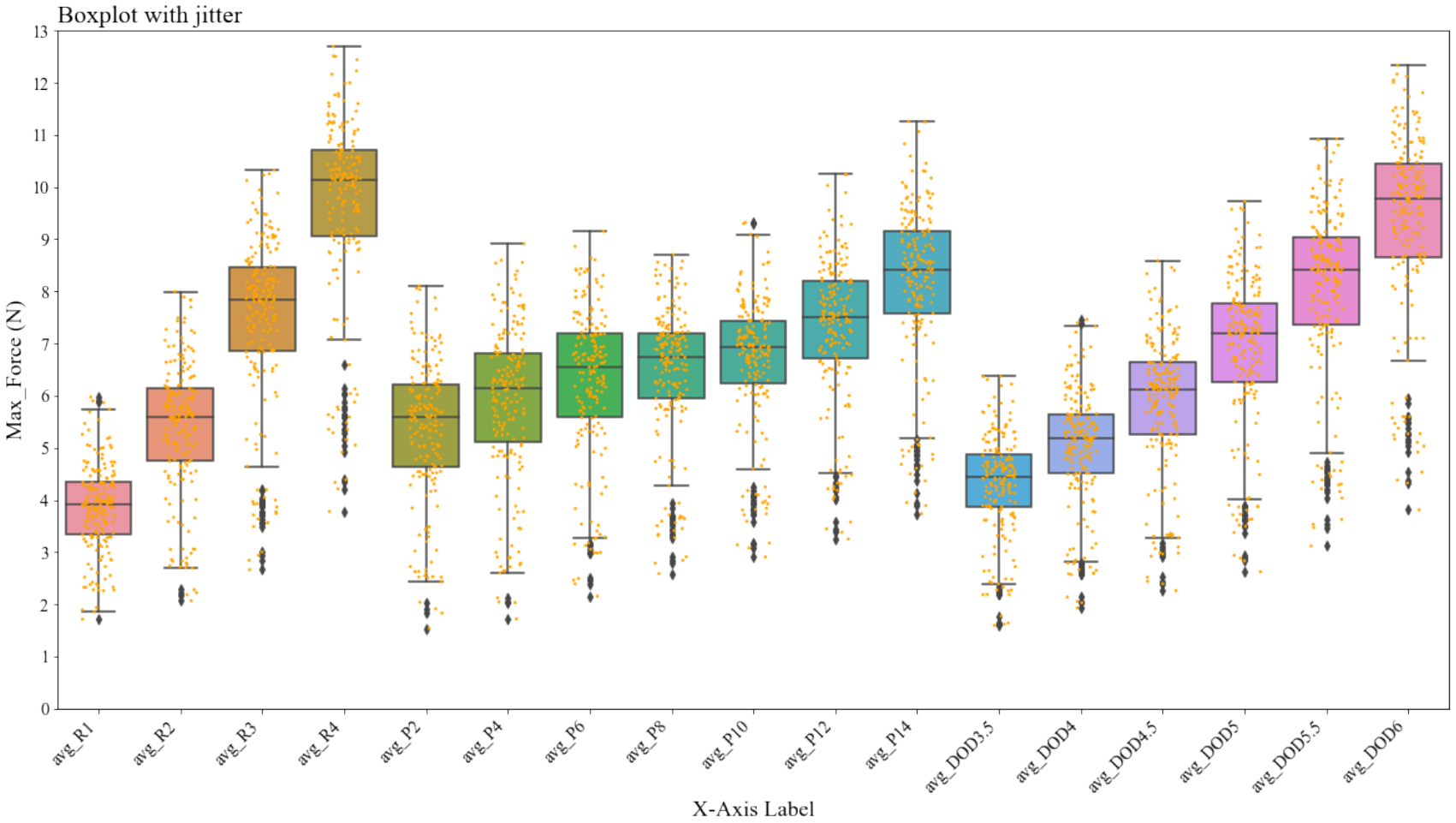

Figure 19. Clustering observations are based on the average number of rows, the average number of prickles per row, and the average value of the outer diameter of the device as one cluster. The figure describes that the distribution for one cluster is very wide.

Figure 19 describes that the distribution for one cluster is very wide. To make the distribution of reaction force less variable, the observations were put into three separate clusters using the k-means method. Figure 20 indicates that the optimal number of clusters is three. The optimal number of clusters was determined by utilizing the Silhouette method and the generation of elbow plots. Figure 21 describes maximum force distributions based on the average number of rows, prickles, and DOD for all 187 observations. Reducing the distributions assists in finding the best optimal configurations for each cluster, which results in a more accurate and robust design.

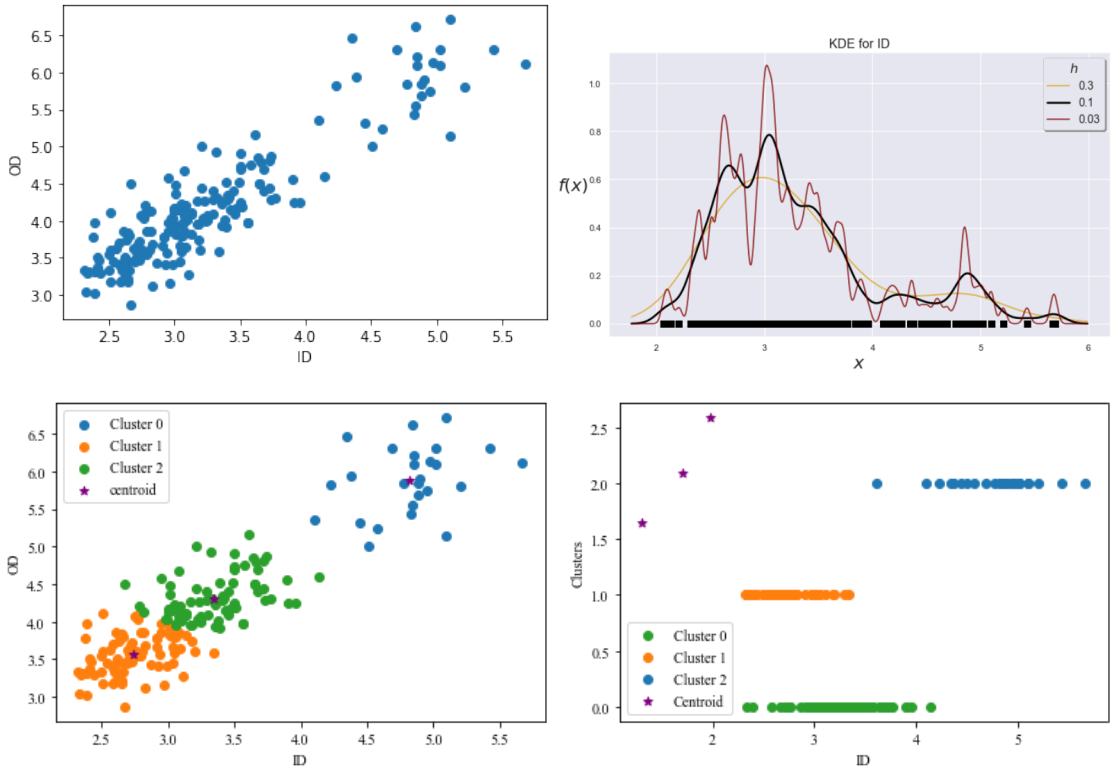

Figure 20. Top left: ID and OD linear relations; top right: ID KDE Gaussian Distribution showing the picks and valleys; bottom left: Define three centroids for three clusters (normalized scale); bottom right: Clustering observations based on ID in three groups.

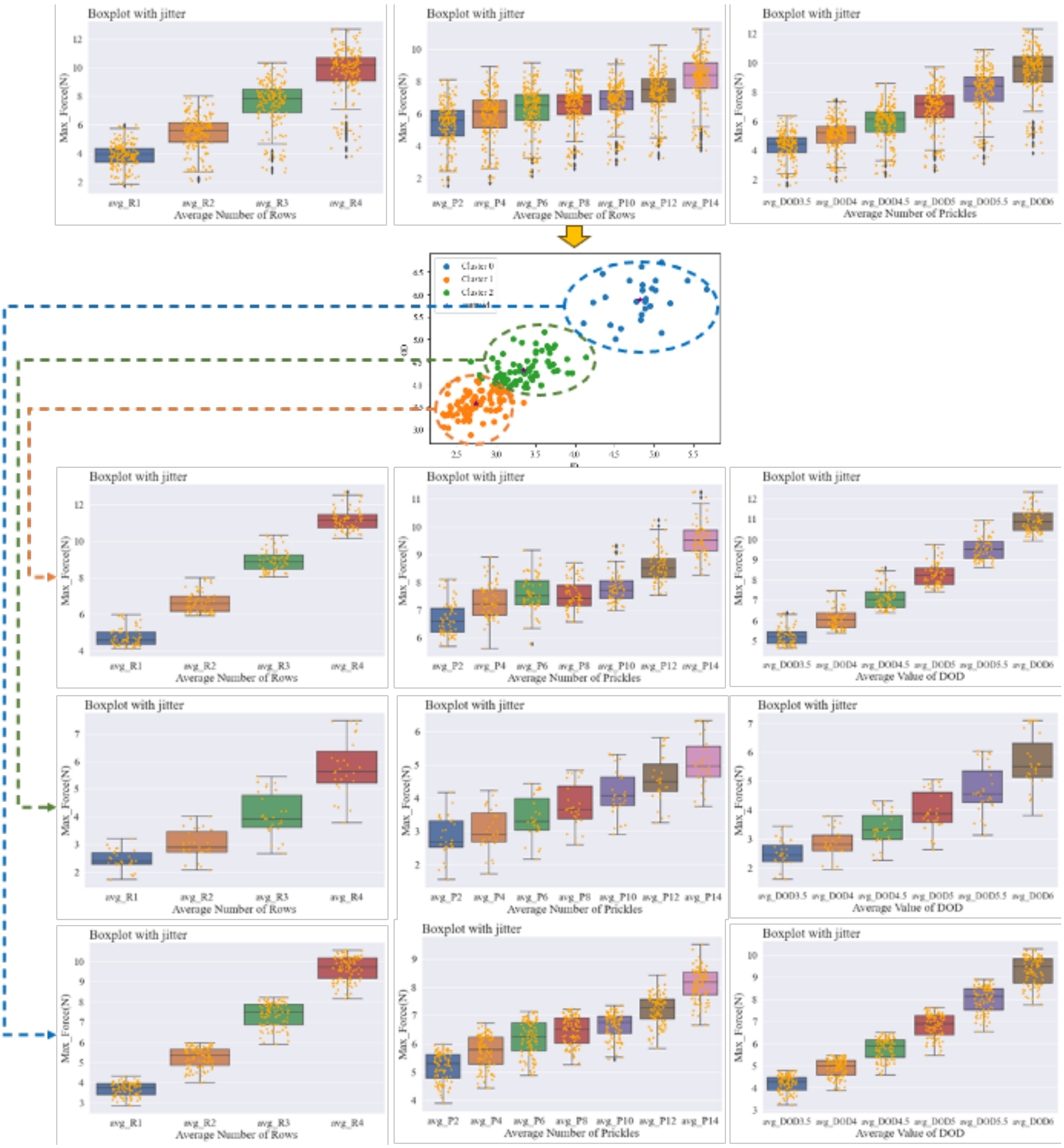

Figure 21. Clustering the observations into three groups. The distributions describe the average distributions for rows, prickles, and DOD for ID. Reducing the distributions assists in finding the best optimal configurations for each cluster, which results in a more accurate and robust design.

5.3.3. Vaso-Lock Design Formulation

The goal of the design optimization process is to find the set of x_d, $x_d = [DOD, R, P]$ that maximize the average response (reaction force/maximum force) of the design, given the variation of the environment and a probability distribution for each variable as:

$$x_e = [\text{OD} \sim \text{p(OD)}, \text{L} \sim \text{p(L)}, \text{WT} \sim \text{p(WT)}, \text{ID} \sim \text{p(ID)}]$$

Additionally, the targeted system is subject to uncertainty due to measurement or manufacturing error (aleatoric or statistical uncertainty) and a lack of data (epistemic uncertainty). According to [92] "aleatoric (aka statistical) uncertainty refers to the notion of randomness, that is, the variability in the outcome of an experiment that is due to inherently random effects." Additionally, [92] defines epistemic (also known as systematic) uncertainty as uncertainty resulting from a lack of knowledge (about the best model). In other words, it refers to the ignorance of the agent or decision-maker and, hence, to the epistemic state of the agent instead of any underlying random phenomenon. "As opposed to uncertainty caused by randomness, uncertainty caused by ignorance can in principle be reduced on the basis of additional information." [92]

The environmental variables, namely OD, ID, WT, and length of the vessels, are classified as epistemic and aleatoric uncertainty. For instance, it creates epistemic uncertainty if the variability in vessel dimensions is poorly understood or if there is insufficient information on the range of dimensions that are anticipated. However, the variability in vessel size falls under aleatoric

uncertainty if it is inherent in the nature of the vessels and cannot be exactly predicted due to natural changes. For example, there may be inherent diversity in the size of vessels among different individuals, and this generates aleatoric uncertainty. Both kinds of uncertainties may be frequently present in real-world engineering situations.

5.3.3.1. RISK-AVERSION OPTIMIZATION

To consider risk aversion in the risk management of Vaso-Lock design, two requirements must be satisfied:

1. **Safety condition,** which refers to the most unfavorable scenario. The performance of Vaso-Lock, specifically the minimum reaction force of a specific design, must exceed a predetermined target value chosen by the user.

2. **The allowable parameter range** is the constraint applied to the value of x_d to ensure the fabricability of the device.

Risk aversion [93,94] optimization problems involve uncertain parameters, which in this case are ID, OD, WT, and length. Equation 8 consists of two mathematical restrictions for an optimization problem. The initial constraint is an inequality that incorporates a function $\mu_F(w \mid x)$, which denotes the mean of a distribution function F with w given x. The outcome of this subtraction is going to be bigger than a target value F_{target}, which is a term that involves a scaler of λ times the standard deviation σ of the same distribution function.

$$\underset{x_c}{argmax} \int f(x_d, x_e)p(x_e)dx_e \qquad x_e \sim p(x_e) \tag{8}$$

$$\underset{x_c}{argmax}\,[F(x_c) - F_{target}]$$

$$s.t.:$$

$$\mu_F(w \mid x) - \lambda\sigma_F(w|x) > F_{target}$$

$$x_{low} \leq x \leq x_{high}$$

The second constraint establishes the boundaries for the variable "x," mandating that it fall between the inclusive range of "x_{low}" and "x_{high}". Figure 22 describes how the illustration computes the average force. The formula in the figure describes that the quantity F is determined by adding up the function values multiplied by their respective probabilities. The weights are represented by w_i, and the summation is performed over N intervals or elements. The indices i correspond to the intervals into which the domain of x_e has been divided. The figure conceptualized the consideration of lower bounds and higher bounds for x_e as a constraint. These constraints are typically used to ensure that the solution to an optimization problem not only seeks to optimize a specific objective function but also satisfies specific performance criteria (like maintaining a value above a target threshold) and respects boundary conditions or limits on the variables involved.

Figure 23 describes the incorporation of controllable and uncontrollable data for the problem formulation.

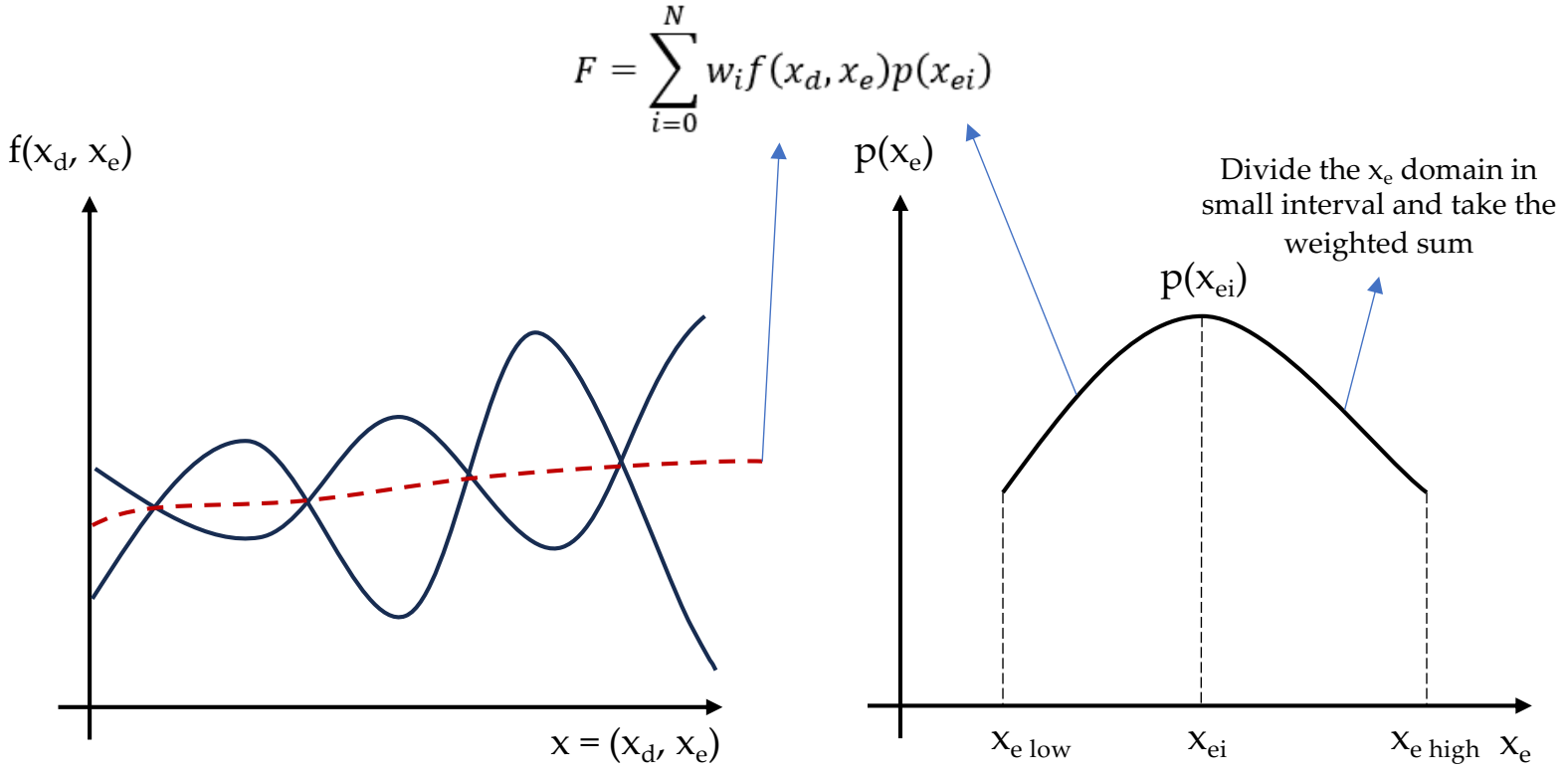

Figure 22. Conceptual average force computation. By multiplying the function values by their respective probabilities, one can calculate the quantity F.

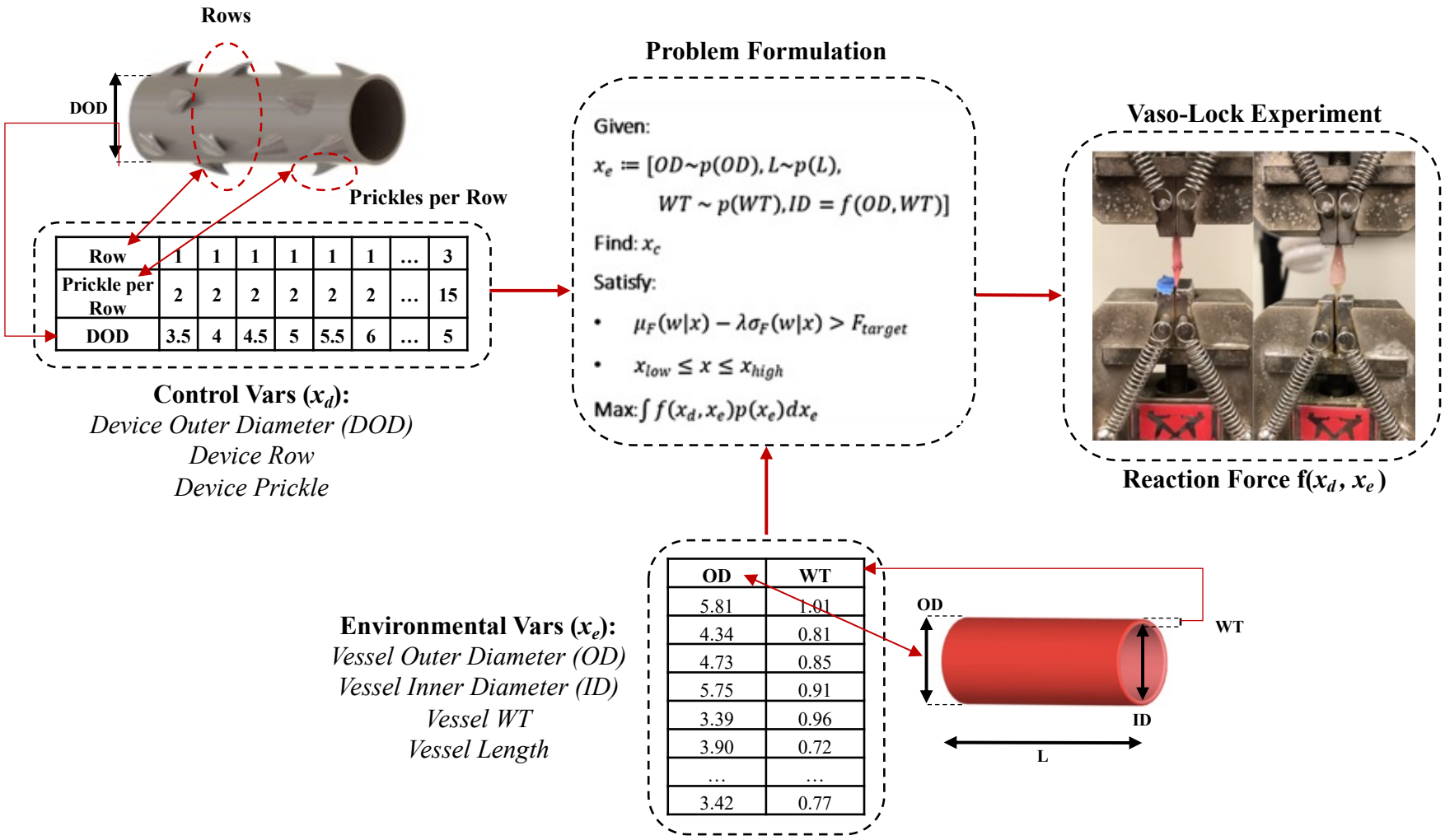

Figure 23. Problem formulation workflow and its extension of the design problem with uncertainty quantification.

5.3.4. Bayesian Optimization

To deal with such a random environment, Bayesian optimization (BO) [95], an AI-assisted design optimization method and a model-based sequential optimization technique [96], capable of accounting for the uncertainty caused by the lack of training data and the randomness in operating environments, are applied. BO uses uncertainty aware exploration and exploitation trade-offs, reducing the required number of iterations [96]. Finding global optimization of non-convex functions is of interest to real-world applications [97]. Bayesian optimization is used to find the global optima when the evaluation is expensive by relying on querying distributions over functions defined by a surrogate model [95,98]. BO was utilized in a random environment to explore the design space. It was used in each of the three clusters to deal with operating conditions and epistemic uncertainty variation, find the best acquisition function, and use a tuned hyperparameter to reach the optimal point in each cluster. Equation 9 shows that the BO formulation identifies maxima for a black-box in a search space $E \subseteq \mathbb{R}^e$.

$$x^+ = \underset{x_e \in E}{argmax} \int f(x_d, x_e) p(x_e) dx_e \qquad x_e \sim p(x_e), \quad f(x_d, x_e) \sim \mathcal{N}(\mu_f, \sigma_f) \tag{9}$$

$p(x_e)$ is the density of the environmental variable. f in finding the best device configuration is expensive, and the Gaussian process (GP) is used as an approximator of the objective function [96]. The Bayesian optimization algorithm comprises two components:

1. A Bayes statistical model for surrogate modeling and uncertainty quantification, and
2. An acquisition function for determining the subsequent design sampling.

5.3.4.1. BAYES STATISTICAL MODEL FOR SURROGATE MODELING AND UNCERTAIN QUANTIFICATION

Figure 24 depicts a conceptual diagram illustrating an active loop constantly updating the surrogate model. The process begins with the modeling of the surrogate model of the GP [99], followed by the selection of the next sample utilizing acquisition functions. Subsequently, additional experiments are conducted for validation purposes, and the convergence of the acquisition functions is assessed.

The Gaussian process is simply defined as a joint distribution of all variables. A distribution model is formulated around functions [96,99], each having a mean $\mu_F(x)$ and covariance function or kernel function $\sigma_F(x, x')$ and integral F as $F \sim \mathcal{N}(\mu_F, \sigma_F)$.

The mean μ_F, the variance σ_F, and the integral F are computed in Equations 10, 11, and 12 using the Bayesian Monte Carlo (BMC) method[9] by using the kernel density estimation integration to compute the integration of kernel density estimation (KDE) as:

[9] Bayesian Monte Carlo refers to the use of Monte Carlo methods within Bayesian inference for probabilistic estimation and inference.

$$\mu_F = \int \mu_F(x_d, x_e) p(x_e) dx_e = \mathbb{E}[f(x)] \tag{10}$$

$$\sigma_F = \int_{x_e \in D_e} \int_{x_e \in D_e} cov[f(x_d, x_e), f(x_d, x'_e)]\, p(x_e) p(x'_e) dx_e\, dx'_e \tag{11}$$

Here μ_F and $cov[f(x_d, x_e), f(x_d, x'_e)]$ are the posterior mean and posterior covariance. Using μ_F and σ_F, the probability distribution of F can be fully defined.

$$f(x) \sim GP\left(\mu_F(x)\right), \sigma_F(x, x')) \tag{12}$$

The updated KDE model in the Bayesian optimization framework is used to select the next best candidate point for evaluation. This process was achieved using the acquisition function, which utilizes the KDE model to maintain a balance between exploration and exploitation. In summary, Bayesian Monte Carlo was used to sample from the KDE model over and over, the objective function was evaluated, the KDE model was updated, and new next points were chosen until convergence criteria were met or a predefined stopping criterion, like budget constraints, was reached.

5.3.4.2. ACQUISITION FUNCTIONS

This section describes the acquisition function for Bayesian optimization. Through scenario-based analysis, identifying the next optimal sampling point to maximize the acquisition functions emerges as a high-ranking initiative. This

section provides optimized sample points for each cluster utilizing an acquisition function.

Some optimization problems have a black-box objective function whose derivatives are unknown. As previously mentioned, to approximate the objective function, surrogate models are used. In Bayesian optimization, the acquisition function was utilized to direct sampling to areas to improve the best observation by exploration and exploitation trade-offs [100]. The goal is to maximize the acquisition function to determine the next sampling point. Figure 25 describes three acquisition functions that were defined and suggests a sample configuration for each cluster. The acquisition functions are:

a) Maximum probability of improvement (MPI) or PI

The next point will be chosen concerning the current maximum point in the probability improvement function. In maximum probability of improvement (MPI), the probability of improvement was considered rather than the magnitude of the next point [101]. Equation 13 shows the maximum probability of improvement formula. $\mu_F(x_d)$ and $\sigma_F(x_d)$ are the mean and variance of the regressor at point x_d. ξ is a parameter controlling the degree of exploration, and $\psi(z)$ denotes the cumulative distribution function of a standard Gaussian distribution. Also, f is the function to be optimized with an estimated maximum at point x_d^+.

$$PI(x_d) = \psi \frac{\mu_F(x_d) - f(x_d^+) - \xi}{\sigma_F(x_d)} \tag{13}$$

b) Upper Confidence Bound (UCB)

Equation 14 describes the upper confidence bound formula. This most straightforward acquisition function used mean and standard deviation linearly for each point x [101]. β is a parameter controlling the degree of exploration.

$$UCB(x_d) = \mu_F(x_d) + \beta\sigma_F(x_d) \tag{14}$$

c) Expected Improvement (EI)

Equation 15 shows the expected improvement formula. $\phi(z)$ denotes the density function of a standard Gaussian distribution. Ultimately, the objective is to find the maximum value of EI, as noted in Equation 16.

$$EI(x_d) = (\mu_F(x_d) - f(x_d^+) - \xi)\,\psi\left(\frac{\mu_F(x_d) - f(x_d^+) - \xi}{\sigma_F(x_d)}\right) + \sigma_F(x_d)\phi\left(\frac{\mu_F(x_d) - f(x_d^+) - \xi}{\sigma_F(x_d)}\right) \tag{15}$$

$$\underset{x_c}{argmax}\ EI\ (x_d) \tag{16}$$

For the sake of this dissertation, the expected improvement (EI) acquisition function was employed for the Vaso-Lock project due to its capability of balancing exploration and exploitation. The upper confidence bound (UCB) and maximum probability of improvement (MPI) acquisition functions are considered for future work.

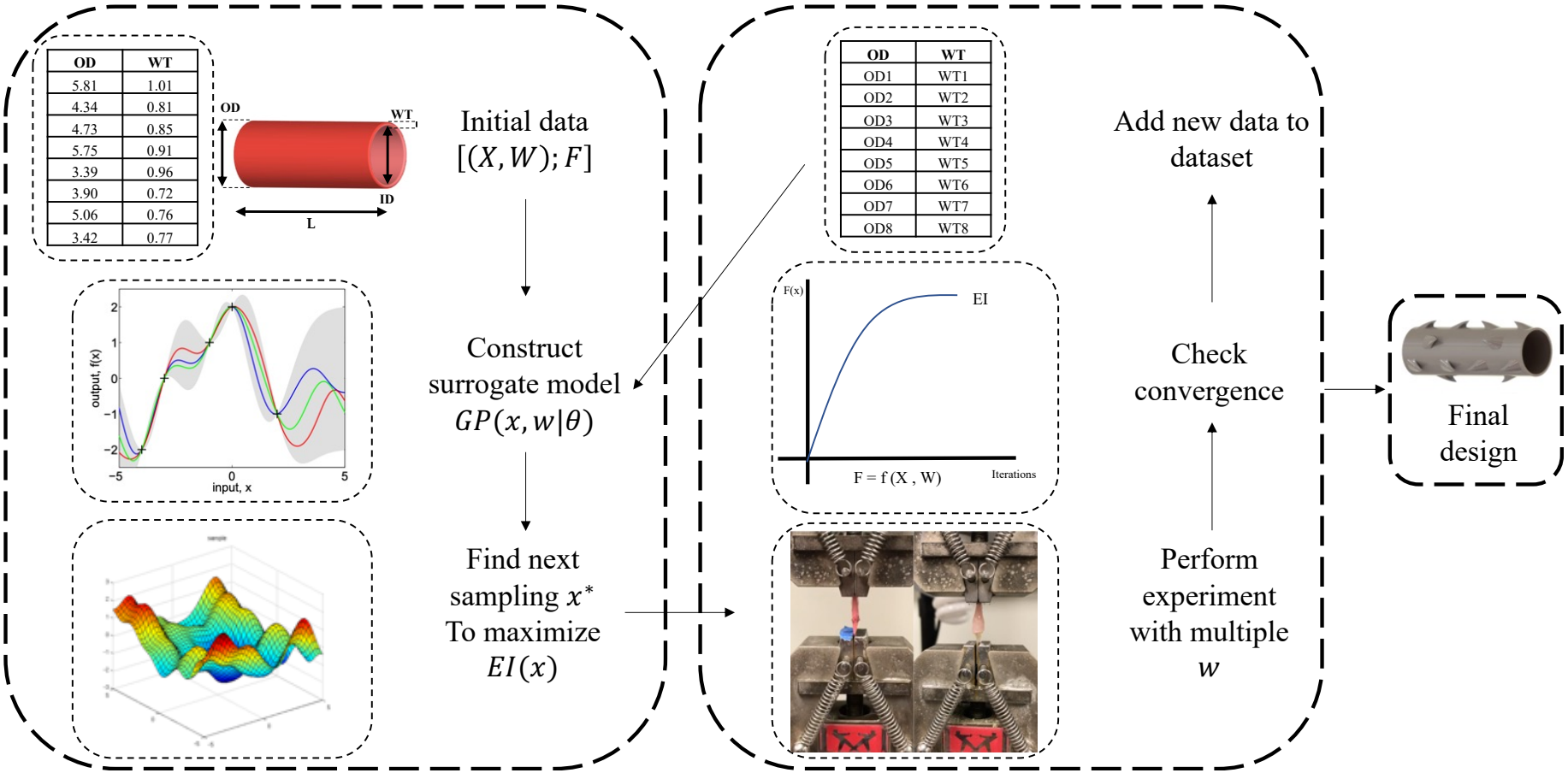

Figure 24. A conceptual diagram of the Bayesian optimization approach used in the demonstration of the *Structure (Sig)* layer. The figure demonstrates the design optimization process, where a surrogate model approximates the acquisition function based on initial data. This function guides the selection of subsequent design examples until convergence or the budget limit is reached.

Figure 24 also describes the process of design optimization. A surrogate model is created based on the initial data sample. Therefore, the surrogate model is employed to approximate the acquisition function. The subsequent design example is determined through the optimization of the provided acquisition function. Using the proposed design sample, physical experiments will be performed, and the design process will continue until the algorithm reaches convergence or the maximum budget is exhausted.

5.3.4.3. CONSTRAINED BAYESIAN OPTIMIZATION

The formulation in the previous section is for an unconstrained optimization problem, yet the scenario involves inequality constraints. However, as mentioned in the Risk-Aversion Optimization section, the process is subjected to lower bound inequality constraints besides the variable bounds. In Bayesian optimization for constrained problems, the acquisition function must be adjusted. Thus, particle swarm optimization (PSO), a gradient-free optimization algorithm, was employed. PSO helps to avoid sensitivity analysis of the optimization problem and guarantees global optima convergence, which is challenging with gradient-based algorithms. Also, to address this problem, EI can be employed to accommodate such constraints. In other words, when a constraint is violated, EI is set to 0. However, given that the constraint is estimated rather than absolute, uncertainty arises. Hence, the constrained EI takes the form below:

$$\alpha_{EIC}(X) = \alpha_{EI}(X) \prod_{k=1}^{K} \Pr(c_k(X) \geq 0 \,|\mathcal{D}) = \alpha_{EIC}(X) \prod_{k=1}^{K} \phi(\frac{\mu_k(X)}{\sigma_k(X)})$$

$\alpha_{EI}(X)$ is the unconstrained EI and $(c_k(X) \geq 0 \,|\mathcal{D})$ is the probability of feasibility. $\mu_k(X)$ is the mean of the estimation, $\sigma_k(X)$ is the variance of the estimation and $\phi(\frac{\mu_k(X)}{\sigma_k(X)})$ is the cumulative density function.

In this case, a two-step approach can be adopted: First, determining the smallest force estimation and its variance concerning changes in the environment, and second, estimating the cumulative density to derive the feasibility weight. This approach ensures the integration of inequality constraints into the optimization process, facilitating more realistic and reliable outcomes.

5.3.4.4. CONSTRAINED AND UNCONSTRAINED NEXT SUGGESTED POINTS

Figure 25 and Figure 26 show the third round of suggested constrained and unconstrained design configurations utilizing the EI acquisition function for three clusters. The unconstrained design uses EI to find the point in the search space that maximizes the expected improvement over the current best value. The optimizer suggests new points for evaluation using EI, evaluates the objective function, updates the GP surrogate model, and repeats the process iteratively. Constrained Bayesian optimization utilizing EI imposes additional constraints on the search space, ensuring the suggested points satisfy the constraints of the problem. In this case, the constraints applied to improve the safety of the device are not to drop below a specific Max_Force threshold. Both aim to efficiently find

the global optimum of the objective function (Max_Focrce). These figures describe that despite the slightly lower force reaction for the constrained design, the worst-case force reaction of the constrained design is always above the Max_Force threshold. However, for the unconstrained design, despite having a higher average force reaction, the worst-case expected force reaction can be lower than the safety value. The suggested next points then went under physical design/production and Max_Force experiments. Then, the results and actual Max_Force were added to the original dataset for more rounds of data training until the results converged.

2mm < ID < 3mm Lower Bound Force: 3 (N)	Bounds	Suggestion EI
Row	[1, 5]	3
Prickle per row	[2,15]	4
DOD	[2.5,4.5]	4.5 (mm)
Pred. force	N/A	3.93 ±0.96 (N)
Current Best		3.97
Lower Bound (Mean)		3.18
Lower Bound (Std)		1.52

3mm < ID < 4mm Lower Bound Force: 3 (N)	Bounds	Suggestion EI
Row	[1, 5]	3
Prickle per row	[2,15]	5
DOD	[3.25,5.0]	4.5 (mm)
Pred. force	N/A	4.05 ±1.05 (N)
Current Best		4.1
Lower Bound (Mean)		3.15
Lower Bound (Std)		1.18

4mm < ID < 5.75mm Lower Bound Force: 3 (N)	Bounds	Suggestion EI
Row	[1, 5]	3
Prickle per row	[2,15]	6
DOD	[3.25,6.5]	6.0 (mm)
Pred. force	N/A	3.66 (N)
Current Best		3.61
Lower Bound (Mean)		3.29
Lower Bound (Std)		1.64

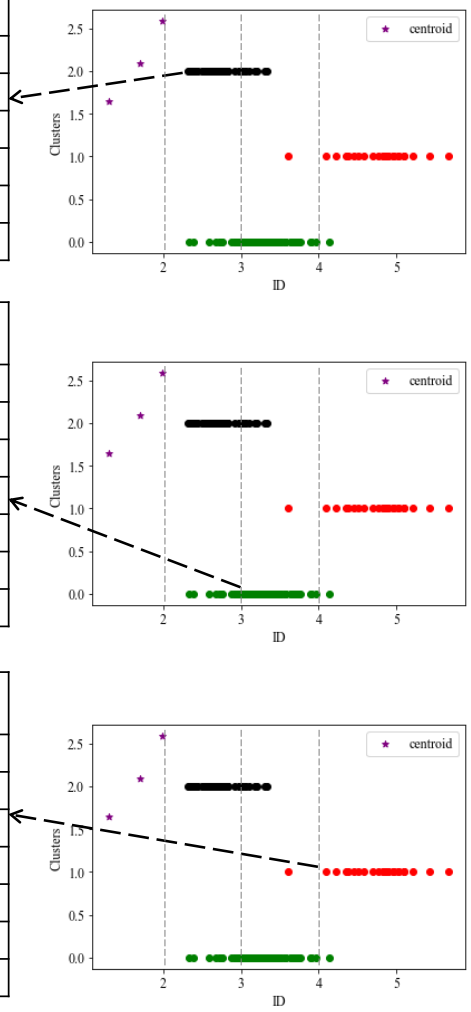

Figure 25. The third round of suggested constrained design configurations utilizes the expected improvement (EI) acquisition function for three clusters.

2mm < ID < 3mm Lower Bound Force: 3 (N)	Bounds	Suggestion EI
Row	[1, 5]	3
Prickle per row	[2,15]	3
DOD	[2.5,4.5]	4.0 (mm)
Pred. force	N/A	4.05 ±1.07 (N)
Current Best		3.91

3mm < ID < 4mm Lower Bound Force: 3 (N)	Bounds	Suggestion EI
Row	[1, 5]	3
Prickle per row	[2,15]	4
DOD	[3.25,5.0]	4.5 (mm)
Pred. force	N/A	4.11 ±1.05 (N)
Current Best		4.07

4mm < ID < 5.75mm Lower Bound Force: 3 (N)	Bounds	Suggestion EI
Row	[1, 5]	3
Prickle per row	[2,15]	5
DOD	[3.25,6.5]	6.0 (mm)
Pred. force	N/A	3.71 (N)
Current Best		3.61

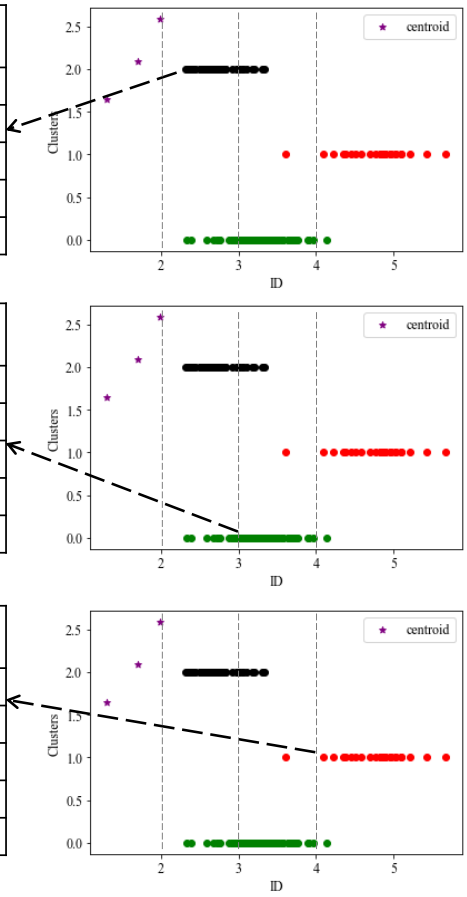

Figure 26. The third round of suggested unconstrained design configurations utilizes the expected improvement (EI) acquisition function for three clusters.

Figure 27 describes the average force convergence based on the number of iterations for both unconstrained and constrained scenarios. The unconstrained average force plot indicates that cluster 2 < ID < 3 reached convergences starting from the second iteration. Cluster 3 < ID < 4 converged in the third iteration, while cluster 4 < ID < 5.75 converged in the second iteration. While the constrained average force plot describes that more iterations are needed to assess the number of iterations so that the average force will converge.

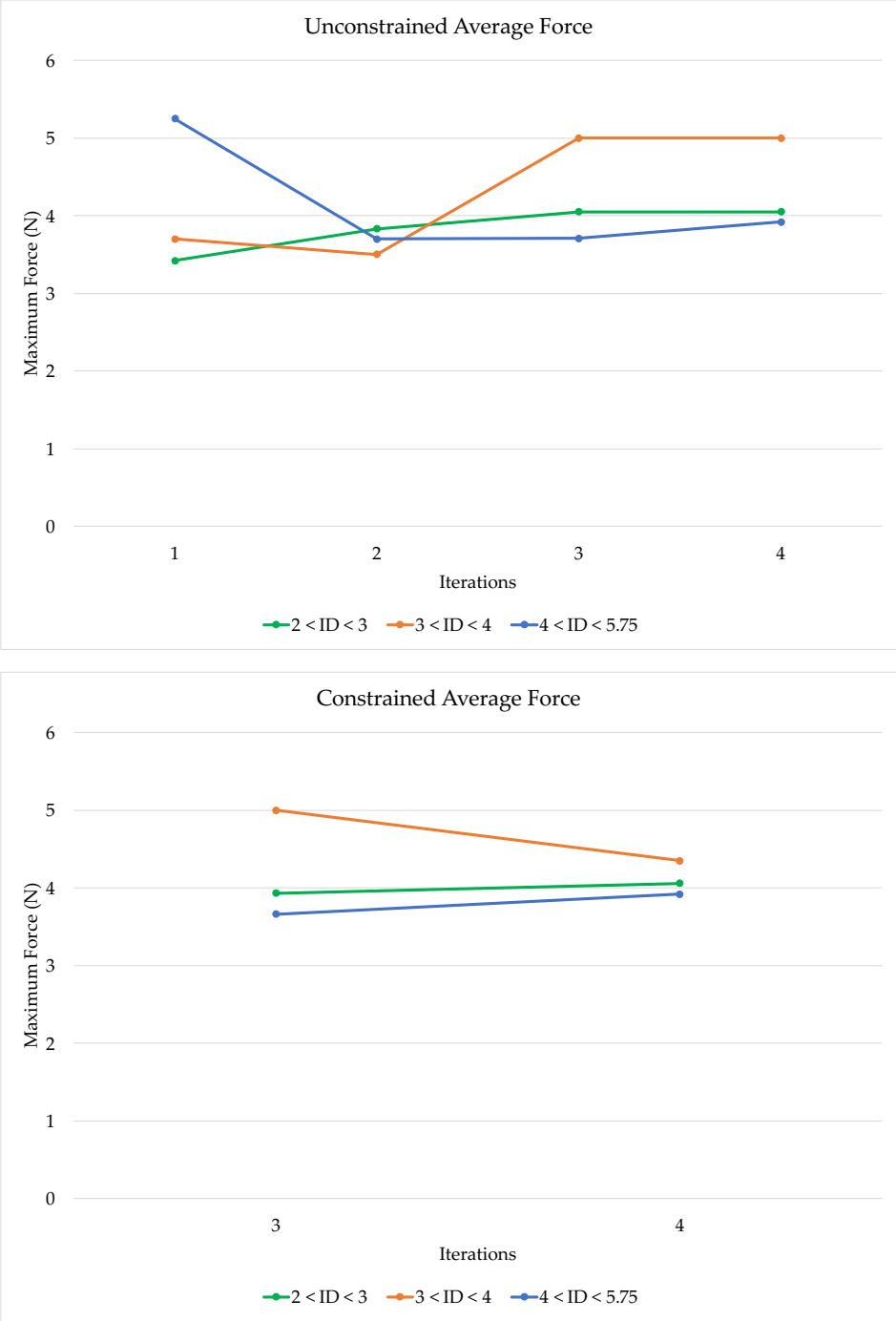

Figure 27. Unconstrained and constrained average force convergence by the number of iterations. According to the unconstrained average force plot, maximum force for clusters 2 < ID < 3 and 4 < ID < 5.75 converged after 2 iterations and cluster 3 < ID < 4 was converged after 3 iterations. More iterations are needed for constrained average force to verify the number of iterations needed for the maximum force to be converged.

In summary, the environmental space was first clustered into smaller subspaces to diminish the variation of reaction force in accordance with variations in the environmental variables. Subsequently, the Bayesian optimization method was utilized to ascertain the optimal device configuration for each environmental variable subspace. Upon receiving design suggestions generated by the algorithm, a prototype of the device was fabricated, and physical tests were conducted. The data was then recorded and transmitted back to the algorithm for updating the surrogate model and proposing new design samples. The design process was sustained until either the convergence criteria were fulfilled, or the maximum number of iterations was attained. With Bayesian optimization, the optimal design configuration for Vaso-Lock can be found with a much smaller number of physical experiments, facilitating the device development process. With the assistance of AI, the Vaso-Lock optimal geometry can be discovered in less time and at lower labor costs. It also has the potential to transition the findings on AI-assisted design optimization to facilitate the development of medical devices in healthcare applications.

In the next section, explainable AI models were used to identify the behavior of each feature and its effect on the Max_Force, assess the most important features in each cluster, validate the suggested points from the third round iteration, and find potential relations between the features that affect the Max_Force the most.

5.4. Explainable AI (XAI) of the Scenarios that are Most Disruptive tothe System Order

This section describes the explainable AI (XAI) of the scenarios that are most disruptive to the system order. The lack of transparency of AI-based models is a significant obstacle to their implementation and is criticized due to their black-box nature. This lack of transparency could deconstruct decision-making processes [102–104].

Figure 11 described that one of the most disruptive scenarios was *s.06 – Non-Interpretable AI and Lack of Human-AI Communications.* Also, interpreting and explaining the suggested sample were defined as some of the highest ranked initiatives in Figure 12 by *x.Sig.40 - Communicating a Description of Why an AI System Made a Particular Prediction or Recommendation*, and *x.Sig.22 - Operate the AI System and Continuously Assessing its Recommendations and Impacts.* Thus, to enhance trustworthy AI by explaining and interpreting the results and answering the question of "how" the system made a decision, Shapely additive explanations (SHAP) [105,106], local interpretable model-agnostic explanations (LIME), partial dependence plots (PDP) with ICE plots, interpret explaining boosting model generalize additive model (GAM), counterfactual analysis, and explain like I am 5 (ELI5) were utilized to interpret the suggested sample points from the acquisition functions and to interpret the overall behavior of the observations in each cluster.

5.4.1. SHAP Analysis of the Scenarios that are Most Disruptive to System Order

The SHAP [39,106–110] method is a highly efficient approach for XAI. It assigns values to the characteristics used for making predictions, indicating their impact on the output of the model. SHAP enables the identification of the factors that impact AI decisions, enhancing their interpretability and reliability. It is an explainer that can be applied to both local and global models. It calculates the importance of features using Shapley values, which are derived from cooperative game theory. The explanations provided by SHAP are based on additive feature attributions [39,107,108].

Figure 28 describes the SHAP global explanation utilizing the random forest model, where red dots indicate the positive impact and contributions of a particular feature to the prediction when it appears on the right side of the SHAP value 0 line. The blue observations on the left side of the SHAP value 0 line indicate negative influences and contributions to the prediction findings. Rows, as a feature, exhibit the most significant impact on the predictions. As the number of rows increases, the prediction value tends to rise, while a decrease in the number of rows leads to a decrease in the prediction value. Conversely, the ID exhibits an inverse relationship, where an increase in the ID value leads to a decrease in the prediction value. This is because the red observations are located on the left side of the SHAP value of 0. Thus, Rows and DOD have the most contributions, and WT and OD have the least contributions to the results.

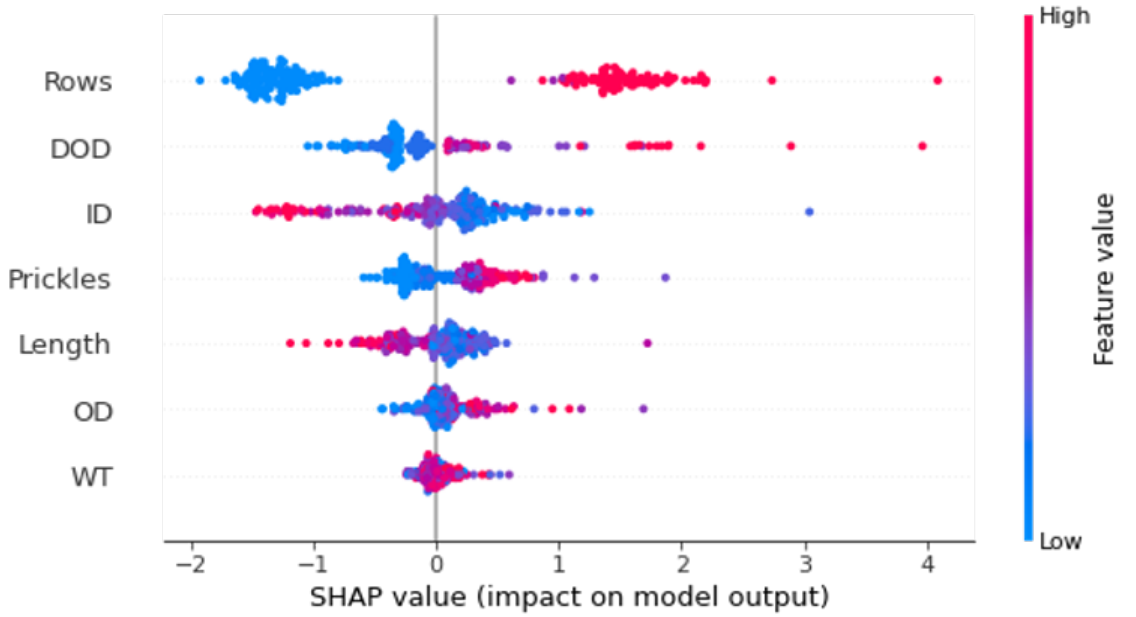

Figure 28. Global SHAP analysis of feature contributions to output predictions for all instances using the random forest model. Higher number of rows and value of DOD corresponds with higher maximum force.

The individual contributions of each feature were evaluated after clustering the observations by OD in this section using SHAP analysis. Three global SHAP analyses were developed for three observation clusters by OD, as follows: 1. OD < 4 mm, 2. 4mm < OD < 5 mm, 3. 5 mm < OD to identify the most and least contributing features to the outcome predictions. As an example,

Figure 29 describes that the number of rows has positively contributed to increasing the maximum force the device could hold the vessels for cluster 4 mm < OD < 5 mm. However, the number of prickles has negatively contributed to the SHAP value. Figure 30 and Figure 31 show the results for the other two clusters of OD < 4 mm and 5 mm < OD, respectively. The results show that each cluster differs in feature contributions to the results; thus, clustering could reduce the

various distributions and result in more accurate suggested configurations of the device.

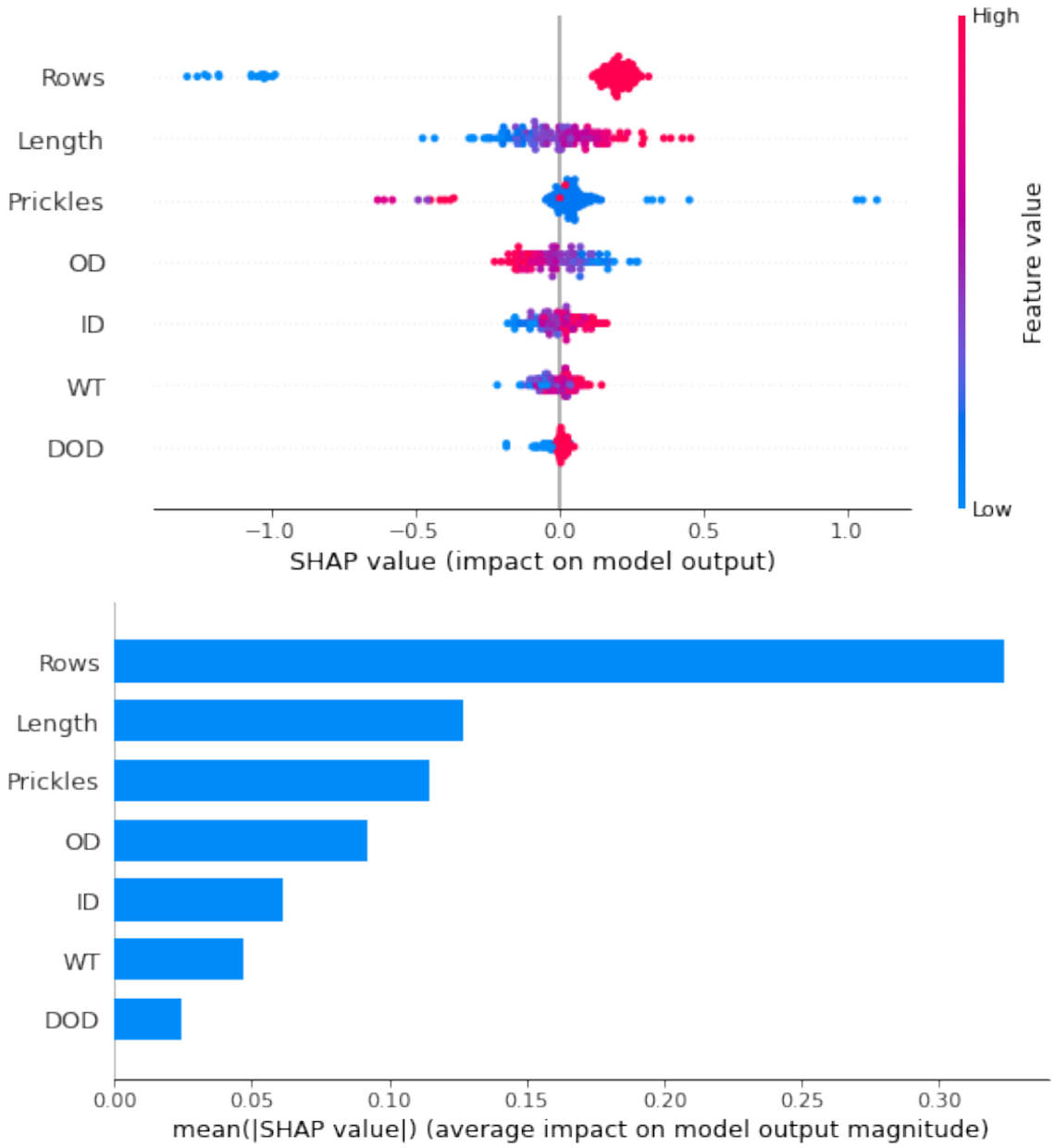

Figure 29. Global SHAP analysis of feature contributions to output predictions for cluster 4mm < OD < 5 mm. Number of rows and length value are most correlated with the value of maximum force.

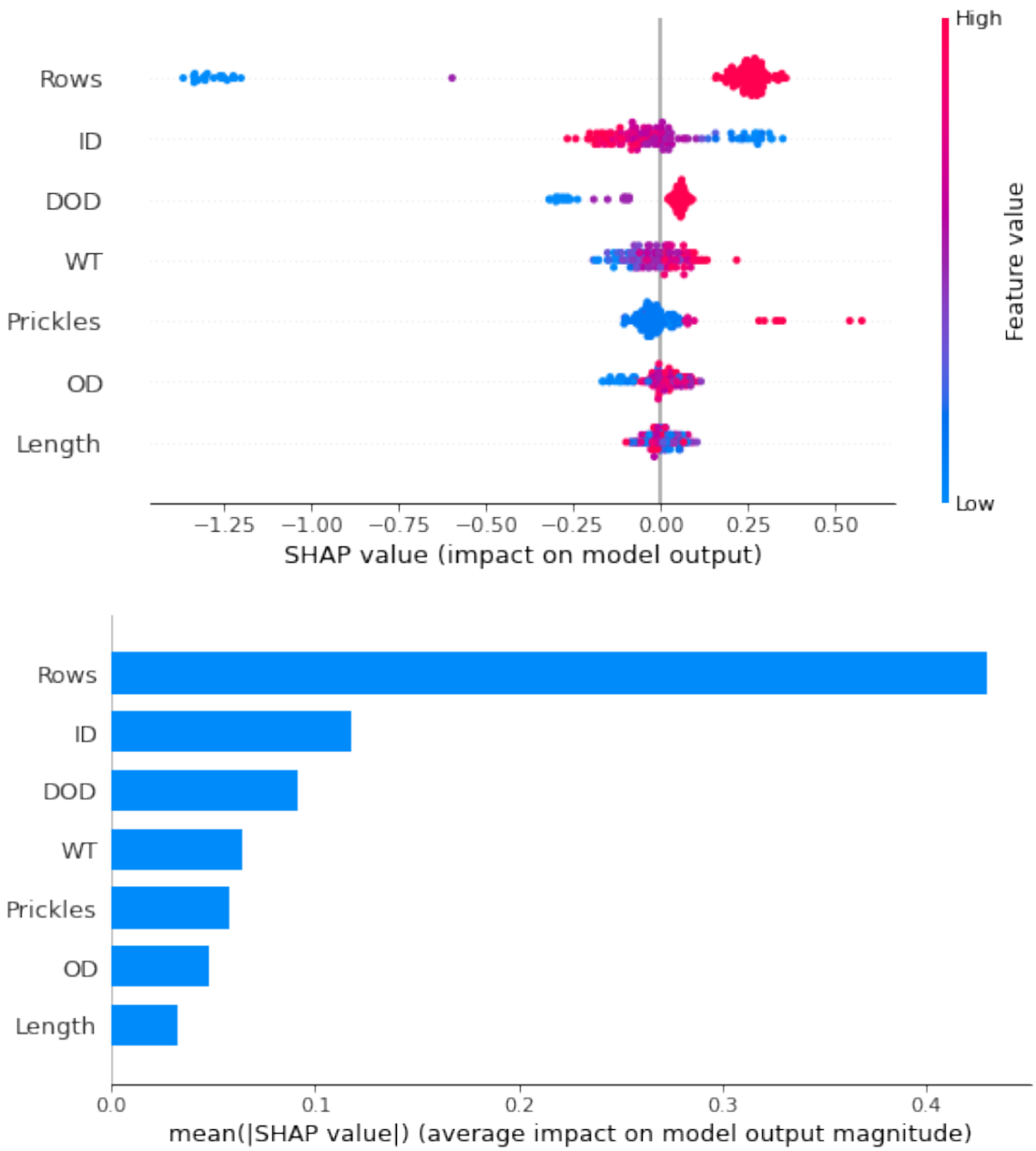

Figure 30. Global SHAP analysis of feature contributions to output predictions for cluster OD < 4mm. Number of rows and ID values are most correlated with the value of maximum force.

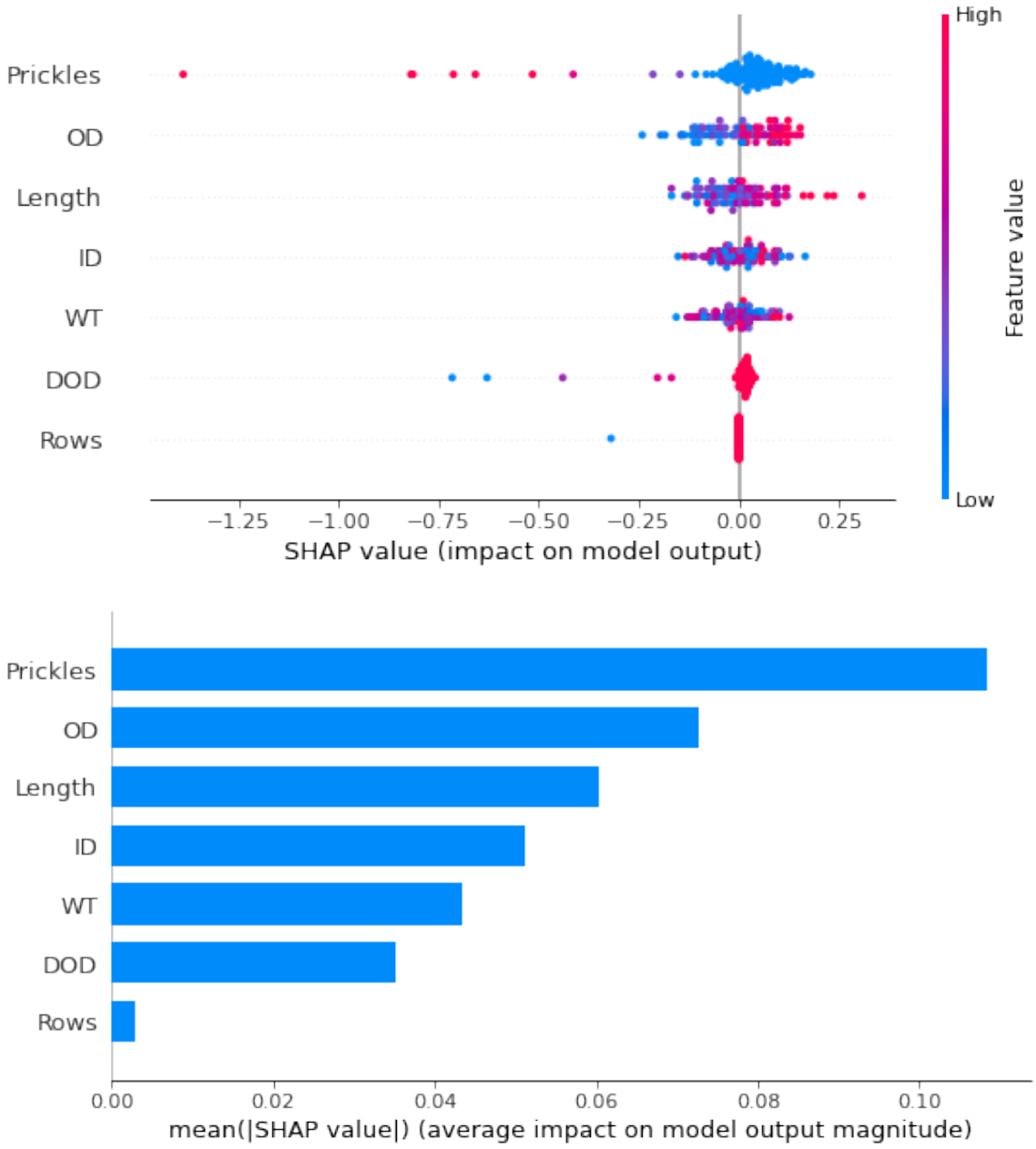

Figure 31. Global SHAP analysis of feature contributions to output predictions for 5mm < OD. Number of prickles and OD values are most correlated with the value of maximum force.

5.4.2. Partial Dependence Plots with ICE Plots for the Scenarios that are Most Disruptive to System Order

The partial dependence plots (PDP) or PD profiles were initially introduced by Friedman in 2000 in the context of gradient boosting machines (GBM) [109]. These visualizations demonstrate the incremental impact that one or two specific features have on the predicted outcome of a machine learning model. The essence of PDP lies in its ability to demonstrate the shifting predicted outcome as specific features shift while holding all other features constant. Comprehending the correlation between input features and predictions, understanding patterns, and detecting interactions [111].

An individual conditional expectation (ICE) plot visually illustrates the variation in the predicted outcome for a single instance as a particular feature change. ICE plots enhance the partial dependence plot by visually representing the specific relationship between the predicted response and the feature for each observation [112]. ICE plots specifically illustrate the differences in the estimated values throughout the entire range of a covariate, indicating the presence and magnitude of heterogeneities. ICE plots involve the analysis of how individual features affect predictions for a particular instance [112].

Figure 33 to Figure 39 describe the PDP with ICE plot using the random forest regression generated in the KNIME[10] interface (Figure 32). The plots were generated for all the features, including device rows, device prickles, device DOD, vessel length, vessel OD, vessel ID, and vessel width.

[10] KNIME is a free and open-source data analytics software.

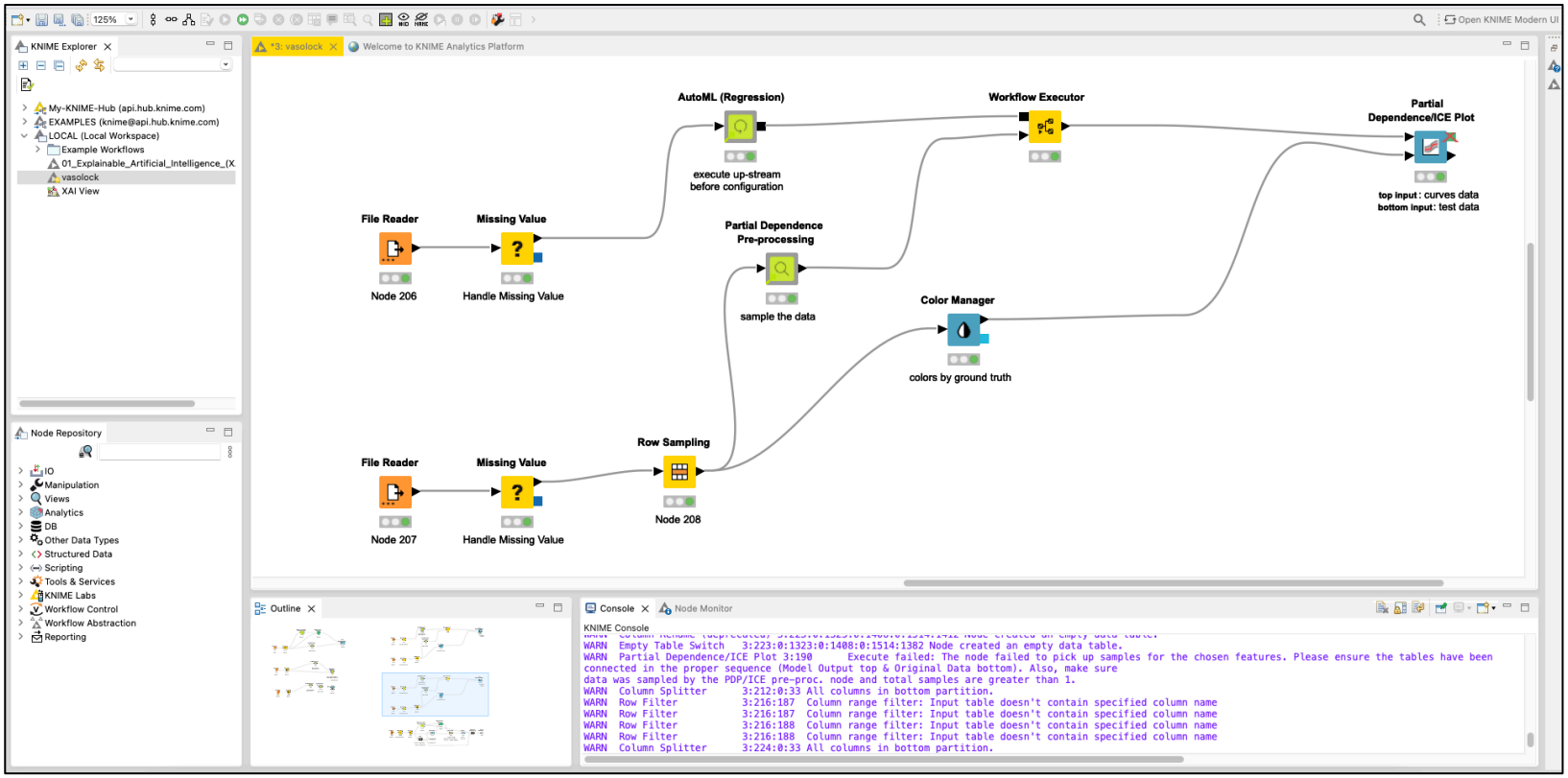

Figure 32. KNIME partial dependence/ICE plot workflow describes the workflow of the steps to generate partial dependence plots.

Device Rows:

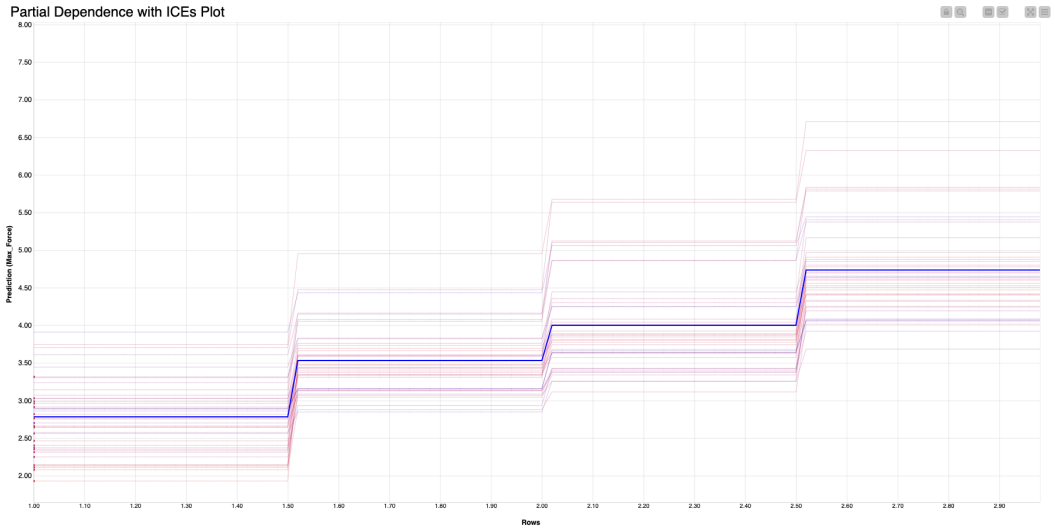

Figure 33. Partial dependence plot (PDP) with ICE plot (KNIME workflow) for device rows using random forest regression. Number of rows corresponds to the Max_Force value.

Figure 33 describes that when the number of device rows is increased while keeping all other features constant, the Max_Force also increases.

Device Prickles:

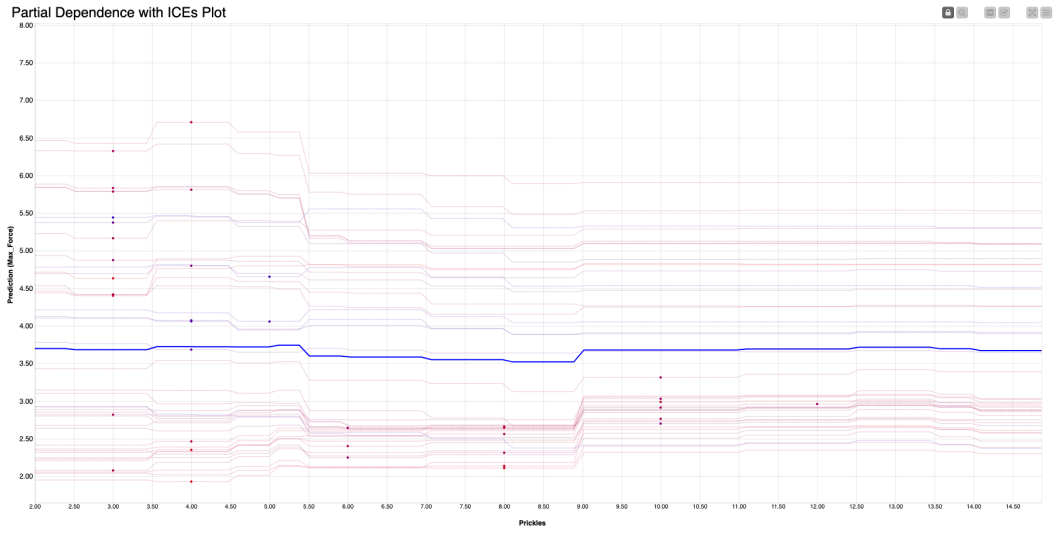

Figure 34. Partial dependence plot (PDP) with ICE plot (KNIME workflow) for device Prickles using random forest regression. Number of Prickles does not correspond to the Max_Force value.

Figure 34 describes that when the quantity of device prickles grows while maintaining the rest of the features unchanged, the Max_Force remains constant, and the line appears nearly horizontal. However, the Max_Force decreased as the number of prickles reached 14. This result is consistent with SHAP analysis for each cluster above, as if the number of prickles grows, they will have a negative impact and contribute to the Max_Force predictions.

Vessel Length (L):

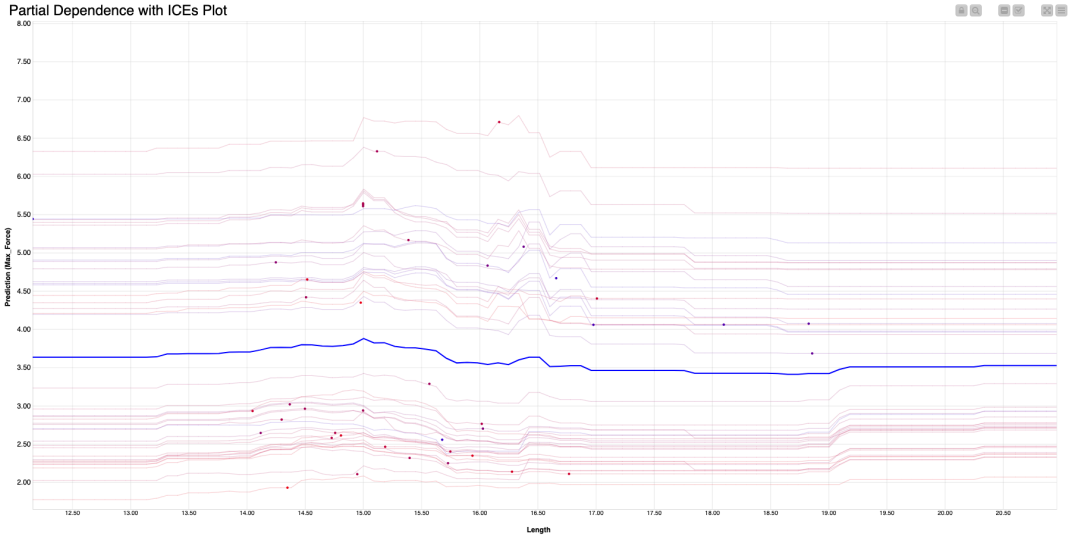

Figure 35. Partial dependence plot (PDP) with ICE plot (KNIME workflow) for vessel length using random forest regression. Vessel length does not correspond to Max_Force value.

Figure 35 describes that as the length of the vessel increases while maintaining all other characteristics constant, the Max_Force remains constant, and the line appears nearly horizontal. This result is in line with the SHAP analysis mentioned earlier, which indicates that vessel length does not have a significant impact on the fluctuations of Max_Force.

Vessel Outer Diameter (OD):

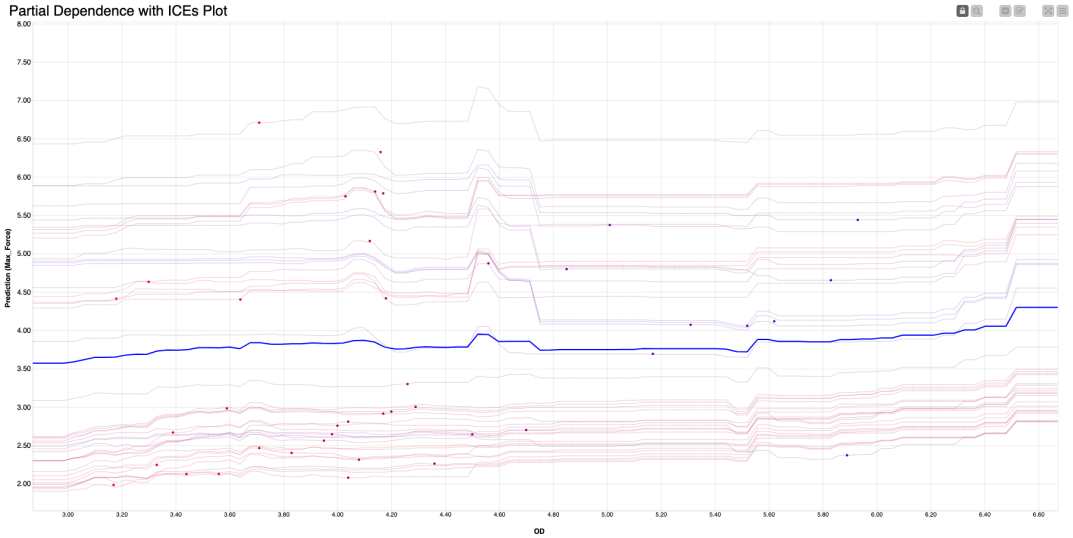

Figure 36. Partial dependence plot (PDP) with ICE plot (KNIME workflow) for vessel outer diameter (OD) using random forest regression. OD value corresponds to the Max_Force value.

Figure 36 describes that as the vessel outer diameter (OD) increases while maintaining all other characteristics constant, the Max_Force exhibits a gradual increase, with a peak occurring at 6.50 mm.

Vessel Inner Diameter (ID):

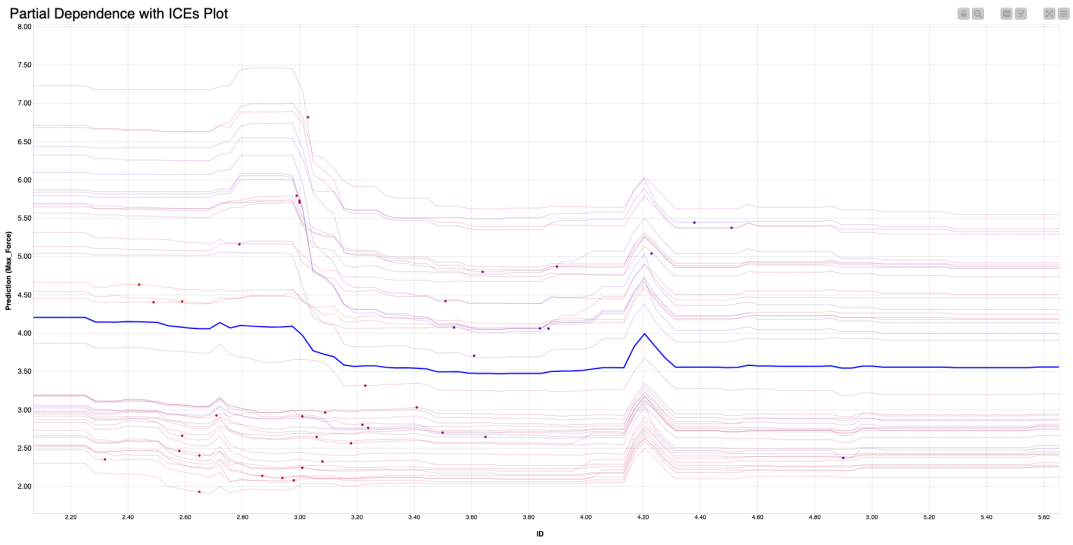

Figure 37. Partial dependence plot (PDP) with ICE plot (KNIME workflow) for vessel inner diameter (ID) using random forest regression. ID value corresponds to the Max_Force value.

Figure 37 describes a consistent decrease in Max_Force as the vessel ID increases, while all other features remain constant. This decrease continues smoothly until ID = 3 mm, at which point the line remains constant.

Vessel Width (WT):

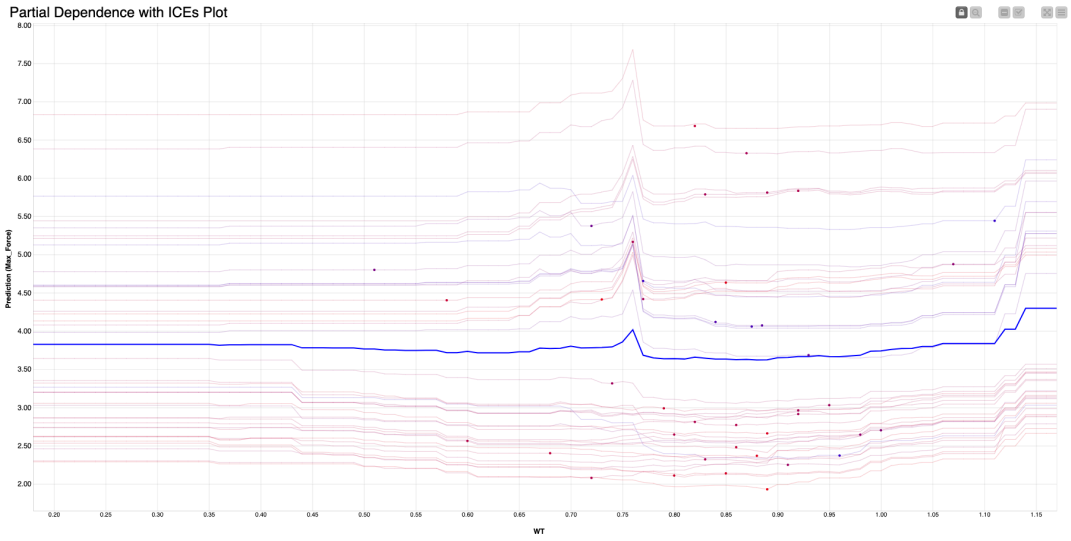

Figure 38. Partial dependence plot (PDP) with ICE plot (KNIME workflow) of vessel width (WT) using random forest regression. Vessel width does not correspond with the Max_Force value.

Figure 38 describes that when the value of vessel width grows while maintaining the rest of the features unchanged, the Max_Force remains constant, and the line appears nearly horizontal. However, the Max_Force increased as the vessel width value reached 1.10 mm.

Device Outer Diameter (DOD):

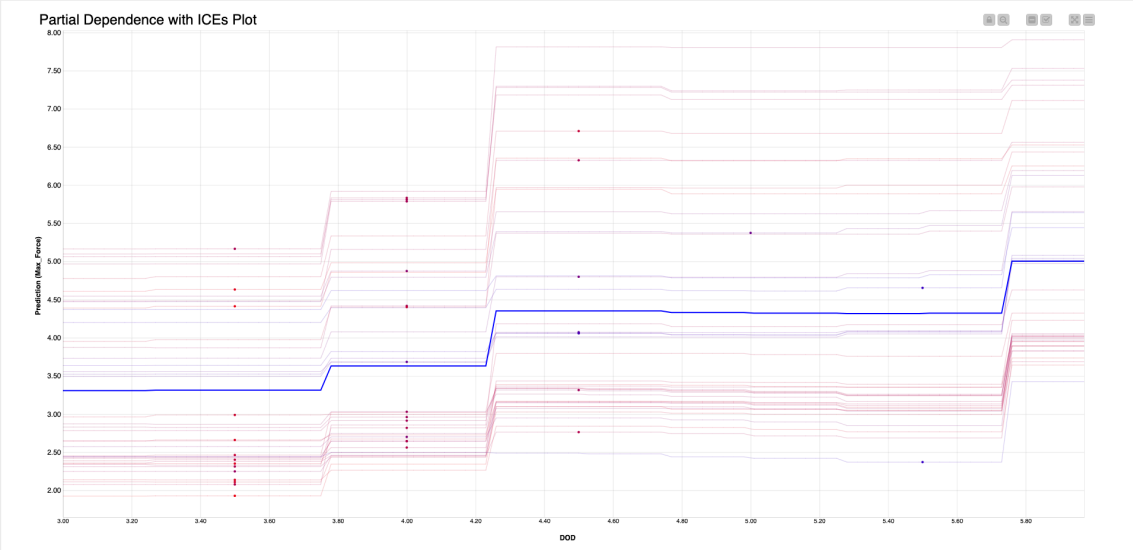

Figure 39. Partial dependence plot (PDP) with ICE plot (KNIME workflow) for device outer diameter (DOD) using random forest regression. Device DOD corresponds to the Max_Force value.

Figure 39 describes that when the value of DOD increases while keeping all other features constant, the Max_Force also increases.

5.4.3. LIME Analysis of the Scenarios that are Most Disruptive to System Order

LIME [39,49,109] explanations are explainers that are not dependent on any specific model and provide explanations in the form of feature importance vectors. The core concept of LIME is that the interpretation can be obtained from nearby data points that are randomly generated in the vicinity of the instance that requires explanation. LIME elucidates individual predictions by approximating the intricate model in a localized manner using a simpler and more comprehensible model. It creates a surrogate model that is easier to comprehend and utilizes it to gain insight into the logic behind the prediction for a particular instance [39,47,49].

To elucidate the proposed subsequent aspects of the BO model in this case, LIME has been employed as a local explainer by utilizing the random forest. Figure 40 describes a detailed description of one of the proposed points, including rows of 3, prickles of 6, length of 16.79, OD of 5.55, ID of 4.2, WT of 0.675, DOD of 6, and a predicted maximum force of 19.7. Rows, DOD, prickles, OD, and WT positively contribute to the predicted result, whereas ID and length negatively contribute to the MAX_Force prediction value. This approach is helpful in validating and verifying the suggested next point from BO. This figure describes that DOD greater than 4.50 has positive contributions to the prediction of Max_Force, and this is in line with the results in the PD plot from Figure 39.

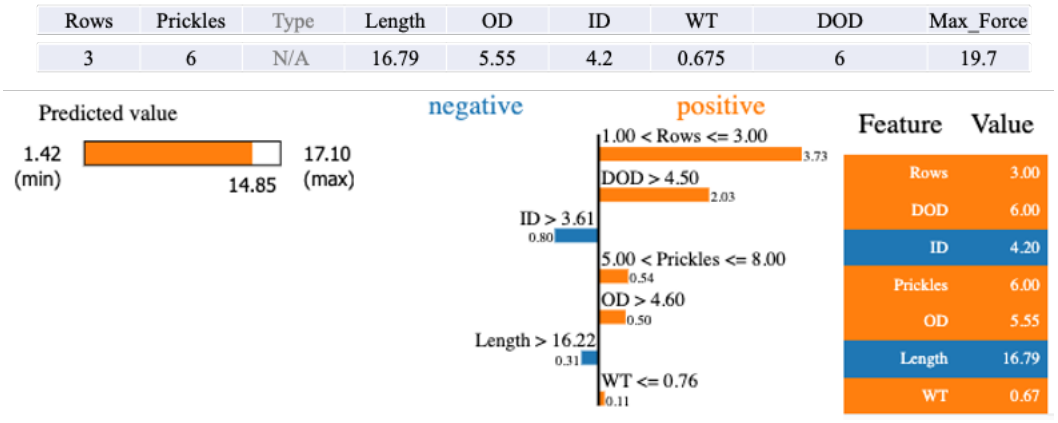

Figure 40. Local LIME explanation of BO suggested points (observation 204) utilizing the random forest model. Increasing the value of Rows, DOD, Prickles, OD, WT and decreasing the value of ID and Length corresponds to the prediction value of the Max_Force. The predicted value of the Max_Force is 14.85 and the actual value of the experiment is 19.7.

5.4.4. GAM Analysis of the Scenarios that are Most Disruptive to System Order

The concept of a generalized additive model (GAM) involves allowing the (generalized) linear model (GLM) to acquire information regarding non-linear

relationships. GAM relaxes the constraint that the relationship between variables must be a simple weighted sum. Instead, they posit that the outcome can be described as a combination of various functions of each feature [111,113].

A GAM remains fundamentally a combination of feature effects, but it offers the flexibility to incorporate non-linear relationships between specific features and the output. In summary, generalized additive models are appropriate for examining the dataset and illustrating the correlation between the dependent variable and the independent variables [111,113]. The simplistic mathematical idea of GAM is to replace $\beta_j x_j$ from GLM by $f_j(x_j)$ which is a more flexible function as shown in Equation 17 [111]:

$$g(\mathrm{E}_Y\ (\mathrm{y}\,|\,\mathrm{x})) = \beta_0 + f_1(x_1) + f_1(x_1) + \cdots + f_p(x_p) \tag{17}$$

Figure 41 describes the combination of all Vaso-Lock feature effects, but it suggests the flexibility to incorporate non-linear relationships between specific features and the Max_Force. This figure shows that the Max_Force value increases proportionally with the number of rows. This finding is consistent with the PDP plots as well. The increase in the value of DOD leads to a corresponding increase in the Max_Force value for DOD. This validation confirms the accuracy of the PDP plots generated previously. The type of GAM model that allows a non-smooth functional form is the Explainable Boosting Machine (EBM). Unlike traditional GAMs which typically use smoothing splines or other smooth functions to model relationships between variables, EBMs use a series of decision trees as base learners, which can capture non-linear and non-smooth relationships in the data.

The dotted lines in the figure describe the average effect of each feature on the predictions of the model, while the red line shows the actual effect of the feature on a specific instance of the data (specified by sample_ind). Essentially, the dotted lines give the overall trend of how changing a particular feature affects the output of the model across the entire dataset, while the red line shows the impact of that feature for a specific data point.

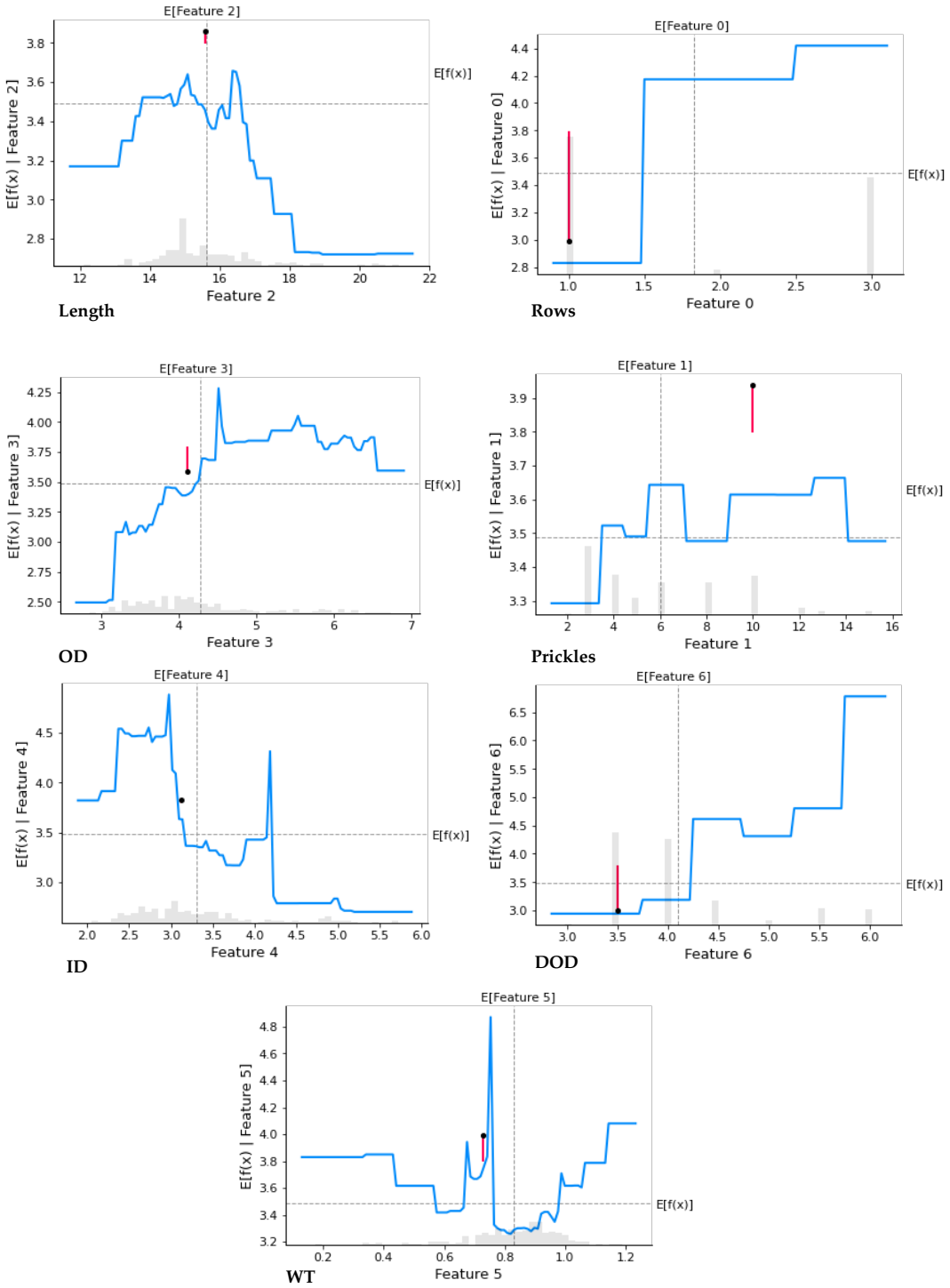

Figure 41. Interpret explaining boosting model, generalize additive model (GAM). Caption: The dotted lines represent the average effect of each feature on the model predictions across the dataset, while the red line illustrates the actual effect of a specific feature on a particular data instance (specified by sample_ind).

5.4.5. Counterfactual Analysis of the Scenarios that are Most Disruptive to System Order

A counterfactual explanation is a description of a causal situation that follows the structure: "If X had not happened, then Y would not have happened" [111]. Counterfactual explanations are employed to clarify the predictions made for specific instances. The term "event" refers to the anticipated result of a specific occurrence, while the "causes" are the specific feature values of this occurrence that were provided as input to the model and resulted in a particular prediction [111,114]. A specific category of explanation shows a connection between the potential outcomes if the input to a model has been changed in a specific manner. To summarize, A counterfactual explanation of a prediction refers to the minimal alteration in the values of the features that results in a change of the prediction to a predefined output [111,114]. Counterfactual analysis can indeed be used as a form of sensitivity analysis. The purpose of employing counterfactuals in the BO model is to assess the sensitivity of the model to minor changes in the input features.

To assess the reliability of the BO suggestions in predicting outcomes, counterfactual inference was employed on observation 204, which represents a sample suggested by the BO. A neural network (NN) model was structured with three hidden layers using the mean_squared_error (MSE) loss function and Adam optimizer. The evaluation involved assessing the shift between the original prediction and the counterfactual prediction by incrementing each feature by a small amount of 0.1 unit. If the predictions made by the model show significant shifts when the input features are slightly modified, it suggests that the decision boundary of the model is highly sensitive to those features.

Conversely, if the predictions remain relatively consistent, it indicates that the decisions of the model are resistant to minor changes in the input, and it may indicate that the suggestions and model provided by the BO are resilient to variations and an unpredictable setting. As stated in Chapter 2, two of the NIST AI Risk Management Framework *principles* include the requirement for the system to be robust and resilient. The counterfactual analysis in this case demonstrated that the BO model is robust and resistant.

Updated Prediction (for the original input): The model forecasts a value of approximately 3.878 for the "Max_Force" variable, considering the original features.

Counterfactual Prediction (perturbed input): The model predicts a marginally increased value of approximately 3.955 for the "Max_Force" when every feature of the original instance changes by a small increase of 0.1. A minor change of 0.077 in the prediction value could imply that the BO model is robust.

Feature Differences:

	Rows	Prickles	Length	OD	ID	WT	DOD
0	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Original Prediction: [[3.878442]]
Counterfactual Prediction: [[3.955788]]

5.4.6. ELI5 Analysis of the Scenarios that are Most Disruptive to System Order

"ELI5" is an abbreviation for "Explain Like I am 5." It refers to the process of simplifying complex concepts or ideas put forth by MIT. ELI5 is commonly used to debug regressor or classifier algorithms. ELI5 is not an appropriate approach for determining the impact of individual features on model performance. The

measure solely indicates the magnitude of changes in feature weights, so it does not consider the weights as feature important on a scale. Basically, it explains their predictions by assigning weights to decisions [110,115,116]. Positive weights imply that a rise in the value of the feature results in a corresponding increase in the predicted target value, whereas negative weights indicate the opposite.

The bias term corresponds to the constant term in the linear model. The intercept represents the predicted target value when all features are set to zero. Put simply, it is the baseline prediction.

For interpreting predictions utilizing an ELI5 interpreter, a variance inflation factor (VIF) was implemented.

VIF is a measure of statistics employed in regression analysis to identify the presence of multicollinearity among predictor variables. Multicollinearity arises when there is a strong correlation between two or more predictor variables in a regression model, leading to challenges in interpreting the coefficients of the model. A VIF value below 10 is considered acceptable, while an increasing VIF indicates decreasing reliability of the regression results. Typically, a VIF exceeding 10 signifies a strong correlation and is considered problematic, so they were removed from the analysis. Table 23 describes the ultimate VIF scores, where the scores of all features are below 10.

Upon implementing VIF, it is observed that two additional features, namely DOD-ID and OD-ID, exhibit stronger correlations with the prediction of Max_Force. These two features were incorporated into the ELI5 analysis.

Table 23. Variance inflation factor (VIF)

	Variables	VIF
0	Rows	4.666047
1	Prickles	3.157200
2	DOD_ID	5.829551
3	OD_ID	8.666678

5.4.6.1. ELI5 GLOBAL INTERPRETATION

After implementing VIF, two new features show more correlations with the prediction of Max_Force, which are DOD-ID and OD-ID. These two features were added to the ELI5 analysis. In Table 24, a global interpretation shows that for rows and DOD-ID, an increase in these values leads to an increase in the predicted target value. This result is also consistent with the global SHAP analysis that has been provided before, which shows that rows have the most contributions among other features.

The bias term, denoted as -0.682, signifies the baseline prediction when all features possess a value of zero.

Table 24. ELI5 global interpretation

Contribution	Feature
+1.382	Rows
+1.242	DOD_ID

Contribution	Feature
+0.125	Prickles
-0.075	OD_ID
-0.682	<BIAS>

5.4.6.2. ELI5 LOCAL INTERPRETATION

On the contrary, a local ELI5 explainer acquired instance number 60 of the dataset to examine the behavior of each design in response to the maximum force. As depicted in Table 25, an increase in both the number of prickles and rows results in a corresponding increase in the predicted target value. The bias term, denoted as -1.461, signifies the baseline prediction when all features possess a value of zero.

Table 25. ELI5 local interpretation for instance number 60.

Contribution	Feature
+1.716	Prickles
+1.496	Rows
+1.371	DOD_ID

Contribution	Feature
+0.011	OD_ID
-1.461	<BIAS>

5.5. GitHub Codes Link for Chapter 5

The following is the link to Chapter 5 codes:

https://github.com/nm2fs/PhD-Dissertation/tree/3c5a70bfffc03100ce5ac284f818f3495a29f6b5/chapter5

5.6. Summary

This chapter has described the risks associated with AI in the *Structure (Sig)* layer and introduced a case study focusing on optimizing the design of Vaso-Lock, a 3D printed prototype aimed at simplifying vascular anastomoses during surgical procedures. The study utilizes scenario-based analysis to identify highly ranked initiatives and the most disruptive scenarios. Bayesian optimization has been incorporated into the optimization process to navigate uncertain environments, followed by the use of explainable AI techniques to enhance interpretability and trust in model predictions.

The next chapter will introduce the *Function (Phi)* layer.

Chapter 6 | Case 3 (*Function (Phi)* Layer)

6.1.Introduction

This chapter describes a case study in the *Function (Phi)* layer. This layer includes a specific operation or a task defined and performed by medical professionals, such as disease diagnosis, particularly in this case, the diagnosis of cardiac sarcoidosis.

Sarcoidosis is an inflammatory, granulomatous systemic disease of unclear etiology with a heterogeneous course primarily affecting the lungs and lymph nodes (90%) [5,6,40,68,117–119] and invading the heart, leading to injury and fibrosis. According to [120] "a granuloma is a focal aggregate of immune cells that forms in response to a persistent inflammatory stimulus." In rare cases, sarcoidosis can be chronic and progress with multiorgan involvement, often associated with extensive scarring, such as in the liver, skin, eyes, central nervous system, and heart. Most sarcoidosis patients have a short, self-limiting disease

course without permanent damage. However, a chronic or relapsing course can also be observed. Therefore, sarcoidosis patients should be screened for multiorgan involvement. Cardiac involvement is statistically relatively rare, with 5% of sarcoidosis patients described [119,121]. Early diagnosis of cardiac sarcoidosis is crucial for timely treatment, prevention of cardiac damage, and management of potentially life-threatening complications. Also, it improves the overall management and long-term prognosis of cardiac sarcoidosis [5,6]. This chapter develops the third layer of the scenario-based disruption of priorities for disease diagnosis or the *Function (Phi)* layer.

6.2. Scenario-Based Disruption of Priorities (Function (Phi) Layer)

In this chapter, cardiac sarcoidosis is a resemblance to the *Function (Phi)* layer in the healthcare system [6,9,34,83–85]. The experts and actors for the *Function (Phi)* layer are director members of the cardiac radiology department, including two radiologists, a cardiologist, and an electrophysiologist who are experts and actors in cardiac sarcoidosis detection at the HDZ-NRW hospital in Germany. Five interview sessions were conducted through an online platform with the director of the cardiac radiology department at HDZ-NRW hospital.

To develop scenario-based analysis for cardiac sarcoidosis, the success criteria list is similar to the *Purpose (Pi)* and *Structure (Sig)* layers, which are the seven *principles* of the NIST AI Risk Management Framework [6,40,68]. Identifying the highest ranked initiatives in diagnosing cardiac sarcoidosis and the most and least disruptive events will assist the decision-makers in determining where to invest more for the most desired outcomes for the experts and actors.

Seven success criteria (Table 6), forty-three initiatives (Table 26), fifty emergent conditions (Table 27), and ten scenarios (Table 28) were identified. Then, baseline relevance (Table 29), criteria-initiative assessment (Table 30), criteria-scenario relevance (Table 31), and initiative-scenario ranking (Table 32) were developed for risk management of AI in cardiac sarcoidosis diagnosis *(Function (Phi)* layer [6,40,68].

Table 26. Initiatives for the *Function (Phi)* layer in enterprise risk management of AI in healthcare [6,40,68]. Abridged from various sources that are identified in the narrative [6,40,68].

Index	Initiative
x.Phi.01	Identify At-Risk Components
x.Phi.02	Understanding ML Tools to Uncover Subtle Patterns in Data
x.Phi.03	Maintaining the Provenance of Training Data
x.Phi.04	Safety/Verifiability of Automated Analyses (Cardiac Region Detection Software)
x.Phi.05	Reproducible Data and Method in Other Health Centers
x.Phi.06	Correctly Labeling the Data
x.Phi.07	Training Data to Follow Application Intellectual Property Rights Laws
x.Phi.08	Informed Consent to Use Data
x.Phi.09	Maintain Organizational Practices Like Implement Risk Management to Reduce Harm Reduction and More Accountable Systems
x.Phi.10	Prioritization Policies and Resources Based on Assesses Risk Levels
x.Phi.11	Safety of Personally Identifiable Information
x.Phi.12	Appropriate Accountability Mechanism, Roles and Responsibilities, Culture, and Incentive Structures for Risk Management to be Effective
x.Phi.13	Avoid Gender and Age Discriminations and Bias in Preparing Data
x.Phi.14	Reducing Unnecessarily Procedures
x.Phi.15	Reducing Costs and Time Consumption
x.Phi.16	Able to Identify Healthy Volunteers before Starting the Procedures
x.Phi.17	Designate Boundaries for AI Operation (Technical, Societal, Legal, and Ethical)
x.Phi.18	To Help Policymakers Ensure that the Moral Demanding Situations Raised by Enforcing AI in Healthcare Settings are Tackled Proactively
x.Phi.19	Articulate and Document the System Concept and Objectives, Underlying Assumptions, and Context in Light of Legal and Regulatory Requirements and Ethical Considerations
x.Phi.20	Gather, Validate, and Clean Data and Document the Metadata and Characteristics of the Dataset, in Light of Objectives, Legal and Ethical Considerations

x.Phi.21	Pilot, Check Compatibility with Legacy Systems, Verify Regulatory Compliance, Manage Organizational Change, and Evaluate User Experience
x.Phi.22	Operate the AI System and Continuously Assess its Recommendations and Impacts
x.Phi.23	Balancing and Tradeoff Each of Trustworthy AI Systems Characteristics Based on the AI System Context of Use
x.Phi.24	Reducing the Hospitalization Time of the Patient by Correct Diagnostics
x.Phi.25	Explain and Identify Most Important Features Using AI Models
x.Phi.26	Measurements Outlier Findings
x.Phi.27	Closeness of Results of Observations, Computations, or Estimates to the True Values or the Values Accepted as being True
x.Phi.28	Human-AI Teaming
x.Phi.29	Demonstrate External Validity or Generalizable Beyond the Training Conditions
x.Phi.30	Ability of a System to Maintain its Level of Performance Under a Variety of Circumstances
x.Phi.31	Minimizing Potential Harms to People if it is Operating in an Unexpected Setting
x.Phi.32	Responsible Design, Development, and Deployment Practices
x.Phi.33	Clear Information to Deployers on Responsible Use of the System
x.Phi.34	Responsible Decision-Making by Deployers and End Users
x.Phi.35	Explanations and Documentation of Risks Based on Empirical Evidence of Incidents
x.Phi.36	Ability to Shut Down, Modify, or Have Human Intervention into Systems that Deviate from Intended or Expected Functionality
x.Phi.37	Human Roles and Responsibilities in Decision Making and Overseeing AI Systems Need to be Clearly Defined and Differentiated
x.Phi.38	AI Systems May Require More Frequent Maintenance and Triggers for Conducting Corrective Maintenance Due to Data, Model, or Concept Drift
x.Phi.39	Managing Risks from Lack of Explainability by Describing How AI Systems Functions Considering Users' Role, Knowledge, and Skill Level
x.Phi.40	Communicating a Description of Why an AI System Made a Particular Prediction or Recommendation
x.Phi.41	Securing Individual Privacy, Anonymity, and Confidentiality
x.Phi.42	De-Identification and Aggregation for Certain Model Outputs
x.Phi.43	Strengthened Engagement with Interested Parties and Relevant AI Actors
x.Phi.i	Others

Table 27. Emergent conditions were used to create sets of scenarios for the *Function (Phi)* layer in enterprise risk management of AI in healthcare [6,40,68]. Abridged from various sources that are identified in the narrative [6,40,68].

Index	Emergent Condition
e.Phi.01	Using Non-Important Features in Sarcoidosis Diagnostics as the Input
e.Phi.02	Improperly Labeling the Data in Surgery-Specific Patient Registries
e.Phi.03	Mis-Identification of Variables Used in Surgery-Specific Patient Registries
e.Phi.04	Misunderstanding AI
e.Phi.05	Limited Generalizability
e.Phi.06	Limitation in Types and Accuracy of Available Data
e.Phi.07	Expensive Data Collection
e.Phi.08	Time Consuming Data Collection
e.Phi.09	Policy and Regulation Changes
e.Phi.10	Difficult and Complex AI Algorithms Interpretability
e.Phi.11	Lack of AI Determination of Casual Relationships in Data at Clinical Implementation Level
e.Phi.12	Inability of AI in Providing an Automated Clinical Interpretation of its Analysis
e.Phi.13	Human Errors in Measurements
e.Phi.14	Abuse or Misuse of the Model or Data
e.Phi.15	Challenges with Training Data to be Subject to Copyright
e.Phi.16	Complicate Risk Measurement by Third Party Software, Hardware, and Data
e.Phi.17	Model Fails to Generalize
e.Phi.18	Lack of Consensus on Robust and Verifiable Measurement Methods for AI Trustworthiness
e.Phi.19	Mis-Identification of Different Risk Perspective in Early or Late Stages of AI Lifecycle
e.Phi.20	Difference Between Controlled Environment vs. Uncontrollable and Real-World Settings
e.Phi.21	Inscrutable Nature of AI Systems in Risk Measurements
e.Phi.22	Systematic Biases in Clinical Data Collection
e.Phi.23	Risk Tolerance Influence by Legal or Regulatory Requirements Changes
e.Phi.24	Unrealistic Expectations About Risk to Misallocate Resources
e.Phi.25	Residual Risk or Risk Remaining after Risk Treatment Directly Impacts End Users
e.Phi.26	Privacy Concerns Related to the Use of Underlying Data to Train AI Systems
e.Phi.27	The Energy and Environmental Implications Associated with Resource-Heavy Computing Demands
e.Phi.28	Security Concerns Related to the Confidentiality, Integrity, and Availability of the System and its Training and Output Data
e.Phi.29	General Security of the Underlying Software and Hardware for AI Systems
e.Phi.30	One-Size-Fits-All Requirements AI Model Challenges
e.Phi.31	Neglecting the Trustworthy AI Characteristics
e.Phi.32	Difficult Decisions in Tradeoff and Balancing Trustworthy AI Characteristics by Organizations

e.Phi.33	Subject Matter Experts and Actors Can Assist in the Evaluation of TEVV Findings and Work with Product and Deployment Teams to Align TEVV Parameters to Requirements and Deployment Conditions
e.Phi.34	Different Perception of the Trustworthy AI Characteristics Between AI Designer than the Deployer
e.Phi.35	Potential Risk of Serious Injury or Death Call
e.Phi.36	Presenting AI System Information to Humans is Complex
e.Phi.37	Data Poisoning
e.Phi.38	Negative Risk Stem from a Lack of Ability to Make Sense of, or Contextualize, System Output Appropriately
e.Phi.39	AI Allowing Inference to Identify Individuals or Previously Private Information About Individuals
e.Phi.40	Privacy Intrusions
e.Phi.41	Data Sparsity
e.Phi.42	Fairness Perceptions Difference Among Cultures and Applications
e.Phi.43	Computational and Statistical Biases Stem from Systematic Errors Due to Non-Representative Samples
e.Phi.44	Human-Cognitive Biases Relates to How the Experts and Actors Perceives AI System Information to Make a Decision
e.Phi.45	Lack of Access to the Ground Truth in the Dataset
e.Phi.46	Intentional or Unintentional Changes During Training
e.Phi.47	Increased Opacity and Concerns About Reproducibility
e.Phi.48	Computational Costs for Developing AI Systems and their Impact on the Environment and Planet
e.Phi.49	Inability to Predict or Detect the Side Effects of AI-Based Systems Beyond Statistical Measures
e.Phi.50	Over-Reliance on AI
e.Phi.i	Others

Table 28. Emergent conditions grouping for the *Function (Phi)* layer in enterprise risk management of AI in healthcare describes which emergent conditions fit in each scenario [6,40,68]. Abridged from various sources that are identified in the narrative [6,40,68].

	s.01 - Funding Decrease	s.02 - Government Regulation and Policy Changes	s.03 - Privacy Attacks	s.04 - Cyber Security Threats	s.05 - Changes in AI RMF	s.06 - Non-Interpretable AI and Lack of Human-AI	s.07 - Global Economic and Societal Crisis	s.08 - Human Errors in Design, Develop, Measurement and	s.09 - Uncontrollable Environment	s.10 - Expensive Design Process
e.Phi.01		✓								
e.Phi.02								✓		
e.Phi.03								✓		
e.Phi.04						✓				
e.Phi.05						✓				
e.Phi.06										✓
e.Phi.07	✓									✓
e.Phi.08										✓
e.Phi.09		✓			✓					
e.Phi.10						✓				
e.Phi.11						✓				
e.Phi.12						✓				
e.Phi.13						✓				
e.Phi.14						✓		✓		
e.Phi.15		✓	✓			✓				
e.Phi.16						✓				✓
e.Phi.17						✓			✓	
e.Phi.18		✓			✓	✓		✓	✓	✓
e.Phi.19					✓	✓		✓		
e.Phi.20									✓	
e.Phi.21						✓				
e.Phi.22						✓				
e.Phi.23		✓								
e.Phi.24						✓		✓		
e.Phi.25						✓		✓	✓	
e.Phi.26			✓							
e.Phi.27		✓					✓			✓
e.Phi.28			✓	✓						
e.Phi.29				✓						

e.Phi.30						✓		✓	✓	✓
e.Phi.31						✓		✓		
e.Phi.32	✓				✓	✓				✓
e.Phi.33						✓				
e.Phi.34						✓				
e.Phi.35								✓	✓	
e.Phi.36									✓	
e.Phi.37			✓	✓				✓		
e.Phi.38						✓				
e.Phi.39			✓							
e.Phi.40			✓							
e.Phi.41								✓	✓	✓
e.Phi.42		✓			✓	✓				
e.Phi.43	✓					✓		✓	✓	✓
e.Phi.44	✓					✓				
e.Phi.45	✓							✓		✓
e.Phi.46					✓	✓		✓		
e.Phi.47						✓		✓		
e.Phi.48	✓						✓			✓
e.Phi.49						✓	✓		✓	
e.Phi.50						✓				

Table 29. Baseline relevance for the *Function (Phi)* layer in enterprise risk management of AI in healthcare [6,40,68].

The criterion c.xx has	s.00 - Baseline	relevance among the other criteria
c.01 - *Safe* has	*high*	relevance
c.02 - *Secure & Resilient* has	*medium*	relevance
c.03 - *Explainable & Interpretable* has	*high*	relevance
c.04 - *Privacy Enhanced* has	*medium*	relevance
c.05 - *Fair - With Harmful Bias Managed* has	*medium*	relevance
c.06 - *Accountable & Transparent* has	*high*	relevance
c.07 - *Valid & Reliable* has	*high*	relevance

Table 30. The criteria-initiative assessment describes how well each initiative addresses the success criteria for the *Function (Phi)* layer in enterprise risk management of AI in healthcare. *Strongly agree* is represented by a filled circle (●), *agree* is represented by a half-filled circle (◐), *somewhat agree* is represented by an unfilled circle (○), and *neutral* is represented by a dash (━) [6,40,68].

	c.01	*c.02*	*c.03*	*c.04*	*c.05*	*c.06*	*c.07*
x.Phi.01	●	◐	○	○	○	○	○
x.Phi.02	○	━	○	━	━	◐	◐
x.Phi.03	●	━	○	◐	◐	◐	●
x.Phi.04	●	◐	●	○	◐	●	●
x.Phi.05	●	●	●	◐	◐	◐	●
x.Phi.06	●	●	◐	◐	◐	●	●
x.Phi.07	○	○	◐	○	○	○	◐
x.Phi.08	●	●	━	●	━	○	○
x.Phi.09	○	◐	◐	○	○	●	○
x.Phi.10	○	○	◐	○	○	◐	◐
x.Phi.11	●	●	━	●	○	◐	━
x.Phi.12	○	○	◐	━	━	○	◐
x.Phi.13	◐	◐	○	○	●	◐	◐
x.Phi.14	●	●	━	━	●	●	●
x.Phi.15	●	●	━	━	●	●	●
x.Phi.16	●	●	◐	○	●	●	●
x.Phi.17	○	○	◐	○	○	◐	◐
x.Phi.18	○	○	○	●	●	◐	○
x.Phi.19	○	○	○	●	●	◐	○
x.Phi.20	●	●	◐	◐	◐	●	●
x.Phi.21	○	○	◐	○	○	◐	◐
x.Phi.22	◐	◐	●	○	○	●	●
x.Phi.23	◐	◐	◐	◐	◐	◐	◐
x.Phi.24	●	●	●	◐	●	●	●
x.Phi.25	●	●	◐	◐	◐	●	●
x.Phi.26	●	●	●	○	○	●	●
x.Phi.27	●	●	◐	◐	◐	●	●
x.Phi.28	●	●	●	◐	◐	●	●
x.Phi.29	◐	●	●	◐	◐	●	●
x.Phi.30	●	●	●	━	━	●	●
x.Phi.31	●	●	◐	━	●	●	●
x.Phi.32	●	●	●	◐	○	●	●
x.Phi.33	○	●	●	○	○	●	●
x.Phi.34	◐	◐	●	○	○	◐	●
x.Phi.35	○	◐	◐	○	○	●	○
x.Phi.36	●	●	●	○	○	●	◐
x.Phi.37	●	●	●	○	○	●	◐

x.Phi.38	◐	●	●	○	○	○	◐
x.Phi.39	●	●	●	−	−	●	●
x.Phi.40	●	●	●	−	−	●	●
x.Phi.41	−	−	−	●	−	−	−
x.Phi.42	○	○	●	○	−	○	○
x.Phi.43	○	◐	●	○	○	◐	◐

Table 31. The criteria-scenario assessment describes how the scenarios influence the relevance of each success criterion for the *Function (Phi)* layer in enterprise risk management of AI in healthcare. *Decrease Somewhat* = DS, *Decrease* = D, *Somewhat Increase* = SI, *Increase* = I [6,40,68].

	s.01	*s.02*	*s.03*	*s.04*	*s.05*	*s.06*	*s.07*	*s.08*	*s.09*	*s.10*
c.01	D	SI	D	-	SI	DS	DS	D	DS	DS
c.02	D	SI	D	-	SI	D	DS	D	DS	DS
c.03	DS	SI	D	-	I	D	-	D	D	-
c.04	-	I	-	DS	-	-	-	-	DS	-
c.05	DS	I	-	-	SI	-	DS	-	DS	DS
c.06	D	SI	D	-	I	D	DS	D	D	DS
c.07	D	SI	D	-	I	D	DS	D	D	DS

Table 32. Initiative-scenario ranking chart. This table describes the ranking of each initiative under each scenario for the *Function (Phi)* layer in enterprise risk management of AI in healthcare. The green filled cells show a higher ranking and the red and orange filled cells indicate a lower ranking [6,40,68].

	s.00 - Baseline	s.01 - Funding Decrease	s.02 - Government Regulation and Policy Changes	s.03 - Privacy Attacks	s.04 - Cyber Security Threats	s.05 - Changes in AI RMF	s.06 - Non-Interpretable AI and Lack of Human-AI Communications	s.07 - Global Economic and Societal Crisis	s.08 - Human Errors in Design, Develop, Measurement and Implementation	s.09 - Uncontrollable Environment	s.10 - Expensive Design Process
x.Phi.01	37	38	38	35	34	35	32	38	35	35	38
x.Phi.02	42	43	42	43	42	41	43	42	43	42	42
x.Phi.03	25	17	25	15	25	25	14	33	15	25	33
x.Phi.04	4	12	11	13	4	4	13	6	13	10	6
x.Phi.05	11	4	4	11	8	12	5	4	11	5	4
x.Phi.06	4	6	4	5	8	12	5	12	5	5	12
x.Phi.07	39	40	39	36	38	34	36	34	36	39	34
x.Phi.08	38	18	37	27	39	42	27	39	27	34	39
x.Phi.09	29	31	31	30	29	29	30	26	30	32	26
x.Phi.10	31	33	34	32	31	31	33	28	32	36	28
x.Phi.11	36	15	33	18	37	40	18	37	18	29	37
x.Phi.12	41	42	41	42	41	39	42	36	42	41	36
x.Phi.13	27	30	27	17	27	28	17	35	17	27	35
x.Phi.14	22	36	21	19	21	22	19	40	19	20	40
x.Phi.15	22	36	21	19	21	22	19	40	19	20	40
x.Phi.16	4	13	4	5	4	6	5	22	5	4	22
x.Phi.17	31	33	34	32	31	31	33	28	32	36	28
x.Phi.18	31	10	29	3	35	36	3	31	3	30	31
x.Phi.19	31	10	29	3	35	36	3	31	3	30	31
x.Phi.20	4	6	4	5	8	12	5	12	5	5	12
x.Phi.21	31	33	34	32	31	31	33	28	32	36	28
x.Phi.22	19	21	19	24	19	19	24	10	24	19	10
x.Phi.23	24	16	24	14	24	24	15	23	14	24	23
x.Phi.24	1	1	1	1	1	1	1	1	1	1	1
x.Phi.25	4	6	4	5	8	12	5	12	5	5	12

x.Phi.26	11	14	12	21	4	4	21	6	21	10	6
x.Phi.27	4	6	4	5	8	12	5	12	5	5	12
x.Phi.28	2	2	2	2	2	2	2	2	2	2	2
x.Phi.29	4	4	4	5	8	7	11	4	5	10	4
x.Phi.30	16	27	16	37	14	9	37	17	37	16	17
x.Phi.31	13	25	13	16	7	8	16	25	16	13	25
x.Phi.32	3	3	3	12	3	3	12	3	12	3	3
x.Phi.33	20	22	20	25	20	20	25	11	25	22	11
x.Phi.34	21	23	23	26	23	21	26	16	26	23	16
x.Phi.35	29	31	31	30	29	29	30	26	30	32	26
x.Phi.36	14	19	14	22	17	17	22	8	22	14	8
x.Phi.37	14	19	14	22	17	17	22	8	22	14	8
x.Phi.38	26	24	26	28	26	26	28	20	28	26	20
x.Phi.39	16	27	16	37	14	9	37	17	37	16	17
x.Phi.40	16	27	16	37	14	9	37	17	37	16	17
x.Phi.41	43	39	43	40	43	43	40	43	40	43	43
x.Phi.42	40	41	40	41	40	38	41	24	41	40	24
x.Phi.43	27	26	28	29	27	27	29	21	29	28	21

Figure 42 describes that *s.06 – Non-Interpretable AI and Lack of Human-AI Communications, s.03– Privacy attacks,* and *08 – Human Errors in Design, Develop, Measurement, and Implementation* have the highest disruption among other scenarios [6,40,68].

Figure 43 describes the variation in the prioritization of initiatives across scenarios. Table 33 describes the highest ranking initiatives in the *Function (Phi)* layer.

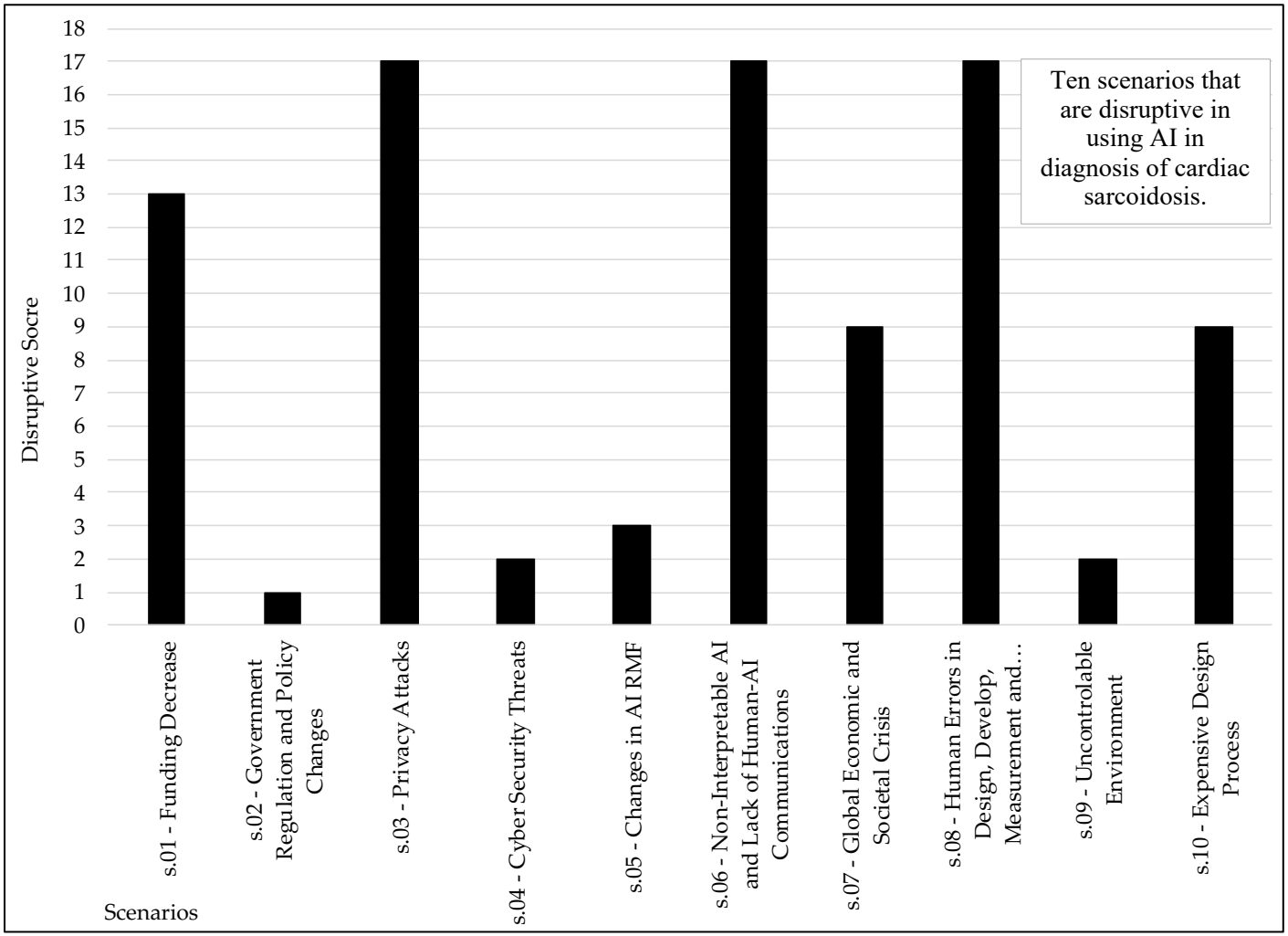

Figure 42. Disruptive score of scenarios is based on the sum of squared differences in the priority of initiatives, relative to the baseline scenario for the *Function (Phi)* layer in enterprise risk management of AI in healthcare. These are scenarios where they caused low levels of trust in AI [6,40,68].

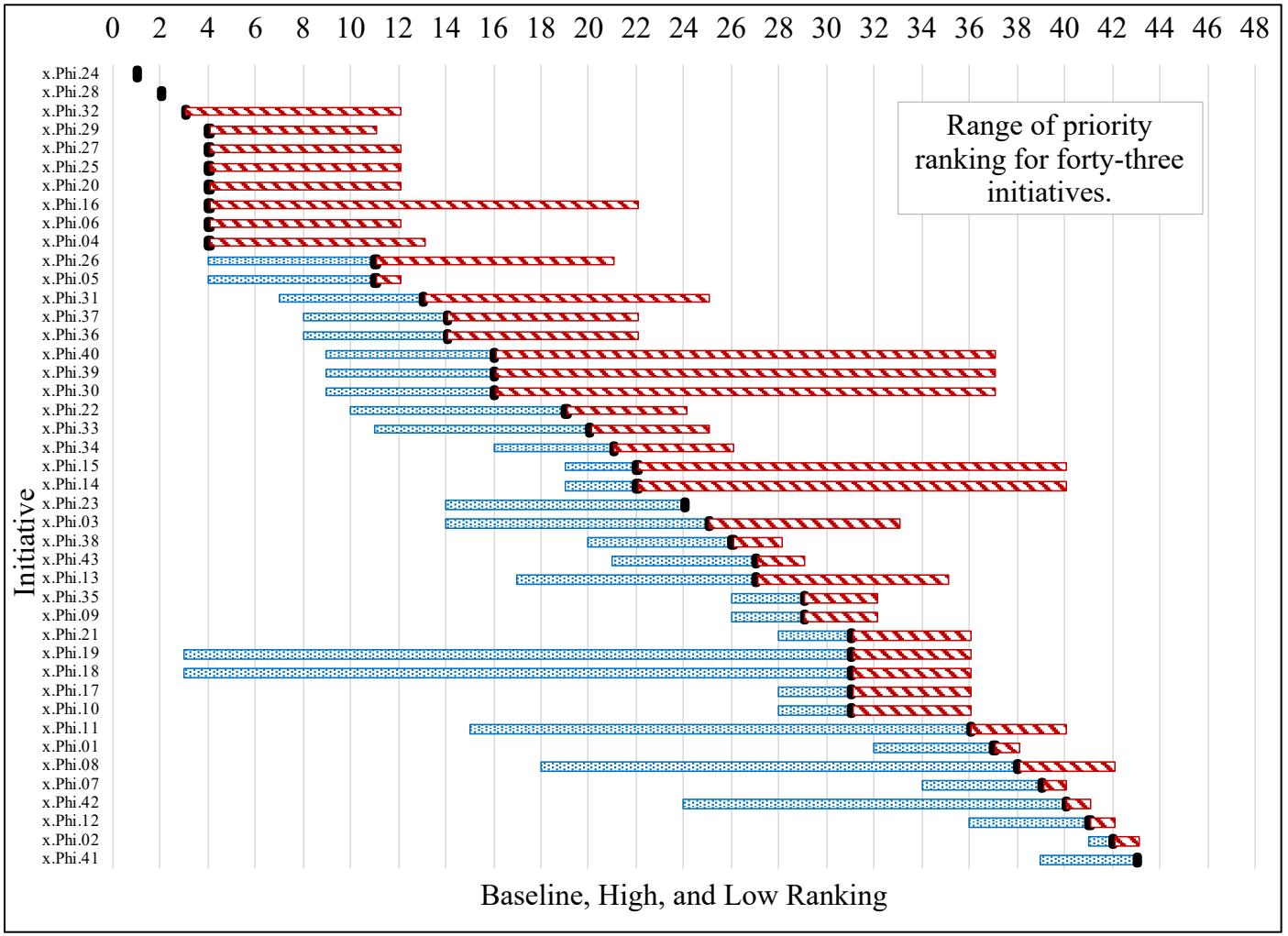

Figure 43. Distributions of initiatives influence rankings are based on which emergent conditions that could arise more often or do not occur for the *Function (Phi)* layer in enterprise risk management of AI in healthcare; blue means promotion in ranking and red means demotion in ranking [6,40,68].

Table 33. The highest ranked initiatives of the *Function (Phi)* layer in enterprise risk management of AI in healthcare [6,40,68].

Index	Most Important Initiative
Function (Phi)	***x.Phi.27 – Closeness of Results of Observations, Computations, or Estimates to the True Values or the Values Accepted as Being True***
	x.Phi.32 – Responsible Design, Development, and Deployment Practices
	x.Phi.29 – Demonstrate External Validity or Generalizable Beyond the Training Conditions
	x.Phi.28 – Human-AI Teaming
	x.Phi.25 – Explain and Identify Most Important Features Using AI Models
	x.Phi.24 – Reducing the Hospitalization Time of the Patient by Correct Diagnostics
	x.Phi.20 – Gather, Validate, and Clean Data and Document the Metadata and Characteristics of the Dataset, in Light of Objectives, Legal and Ethical Considerations
	x.Phi.16 – Able to Identify Healthy Volunteers Before Starting the Procedures
	x.Phi.06 – Correctly Labeling the Data
	x.Phi.04 – Safety/Verifiability of Automated Analyses (Cardiac region detection software)

In the next section, sensitivity analysis is conducted to show how robust the mathematical framework is for the *Function (Phi)* layer outcomes.

6.2.1. Sensitivity Analysis of Scenario-Based Disruption of Priorities (Function (Phi) Layer) Mathematical Framework

This section introduces a sensitivity analysis of the mathematical framework used in the *Function (Phi)* layer. Sensitivity analysis can be used to identify and assess the most significant exposure or risk factors, as well as the most resilient initiatives and priorities for risk mitigation [122]. Sensitivity analysis is an approach in this case to validate the results. This analysis is significant as it may be of interest to experts and actors who are concerned about the potential influence of different weighting scores on the outcomes. Two sets of success criterion relevance weightings were evaluated for two decision-makers: 1. Aggressive and assertive decision-makers. 2. Cautious decision-makers. The objective of this section is to assess how the rating of the relevance of success criteria will impact the baseline rankings of the initiatives.

6.2.1.1. AGGRESSIVE DECISION-MAKER

Table 34 describes that the success criterion relevance importance weight for *"High"* was to be changed by aggressive decision-makers from 4 to 10. Figure 44 describes how the initiatives baseline rankings were altered in accordance with the changes in the relative importance weights. The green line shows the initial weighting with *"High"* weights of 4, and the red line shows the weight changes of *"High"* from 4 to 10 by the aggressive decision-makers. As an example, initiatives *x.Phi.24 – Reducing the Hospitalization Time of the Patient by Correct Diagnostics*, *x.Phi.28 – Human-AI Teaming*, and *x.Phi.32 – Responsible Design*,

Development, and Deployment Practices are resistant to ranking shifts; while some initiatives such as *x.Phi.29 – Demonstrate External Validity or Generalizable Beyond the Training Conditions,* and *x.Phi.27 – Closeness of Results of Observations, Computations, or Estimates to the True Values or the Values Accepted as Being True* dropped in baseline rankings. This analysis could assist in identifying initiatives that are more resistant to increasing success criteria relative to *"High"* weight. Overall, the majority of the rankings of the initiatives remained consistent, indicating the robustness of the framework.

Table 34. Criteria-scenario relative importance weights.

Criteria Scenario Relative Importance	Weights
High	**10**
Medium	2
Low	1
-	0

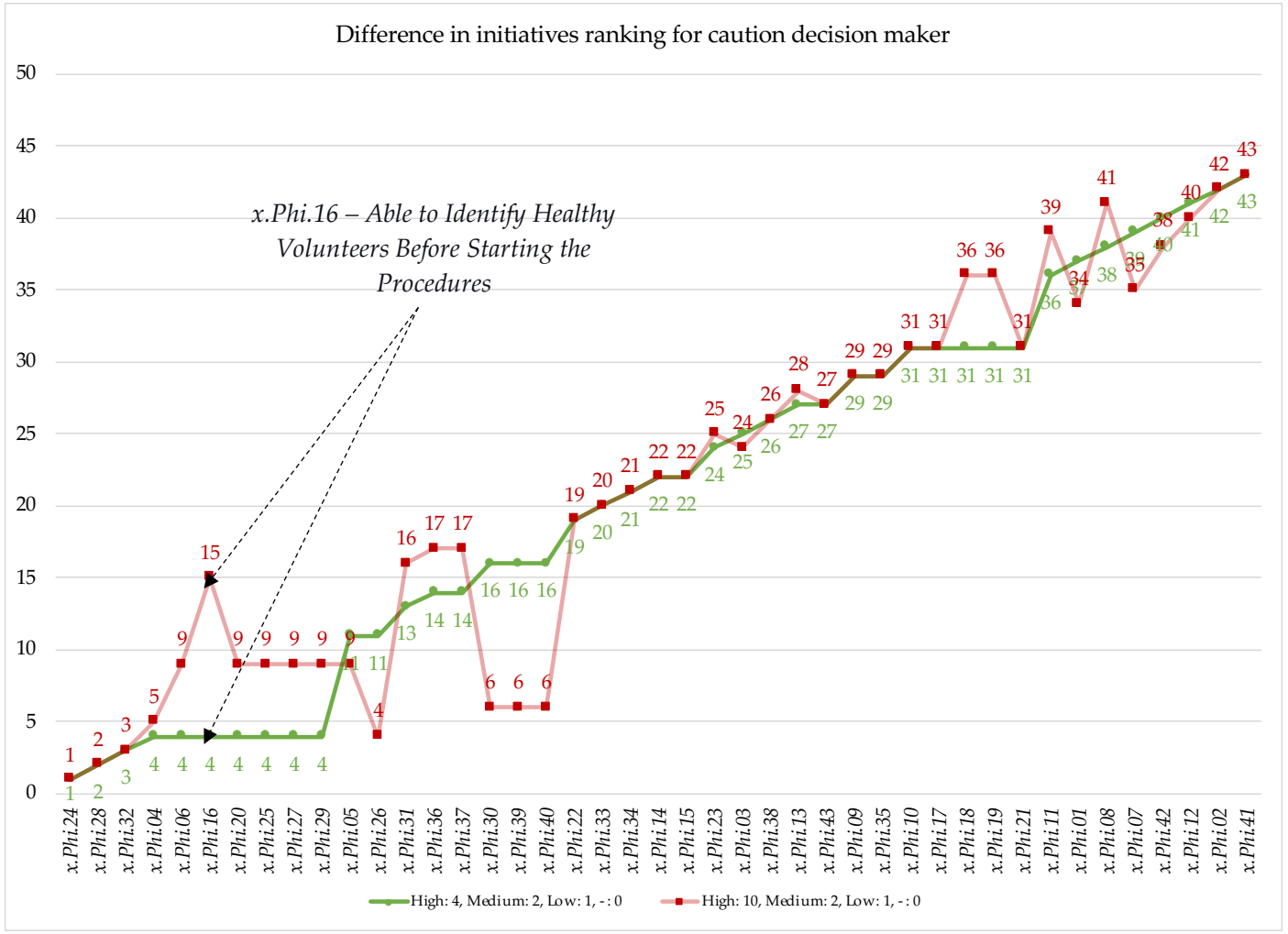

Figure 44. Change in priority of initiatives for the *Function (Phi)* layer for an aggressive analyst. The framework is robust as most of the initiatives ranking did not change by increasing the criteria-scenario relative importance weights.

6.2.1.2. CAUTIOUS DECISION-MAKER

Following the section above, the same analysis assesses initiative ranking shifts for cautious decision-makers. Table 35 describes the success criterion relevance importance weight for *"High"* that was changed by cautious decision-makers from 4 to 3. Figure 45 describes how the baseline initiatives rankings were altered in accordance with the changes in the relative importance weights. The green line shows the initial weighting with *"High"* weights of 4, and the blue line shows the weight changes of *"High"* from 4 to 3 by the cautious decision-makers. Sensitivity analysis confirmed that most of the initiatives did not shift in ranking. However, initiatives *x.Phi.04 – Safety/Verifiability of Automated Analyses (Cardiac region detection software)* and *x.Phi.16 – Able to Identify Healthy Volunteers Before Starting the Procedures* were most vulnerable to the changes in the criteria relative importance of *"High"* weight.

Table 35. Criteria-scenario relative importance weights.

Criteria Scenario Relative Importance	Weights
High	**3**
Medium	2
Low	1
-	0

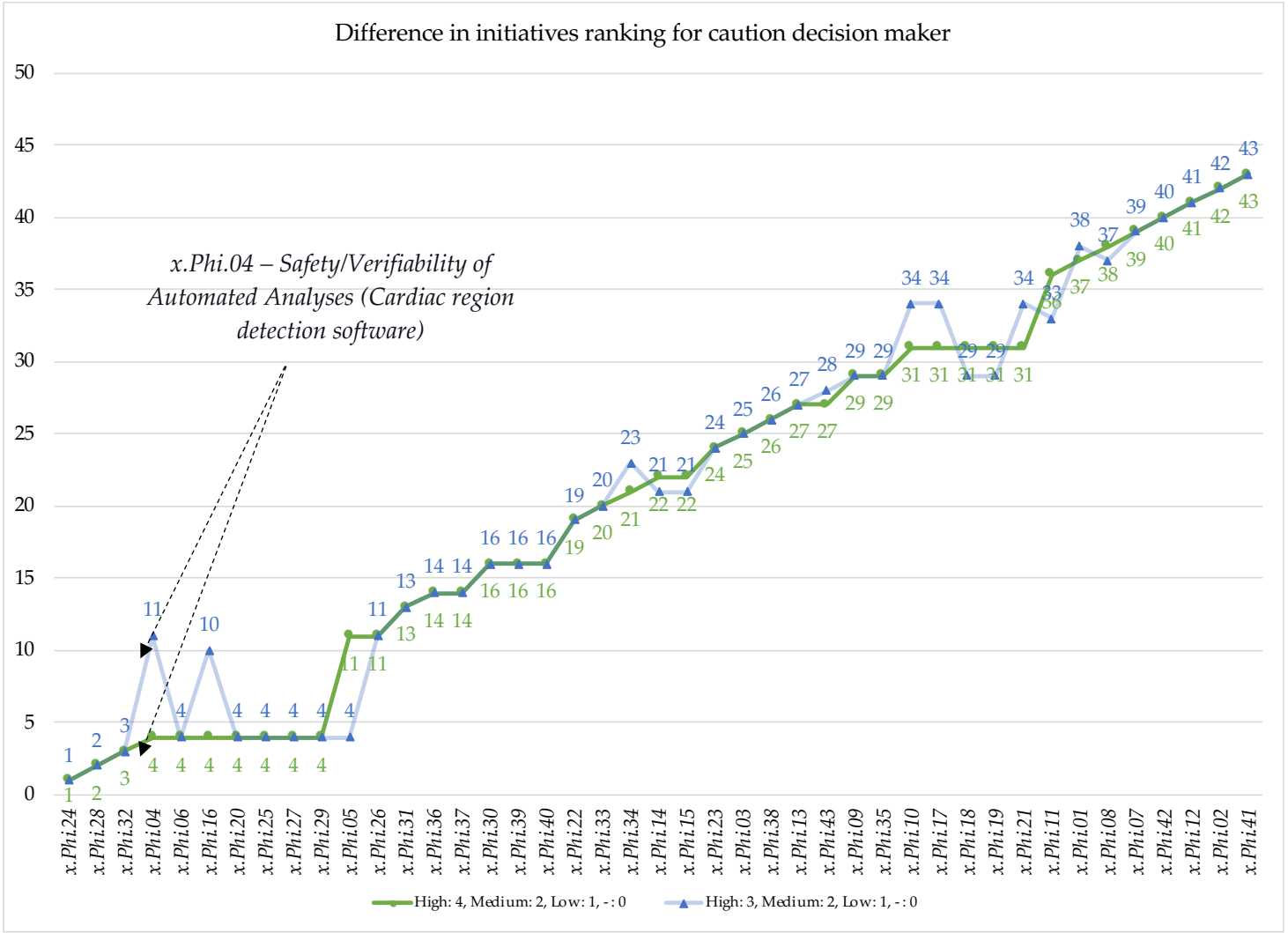

Figure 45. Change in priority of initiatives for the *Function (Phi)* layer for a cautious analyst. The framework is robust as most of the initiatives ranking did not change by decreasing the criteria-scenario relative importance weights.

It could be implied from the sensitivity analysis that the multicriteria analysis, in this case, is robust as most of the initiatives did not shift in ranking. As described in Chapter 2, two of the NIST AI Risk Management Framework *principles* include the requirement for the system to be robust and resilient. The sensitivity analysis in this case demonstrated that it is robust and resistant to risks, which was consistent with these *principles*.

The next section addresses the highest ranked initiatives identified in the previous section for the *Function (Phi)* layer using AI classification models for early detection of cardiac sarcoidosis.

6.3. Diagnosis of Cardiac Sarcoidosis Using AI Classification Models

Forty-five non-cardiac (NC, 56.5 (53.0; 63.0) years) sarcoidosis, eighteen cardiac sarcoidosis patients (CS, 64.0 (57.8; 67.0) years) patients, and forty-four healthy controls volunteers (CTRL, 56.5 (53.0; 63.0) years) underwent contrast-enhanced cardiac magnetic resonance (CMR) examination. Bi-Atrial and left ventricular strains and volumetrics of all cardiac chambers were assessed by algorithmic processing using classifiers such as support vector machine (SVM), K-Nearest neighbour (KNN), decision tree (DT), random forest (RF) logistic regression (LR), GBoost, XGBoost, and Voting [5].

Figure 43 described that initiative *x.Phi.16 – Able to Identify Healthy Volunteers Before Starting the Procedures* was one of the most identified initiatives. Table 37 describes competitive prediction rates achieved for discriminating between CTRL and all sarcoidosis patients (S) via 37 features, including age. Figure 47 describes that logistic regression, random forest, and SVM yielded the highest prediction

rates in the two-cluster (96.97%) model for CTRL and S diagnosis discrimination, further consolidated by confusion matrix data (e.g., two-cluster model; precision = 97%; recall = 97%; F1-score = 97% for CTRL and S). Random forest and Voting yielded the highest prediction rates in a three-cluster (81.82%) of CTRL, NC, and CS (Figure 46). Figure 48 describes poor algorithmic discrimination between NC and CS, with the highest prediction rates for decision trees, GBoost, and XGBoost (68.42%). To enhance the prediction rates between NC and CS, feature selection is used to reduce the complexity of the data and select the most contributing features to the prediction of the output. Thus, the random forest classifier selects the five most important features. Table 36 describes the five most important features with the highest algorithmic impact for NC versus CS. Features with higher importance scores are considered more influential in making predictions, while those with lower scores have less impact. All five parameters were based on left ventricular cardiac motion or volumetrics. Mainly, longitudinal strain rates of the left ventricle appeared to have a high discriminative value for machine learning algorithms. Among the most important features for both analyses was the routinely available indexed left ventricular end systolic volume [5]. The full names of the five features below are as follows:

- LV_ESVi – indexed left ventricular end systolic volume,
- LA_syst_radial_LAX_SR – left atrial systolic radial strain rate in longitudinal axis,
- LA_diast_radial_LAX_SR – left atrial diastolic radial strain rate in longitudinal axis,
- LV_radial_LAX_S – left ventricular radial strain in longitudinal axis,

- LA_syst_long_LAX_SR – left atrial systolic longitudinal strain rate in the longitudinal axis [5,40].

Table 36. Feature rates of the five most important parameters for machine learning discrimination generated by the random forest classifier [5].

Parameter	Feature Rates
LV_ESVi	0.045
LV_syst_radial_LAX_SR	0.045
LV_diast_radial_LAX_SR	0.049
LV_radial_LAX_S	0.056
LA_syst_long_LAX_SR	0.060

A next step is the eight classifying ML models that predict the output using the reduced features dataset. The results show an enhancement. Figure 49 shows that logistic regression yielded the highest prediction rates in the two-cluster (89.47%) model for NC and CS diagnosis discrimination.

Table 37. Eight classification algorithm performance on diagnosis of cardiac sarcoidosis [5,40].

		Logistic Regression	*KNN*	*Decision Tree*	*Random Forest*	*SVM*	*GBoost*	*XGBoost*	*Voting*
Three clusters CTRL, NC Sarc., Sarc. (with age)	*Accuracy%*	69.70%	69.70%	75.76%	**81.82%**	75.76%	78.79%	78.79%	**81.82%**
	Precision Weighted-Avg	0.70	0.77	0.76	**0.86**	0.64	0.81	0.84	**0.86**
	Recall Weighted-Avg	0.70	0.70	0.76	**0.82**	0.76	0.79	0.79	**0.82**
	F1-Score Weighted-Avg	0.67	0.65	0.75	**0.79**	0.68	0.78	0.76	**0.79**
Two clusters CTRL versus all Sarc. (with age)	*Accuracy%*	**96.97%**	87.88%	87.88%	**96.97%**	**96.97%**	87.88%	90.91%	93.94%
	Precision Weighted-Avg	**0.97**	0.88	0.89	**0.97**	**0.97**	0.89	0.91	0.94
	Recall Weighted-Avg	**0.97**	0.88	0.88	**0.97**	**0.97**	0.88	0.91	0.94
	F1-Score Weighted-Avg	**0.97**	0.88	0.88	**0.97**	**0.97**	0.88	0.91	0.94
Two clusters NC Sarc. versus Sarc. (with age)	*Accuracy%*	52.63%	47.37%	**68.42%**	57.89%	63.16%	**68.42%**	**68.42%**	63.16%
	Precision Weighted-Avg	0.60	0.53	**0.68**	0.63	0.61	**0.73**	**0.68**	0.65
	Recall Weighted-Avg	0.53	0.47	**0.68**	0.58	0.63	**0.68**	**0.68**	0.63
	F1-Score Weighted-Avg	0.55	0.50	**0.68**	0.60	0.62	**0.70**	**0.68**	0.64

Two clusters NC Sarc. versus Sarc. (with age) (Enhanced with Feature Selection)	*Accuracy%*	**89.47%**	78.95%	68.42%	78.95%	84.21%	73.68%	68.42%	78.95%
	Precision Weighted-Avg	**0.89**	0.83	0.68	0.83	0.86	0.80	0.78	0.83
	Recall Weighted-Avg	**0.89**	0.79	0.68	0.79	0.84	0.74	0.68	0.79
	F1-Score Weighted-Avg	**0.89**	0.80	0.68	0.80	0.85	0.75	0.70	0.80

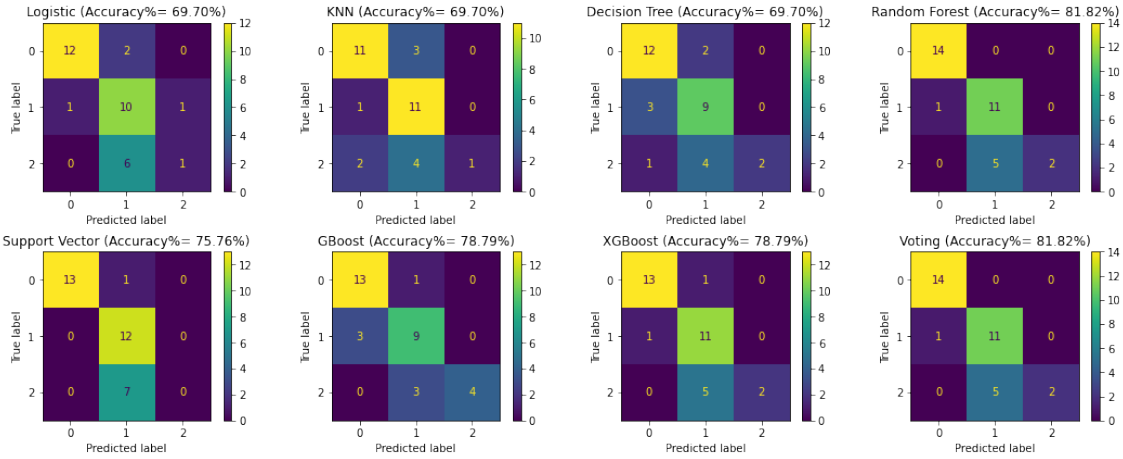

Figure 46. Three clusters: CTRL, NC Sarc., Sarc. (with age) confusion matrix.

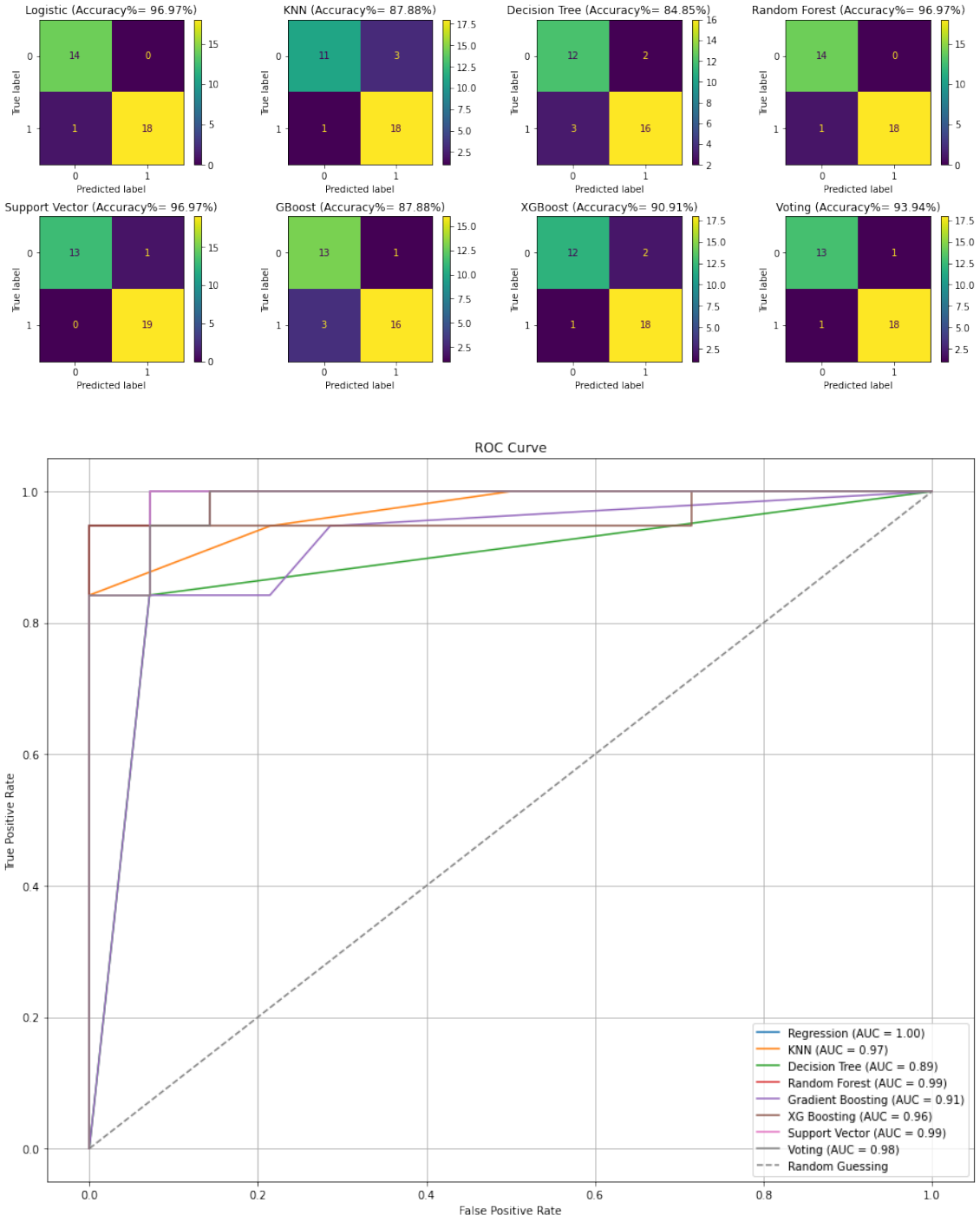

Figure 47. Two clusters: CTRL versus all Sarc. (with age) confusion matrices and ROC curve.

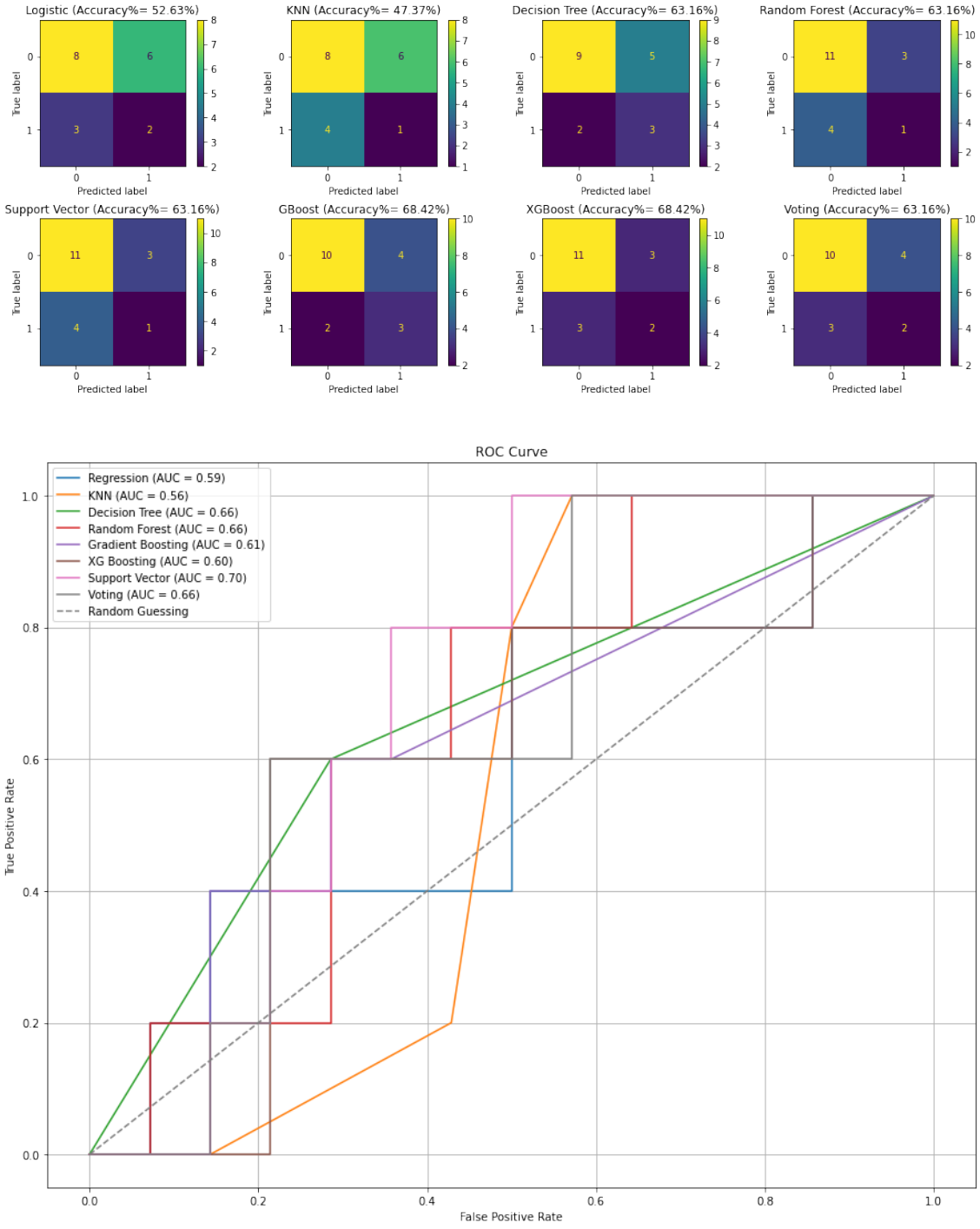

Figure 48. Two clusters: NC Sarc. versus Sarc. (with age) confusion matrices and ROC curve.

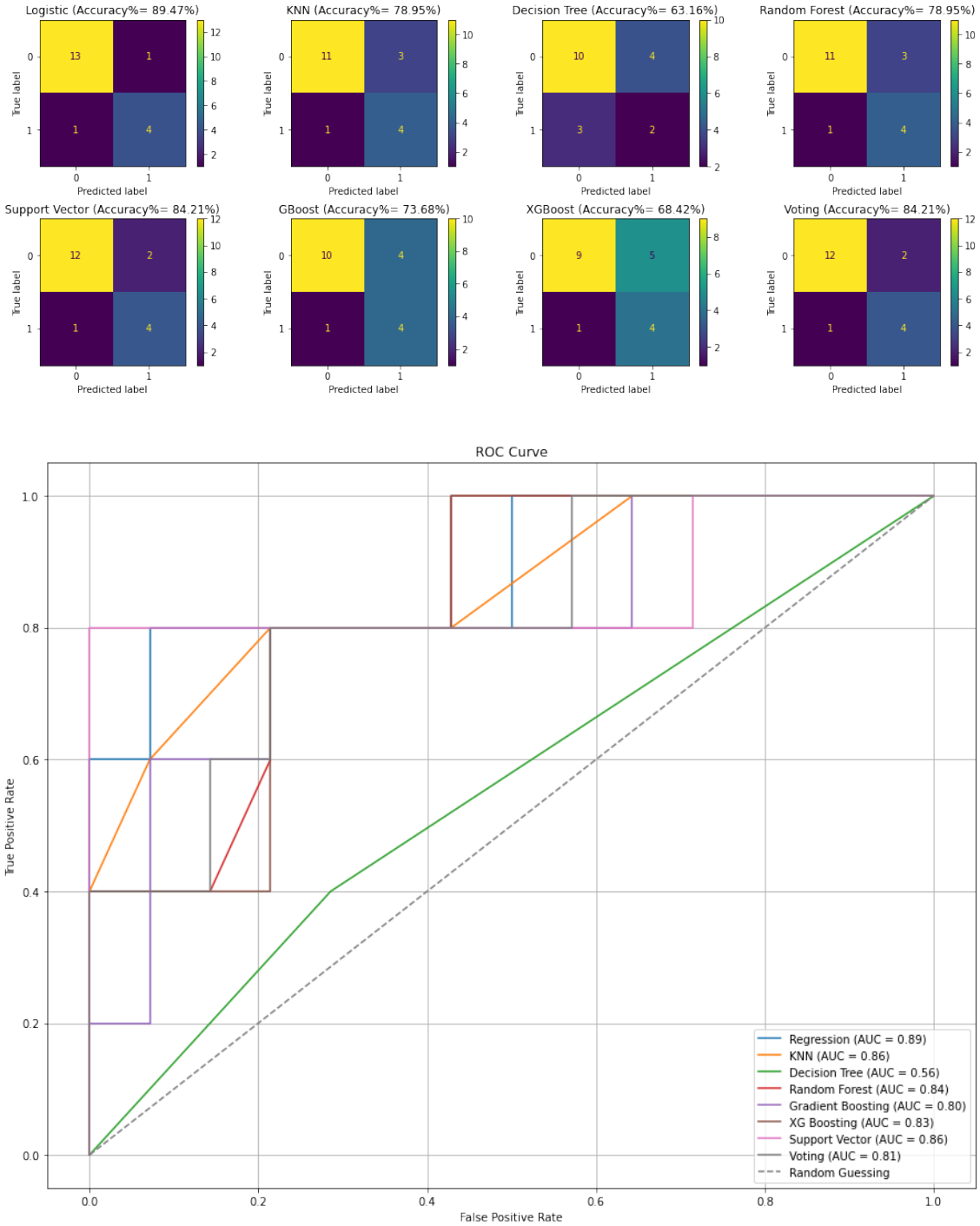

Figure 49. Two clusters: NC Sarc. versus Sarc. (with age) (enhanced with feature selection) confusion matrices and ROC curve. AUC score increases after reducing the complexity of the data by selecting five most correspondent variables to the classification predictions.

It is important to recognize that the effectiveness of machine learning models, like K-Nearest Neighbors (KNN), can be heavily influenced by factors such as hyperparameter configurations, feature selection, data preprocessing techniques, and the choice of evaluation metrics. If the performance of KNN, as depicted in Figure 48 , is lower than expected, it may be due to insufficient optimization of its hyperparameters, resulting in subpar performance. Nevertheless, it is worth noting that fine-tuning models can be resource-intensive and may not always yield substantial improvements. Therefore, refining the model through hyperparameter tuning in future work is a logical step towards potentially enhancing performance. It is essential to interpret the current findings with an awareness of this limitation.

Employing hard classification metrics in Figure 47, Figure 48, and Figure 49 can present ethical intricacies and necessitates meticulous deliberation. Hard classification is the process of directly assigning instances to specific classes without including any additional context or uncertainty estimates. When faced with situations where decisions made using these classifications can have substantial real-world effects, it is crucial to evaluate the ethical ramifications. Below are several crucial factors to take into account:

1. Transparency and explainability: The provision of clear and understandable information about the decision-making process is essential for establishing confidence in machine learning systems. The use of hard classification may not offer adequate elucidation regarding the rationale behind a specific decision, resulting in a lack of comprehension and potential mistrust among users.

2. Uncertainty estimation: Rather than solely depending on hard classifications, providing uncertainty estimates or probability scores can

provide more informative insights. Decision-makers can utilize these probabilities in conjunction with their expertise in the field to make more knowledgeable decisions and evaluate the corresponding risks.

3. Distrust caused by eisclassification: Hard classification metrics may not effectively capture the actual costs in the real world that are linked to misclassifications. Various categories of misclassifications can result in different outcomes, and making the assumption that all misclassifications have the same costs can result in decisions that are not optimal.

4. User involvement: When employing machine learning algorithms to facilitate decision-making, it is crucial to engage users and stakeholders in the process. Gaining insight into their preferences, concerns, and the possible consequences of misclassifications can assist in customizing the system to more effectively fulfill their requirements.

5. Ethical considerations: The design and deployment of machine learning systems should be guided by ethical considerations. This encompasses the guarantee of equity, responsibility, and openness at every stage of the process. Engaging in hard classification without taking into account the wider circumstances and possible repercussions can result in unethical results.

Although hard classification metrics offer a straightforward method for assessing machine learning algorithms, their application in decision-making should be exercised with caution. To address ethical concerns and foster trust in machine learning systems, it is beneficial to provide decision-makers with supplementary context, uncertainty estimates, and involve them in the decision-making process.

To our knowledge, this dissertation is the first to demonstrate diagnostic prediction rates for cardiac sarcoidosis based on CS acquired multi-chamber wall motion and volumetrics analyses using supervised machine learning algorithms.

The case study of the *Function (Pi)* layer thus has the following innovations:

1. Accurate algorithmic discrimination is achieved between healthy subjects and all sarcoidosis patients, particularly with Voting and RF classifiers.
2. Poor algorithmic discrimination of NC and CS patients is improved to accurate levels via algorithmic feature selection application, mainly using logistic regression and SVM classifiers.
3. The algorithmic challenge associated with discrimination between both patient groups implies cardiac involvement may be more prevalent than anticipated, potentially evading CS detection [5].

The next section describes how to utilize XAI to address the most disruptive scenarios identified by the risk register for the *Function (Phi)* Layer.

6.4. Explainable AI (XAI) of the Scenarios that are Most Disruptive to System Order

This section describes explainable AI (XAI) techniques used for the risks of AI in the healthcare *Function (Pi)* layer. Previously, Figure 43 described that one of the highest ranked initiatives was *x.Phi.25 – Explain and Identify Most Important Features Using AI Models.* Also, in Figure 42, one of the most disruptive scenarios was *s.06 – Non-Interpretable AI and Lack of Human-AI Communications.* Thus, to

interpret the results for the AI users, perform feature importance, enhance understanding of AI outputs, mitigate distrust in the AI outputs, and facilitate data evaluation, Shapley additive explanations (SHAP), local interpretable model-agnostic explanations (LIME), and Anchors models were implemented, and individual and global contributions of each feature were evaluated. These XAI methods were used to find the most important features that mainly corresponded with the prediction, and the results also explain how the AI model came up with this decision to classify the patient as if they were diagnosed with cardiac sarcoidosis.

6.4.1. SHAP Analysis of the Scenarios that are Most Disruptive to System Order

In Chapter 5, SHAP was introduced. A positive SHAP value implies a positive impact on prediction. For instance, if the SHAP value in the analysis of cardiac sarcoidosis as a continuous target variable in the NC versus CS analysis approaches 1, it suggests that these features are strong positive predictors of cardiac sarcoidosis. As logistic regression yielded the highest prediction rates in the two-cluster (89.47%) model for NC and CS diagnosis discrimination, a global SHAP plot in

Figure 50 was developed to understand the importance or contribution of the selected features. The results explain five features that have the most contribution to NC and CS diagnosis discrimination. This figure describes that LA_diast_radial_LAX_SR has the highest contribution to the predictions, followed by LV_radial_LAX_S and LA_syst_long_LAX_SR. In this figure, red dots implied higher values, and blue dots implied lower values for the features.

Thus, when LA_diast_radial_LAX_SR feature values are high (red), they have a strong positive impact on the prediction. On the contrary, when LV_radial_LAX_S and LA_syst_long_LAX_SR feature values are low (blue), they have a strong negative impact on the prediction [5].

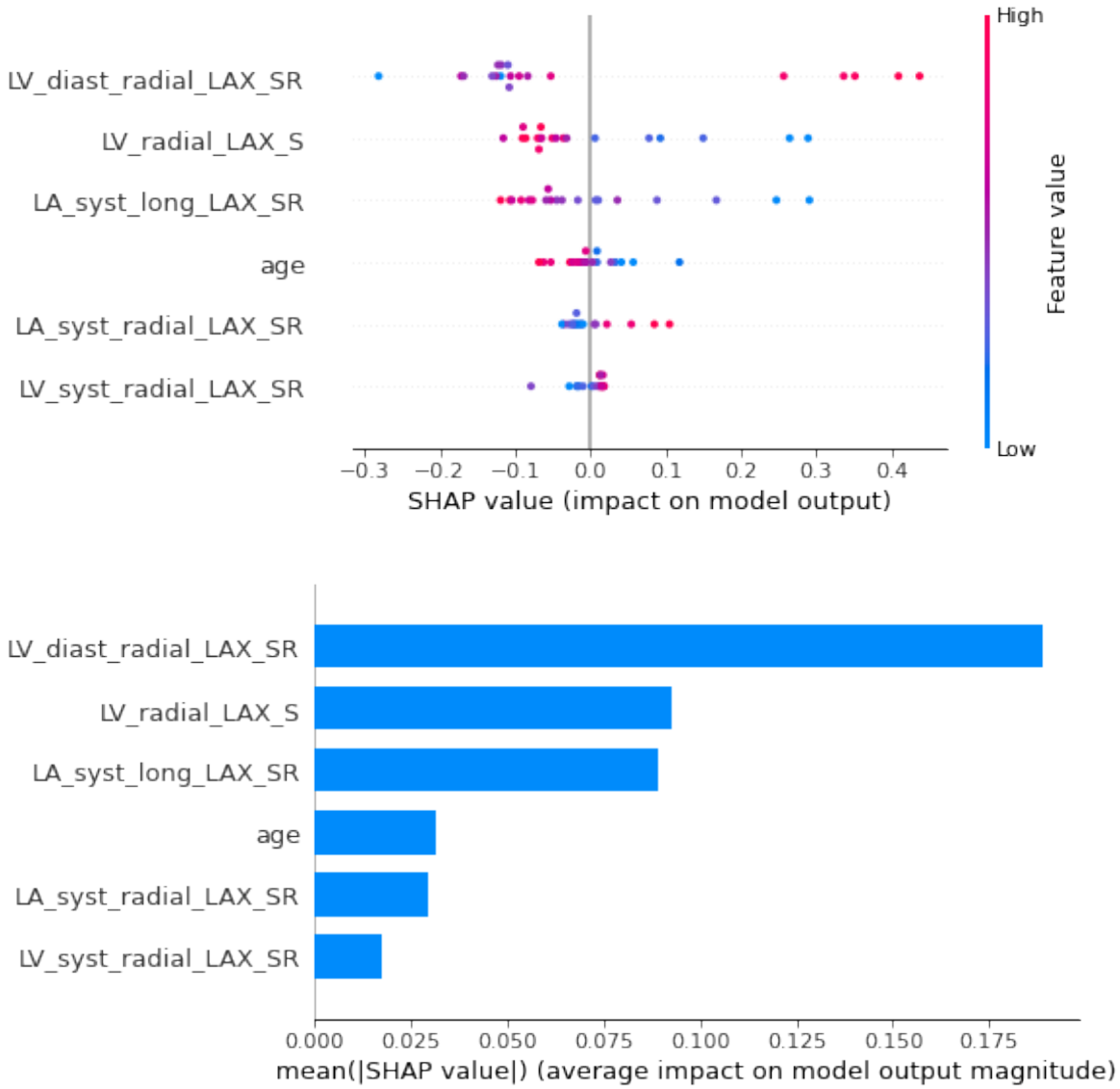

Figure 50. Global SHAP analysis of feature contributions of NC and CS diagnosis discrimination [5]. Top six most corresponding variables to the prediction of the cardiac sarcoidosis.

6.4.2. LIME Explainer of the Scenarios that are Most Disruptive to System Order

Following SHAP, which was utilized for global explanation in the previous section, LIME is helpful in interpreting the predictions and disease outcomes for each patient as an individual case. The LIME [40,49,123] plot describes how various features influence the predicted disease diagnosis of the patient. Each feature is depicted by a colored bar, where red signifies a negative impact (class 0: CTRL, class 1: NC, and class 2: CS), and green indicates a positive impact. The length of the bar corresponds to the strength of the impact, with longer bars indicating a more pronounced influence. For instance, Figure 51 (top figure) is the local interpretation for a patient with an instance number of 30, which describes that RV_EF has the most negative impact and RA_long_LAX has the most positive impact on the prediction of class 1, which indicates that the patient is not diagnosed with cardiac sarcoidosis. This result addresses the objective of *x.Phi.16 – Able to Identify Healthy Volunteers Before Starting the Procedures*.

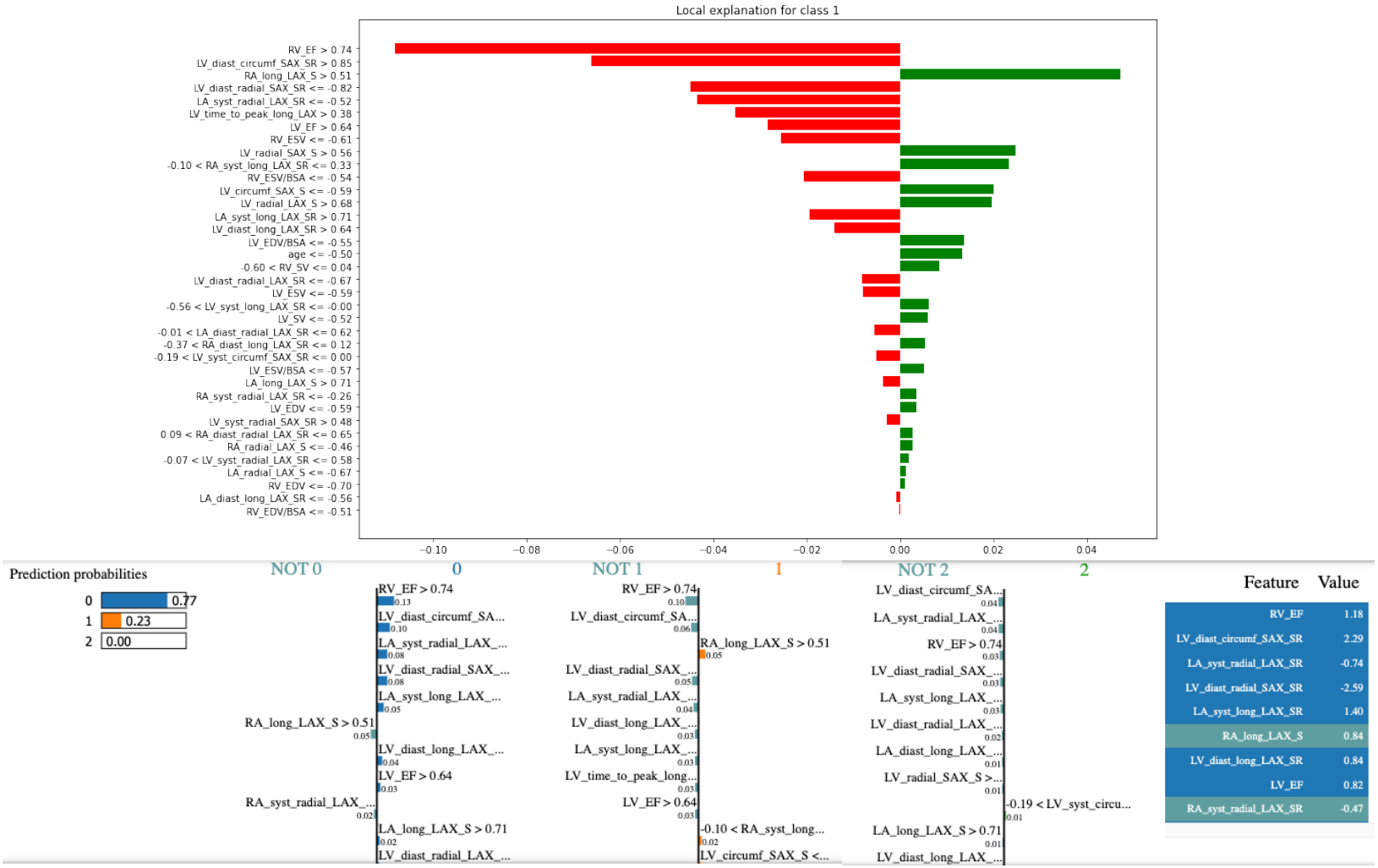

Figure 51. Local explanation of a patient with index number 30 prediction class using LIME; three clusters CTRL, NC Sarc., Sarc. (with age). RV_EF and LV_diast_circumf_SAX_SR are the most negative correspondence to the prediction of patient 30 cardiac sarcoidosis status.

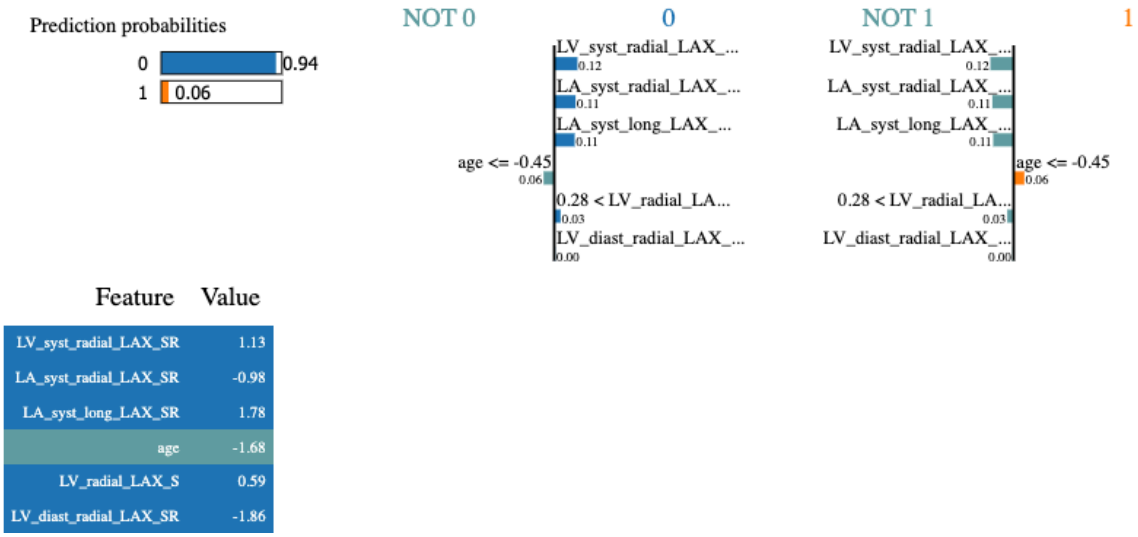

Figure 52. Local explanation of a patient with index number 10 prediction class using LIME; two clusters NC Sarc. versus Sarc. (with age) (enhanced with feature selection). Higher value of the most important features (except age) correspond to higher prediction probability that the patient does not diagnose with cardiac sarcoidosis.

Figure 52 describes the LIME explanation for patients with index number 10. The probability of the patient being diagnosed with class 0 (CTRL) is 94%. The figure describes which of the five features positively corresponds with the class 0 prediction of this instance, such as the higher value of LV_syst_radial_LAX_SR and the lower value for age. This analysis assists in explaining which features impact the prediction results [40].

6.4.3. Anchors Explainer of the Scenarios that are Most Disruptive to System Order

As discussed earlier, SHAP is both a global and local explainer, while LIME is a local explainer. LIME has the disadvantage of being incapable of explaining models with non-decision boundaries, and it is not capable of explaining surrounding instances [124]. However, SHAP computation is costly and has the

disadvantage of misinterpreting SHAP values[11] [124]. Considering all the disadvantages of SHAP and LIME, Anchors, as another explainer, has been developed for this case. Anchors [125] are independent of the underlying model (model agnostic) that uses reinforcement learning methods to calculate the set of feature conditions [124]. Unlike SHAP, it does not impose a significant computational burden. It exhibits superior generalizability compared to LIME and can elucidate non-linear decision boundaries by operating on feature predicates instead of attempting to fit a linear model to the data [124].

Figure 53 describes an anchor local explanation for a patient with instance number 30. The length of the bars represents the values of individual features for the selected instance, offering insight into the significance of each feature concerning that instance and its predicted class. Features exhibiting noticeable differences across classes are likely pivotal factors influencing the decision of the model. The figure depicts the local explanation utilizing Anchors with fitting on the random forest classifier, which is classified into three categories: CTRL, NC, and CS. The bar chart depicts the significance or impact of 37 individual features. The objective is to elucidate the decision-making process of the model by emphasizing the important features that provide the most significant impact on the outcome of the model. All 37 features were scaled to have the same range for all features. The graph describes that LV_diast_radial_SAX_SR and LA_diast_long_LAX_SR have the most negative impacts on the outcome predictions, while LV_diast_circumf_SAX_SR and LA_long_LAX_S have the

[11] SHAP values explain the deviation from mean prediction and not the prediction [124].

most positive impact and contributions to the outcome prediction results. These results will be assessed by physicians and medical experts for validation.

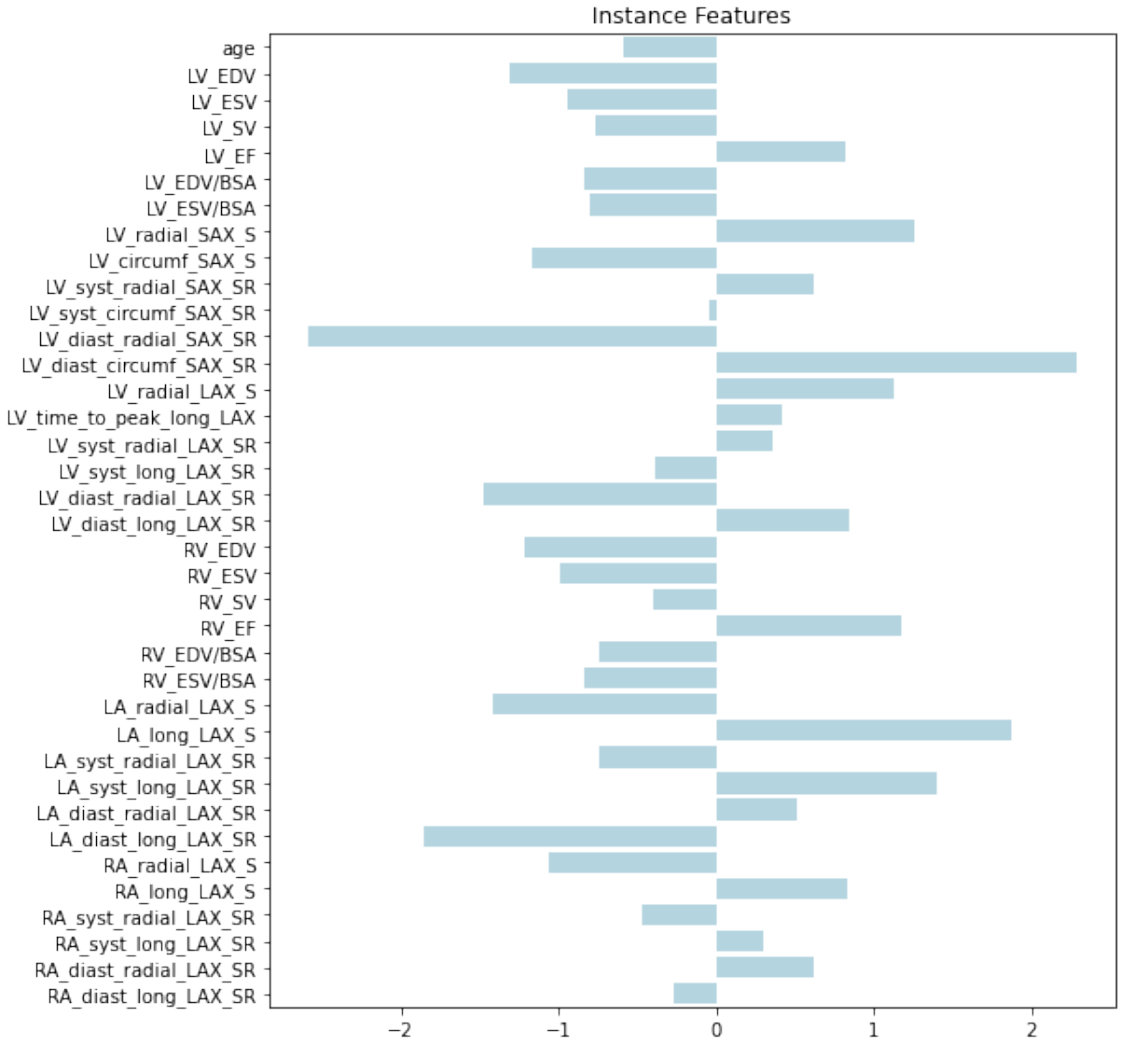

Figure 53. Local explanation of a patient with index number 30 prediction class using Anchors; three clusters CTRL, NC Sarc., and Sarc. (with age). LV_diast_radial_SAX_SR and LV_diast_circumf_SAX_SR are most correspondence to the prediction of the patient 30 cardiac sarcoidosis status.

The only way to detect cardiac sarcoidosis is by using contrast agents during imaging. However, this chapter describes that a non-contrast diagnosis approach using machine learning has the potential to detect subclinical cardiac involvement. This approach can help identify early signs of cardiac sarcoidosis in patients who do not yet show symptoms. This is particularly beneficial for patients who cannot tolerate contrast agents or for whom they are contraindicated. While contrast agents remain an important tool in detecting cardiac sarcoidosis, the non-contrast diagnosis approach using machine learning is a promising development in the diagnosis and management of this condition.

6.5. GitHub Codes Link for Chapter 6

The following is the link to Chapter 6 codes:

https://github.com/nm2fs/PhD-Dissertation/tree/3c5a70bfffc03100ce5ac284f818f3495a29f6b5/chapter6

6.6. Summary

In summary, this chapter describes a comprehensive systems modeling framework aimed at enhancing trust in AI-assisted medical diagnosis, with a specific focus on the diagnosis of cardiac sarcoidosis and utilizing XAI techniques. The design includes two primary sections: 1. Identifying the most and least disruptive scenarios to the system, as well as the highest ranked initiatives for the system. 2. Utilizing XAI techniques such as SHAP, LIME, and Anchors models to provide explanations on how machine learning models justify

their outcomes. The findings indicate the significance of employing XAI in critical domains like healthcare, where the lives of the patients are in jeopardy. XAI can be employed to analyze the outcomes for AI users, determine the significance of features, improve comprehension of AI outputs, reduce skepticism towards AI outputs, and facilitate data assessment.

Chapter 7 | Case 3: *Function (Phi)* Layer – Perspectives Comparison

7.1. Introduction

This chapter describes a scenario-based preference risk register framework for quantifying AI-related risks in disease diagnosis from the perspectives of two expert groups. Two case studies, one involving physicians and the other involving patients as the leading experts and actors, are compared to evaluate the effectiveness of the framework and to evaluate how the initiatives ranking orders will evolve based on the most and least disruptive scenarios for each case. The framework is applied to realistic case studies on cardiac sarcoidosis, identifying success criteria, initiatives, emergent conditions, and the most and least disruptive scenarios. The success criteria align with the NIST AI Risk Management Framework seven trustworthy AI *principles*. Finally, the framework

identifies the most and least disruptive scenarios and the highest ranked initiatives for both cases. Results from Chapter 6 multicriteria decision analysis will be compared to those obtained in this chapter to show how changing perspectives and the involvement of other experts and actors will change initiatives, order disruptions, and scenarios.

The introduction of AI in the healthcare domain holds great promise for improving medical research. However, it is essential to incorporate patient values, needs, and perspectives when implementing AI to meet their preferences and concerns. Below are some related studies by other researchers that consider the priorities and concerns of patients regarding the use of AI in healthcare systems [68].

Macri and Roberts stated that previous studies have explored various aspects of the views of patients, values, and worries related to the use of AI in healthcare. These concerns are comprised of trust, compassion, privacy [126], safety, autonomy, and fairness. To navigate these complexities, adopting a proposed values-based framework by the authors is essential. These frameworks assist healthcare providers and AI designers in addressing patient questions and designing AI systems that align with patients identified values. Considering the priorities of patients can support collaborative decision-making across a wide range of clinical AI applications, ensuring that choices are aligned with patient values and expectations [68,127].

Al Kuweiti et al. mentioned that healthcare innovations are derived from the experiences and needs of patients. Patients seek to have empowered digital interactions with their healthcare providers and access patient centered services on a global scale [128]. Ethical and societal issues related to AI intersect with

those arising from technology, automation, data use, and the growing use of telehealth and other technologies. As AI grows, ethical concerns [126,129,130] become an important problem that needs careful attention. These include accountability in AI driven decision making, the potential for AI to make incorrect judgments, issues with validation of AI results, the protection of sensitive data, systemic data, and algorithmic biases in AI training data [18], maintaining public trust in the development of AI and its benefits, its impact on privacy and liberty of each individual, and social isolation in healthcare settings, implications for healthcare roles of professionals and skills, and the potential misuse of AI. In addition, using AI for treatment, decision making, and managing healthcare devices introduces safety and reliability concerns. Despite its potential, AI is not resistant to errors, which can be challenging to detect and may have serious consequences. The lack of "explainability" [126] poses a significant challenge for AI, particularly in its practical applications across various fields, specifically in high-risk domains such as healthcare and national security [68].

Aggarwal et al. mentioned the issue of bias in AI algorithms used in healthcare [131]. They demonstrate how using biased or unsuitable data can exacerbate healthcare disparities and harm people. They also emphasize the importance of diversity, transparency, and accountability in developing AI algorithms, with a particular focus on addressing bias [126] and promoting fairness through AI in healthcare [132]. AI algorithms learn from data through a computational process. This can inadvertently perpetuate historical biases, such as those related to race, gender, or any demographic information, often without the awareness or intent of developers or practitioners. This issue is exacerbated when a diverse group of experts and actors, including patients and healthcare professionals,

characterizing the diversity of the populations served by AI solutions, are not involved in the decision making processes that shape and implement AI solutions [132]. Bias could be in various forms, such as data bias and algorithmic bias. Algorithmic bias could result from data bias [133]. Thus, it is important not to use biased data in training AI models, as garbage in leads to garbage out [18,68].

Moreover, Agarwal et al. emphasize the important role of early and accurate disease detection using AI models. Such methods can significantly reduce mortality rates, enhance disease prognosis, and identify risk factors [126] that contribute to complications, challenges, healthcare costs, and time [134]. Similarly, Kasthuri and Meeradevi describe the significance of early detection in healthcare AI diagnosis, especially for diseases that require immediate attention, such as breast cancer. Early detection plays a critical role in addressing patient concerns and ensuring timely interventions [68,135].

Khedkar et al. discuss that the concern for patients in healthcare AI diagnosis is their ability to understand and trust the predictions made by AI models [136]. This result aligns with the results from [6] that the lack of explainability and interpretability of AI in healthcare is the most disruptive scenario in ordering the initiatives and objectives of the experts and actors in healthcare [68].

Moghadasi et al. note that the priorities of the patients regarding the impact and the risks of AI in medical diagnosis depend on the specific context. This highlights the importance of not making broad generalizations about patient perspectives across all contexts and healthcare settings [6,18].

This chapter is an extension to the papers titled “Systems Analysis of Bias and Risk in AI-enabled Medical Diagnosis” and “Risk Analysis of Artificial

Intelligence in Healthcare with Multilayer Concept of System Order" by Moghadasi et al. [6,68] to involve patients as the leading experts and actors to quantify risk as the disruption of the order of initiatives in healthcare systems with the focus on diagnosing cardiac sarcoidosis.

As mentioned in Chapter 6, sarcoidosis is a systemic, inflammatory, and granulomatous disease of uncertain etiology. Timely detection of cardiac sarcoidosis is crucial for patients to prevent further harm to the heart and other organs [18,68].

Experts and actors in Chapter 6 are limited to physicians with specialties in radiology, cardiac imaging, and cardiology. It is essential to acknowledge that in the healthcare sector, individuals such as patients, caregiving partners, and community entities are taking on more prominent roles as experts and actors. Their expertise is supported by their personal experiences, which is a form of knowledge gaining recognition in a balance with established sources in various contexts. Consequently, their active participation is essential at every phase, beginning with the initial conceptualization and preliminary analysis of AI applications in healthcare. Thus, the results from Chapter 6 will be compared with the results from the most and least disruptive scenarios in this chapter and observe how the ranking orders will change by involving different experts and actors in the system.

7.2. Scenario-Based Disruption of Priorities (Function (Phi) Layer)

This section develops a mathematical framework to help analysts comprehend which potential future conditions have the most significant disruptive impact on system priorities and how these priorities evolve in response to such disruptions. The framework aids experts and actors in selecting resilience measures based on evolving priorities in the face of disruptions. The results show what initiatives are most important for the patients and for the physicians and what scenarios disrupt these initiatives.

For the sake of illustration, experts and actors are represented as two patients suspected to have cardiac sarcoidosis symptoms and signs. The engagement of patients commenced at an early stage of the study, encompassing activities such as identifying initiatives, emergent conditions, and scenarios and participating in the scoring and ranking assessments. These interactions were conducted through oral interviews to gather patient evidence by HDZ_NRW medical experts and actors[12]. In addition, medical experts and actors were invited for interviews to identify further initiatives and emerging factors, as well as to evaluate and assess them. Thus, to ensure a comprehensive understanding of the risks involved and to capture the nuances of patient experiences, medical experts

[12] Due to the privacy protection of the patients, all the patients' information is protected and will not be shared with external sources. The medical experts and actors conducted all the interviews with the patients.

One limitation of this chapter was the scarcity of patients suspected of having cardiac sarcoidosis. This was due to restrictions in Germany that limited direct access to patients for privacy and information protection and the scarcity of cardiac sarcoidosis disease. As a result, the number of patients included in the case study was limited to two. However, to obtain more precise results, a larger number of patients are required for the case study.

collaborated closely with patients throughout the risk assessment process. Patients were actively involved in identifying potential risks, providing insights into their perceptions and concerns, and contributing to the development of mitigation strategies. Patients suspected to have cardiac sarcoidosis symptoms and signs were asked to complete risk register tables, as discussed in Chapter 3. These tables include prompts to identify potential risks associated with their condition and assess their severity and likelihood. Thus, the inputs from the patients were gathered not through a formal survey but rather through informal interviews conducted by the medical experts.

Patients formulate sets of initiatives, emergent conditions, and scenarios for analysis. They also derived from various sources, including third-party program analysis, literature reviews, established standards, internal expertise, and other references.

Table 38 describes that all seven criteria have a high relevance among the other criteria. For instance, *c.04 – Privacy Enhanced* has a high relevance among the other criteria.

Table 38. Baseline relevance for cardiac sarcoidosis diagnosis in enterprise risk management of AI in healthcare [68].

The criterion c.xx has	s.00 - Baseline	relevance among the other criteria
c.01 - *Safe* has	*high*	relevance
c.02 - *Secure & Resilient* has	*high*	relevance
c.03 - *Explainable & Interpretable* has	*high*	relevance
c.04 - *Privacy Enhanced* has	*high*	relevance
c.05 - *Fair - With Harmful Bias Managed* has	*high*	relevance
c.06 - *Accountable & Transparent* has	*medium*	relevance
c.07 - *Valid & Reliable* has	*high*	relevance

Table 39 describes forty-three initiatives that the patients and the analysts identified.

Table 39. Initiatives address one or more of the success criteria for the risk of AI in cardiac sarcoidosis diagnosis. Abridged and adapted from various sources that are identified in the narrative [34,68,129,137–150].

Index	Initiative
x.01	Identify At-Risk Components
x.02	Understanding ML Tools to Uncover Any Patterns in Data
x.03	Maintaining the Provenance of Training Data
x.04	Safety/Verifiability of Automated Analyses (Cardiac Region Detection Software)
x.05	Reproducible Data and Method in Other Health Centers
x.06	Correctly Labeling the Data
x.07	Training Data to Follow Application Intellectual Property Rights Laws
x.08	Informed Consent to Use Data
x.09	Maintain Organizational Practices Like Implement Risk Management to Reduce Harm Reduction and More Accountable Systems
x.10	Prioritization Policies and Resources Based on Assesses Risk Levels
x.11	Safety of Personally Identifiable Information
x.12	Effective Risk Management by Appropriate Accountability Mechanism, Roles and Responsibilities, and Incentive Structures for Risk Management to be Effective
x.13	Avoid Gender and Age Discriminations and Bias in Preparing Data
x.14	Reducing Unnecessarily Procedures
x.15	Reducing Costs and Time Consumption
x.16	Able to Identify Healthy Volunteers before Starting the Procedures
x.17	Designate Ethical, Legal, Societal and Technical Boundaries for AI Operation
x.18	Policymakers to Ensure the Moral Demanding Situations are Tackled Proactively
x.19	Articulate and Document the Concept and Objectives of the System Considering Legal, Regulatory and Ethical Requirements
x.20	Gather, Clean and Validate Data and Document the Metadata, Also Characteristics of the Dataset Considering Legal, Regulatory and Ethical Requirements
x.21	- Key steps for implementing a new software system: Pilot, Compatibility with Legacy Systems, Regulatory Compliance, Organizational Change Management, and User Experience Evaluation
x.22	Continuously Assess AI System Recommendations and Impacts
x.23	Balancing and Trade Off of Trustworthy AI System Characteristics Based on Context
x.24	Reducing the Hospitalization Time of the Patient by Correct Diagnostics
x.25	Explain and Identify Most Important Features Using AI Models
x.26	Measurements Outlier Findings
x.27	Closeness of Results of Estimates, Observations and Computations to the Ground True (True Values)
x.28	Human-AI Teaming
x.29	Demonstrate Validity or Generalizability Beyond the Training Conditions
x.30	System Ability to Maintain its Performance Under an Uncertain Circumstances
x.31	Minimizing Potential Harms to People Under Unexpected Operating Settings
x.32	Responsible AI System Design, Development and Deployment Practices

x.33	Clear Information to the Users on Responsible Use of the AI System
x.34	Deployers and End Users to Make Responsible Decisions
x.35	Documentation and Explanation of Risks, Grounded in Empirical Evidence from Past Incidents
x.36	The Ability to Control, Adjust, or Involve Humans in Systems When They Do Not Perform as Intended or Expected
x.37	Clear and Distinct Definitions of Human Roles and Responsibilities Are Essential for Decision Making and Oversight in the Context of AI Systems
x.38	AI Systems May Need More Frequent Maintenance and Triggers for Corrective Maintenance Because of Data, Model, or Concept Drift
x.39	Managing Risks from Lack of Explainability by Defining the AI Systems Functions Considering Users' Role, Knowledge and Skill Level
x.40	The Ability to Describe Why an AI System Made a Specific Prediction or Recommendation
x.41	Securing Individual Privacy, Anonymity and Confidentiality
x.42	The Process of Removing Identifying Information and Combining Specific Model Results to Maintain Privacy and Confidentiality in Certain Model Outputs
x.43	Strengthened Engagement with Relevant AI Actors and Interested Experts and Actors
x.i	Others

Analysts, acting on behalf of the patients, assess the relevance of each initiative to a specific criterion through a response scale, including *"strongly agree," "agree," "somewhat agree,"* or *"neutral,"* corresponding to initial even split weights of 1, 2/3, 1/3, and 0, respectively. For instance, in the case of the success criterion *c.04 - Privacy Enhanced* and an initiative such as *x.41 - Securing Individual Privacy, Anonymity and Confidentiality* impacts *Privacy Enhanced*, the evaluation would be *"strongly agree"* as shown in Table 40.

Table 40. The criteria-initiative assessment describes how well each initiative addresses the success criteria for the risk of AI in cardiac sarcoidosis diagnosis. *Strongly agree* is represented by a filled circle *(●), agree* is represented by a half-filled circle (◐), *somewhat agree* is represented by an unfilled circle (○), and *neutral* is represented by a dash (━) [68].

	c.01	*c.02*	*c.03*	*c.04*	*c.05*	*c.06*	*c.07*
x.01	●	◐	○	○	○	○	○
x.02	━	━	○	━	━	○	○
x.03	━	━	◐	○	◐	◐	○
x.04	●	●	◐	◐	●	●	●
x.05	●	●	●	●	●	●	●
x.06	◐	◐	━	◐	●	●	●
x.07	━	○	━	●	◐	◐	◐
x.08	●	●	◐	●	◐	●	◐
x.09	━	○	━	●	◐	◐	◐
x.10	━	○	━	●	◐	◐	◐
x.11	●	●	◐	●	◐	●	◐
x.12	○	○	◐	━	━	○	◐
x.13	●	●	◐	◐	●	●	●
x.14	●	●	◐	○	●	●	●
x.15	●	●	○	○	●	●	●
x.16	○	○	○	○	○	○	●
x.17	○	○	◐	○	○	◐	◐
x.18	○	○	◐	○	○	○	◐
x.19	○	○	◐	○	○	○	◐
x.20	◐	◐	◐	○	○	○	◐
x.21	○	○	○	○	○	○	○
x.22	○	○	○	○	○	○	○
x.23	○	○	◐	○	◐	◐	◐
x.24	●	●	●	◐	●	●	●
x.25	○	○	◐	○	○	◐	○
x.26	━	━	━	━	━	━	━
x.27	●	●	◐	◐	◐	●	●
x.28	◐	○	○	○	○	○	○
x.29	○	○	○	◐	◐	━	◐
x.30	○	○	○	◐	◐	━	◐
x.31	●	●	◐	◐	●	●	●
x.32	◐	◐	━	●	●	○	◐
x.33	○	○	○	◐	◐	◐	○
x.34	◐	◐	━	●	●	○	◐
x.35	○	○	◐	○	○	◐	○
x.36	○	○	●	○	○	○	○
x.37	○	○	◐	○	○	◐	○
x.38	◐	◐	━	━	○	━	◐
x.39	○	○	●	━	━	◐	○
x.40	●	●	●	○	●	●	●
x.41	●	●	○	●	●	●	●
x.42	◐	◐	●	○	◐	◐	◐
x.43	○	○	○	○	○	○	○

Table 41 describes the emergent conditions that were identified, and Table 42 describes the scenarios that were developed by grouping the emergent conditions.

Table 41. Emergent conditions were used to create sets of scenarios for the risk of AI in cardiac sarcoidosis diagnosis. Abridged and adapted from various sources that are identified in the narrative [6,11,34,68,137–139,143,150,151].

Index	Emergent Condition
e.01	Using Non-Important Features in Sarcoidosis Diagnostics as the Input
e.02	Improperly Labeling the Data in Surgery-Specific Patient Registries
e.03	Issue of Incorrect Identification and Labeling of Variables in Registries Used for Surgery-Related Patient Data, Highlighting the Potential Consequences of Such Misidentification
e.04	Misunderstanding AI
e.05	Limited Generalizability
e.06	Limitation in Types and Performance of Available Data
e.07	Expensive Data Collection
e.08	Time Consuming Data Collection
e.09	Policy and Regulation Changes
e.10	Difficult and Complex AI Algorithms Interpretability
e.11	Lack of AI Determination of Casual Relationships in Data at Clinical Implementation Level
e.12	Inability of AI in Providing an Automated Clinical Interpretation of its Analysis
e.13	Human Errors in Measurements
e.14	Abuse or Misuse of the AI Model or Data
e.15	Challenges with Training Data to be Subject to Copyright
e.16	Complicate Risk Measurement by Third Party Software, Hardware and Data
e.17	Model Fails to Generalize
e.18	Lack of Robustness and Verifiable Methods for AI Trustworthiness
e.19	Mis-Identification of Different Risk Perspective in Early or Late Stages of AI Lifecycle
e.20	Difference Between Controlled Environment vs. Uncontrollable and Real-World Settings
e.21	Inscrutable Nature of AI Systems in Risk Measurements
e.22	Systematic Biases in Collecting Clinical Data
e.23	Risk Tolerance Influence by Legal or Regulatory Requirements Changes
e.24	Unrealistic Expectations About Risk to Misallocate Resources
e.25	Residual Risk after Risk Treatment Directly Impacts Healthcare Deployers
e.26	Privacy Concerns Regarding Using Underlying Data to Train the Systems
e.27	The Energy and Environmental Implications Causing from Resource Heavy Computing Demands

e.28	Security Concerns Related to the Confidentiality of the System Training and Output
e.29	Security of the System Underlying Software and Hardware
e.30	One-Size-Fits-All Requirements AI Model Challenges
e.31	Neglecting the Trustworthy AI Characteristics
e.32	Difficult Decisions in Tradeoff and Balancing Trustworthy AI Characteristics by Organizations
e.33	Subject Matter Experts and Actors Collaborate to Evaluate TEVV Findings, Aligning Parameters with Project Requirements and Deployment Conditions
e.34	Different Perception of the Trustworthy AI Characteristics Between AI Designer than the Deployer
e.35	Potential Risk of Serious Injury to the Patients
e.36	Complexity of Explaining AI System to End Users
e.37	Data Poisoning
e.38	Negative Risks Result from an Inability to Appropriately Understand or Contextualize System Output
e.39	AI Allowing Inference to Identify Individuals or their Private Information
e.40	Privacy Intrusions
e.41	Data Sparsity
e.42	Fairness Perceptions Difference Among Cultures and Applications
e.43	Computational and Statistical Biases Stem from Systematic Errors Due to Limited and Non-Representative Samples
e.44	Human-Cognitive Biases Relates to How the Experts and Actors Perceives AI System Information and Use them to Make Decisions
e.45	Lack of Access to the Ground Truth in the Dataset
e.46	Intentional or Unintentional Changes During Training
e.47	Increased Opacity and Concerns About Reproducibility
e.48	Impacts of Computational Costs on the Environment and Planet
e.49	Incapacity to Anticipate or Identify the Adverse Effects of AI-Driven Systems Beyond Statistical Metrics
e.50	Over-Reliance on AI
e.i	Others

Emergent conditions and scenarios do not aim to encompass all conceivable future states or disruptions but focus on the concerns of system experts, actors, and analysts.

Table 42 also describes ten scenarios as listed in Chapter 6 which are *s.01 – Funding Decrease, s.02 – Government Regulation and Policy Changes, s.03 – Privacy Attacks, s.04 – Cyber Security Threats, s.05 – Changes in AI RMF, s.06 – Non-Interpretable AI and Lack of Human-AI Communications, s.07 – Global Economic and Societal Crisis, s.08 – Human Errors in Design, Develop, Measurement and Implementation, s.09 – Uncontrollable Environment,* and *s.10 – Expensive Design Process.*

Table 42. Emergent conditions grouping in the risk of AI in the diagnosis of cardiac sarcoidosis, identifying which conditions fit in each scenario, from various sources that are identified in the narrative [6,68].

	s.01 - Historical Human Biases	s.02 - Misclassification or Measurement Error	s.03 - Privacy Attacks	s.04 - Cyber Security Threats	s.05 - Conflict of Interest	s.06 - Lack of Ethical Considerations and Oversight Policies	s.07 - Socioeconomic Status	s.08 - Sample Size and Missing Data	s.09 - Global Crisis and Immigrations	s.10 - Lack of Healthcare Resource Allocation and Access to Healthcare
e.01		✓						✓		
e.02		✓	✓		✓		✓			
e.03		✓			✓	✓		✓		
e.04		✓								
e.05	✓				✓		✓	✓	✓	✓
e.06							✓	✓	✓	✓
e.07	✓						✓	✓	✓	✓
e.08	✓									✓
e.09						✓				
e.10		✓								
e.11						✓				
e.12		✓								
e.13		✓						✓		
e.14	✓	✓			✓	✓		✓		
e.15					✓					
e.16		✓								
e.17	✓				✓		✓	✓	✓	✓
e.18	✓	✓			✓		✓	✓	✓	✓
e.19		✓								
e.20	✓	✓			✓		✓		✓	✓
e.21		✓								
e.22	✓				✓		✓	✓	✓	✓
e.23						✓				
e.24		✓					✓	✓	✓	✓
e.25						✓				
e.26			✓							
e.27									✓	
e.28			✓	✓						
e.29				✓						

e.30							✓	✓	✓	✓
e.31	✓	✓	✓		✓					
e.32		✓				✓				
e.33		✓				✓				
e.34	✓	✓			✓			✓		
e.35		✓								
e.36		✓								
e.37	✓	✓			✓	✓		✓		
e.38		✓								
e.39		✓	✓					✓		
e.40			✓							
e.41							✓	✓		✓
e.42	✓									
e.43	✓						✓	✓	✓	✓
e.44					✓					
e.45							✓	✓	✓	✓
e.46	✓	✓			✓					
e.47							✓	✓	✓	✓
e.48							✓	✓	✓	✓
e.49		✓								
e.50		✓								

For comparison, Table 39 and Table 41 are similar to the initiatives and emergent conditions identified in Chapter 6. The impact of scenarios on priorities is observed in terms of how scenarios influence the relative importance of success criteria. In other words, system priorities change when the system is exposed to disruptions. Patients then adjust the importance of the criteria for each scenario. Analysts evaluate if the relative importance of criterion c.j, *increases, increases somewhat, remains unchanged (-), decreases somewhat,* or *decreases* for scenario s.p compared to the baseline s.00.

Table 43 describes the criteria_scenario relevance, and Table 44 describes the initiative-scenario ranking chart.

Table 43. The criteria-scenario relevance describes how well each scenario fits the success criterion for cardiac sarcoidosis diagnosis in the risk of AI in cardiac sarcoidosis diagnosis. *Decrease Somewhat* = DS, *Decrease* = D, *Somewhat Increase* = SI, *Increase* = I [10,68].

	s.01	*s.02*	*s.03*	*s.04*	*s.05*	*s.06*	*s.07*	*s.08*	*s.09*	*s.10*
c.01	DS	IS	D	D	IS	DS	-	D	DS	DS
c.02	DS	IS	D	D	IS	DS	-	-	DS	DS
c.03	DS	IS	-	-	IS	D	-	-	D	-
c.04	-	I	D	D	-	-	-	D	DS	-
c.05	DS	I	DS	DS	IS	DS	DS	D	DS	DS
c.06	DS	IS	D	D	I	D	-	D	D	DS
c.07	DS	IS	D	D	I	D	-	D	D	DS

Table 44. Initiative-scenario ranking chart. This table describes the ranking of each initiative under each scenario. The green filled cells show a higher ranking and the red and orange filled cells indicate a lower ranking.

	s.00 - Baseline	s.01 - Funding Decrease	s.02 - Government Regulation and Policy Changes	s.03 - Privacy Attacks	s.04 - Cyber Security Threats	s.05 - Changes in AI RMF	s.06 - Non-Interpretable AI and Lack of Human-AI Communications	s.07 - Global Economic and Societal Crisis	s.08 - Human Errors in Design, Develop, Measurement and Implementation	s.09 - Uncontrollable Environment	s.10 - Expensive Design Process
x.01	19	25	22	26	26	19	24	18	18	18	34
x.02	42	42	42	41	41	42	42	42	42	42	41
x.03	35	35	35	24	24	38	35	41	34	36	32
x.04	4	6	4	6	6	5	8	4	5	4	7
x.05	1	1	1	1	1	1	1	1	1	1	1
x.06	13	15	13	32	32	13	15	14	26	13	28
x.07	24	12	19	38	38	33	12	28	39	23	23
x.08	8	3	8	10	10	11	3	8	8	8	4
x.09	24	12	19	38	38	33	12	28	39	23	23
x.10	24	12	19	38	38	33	12	28	39	23	23
x.11	8	3	8	10	10	11	3	8	8	8	4
x.12	41	41	41	25	25	36	41	36	25	41	40
x.13	4	6	4	6	6	5	8	4	5	4	7
x.14	10	17	11	9	9	8	17	11	8	10	12
x.15	12	18	12	17	17	10	18	12	13	12	16
x.16	27	29	27	30	30	21	29	24	32	27	35
x.17	21	26	26	18	18	20	26	20	19	26	19
x.18	29	27	29	19	19	22	27	21	20	27	26
x.19	29	27	29	19	19	22	27	21	20	27	26
x.20	17	23	17	15	15	17	23	17	14	17	17
x.21	38	36	37	33	33	39	36	38	36	37	37
x.22	38	36	37	33	33	39	36	38	36	37	37
x.23	18	24	18	16	16	18	25	19	17	19	18
x.24	2	5	3	2	2	2	7	2	2	3	2
x.25	31	31	31	21	21	28	31	31	22	31	29
x.26	43	43	43	43	43	43	43	43	43	43	43
x.27	10	11	10	10	10	9	11	10	8	10	10

x.28	34	34	34	31	31	37	34	35	33	34	36
x.29	22	20	24	28	28	26	20	26	30	21	21
x.30	22	20	24	28	28	26	20	26	30	21	21
x.31	4	6	4	6	6	5	8	4	5	4	7
x.32	15	9	14	36	36	15	5	15	27	14	14
x.33	20	19	23	27	27	25	19	25	29	20	19
x.34	15	9	14	36	36	15	5	15	27	14	14
x.35	31	31	31	21	21	28	31	31	22	31	29
x.36	27	29	27	13	13	24	30	21	15	27	13
x.37	31	31	31	21	21	28	31	31	22	31	29
x.38	37	40	36	42	42	32	39	37	35	35	42
x.39	35	39	40	14	14	31	40	34	16	40	33
x.40	4	16	7	3	3	3	16	4	3	7	6
x.41	2	2	2	5	5	4	2	2	4	2	2
x.42	14	22	16	4	4	14	22	13	12	16	11
x.43	38	36	37	33	33	39	36	38	36	37	37

These assessments rely on expertise, institutional knowledge, and iteration with other experts and actors.

In Figure 54, each scenario is given a disruptiveness score; the higher the score, the more disruptive the scenario will be to the system [10]. This figure describes that *s.03 – Privacy Attacks, s.04 - Cyber Security Threats, s.06 – Non-Interpretable AI and Lack of Human-AI Communications,* and *s.08 – Human Errors in Design, Develop, Measurement and Implementation* are predicted to have the highest disruption among other scenarios in the realistic case study of the diagnosis of cardiac sarcoidosis. Features are drawn from the experience of the authors.

In Figure 55, the chart describes the fluctuation in the prioritization of initiatives across different scenarios. The ranking of initiatives offers a holistic view of their overall performance. The highest ranked initiatives are *x.04 - Safety/Verifiability of Automated Analyses (Cardiac Region Detection Software), x.05 - Reproducible Data and Method in Other Health Centers, x.08 - Informed Consent to Use*

Data, x.11 - Safety of Personally Identifiable Information, x.13 - Avoid Gender and Age Discriminations and Bias in Preparing Data, x.14 - Reducing Unnecessarily Procedures, x.15 - Reducing Costs and Time Consumption, x.24 - Reducing the Hospitalization Time of the Patient by Correct Diagnostics, x.27 - Closeness of Results of Observations, Computations, or Estimates to the True Values or the Values Accepted as Being True, x.31 - Minimizing Potential Harms to People if it is Operating in an Unexpected Setting, x.40 - Communicating a Description of Why an AI System Made a Particular Prediction or Recommendation, and *x.41 - Securing Individual Privacy, Anonymity, and Confidentiality.*

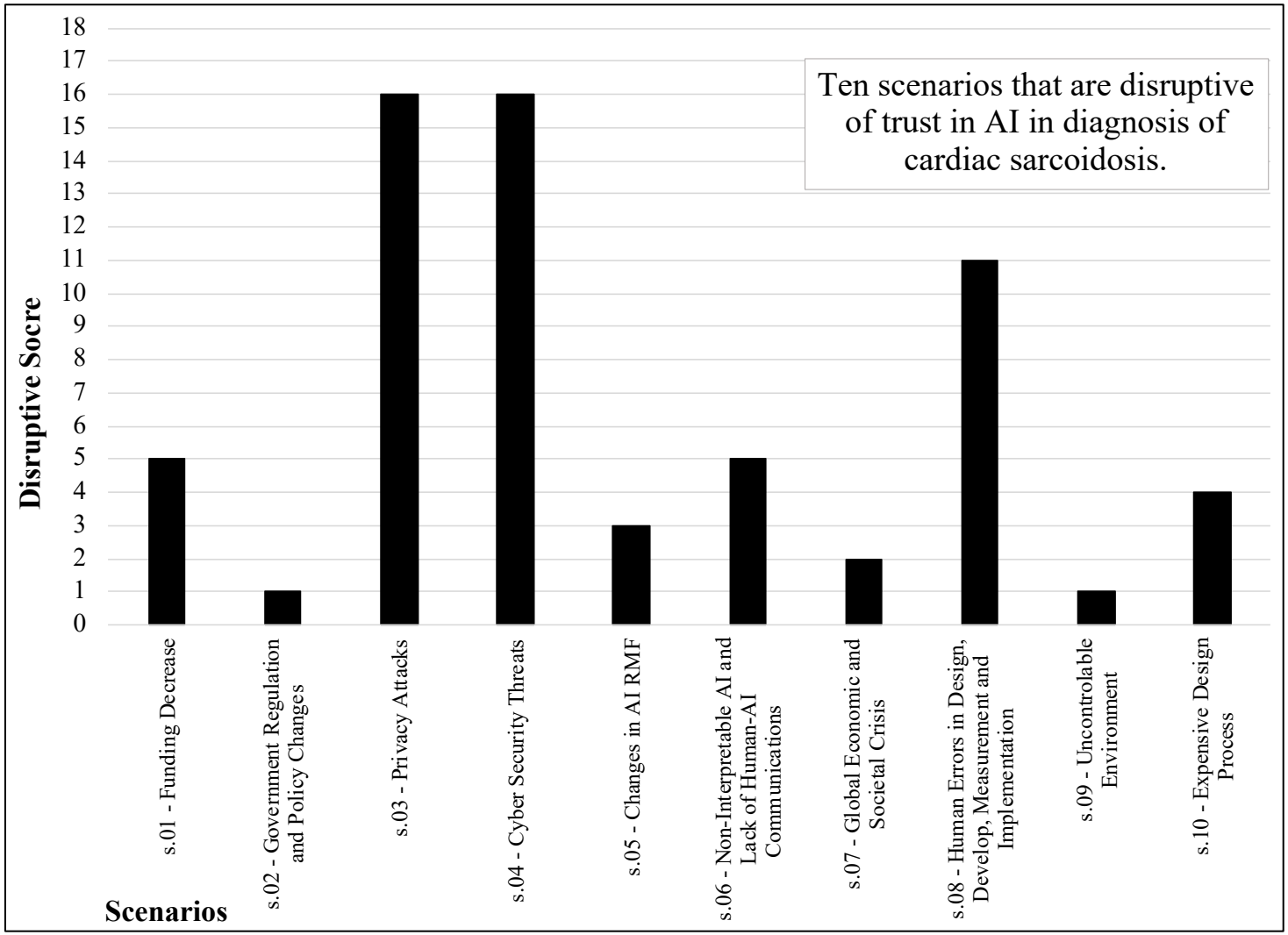

Figure 54. The disruptive score of scenarios is based on the sum of squared differences in the priority of initiatives relative to the baseline scenario in the risk analysis of AI in cardiac sarcoidosis diagnosis.

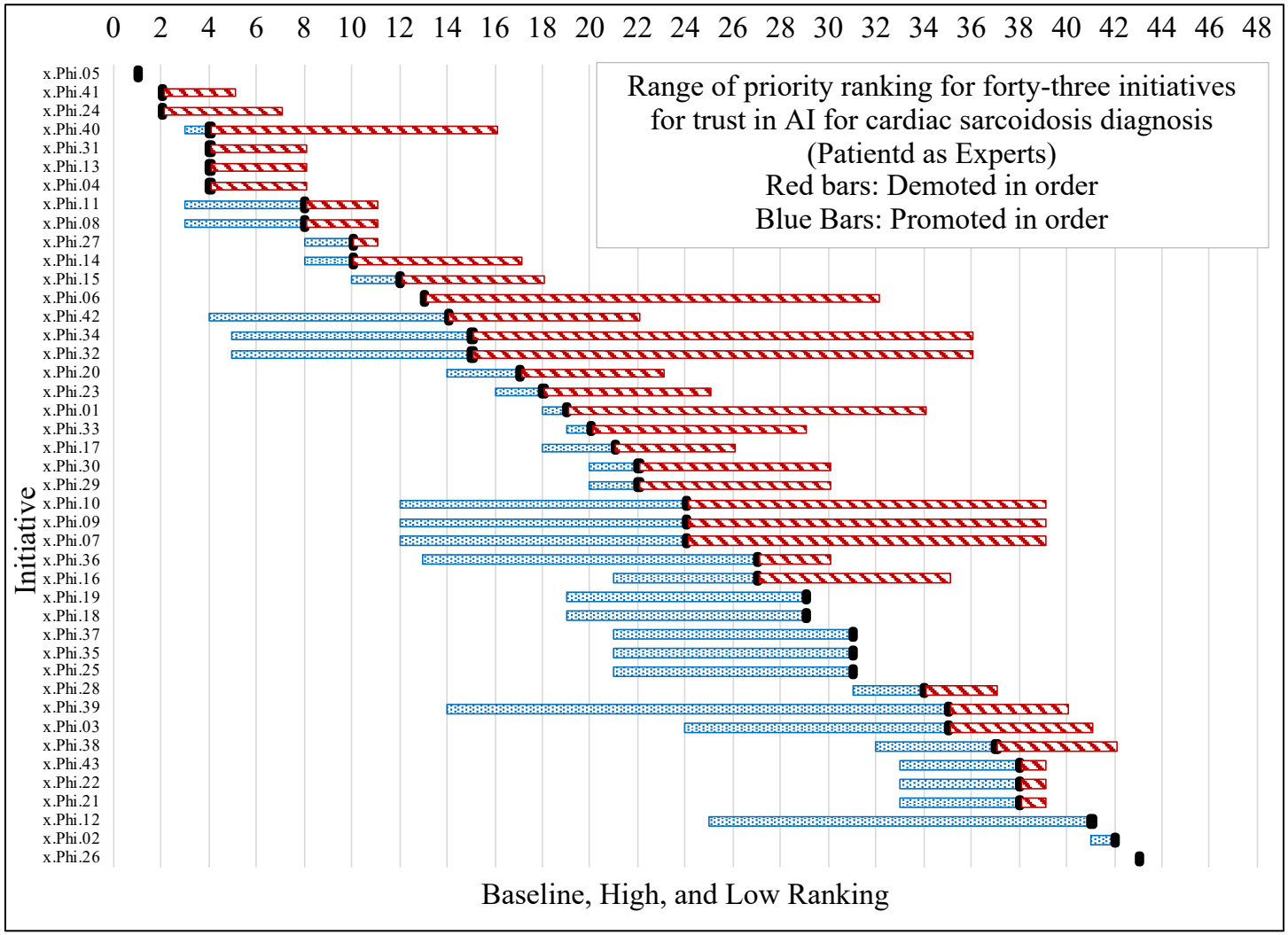

Figure 55. Distributions of initiatives that influence rankings are based on which emergent conditions could arise more often or do not occur in the risk of AI in cardiac sarcoidosis diagnosis; blue means promotion in ranking and red means demotion in ranking.

The findings may not be extrapolated to other clinical centers in different geographical locations or even other departments within the same hospital. So, the framework should be generated for each case study individually. However, the framework is generalizable beyond the diagnosis of sarcoidosis and could be applied to any medical diagnosis [152].

Figure 56, derived from Chapter 6, involving physicians as the leading experts and actors, reveals that several scenarios hold the highest potential for disruption in the real-world diagnosis of cardiac sarcoidosis. Specifically, these scenarios include *s.06 – Non-Interpretable AI and Lack of Human-AI Communications, s.03– Privacy attacks,* and *08 – Human Errors in Design, Develop, Measurement and Implementation* have the highest disruption among other scenarios.

Additionally, Figure 57, also sourced from Chapter 6, highlights the most significant initiatives in this context by involving physicians (medical experts and actors) as the leading experts and actors. These initiatives are as follows: *x.Phi.24 – Reducing the Hospitalization Time of the Patient by Correct Diagnostics, x.Phi.28 – Human-AI Teaming, x.Phi.32 – Responsible AI System Design, Development and Deployment Practices, x.Phi.29 – Demonstrate Validity or Generalizability Beyond the Training Conditions, x.Phi.27 – Closeness of Results of Estimates, Observations and Computations to the Ground True (True Values), x.Phi.25 – Explain and Identify Most Important Features Using AI Models, x.Phi.20 – Gather, Clean and Validate Data and Document the Metadata, Also Characteristics of the Dataset Considering Legal, Regulatory and Ethical Requirements, x.Phi.16 – Able to Identify Healthy Volunteers before Starting the Procedures, x.Phi.06 – Correctly Labeling the Data and x.Phi.04 – Safety/Verifiability of Automated Analyses (Cardiac region detection software).*

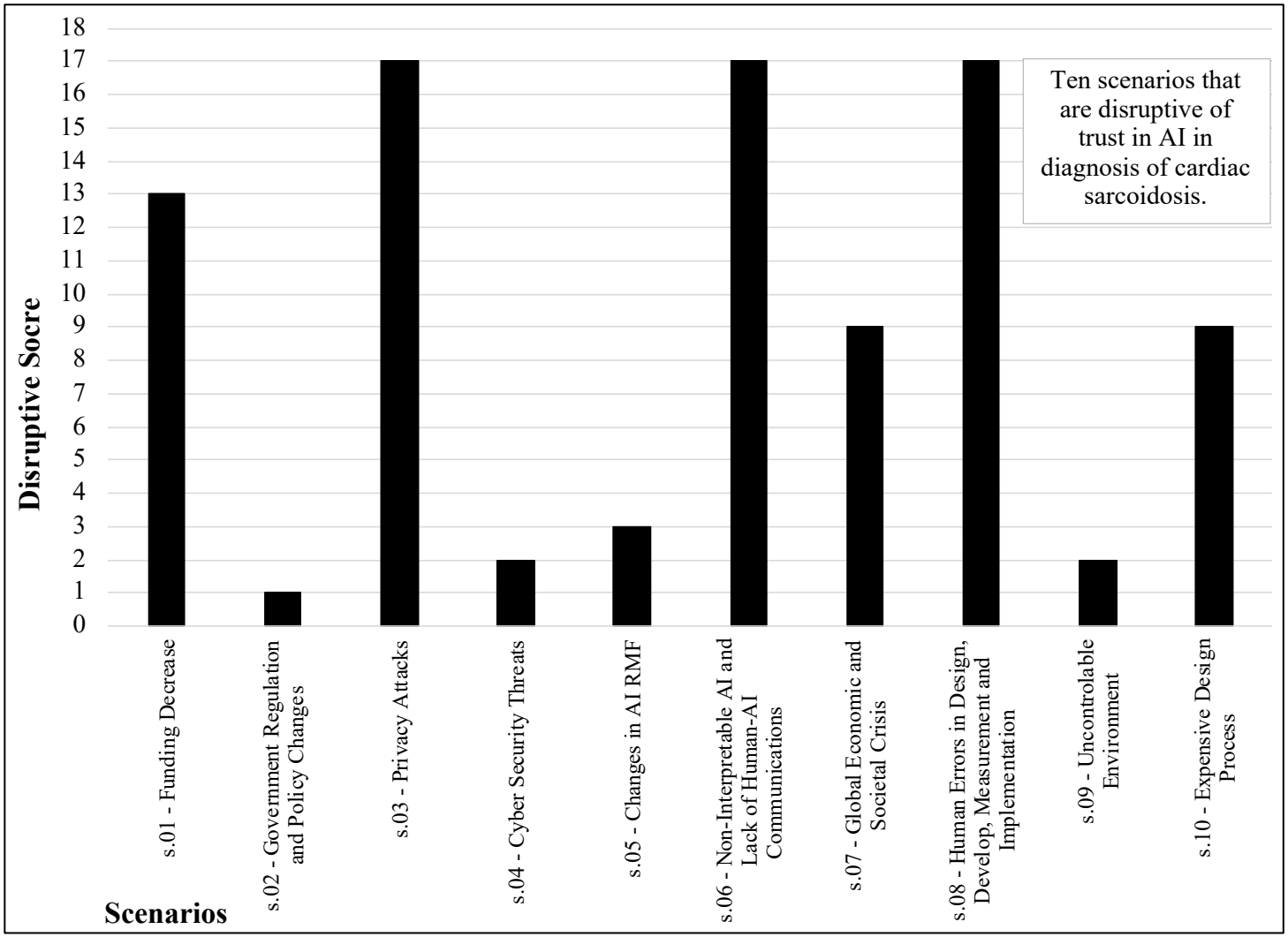

Figure 56. The disruptive score of scenarios is based on the sum of squared differences in the priority of initiatives relative to the baseline scenario in the risk of AI in cardiac sarcoidosis diagnosis.

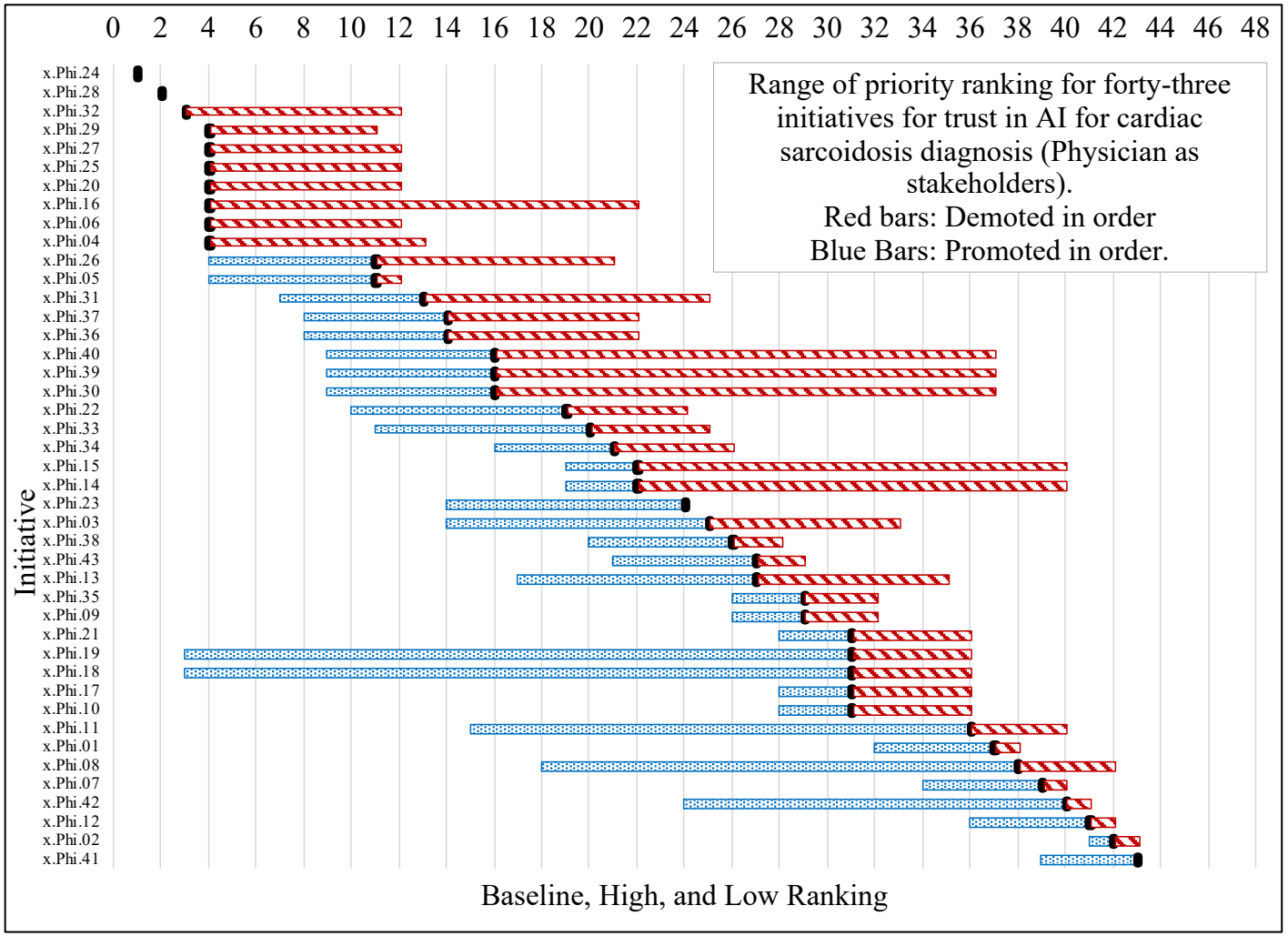

Figure 57. Distributions of initiatives that influence rankings are based on which emergent conditions that could arise more often or do not occur in the risk analysis of AI in cardiac sarcoidosis diagnosis; blue means promotion in ranking and red means demotion in ranking.

The methods employed in this chapter serve to enhance transparency and, by involving patients, identify the unintended adverse consequences of AI applications within healthcare systems. The initiatives and emergent conditions will continue to evolve with additional findings and are not confined to the lists provided above.

On the one hand, Table 45 described that the most disruptive scenarios, when patients are involved in the framework design, encompass *s.03 – Privacy Attacks, s.04 - Cyber Security Threats, s.06 – Non-Interpretable AI and Lack of Human-AI Communications,* and *s.08 – Human Errors in Design, Development, Measurement, and Implementation.* On the other hand, the table describes that the most disruptive scenarios involving physicians in the framework design include *s.06 – Non-Interpretable AI and Lack of Human-AI Communications, s.03 – Privacy Attacks,* and *s.08 – Human Errors in Design, Development, Measurement, and Implementation*. The outcomes underscore that patients prioritize early disease detection, cost and time savings in hospitalization and procedures, information privacy, and protection against cyberattacks. Physicians, conversely, emphasize the interpretability and explainability of AI models, the validity of AI models, reduced hospitalization time, and the minimization of human error in the diagnostic process. The table also describes that *s.03 – Privacy Attacks, s.06 – Non-Interpretable AI and Lack of Human-AI Communications,* and *s.08 – Human Errors in Design, Development, Measurement, and Implementation* are mutually the most disruptive scenarios for both sets of experts and actors.

Table 45. Scenarios of the risk of AI that are most disruptive to the system order of technologies in medical diagnosis. The most disruptive scenarios were shown with +++ and the least disruptive scenarios were shown with + [68].

Disruptive Scenarios	*Experts and Actors (Patients)*	*Experts and Actors (Physicians)*
s.01 – Funding Decrease	+	+
s.02 – Government Regulation and Policy Changes	+	+
s.03 – Privacy Attacks	**+++**	**+++**
s.04 – Cyber Security Threats	**+++**	+
s.05 – Changes in AI RMF	+	+
s.06 – Non-Interpretable AI and Lack of Human-AI Communications	**+++**	**+++**
s.07 – Global Economic and Societal Crisis	+	+
s.08 – Human Errors in Design, Develop, Measurement and Implementation	**+++**	**+++**
s.09 – Uncontrollable Environment	+	+
s.10 – Expensive Design Process	+	+

Both patient and physician involvement are essential throughout the entire process, commencing with the initial conceptualization of the AI objectives of the application within healthcare. This framework possesses the potential to not only facilitate the transfer of findings to healthcare systems globally but also extend its applicability to other domains such as transportation, finance, design, and beyond.

7.3. Summary

This chapter has described a scenario-based preference risk register framework to assess AI-related risks in disease diagnosis, specifically focusing on cardiac sarcoidosis, from the viewpoints of both physicians and patients. By comparing two case studies, the framework evaluates the ranking of initiatives in response to disruptive scenarios, revealing that system priorities shift

accordingly. The most disruptive scenarios identified, particularly when patients are involved in framework design, include privacy attacks, cybersecurity threats, non-interpretable AI, and human errors. Conversely, scenarios prioritized by physicians emphasize AI model interpretability, validity, minimizing human error, and reduced hospitalization time. Despite these differences, certain disruptive scenarios overlap between patients and physicians. Involving both parties in the framework design ensures alignment with priorities such as early detection, cost savings, and privacy protection. The framework intends to broaden its application beyond healthcare to various domains such as transportation and finance.

Chapter 8 | Synthesis and Comparison of Cases

8.1. Introduction

This chapter describes the highlights of the cases, additional limitations, and lessons learned. This section also compares the highest ranking initiatives and the most disruptive scenarios for three cases of the healthcare system layers: *Purpose (Pi), Structure (Sig),* and *Function (Phi).*

Table 46 and Table 47 suggest that the topic of the scenarios should be used to describe the scope of the tentative project, which shapes and guides the input of the R&D portfolio. This information allows investors and R&D managers to make informed decisions regarding resource allocation. Specifically, they can focus their investments on the most critical initiatives related to the risk analysis of AI [153] in healthcare applications. Additionally, they can consider the various scenarios presented in Table 46 ranging from the most disruptive to the least

disruptive. The dissertation recommends the following methods for user education about safe AI usage based on the results in this table: Informing users about why and how the benefits of using the AI system outweigh its risks compared to other technologies on the market, convincing clinicians that specific AI system outcomes are safe, providing information to users on what data to use for training, validating, and testing AI models, including potential changes due to various input data, highlighting that AI systems may require more frequent maintenance and triggers for corrective maintenance due to data, model, or concept drift, demonstrating the validity or generalizability of AI systems beyond the training conditions, emphasizing the closeness of results of estimates, observations, and computations to the ground truth (true values), and advocating for responsible AI system design, development, and deployment practices [6]. This analysis enables the identification of new topics that warrant additional resources and time, with the goal of improving the overall success of the system. For instance, Table 46 highlights scenario *s.06 - Non-Interpretable AI and Lack of Human-AI Communications* as the most disruptive scenario across all three layers of healthcare systems. Although the results from this pilot must be interpreted with caution and validated in a larger sample, this observation is consistent with the findings of [27,154], which indicate that AI transparency solutions primarily target domain experts and actors. Given the emphasis on "high-risk" AI systems, particularly in healthcare, this inclination is reasonable. Thus, optimizing trustworthy AI properties is recommended in these situations where scenarios involving the handling of sensitive and private data of individuals are present. By prioritizing the highest ranked initiatives and investing in identifying the most disruptive scenarios for the system, the full

potential of AI will be unlocked while responsibly integrating it into healthcare practices, benefiting both patients and the healthcare industry [6], although daily-based tasks that involve AI are less important for assessing the risks of AI in the domain, such as suggesting movies in online streaming or suggesting other items in online shopping systems. The necessity of AI interpretability and human-AI communications in everyday contexts for end users remains poorly understood. Existing research on this topic is limited, but the available findings suggest that this form of transparency may be insignificant to users in their everyday experiences [27].

Another observation is that the risks of AI should be context-based [154] and it should consider all the participants, experts, and actors in the study for more comprehensive findings. One explanation does not fit all [155]. Moreover, having a human-in-the-loop [156] is important for AI prediction verification and facilitates effective collaboration and partnership between humans and AI [6].

Table 46. Most and least disruptive scenarios with respect to rankings of the initiatives for systems *characteristic* layers in enterprise risk management of AI in healthcare. Most Disruptive Scenarios = (+++), Least Disruptive Scenarios = (+) [68].

Scenarios	***Purpose (Pi)***	***Structure (Sig)***	***Function (Phi)***	*Boundary (Bet)*	*Environment (Eps)*	*Interconnections (Iot)*
s.01 - Funding Decrease						
s.02 - Government Regulation and Policy Changes	+	+	+			
s.03 - Privacy Attacks		+	+++			
s.04 - Cyber Security Threats		+				
s.05 - Changes in AI RMF	+	+				
s.06 - Non-Interpretable AI and Lack of Human-AI Communications	**+++**	**+++**	**+++**	—	tbd in future effort	—
s.07 - Global Economic and Societal Crisis	+	+				
s.08 - Human Errors in Design, Develop, Measurement and Implementation			+++			
s.09 - Uncontrollable Environment		+++				
s.10 - Expensive Design Process		+++				

Table 47. The highest ranked initiatives of the systems *characteristic* layer in enterprise risk management of AI in healthcare [6].

Index	Initiative
Purpose (Pi)	x.Pi.35 - Inform Users on Why and How the Benefits of the Use of AI System Overweigh its Risks Compare to Other Technologies on the Market x.Pi.23 - Clinicians to be Convinced that Specific AI System Outcome is Safe x.Pi.33 - Users to be Informed that What Data to Use for Training, Validating and Testing the AI Models; Also, any Potential Changes Due to Various Input Data
Structure (Sig)	x.Sig.40 - The Ability to Describe Why an AI System Made a Specific Prediction or Recommendation x.Sig.44 - AI Systems May Need More Frequent Maintenance and Triggers for Corrective Maintenance Because of Data, Model, or Concept Drift x.Sig.24 - Reduce the Number of Experiments to be Cost and Time Effective by Optimizing the Configurations
Function (Phi)	x.Phi.29 – Demonstrate Validity or Generalizability Beyond the Training Conditions x.Phi.27 – Closeness of Results of Estimates, Observations and Computations to the Ground True (True Values) x.Phi.32 – Responsible AI System Design, Development and Deployment Practices
Boundary (Bet)	tbd
Environment (Eps)	tbd
Interconnections (Iot)	tbd

In healthcare, experts and actors typically use AI as a decision-support system. Consequently, the development of solutions prioritizes the needs and requirements of these knowledgeable professionals. Recognizing this context, it becomes evident that addressing the issues of non-interpretable AI and a lack of human-AI communications is crucial within healthcare systems. This is essential not only to ensure patient safety but also to foster trust, consider ethical implications, promote continuous learning, and ensure compliance with legal and regulatory frameworks. The implementation of artificial intelligence in healthcare comes with more human risks than in other sectors due to its unique capacity to directly impact the quality of care and healthcare outcomes [6].

Some methods are advised for confirming the efficacy of AI systems after training the dataset, such as confusion matrix analysis, using XAI techniques, having experts and actors in the loop to validate the outcome, continuous iteration and training monitoring, validation and testing assessments, bias, and fairness assessment, and more. Fairness and bias are also critical issues to understand and assess in AI applied to or used in the healthcare sector. For example, AI requires large, robust "training" databases, but many of the databases used for healthcare and medical datasets are limited. These data sets can perpetuate biases that exist in society and cause further health disparities and inequities. It is critical to have a clear understanding of possible biases that could exist in AI systems, as well as how choosing specific outcome variables and labels can impact predictions [157]. Moreover, studies have found that patients have concerns related to AI use in healthcare, including threats to patient choice, increased costs of healthcare, patient privacy and data security, and biases in the data sources used to train AI [158,159]. By involving patients and care partners,

it identifies and mitigates the risks of bias and unintended adverse consequences in AI applications within healthcare systems. Successfully using and implementing AI in healthcare settings will require a thoughtful understanding of social determinants of health, health equity, and ethics. The data used in the dissertation was collected to safeguard the privacy rights of individuals by implementing robust data collection measures, such as data quality assessments and validations by experts and actors, standard data collection procedures, clinic data security measures, and more. Improving data management procedures, including metadata documentation, collection, cleansing, and validation, is crucial for ensuring the quality, reliability, and usefulness of data. Integrating new software into an existing system requires careful planning to ensure compatibility, compliance with regulations, and a positive user experience by training on balanced datasets, performing risk analysis and assessment to find potential abnormalities in the dataset, enhancing data protection, and more [6].

The methods of this dissertation serve as a demonstration and emphasize the constraints associated with each disruptive scenario in tandem with the partial consideration of system layers. The dissertation serves to enhance transparency.

The scope of initiatives and emerging conditions extends beyond the lists in this dissertation and will be further elaborated upon. While the dissertation primarily focuses on socioeconomic status, it is important to note that future endeavors will encompass other demographic factors linked to health disparities, such as race/ethnicity, sexual orientation, geographic location, and disability status. As an extension to this dissertation, the study by [160] demonstrated how developing plans with diverse participants in terms of expertise, aptitude, and background changes the most and least disruptive scenarios in the system [6].

The upcoming interviews will encompass patients, care partners, and community-based organizations that work with populations affected by health disparities. It is crucial to recognize that individuals, including patients, caregiving partners, and community entities, are assuming increasingly important roles. These entities are acknowledged as authoritative sources due to their personal experiences, a form of knowledge gaining equitable recognition in various national contexts. Consequently, their involvement is important across all stages, starting from the initial conceptualization of AI application goals in healthcare [6].

8.2. Framework Scoring

This section discusses various notes on the framework scoring that was introduced in Chapter 3.

This dissertation framework is well-suited for use by healthcare professionals who lack the background necessary to comprehend and employ more complex methodologies that capture the intricacies of artificial intelligence. This argument acknowledges the limitations of the method and provides a clear explanation of why these limitations render it fit for its intended purpose.

The advantage of ordinal over cardinal ratings that were utilized in Chapter 3 is an improvement in ease of elicitation. The ratings in this dissertation are used as a measurement scale and are not vulnerable to ordinal disadvantages. Cox points out the subjectivity, loss of granularity, and challenges in prioritization associated with these matrices. Cox suggests the need for more robust, data-driven approaches to improve the accuracy and reliability of risk assessments

through methods such as probabilistic risk assessment (PRA), Bayesian networks, or other quantitative methods [6,161,162]. To overcome this challenge, Krisper introduces different kinds of distributions, both numerically and graphically. Some common distributions of ranks are *linear, logarithmic, normally distributed (Gaussian),* and *arbitrary (fitted)* [163]. For instance, for each scenario in this dissertation, *linear* distributions of ranks were used. That is, the scales split a value range into equally distributed ranges of {8, 6, 1, 1/6, 1/8}. It is crucial to interpret these results thoughtfully before engaging in further discussions on alternatives, including non-linear combinations of statements within multi-criteria decision analysis frameworks. The interpretation should be undertaken by principals and managers, considering the context of different systems.

Rozell describes the challenges of using qualitative and semi-qualitative risk ranking systems. When time and resources are limited, obtaining a simple, fully quantitative risk assessment or an informal expert managerial review and judgment are considered better approaches [164]. In this dissertation, expert managerial review and judgment are the core of the risk registers across all three layers [6].

The innovation of the dissertation is not in the scoring but instead in the measurement of risk via disruptions of system order by scenarios. The readers are encouraged to select their ways of ordering and re-ordering the initiatives. The identification of scenarios that most disrupt the system order helps healthcare professionals in the characterization of AI-related risks. This characterization occurs in parallel across various system layers: *Purpose, Structure,* and *Function*. The method contributes to the reduction of errors, offering a user-friendly interface that enhances accessibility and ease of use. It

promotes adaptability, providing flexibility to accommodate diverse healthcare settings and contexts. This usability fosters increased engagement from both experts and actors, facilitating a more inclusive and comprehensive analysis of AI-related risks within the healthcare sector.

8.3. Additional Limitations

This section will describe the additional limitations of this dissertation. As a scenario-based methodology, this dissertation identifies the least and most disruptive scenarios within the context of the identified scenarios based on the available sources and data. Limited access to additional data and documents, as well as restricted expert engagement, are additional limitations. As an example, a limitation within the context of the *Function (Phi)* layer was that two patients were subjected to interviews for illustration due to the scarcity of patients suspected of having cardiac sarcoidosis and restrictions policies in Germany that limited direct access to patients for privacy and information protection; also, the scarcity of cardiac sarcoidosis disease. As a result, the number of patients included in the case study was limited to two. However, to obtain more precise results, a larger number of patients are required for the case study as a sample.

It is important to consider the potential for biases among experts and actors during the interview process, given their diverse motivations. To mitigate any strategic or manipulative behavior that might affect the analysis results, conducting an investigation focused on identifying the most disruptive scenarios could be beneficial. The primary aim was not solely to aggregate expert inputs

but also to identify areas requiring further examination, preserving the unique influences of individual experts and actors.

By acknowledging the biases and perspectives of individuals and communities, the proposed scenarios can effectively capture the diverse weights assigned by different experts and actors [58]. The matter of expert bias is of concern, not only in this context but also across the broader field. Various approaches could be employed to alleviate such biases. These methods include techniques such as simple averaging, assigning importance weights to experts and actors, employing the analytic hierarchy process (AHP), the fuzzy analytic hierarchy process (FAHP), decomposing complex problems into multiple layers, and others. Experts and actors could be weighted in future efforts according to their level of expertise in the field [10,11].

8.4. Summary

The chapter has described the highlights of the three cases: *Purpose (Pi), Structure (Sig),* and *Function (Phi)*, alongside discussing additional limitations and lessons learned. Furthermore, it has provided a comparison between the high ranked initiatives and the most disruptive scenarios across different layers of the healthcare system.

Chapter 9 | Discussion of Research Opportunities

9.1. Introduction

This chapter describes a complementary introduction to a fundamental methodology for comparing the order disruption of unrelated systems with different performance metrics and how it could assist managers, experts, and actors in making decisions for each system.

9.2. On Evaluating System Resilience by the Trajectory of Order Disruption

Overview

This section is an overview of evaluating system resilience by the trajectory of order disruption that will be used to compare various healthcare *Purpose (Pi),*

Structure (Sig), and *Function (Phi)* layers in future work, as more information and data will be available. For a demonstration of the method, three unrelated systems with different performance metrics are described below.

Systems operate differently in nature, and disruption of order in the system is inevitable in the operation of the systems. Comparing the order disruption of unrelated systems with different performance metrics could assist managers, experts, and actors in making decisions for each system [13].

For instance, in a power system, the lost power generation is typically measured in megawatts, and the persistence of the disruption is measured in megawatt-hours, as the area over the resilience curve. In a communications system, the performance in gigabits per second can be integrated over time as gigabits. The metrics of persistence in the two examples above are different. How to compare the resilience of different systems, such as:

- Diminished performance of immigration to the U.S. based on the number of immigrants,
- Diminished performance of technology companies based on generated revenue and
- Diminished performance of pharmaceutical companies based on generated revenue [13].

9.2.1. Order Disruption

Figure 58 describes a simple conceptual example of a disruption order of tokens. The top figure describes a baseline for disruption order analysis by

sorting tokens by their size. However, if the order of the tokens changes, it is considered a disruption to the order of the tokens. The bottom figure describes the disruption in the order of the tokens by their size.

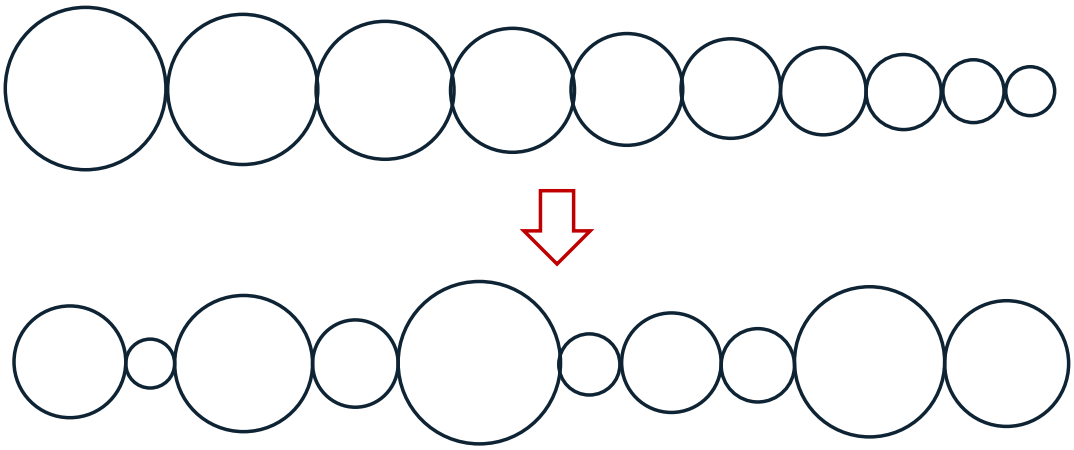

Figure 58. Top figure: Ordered tokens by size. Bottom Figure: Disrupted order tokens [13].

Considering the introduction to order disruption, the following section introduces the methodology for evaluating system resilience by the trajectory of order disruption.

9.2.2. Reilience Curve

Resilience curves represent the level of performance loss after a disruptive event over time. Resiliency curves are used to evaluate the behavior of the system and its resilience to critical scenarios.

Figure 59 describes the disruption of performance or the system resilience curve (top figure) and the disruption of order (bottom figure) which starts with a steady state. This state describes the normal operation and performance of the

system before the disruptive event [165] causes performance loss. Disruption of performance ranges between 0% and infinity since sometimes the performance recovery could exceed the original performance of the system. In other words, the system could recover to more than 100% of its performance when it is capable of enhancing the level of functionality compared to the level before the disruption [166,167]. 100% means the system is fully operational, and 0% means that the system is not operational. The slope of the performance drop, or failure rate, depends on the resistance and adaptive capability of the disrupted system [168]. As the system improves its performance, it is in a recovery state. In some scenarios, the system cannot recover from the disruptive event, which results in performance collapse. After the recovery state, system performance reaches a new recovered steady state called a recovered steady state [15]. During the reaction period, the system is in a disrupted state [13].

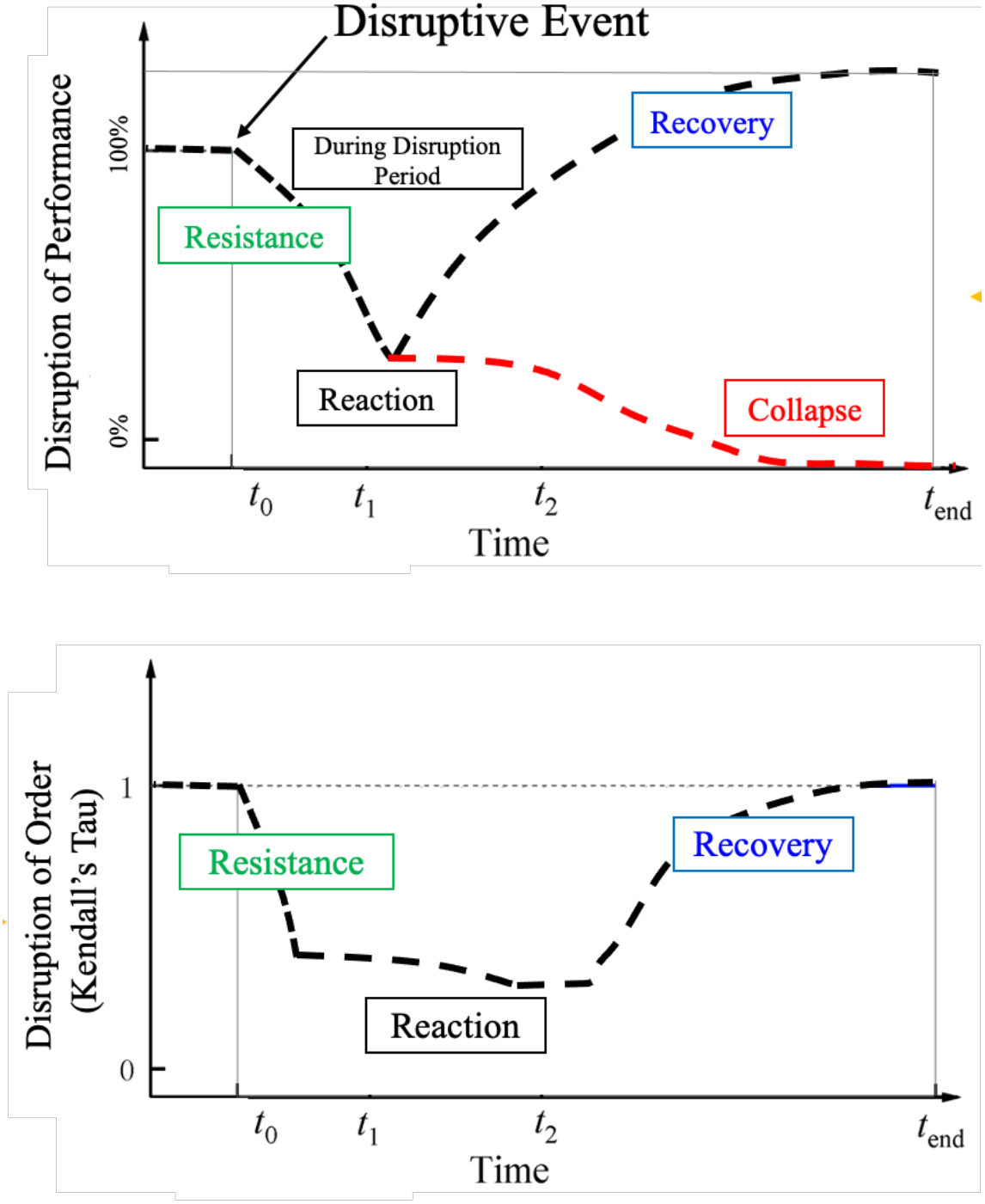

Figure 59. The concept for extending resilience analytics to systems order. The top figure describes a traditional systems resilience curve, and the bottom figure describes the disruption of the order curve [13].

Disruption duration [168], or recovery period, is the linear distance between the occurrence of the disruptive event and the performance reaching the steady recovery point. The area integral above the curve shows the persistence of the lost performance. The larger the distance, the longer the time needed for the system to recover fully [13].

Disruption of performance shows the evolution of system performance by a disrupted event [168,169]. Disruption of order, however, shows the evolution of systems ordered by a disruptive event. Kendall's tau matric is used to calculate the disruption of order, which will be defined in detail below [13].

Figure 60 characterizes different stages of resilience as a function of time. The left subfigure describes a conceptual diagram of a resilience curve for a scenario that disrupts the system plan. The right subfigure describes how a scenario affects the performance of the system, i.e., considering scenario S2, the performance of the system diverges from the plan line and adjusts to a new ending point. This divergence can be obtained from various emergent and future conditions. Emergent conditions are disruptive future events, trends, and other uncertainties that can affect a project, system, schedule, and budget [13]. Scenarios consist of one or more emergent conditions [9,10].

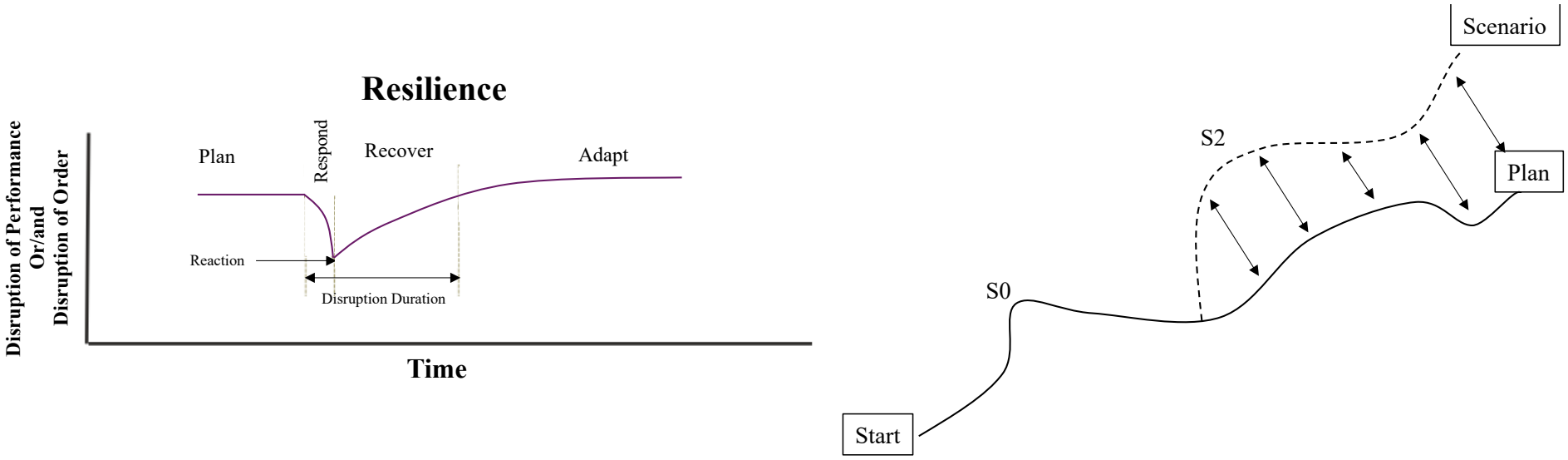

Figure 60. Different stages of resilience as a function of time.

9.2.3. Kendall's Tau

Kendall's tau statistic, τ, in this chapter compares an initial, pre-shock set of priorities with the priorities after the shock. Priorities evolve with time after the shock.

Munoz-Pichardo et al. define Kendall's tau as "the proportional reduction in prediction errors obtained by predicting the ordering of pairs of observations on the objective variable based on the orderings of the pairs on the explanatory variables" [170,171]. Kendall's tau is defined in Equation 18:

$$\tau = \frac{|Pairs_{Condordant}| - |Pairs_{Discondordant}|}{|Pairs_{Condordant}| + |Pairs_{Discondordant}|} \quad (18)$$

Assuming x and y are a pair point one, one point is xi, yi and the other point is xj, yj that $i \neq j$. Concordant pair definition is $xi > xj$ and $yi > yj$ or $xi < xj$ and $yi < yj$. Discordant pair definition is $xi > xj$ and $yi < yj$ or $xi < xj$ and $yi > yj$ [172].

This section aims to define a framework to compare countermeasures that improve resilience across different systems, focusing on restoring the disruption of order. The section describes scale-free comparisons of system resilience by the degree of order disruption [13].

Below are examples of unrelated systems that are reviewed as a demonstration in this chapter:

Table 48 describes the top ten countries of origin for immigrants to the U.S. from 2013 to 2020 [173]. The order of countries of origin for immigrants to the U.S. changed yearly. Various pull and push factors influence the order of countries in migration. On the one hand, the pull factors could be employment opportunities, better shelter, and higher standards of living; they also include social and political factors such as better healthcare facilities, religious tolerance, freedom from persecution, and more; on the other hand, the push factors could be economic factors such as lack of employment, low standards of living, and lack of food and shelter [174]. A pandemic as a disruptive scenario could be one of the factors that push immigrants from countries with weak healthcare facilities [13].

Table 49 describes the top ten technology companies based on revenue generation from 2015 to 2022 [175]. The order of technology companies was changing due to multiple disruptions. For instance, in 2020, as COVID-19 spread globally and caused major lockdowns, internet applications became more popular among users. The growth in the use of Internet applications has become a supercharging business for tech companies. They started to hire more employees as the profit grew, but after the pandemic was over, the yields dropped [176]. The pandemic could be named as one of the disruptive scenarios for order changing [13].

Table 50 describes the top ten valuable brands in the pharmaceutical industry from 2015 to 2022 [175]. The order of pharmaceutical companies was another domain affected by a disruptive scenario like the pandemic. For instance, Pfizer won the pandemic with its mRNA vaccine, which holds 70% of the U.S. and

European markets, and the antiviral Paxlovid pills to treat early symptoms of COVID-19 [13]. The profit of Pfizer has roughly doubled from 2020 to 2021 [177].

Table 48. Top ten countries of origin for immigrants to the U.S. from 2013 to 2020 [13].

	2020	**2019**	**2018**	**2017**	**2016**	**2015**	**2014**	**2013 (BASELINE)**
1	Mexico	Mexico	Mexico	Mexico	Mexico	Mexico	Mexico	Mexico
2	India	China	Cuba	China	China	China	India	China
3	China	India	China	Cuba	Cuba	India	China	India
4	Dominican Republic	Dominican Republic	India	India	India	Philippines	Philippines	Philippines
5	Vietnam	Philippines	Dominican Republic	Dominican Republic	Dominican Republic	Cuba	Cuba	Dominican Republic
6	Philippines	Cuba	Philippines	Philippines	Philippines	Dominican Republic	Dominican Republic	Cuba
7	El Salvador	Vietnam	Vietnam	Vietnam	Vietnam	Vietnam	Vietnam	Vietnam
8	Brazil	El Salvador	El Salvador	El Salvador	Haiti	Iraq	South Korea	South Korea
9	Cuba	Jamaica	Haiti	Jamaica	El Salvador	El Salvador	El Salvador	Colombia
10	South Korea	Colombia	Jamaica	Haiti	Jamaica	Pakistan	Iraq	Haiti

Table 49. Top ten technology companies from 2015 to 2022 [13].

	2022	2021	2020	2019	2018	2017	2016	2015 (BASELINE)
1	Apple	Apple	Amazon	Amazon	Amazon	Google	Apple	Apple
2	Amazon	Amazon	Google	Apple	Apple	Apple	Google	Google
3	Google	Google	Apple	Google	Google	Amazon	Amazon	Microsoft
4	Microsoft	Microsoft	Microsoft	Microsoft	Microsoft	Microsoft	Microsoft	Samsung
5	Facebook	Samsung	Facebook	Facebook	Samsung	Facebook	Facebook	Amazon
6	Samsung	Facebook	Samsung	Samsung	Facebook	Samsung	IBM	General Electric
7	Huawei	WeChat	Huawei	Huawei	Tencent	IBM	General Electric	IBM
8	WeChat	Tencent	WeChat	WeChat	Huawei	Alibaba	Intel	Intel
9	TikTok	Huawei	Tencent	Tencent	IBM	Oracle	Oracle	Facebook
10	Taobao	Taobao	Taobao	Taobao	Oracle	Huawei	Huawei	Oracle

Table 50. Top ten pharmaceutical brands from 2015 to 2022 [13].

	2022	**2021**	**2020**	**2019**	**2018**	**2017**	**2016**	**2015 (BASELINE)**
1	Johnson & Johnson	Johnson & Johnson	Johnson & Johnson	Johnson & Johnson	Roche	Roche	Pfizer	Bayer
2	Roche	Roche	Roche	Roche	Bayer	Pfizer	Bayer	Roche
3	Pfizer	AbbVie	Bayer	Bayer	Pfizer	Bayer	Novartis	Novartis
4	AstraZeneca	Bayer	Abbott	Pfizer	Abbott	Novartis	Roche	Pfizer
5	Bayer	Bristol-Myers Squibb	Merck & Co	Abbott	Novartis	Merck & Co	Merck & Co	Merck & Co
6	AbbVie	Merck & Co	Pfizer	Merck & Co	Sanofi	Celgene	GlaxoSmith Kline	Sanofi
7	Merck & Co	Pfizer	Celgene	Sanofi	Merck & Co	Sanofi	Sanofi	GlaxoSmith Kline
8	Bristol-Myers Squibb	GSK	GSK	Celgene	Celgene	GlaxoSmith Kline	Celgene	Biogen
9	GSK	Novartis	Sanofi	GlaxoSmith Kline	GlaxoSmith Kline	Abbvie	Valeant Pharmace	Valeant Pharmace
10	Sanofi	Sanofi	AbbVie	Novartis	Biogen	Biogen	Biogen	Celgene

Considering the cases have introduced above, the metrics of disruptions are not consistent between systems. The metrics of disruptions in Table 48 are based on the number of immigrants; while those in Table 49, and Table 50 are based on generated revenue. To find how a disruption scenario affects the order of a system each year, Kendall's tau matric is implemented. Kendall's tau is used to measure the disruption of systems perspective [12].

To find the Kendall's tau [178] statistic for each system, a number has been randomly assigned to each name in Table 48, Table 49, and Table 50. For instance, in Table 48, numeric 1 has been given to Mexico, 2 to India, 3 to China, and so on. Fifteen values have been given to fifteen countries, sixteen to technology companies, and sixteen to pharmaceutical brands. As seen, the order of the numbers changes each year. Then pairs of years are compared with the baseline year using Equation 18. Most similar pairs receive a Kendall's tau score closer to 1, and the least similar pairs receive a Kendall's tau score closer to 0. The baseline year in Table 48 is 2013, 2015 in Table 49, and 2015 in Table 50. For instance, in Table 50, Kendall's tau score between pairs of 2022 and 2015 is 0.68. Table 51 shows seven calculated Kendall's tau scores from Table 48, Table 49, and Table 50 [13].

Table 51. Calculated Kendall's tau scores for the three cases to estimate disruption of order [179].

Time	Top Ten Pharmaceutical Companies	Top Ten Tech Companies	Top Ten Immigrants to the US
t0	-	-	-
t1	0.64	0.51	0.64
t2	0.29	0.38	0.69
t3	0.38	0.38	0.51
t4	0.69	0.42	0.60
t5	0.42	0.38	0.60
t6	0.51	0.60	0.82
t7	0.68	0.51	0.82

Figure 61 describes a scale-free disruption order chart comparing three different systems with different units and timelines. Considering various emergent conditions and scenarios, the chart describes how a system is disrupted over time. For example, using Kendall's tau measurement for the top ten pharmaceutical companies, pairs of years 2016 to 2022 are compared with the baseline year 2015, and the score is generated. Time 0 (t0, baseline) starts with a steady state, but at t1, a disruptive event causes performance loss and a degraded order slope. The disruptive order of the system fluctuated between t1 and t7. Considering various disruptive events, the system disruptive order may recover its pre-shock level or not at the resilience adaptation stage. Accordingly, Kendall's tau statistic compares the countermeasures that improve resilience across three different systems [179].

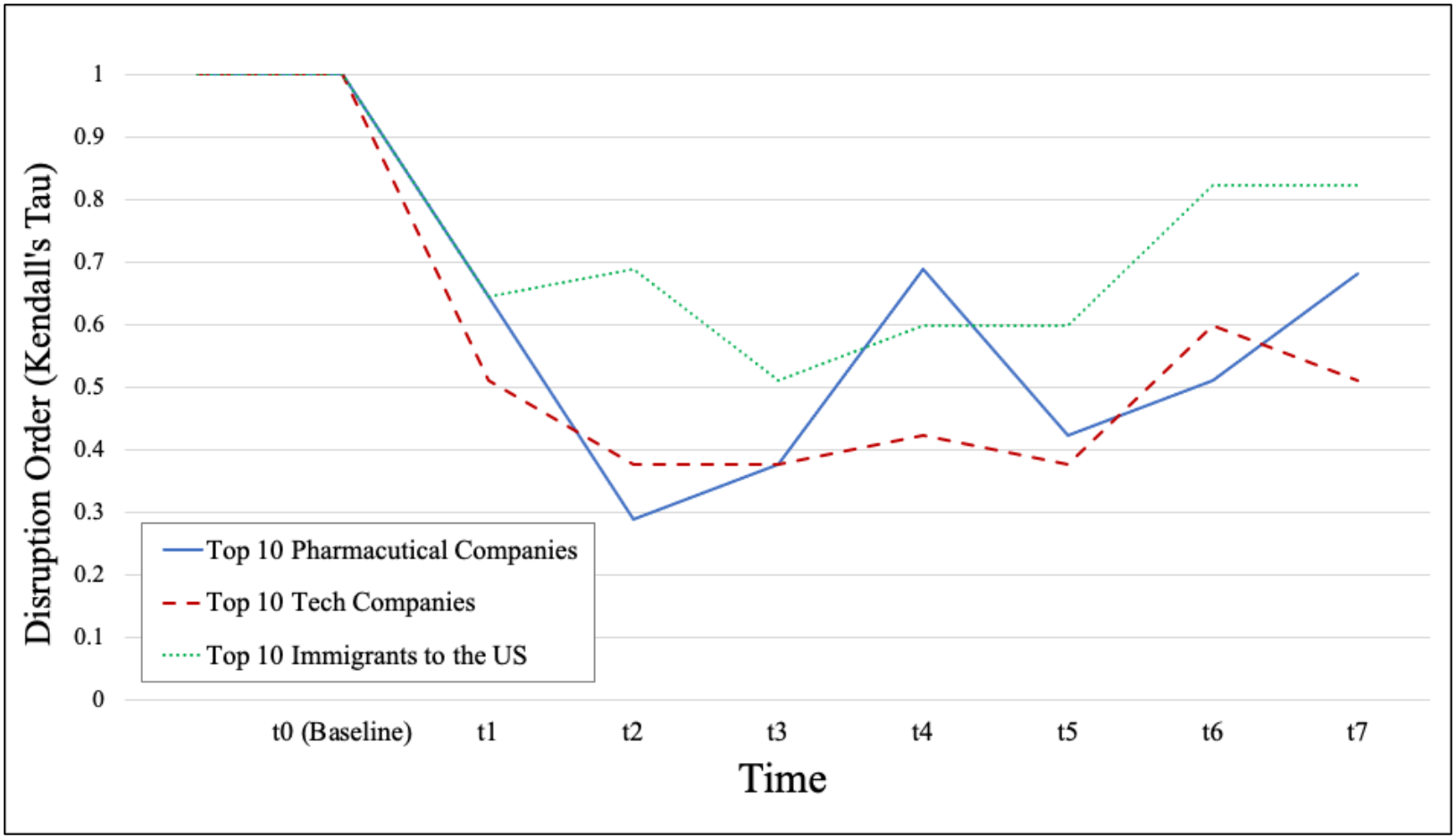

Figure 61. Compare the resilience between three different systems using Kendall's tau [13].

Of course, the readers of the chapter should be interested in how the resilience of system orders matters for engineering design. A key contribution of this chapter is how engineering design can give attention to the scenarios (risks) that are most disruptive to system order [11]. Furthermore, engineering design pays attention to resilience by focusing on the trajectories of system order [180], in which the disruption of order is persistent [179]. The orders can be for system assets, projects, policies, products, processes, etc. The system thus consists of the sets of entities to be ordered and a mechanism for doing the ordering, for example, a figure of merit. The mechanism may be unknown, as in these examples, where only the ordering over time is observable. If it were necessary to know the mechanism of ordering, the usual methods of system identification [155] and estimation are available. To demonstrate the evaluation of resilience in terms of system orders, it is not necessary to know the mechanisms of ordering. Replacing system function with system order in the estimation and integration of resilience curves using Kendall's tau statistic enables a new quantification of resilience as a disruption of system order that can be used across application domains of risk analysis. A reason for scale-free metrics of disruption (for example, Kendall's tau) is that an owner or regulator of multiple systems may be allocating shared and common resources to enhance resilience or identify risk. Ongoing work in the field of engineering systems is focused on disrupted order rather than disrupted performance [179].

Readers should understand the technical mechanism of disruption that is underlying each of the examples. That is not a topic of this chapter, though it would be a worthwhile focus of system identification [54,181].

The following steps of this effort are to study various mechanisms that underlie order disruption and to characterize how understanding those mechanisms is important to engineering design. While this chapter is an introduction to the shift in resiliency focus from performance to order, more studies could be provided on order and the implications for engineered systems. For example, how could one use Kendall's tau to engineer more resilient systems, recover from a disruptive event, or predict future events [179]. Examples of applications to be described include energy systems, communications systems, freight logistics, and cybersecurity.

9.3. Summary

This chapter has described an introduction to the research opportunity for evaluating system resilience by the trajectory of order disruption. This complementary model of a system as priority orders will be implemented to compare various healthcare *Purpose (Pi), Structure (Sig)*, and *Function (Phi)* layers as more information and data become available.

Chapter 10 | Conclusions and Future Work

10.1. Introduction

This chapter describes the summary of contributions, summary conclusions, dissertation schedule and timeline, and other future works.

10.2. Summary of Contributions

This dissertation makes eight contributions at the intersection of risk analysis, systems modeling, and AI applications in healthcare. The seven contributions to this dissertation are as follows:

Contribution 1. Development of a Mathematical Framework for the Risks of AI in Healthcare

This dissertation is the first to develop a mathematical framework to assess AI risks in healthcare across several system layers and to explain how disruptive scenarios influence system priorities. This framework aids in understanding and identifying potential disruptions. The objective is to identify scenarios that are most and least disruptive to the system order. The framework identifies various biases, focusing on those with the most disruptive impact on order. Rather than eliminate bias, the framework accounts for potential biases (scenarios) within each of the layers [6,10,11,18,51].

Contribution 2. Healthcare Center-Level Demonstration (*Purpose* Layer)

This dissertation is the first to do the case demonstration, which describes the applicability of the framework at the operational level within healthcare centers that underlies its effectiveness in real-world settings [6].

Contribution 3: Healthcare Device-Level Demonstration (*Structure* Layer)

This dissertation is the first to do the case demonstration, which describes the applicability of the framework at the device level, specifically its utility in analyzing the risks of AI in the Vaso-Lock optimization design [6].

Contribution 4. Healthcare Process-Level Demonstration (*Function* Layer)

This dissertation is the first to do the case demonstration, which describes the applicability of the framework at the process level or disease diagnosis,

specifically its utility in analyzing the risks of AI in cardiac sarcoidosis diagnosis [5,6,18].

Contribution 5. Systems Simulation

This dissertation is the first to do a machine learning-based diagnosis of cardiac sarcoidosis using multi-chamber wall motion analysis at the *Function (Phi)* layer. Explainable AI (XAI) techniques describe the validity of the predictions and explain and interpret the outcomes. Additionally, XAI techniques in a design optimization framework determine the optimal geometry of a vascular anastomosis device at the *Structure (Sig)* layer [5,6,18,40].

Contribution 6. Priority in Experts and Actors[13] Involvement Effect

This dissertation is the first to do comparative analysis and introduces a multi-layered scenario-based disruption of priorities for the risk of AI in healthcare at the *Function (Phi)* layer, which involves two expert groups: Patients versus medical experts/physicians [6,68].

Contribution 7. Evaluation of System Resilience

This dissertation is the first to evaluate system resilience and not only synthesizes lessons learned for the NIST and other practitioners but also provides actionable insights for enhancing future risk management strategies in healthcare [6,18,40,68].

[13] Actors could refer to experts and stakeholders through the Dissertation.

Contribution 8. System Resilience Evaluation Based on the Degree of Order Disruption

This dissertation is the first to introduce an approach for evaluating system resilience based on the degree of order disruption in various systems [13].

Seventeen publications from 2020 to 2024 document the previously mentioned contributions.

10.3. Summary of Conclusions

This section describes the summary conclusions of the chapters above. This dissertation focuses on research and development priorities for managing the risks associated with the risk of AI in health applications [182,183]. The methodology serves as a demonstration and emphasizes the constraints associated with the chosen scenarios and the partial consideration of system layers. The methodology identifies success criteria, R&D initiatives, and emergent conditions across multiple layers of the healthcare system, including the *Purpose (Pi)* layer, implant/device or the *Structure (Sig)* layer, and disease diagnosis or the *Function (Phi)* layer [6].

Figure 62 describes the dissertation theory and method conceptual diagram of systems modeling for enterprise risk management of AI in healthcare. Each related chapter of the dissertation is noted for each section.

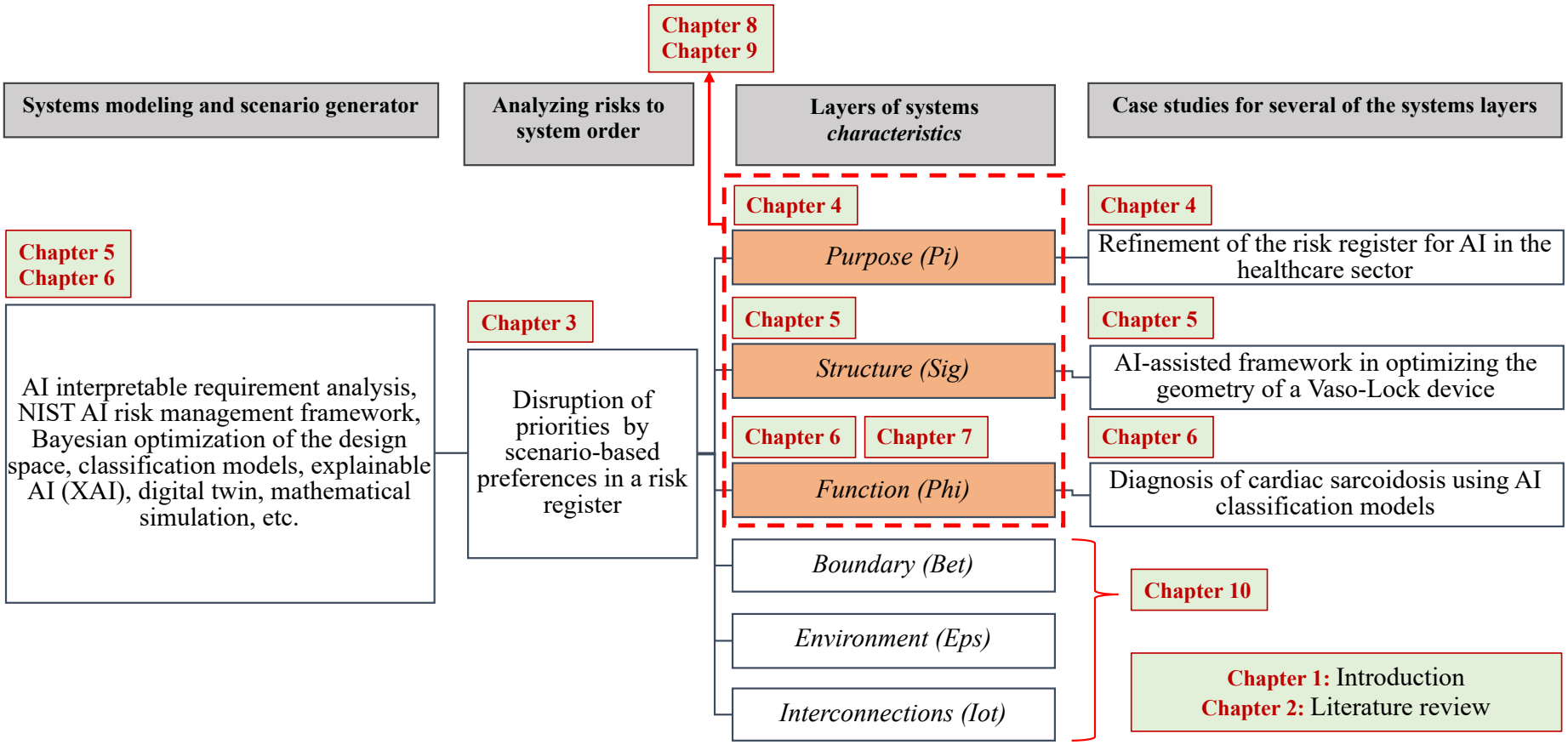

Figure 62. The dissertation theory and method conceptual diagram of systems modeling for enterprise risk management of AI in healthcare, with the notation of related chapters to each section [5,6].

The core concept of the dissertation is not to make the judgments required by the model; instead, the focus is on measuring disruptive order. In other words, the emphasis is on adapting a figure of merit to score the initiatives and rank them rather than performing decision analysis [6].

This dissertation achieves a balance between the goals of AI, human rights, and societal values by considering seven *principles* of the NIST AI risk management framework as the success criteria for all layers, and involving a variety of perspectives, managers, experts, and actors in each system layer in the process. By analyzing these initiatives, emergent conditions, and scenarios within the healthcare system layers, the dissertation identifies the most and least disruptive scenarios based on expert preferences [10]. This information allows experts, actors, and managers to make informed decisions regarding resource allocation and prioritize specific initiatives over others [6].

The initiatives outlined in this dissertation hold promise for improving communication and identifying the risks associated with AI in healthcare applications involving various experts and actors. Moving forward, it is important to incorporate the viewpoints of healthcare practitioners and patients who are directly impacted by these approaches [6].

Notably, the methods described in this dissertation can offer patients insights into the relevance of AI applications in their treatment plans, promoting transparency for both patients and care partners. The initiatives and emergent conditions discussed in this dissertation provide a foundation for future research, which will build upon these findings to delve deeper into the subject. Further investigations will expand the analysis to encompass additional layers, such as the *Boundary (Bet)* that exists between patients and society. The extended scope

of this research endeavor will delve into the wider ramifications of AI within healthcare systems, elucidating its effects on diverse facets of society [6].

In summary, addressing the major challenge of risk assessment for AI tools, this dissertation introduces a context-specific approach to understanding the risks associated with AI, emphasizing that these risks cannot be universally applied. The proposed AI risk framework in this dissertation recognizes insights into quantifying risk by assessing the disturbance to the order of AI initiatives in healthcare systems. Additionally, the dissertation highlights the significant role of the human-in-the-loop in identifying the risks associated with AI in healthcare and evaluating and improving the suggestions and outcomes of AI systems [6].

There are additional components of a practical AI risk management framework that may improve the accuracy and consistency of outputs produced by AI. These include fostering diversity among participants [160], identifying AI effects in terms of ethics [184], law, society, and technology, seeking official guidelines from experts and actors, considering various social values, enhancing and improving unbiased algorithms and data quality by prioritizing privacy and security, and regular maintenance of AI systems [34]. Moreover, identifying and minimizing uncertainties and unexpected scenarios, adhering to ethical and legal standards, ensuring the correctness of AI outputs and predictions through various validation and assessment practices, such as employing XAI techniques [41], ensuring human-AI teaming [160] and collaboration, and optimizing AI features and performance during design and implementation, among other aspects, are more components of a practical AI risk management framework. Given different business sizes and resource availability, and based on the

experience mentioned above, there is a need and opportunity for each system principal to determine appropriate AI risk management frameworks [6].

There are several potential methods for identifying reliable and trustworthy formal guidance for AI risk management. Seeking government guidance and guidelines from officials, R&D findings from industry and academia, verifying compliance with standard and legal protocols, and more could be some of the sources for risk management with AI. Several safeguards and security measures can be implemented to assist the dependability and more accurate operation of AI systems, such as validating the results by engaging the patients, medical professionals, and system designers in the loop, identifying and mitigating the risks of uncertain scenarios to the system, regular monitoring, and updating and training the system to adhere to ethical and lawful standards and protocols [6].

The methods outlined in the dissertation hold potential for cross-domain applicability beyond the healthcare sector. They can be adapted and applied to diverse fields such as transportation, finance, design, risk analysis of quantum technologies in medicine, and more [185]. By enhancing transparency and addressing the associated risks of AI, the research benefits not only healthcare systems globally but also various other applications and industries. The findings and insights gained from this dissertation can inform and guide the development and implementation of AI systems in a wide range of domains, such as supply chains, disaster management, emergency response, and more, fostering responsible and effective use of this technology. Also, the method and its rubrics have general relevance to a variety of life science topics across medical diagnosis, epidemiology, pathology, pharmacology, toxicology, microbiology, immunology, and more [6].

10.4. Schedule and Timeline

Figure 63 describes the degree milestones. Figure 64 describes the journal and conference publications and conference presentations from 2020 to 2024.

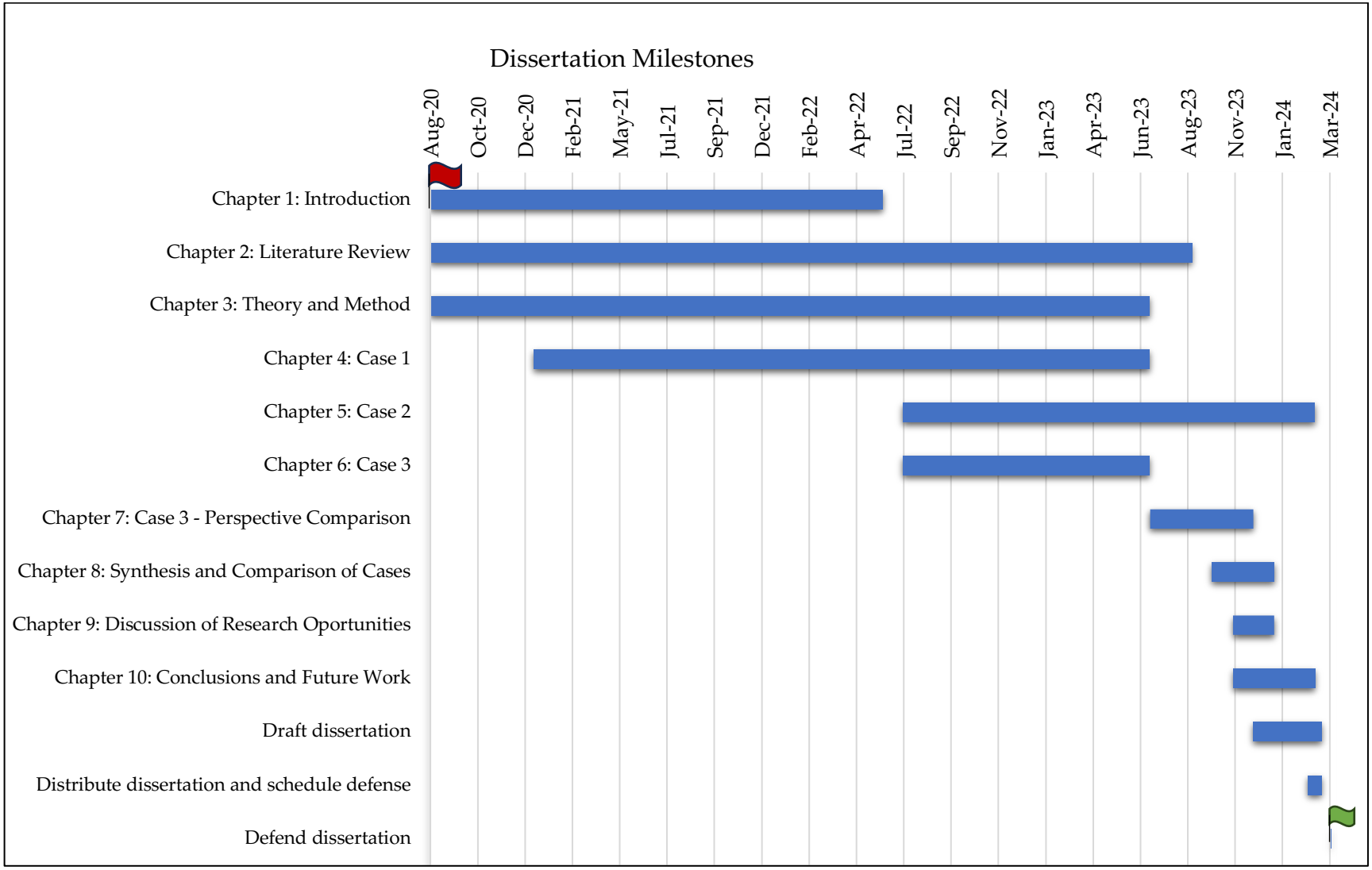

Figure 63. Schedule of dissertation milestones from August 2022 to March 2024.

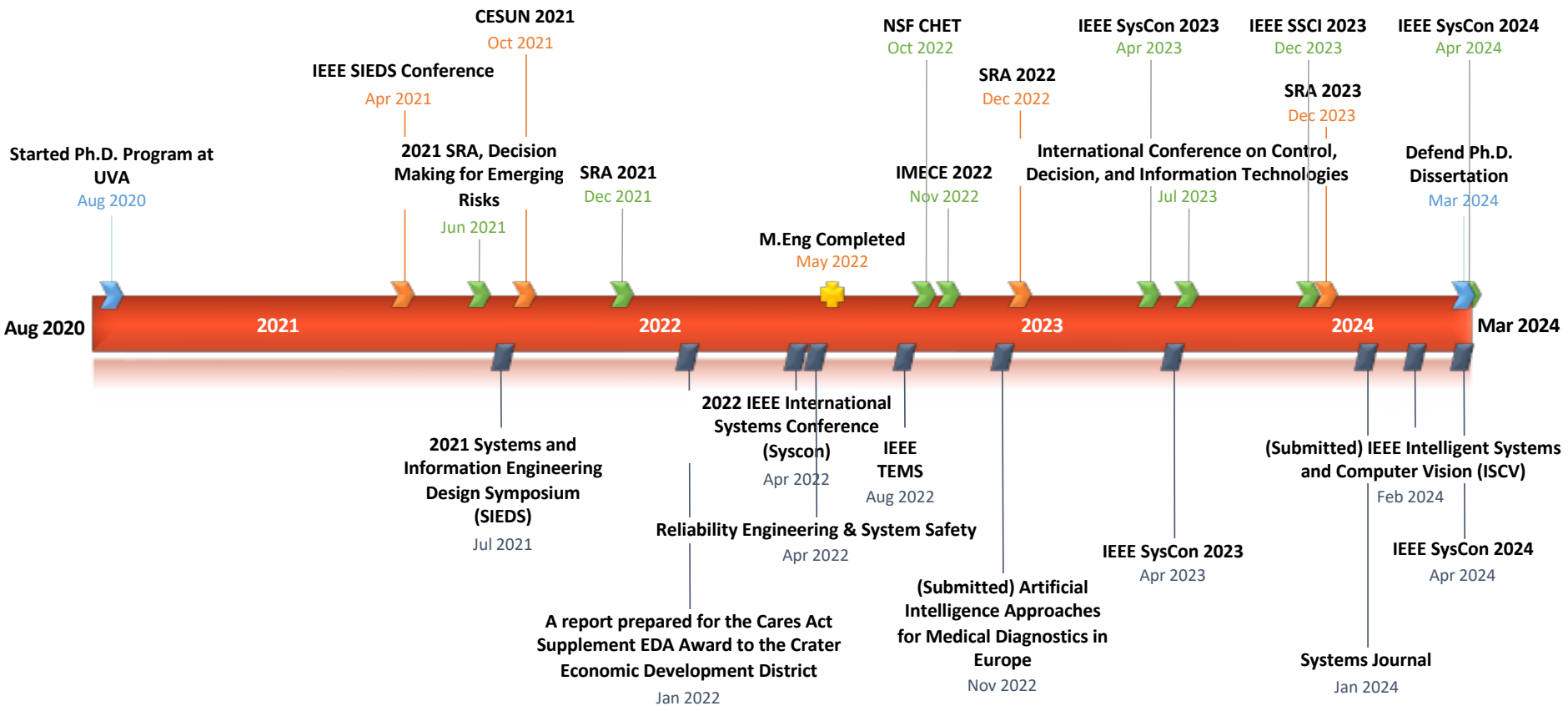

Figure 64. Timeline of conference presentations and publications. Annotations above the timeline represent conference presentations, and annotations below the timeline shows journal and conference publications.

10.5. Other Future Works

This section describes several other future works, including: 1. Other system layers introduced in Figure 1 will be reviewed, such as *Boundary (Bet)*, *Environment (Eps)*, and *Interconnection (Iot)*. For instance, a case study of the *Boundary (Bet)* could be the relations between patients and society. 2. Also, evaluating system resilience by the degree of order disruption for all layers will be addressed as a research opportunity to compare various layers with uncommon units as more information and data become available. 3. A demonstration of various systems with emerging AI applications, including cybersecurity of electric vehicle charging systems and supply chains, transportation, infrastructure, architecture, design, etc.

10.6. Summary

This chapter has described the summary of contributions, summary conclusions, dissertation schedule and timeline, and other future works.

In summary, deciding whether to commission or deploy an AI system should involve evaluating its trustworthiness *characteristics* in the given context, along with weighing the relative risks, impacts, costs, and benefits, while considering the opinions of various experts and actors. As mentioned earlier in the dissertation, in situations where the consequences of the actions of the system could be severe, such as when human life or liberty are at risk, AI developers and deployers should take proactive measures to adjust their transparency and accountability practices proportionally. Organizations need to have practices and

governance structures in place that focus on reducing harm, such as risk management, to ensure that their systems are held accountable.

References

1. Hagras, H. Toward Human-Understandable, Explainable AI. *Computer* **2018**, *51*, 28–36, doi:10.1109/MC.2018.3620965.
2. FDA Statement Statement from FDA Commissioner Scott Gottlieb, M.D. on Steps toward a New, Tailored Review Framework for Artificial Intelligence-Based Medical Devices. *U.S. Food & Drug Administration* 2019.
3. Suresh, H. Investigating Physician Trust in AI Systems: Where Is the Line between Effective Collaboration and over-Reliance? *HEALTH MIT* 2021.
4. Hasani, N.; Morris, M.A.; Rahmim, A.; Summers, R.M.; Jones, E.; Siegel, E.; Saboury, B. Trustworthy Artificial Intelligence in Medical Imaging. *PET Clinics* **2022**, *17*, 1–12, doi:10.1016/j.cpet.2021.09.007.
5. Eckstein, J.; Moghadasi, N.; Körperich, H.; Akkuzu, R.; Sciacca, V.; Sohns, C.; Sommer, P.; Berg, J.; Paluszkiewicz, J.; Burchert, W.; et al. Machine Learning-Based Diagnostics of Cardiac Sarcoidosis Using Multi-Chamber Wall Motion Analyses. *Diagnostics* **2023**.
6. Moghadasi, N.; Valdez, R.S.; Piran, M.; Moghaddasi, N.; Linkov, I.; Polmateer, T.L.; Loose, D.C.; Lambert, J.H. Risk Analysis of Artificial Intelligence in Healthcare with Disruption of System Order. *MDPI Systems Journal* **2024**.

7. Ligo, A.K.; Kott, A.; Linkov, I. Autonomous Cyberdefense Introduces Risk: Can We Manage the Risk? *Computer* **2021**, *54*, 106–110, doi:10.1109/MC.2021.3099042.
8. Hassler, M.L.; Andrews, D.J.; Ezell, B.C.; Polmateer, T.L.; Lambert, J.H. Multi-Perspective Scenario-Based Preferences in Enterprise Risk Analysis of Public Safety Wireless Broadband Network. *Reliability Engineering & System Safety* **2020**, *197*, 106775, doi:10.1016/j.ress.2019.106775.
9. Loose, D.C.; Eddy, T.L.; Polmateer, T.L.; Manasco, M.C.; Moghadasi, N.; Lambert, J.H. Managing Pandemic Resilience with Other Cascading Disruptions of a Socio-Technical System. In Proceedings of the 2022 IEEE International Systems Conference (SysCon); IEEE: Montreal, QC, Canada, April 25 2022; pp. 1–6.
10. Moghadasi, N.; Collier, Z.A.; Koch, A.; Slutzky, D.L.; Polmateer, T.L.; Manasco, M.C.; Lambert, J.H. Trust and Security of Electric Vehicle-to-Grid Systems and Hardware Supply Chains. *Reliability Engineering & System Safety* **2022**, *225*, 108565, doi:10.1016/j.ress.2022.108565.
11. Moghadasi, N.; Luu, M.; Adekunle, R.O.; Polmateer, T.L.; Manasco, M.C.; Emmert, J.M.; Lambert, J.H. Research and Development Priorities for Security of Embedded Hardware Devices. *IEEE Trans. Eng. Manage.* **2022**, 1–12, doi:10.1109/TEM.2022.3197240.
12. Quenum, A.; Thorisson, H.; Wu, D.; Lambert, J.H. Resilience of Business Strategy to Emergent and Future Conditions. *null* **2019**, 1–19, doi:10.1080/13669877.2018.1485172.
13. Moghadasi, N.; Lambert, J.H. On Evaluating System Resilience by the Degree of Order Disruption.; July 19 2023.
14. Gratt, L.B. The Definition of Risk and Associated Terminology for Risk Analysis. In *Risk Assessment in Setting National Priorities*; Bonin, J.J., Stevenson, D.E., Eds.; Springer US: Boston, MA, 1989; pp. 675–680 ISBN 978-1-4684-5684-4.
15. Yodo, N.; Wang, P. Engineering Resilience Quantification and System Design Implications: A Literature Survey. *Journal of Mechanical Design* **2016**, *138*, doi:10.1115/1.4034223.

16. Collier, Z.A.; DiMase, D.; Walters, S.; Tehranipoor, M.M.; Lambert, J.H.; Linkov, I. Cybersecurity Standards: Managing Risk and Creating Resilience. *Computer* **2014**, *47*, 70–76, doi:10.1109/MC.2013.448.
17. Haimes, Y.Y.; Kaplan, S.; Lambert, J.H. Risk Filtering, Ranking, and Management Framework Using Hierarchical Holographic Modeling - Haimes - 2002 - Risk Analysis - Wiley Online Library Available online: https://onlinelibrary.wiley.com/doi/abs/10.1111/0272-4332.00020 (accessed on 13 June 2022).
18. Moghadasi, N.; Piran, M.; Baek, S.; Valdez, R.S.; Porter, M.D.; Johnson, D.; Lambert, J.H. Systems Analysis of Bias and Risk in AI-Enabled Medical Diagnosis.; Mexico City, Mexico, September 2023.
19. Bohr, A.; Memarzadeh, K. The Rise of Artificial Intelligence in Healthcare Applications. In *Artificial Intelligence in Healthcare*; Elsevier, 2020; pp. 25–60 ISBN 978-0-12-818438-7.
20. Gainey, J.C.; He, Y.; Zhu, R.; Baek, S.S.; Wu, X.; Buatti, J.M.; Allen, B.G.; Smith, B.J.; Kim, Y. Predictive Power of Deep-Learning Segmentation Based Prognostication Model in Non-Small Cell Lung Cancer. *Front. Oncol.* **2023**, *13*, 868471, doi:10.3389/fonc.2023.868471.
21. Bullock, J.; Grieco, M.; Liu, Y.; Pedersen, I.; Roberson, W.; Wright, G.; Alonzi, P.; McCulloch, M.A.; Porter, M.D. Determining Factors of Heart Quality and Donor Acceptance in Pediatric Heart Transplants. In Proceedings of the 2021 Systems and Information Engineering Design Symposium (SIEDS); IEEE: Charlottesville, VA, USA, April 30 2021; pp. 1–6.
22. Pati, S.; Baid, U.; Edwards, B.; Sheller, M.; Wang, S.-H.; Reina, G.A.; Foley, P.; Gruzdev, A.; Karkada, D.; Davatzikos, C.; et al. Federated Learning Enables Big Data for Rare Cancer Boundary Detection. **2022**, doi:10.48550/ARXIV.2204.10836.
23. Dicuonzo, G.; Donofrio, F.; Fusco, A.; Shini, M. Healthcare System: Moving Forward with Artificial Intelligence. *Technovation* **2023**, *120*, 102510, doi:10.1016/j.technovation.2022.102510.
24. Dauda, O.I.; Awotunde, J.B.; Muyideen AbdulRaheem; Salihu, S.A. Basic Issues and Challenges on Explainable Artificial Intelligence (XAI) in

Healthcare Systems: In *Advances in Medical Technologies and Clinical Practice*; Albuquerque, V.H.C. de, Srinivasu, P.N., Bhoi, A.K., Briones, A.G., Eds.; IGI Global, 2022; pp. 248–271 ISBN 978-1-66843-791-9.

25. Grodzinsky, F.S.; Wolf, M.J.; Miller, K.W. Ethical Issues From Emerging AI Applications: Harms Are Happening. *Computer* **2024**, *57*, 44–52, doi:10.1109/MC.2023.3332850.
26. Khan, B.; Fatima, H.; Qureshi, A.; Kumar, S.; Hanan, A.; Hussain, J.; Abdullah, S. Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector. *Biomedical Materials & Devices* **2023**, *1*, 731–738, doi:10.1007/s44174-023-00063-2.
27. Haresamudram, K.; Larsson, S.; Heintz, F. Three Levels of AI Transparency. *Computer* **2023**, *56*, 93–100, doi:10.1109/MC.2022.3213181.
28. Mariani, R.; Rossi, F.; Cucchiara, R.; Pavone, M.; Simkin, B.; Koene, A.; Papenbrock, J. Trustworthy AI—Part 1. *Computer* **2023**, *56*, 14–18, doi:10.1109/MC.2022.3227683.
29. Mohler, G.; Porter, M.D. A Note on the Multiplicative Fairness Score in the NIJ Recidivism Forecasting Challenge. *Crime Sci* **2021**, *10*, 17, doi:10.1186/s40163-021-00152-x.
30. Valdez, R.S.; Detmer, D.E.; Bourne, P.; Kim, K.K.; Austin, R.; McCollister, A.; Rogers, C.C.; Waters-Wicks, K.C. Informatics-Enabled Citizen Science to Advance Health Equity. *Journal of the American Medical Informatics Association* **2021**, *28*, 2009–2012, doi:10.1093/jamia/ocab088.
31. Valdez, R.S.; Ancker, J.S.; Veinot, T.C. Provocations for Reimagining Informatics Approaches to Health Equity. *Yearb Med Inform* **2022**, *31*, 015–019, doi:10.1055/s-0042-1742514.
32. Avin, S.; Belfield, H.; Brundage, M.; Krueger, G.; Wang, J.; Weller, A.; Anderljung, M.; Krawczuk, I.; Krueger, D.; Lebensold, J.; et al. Filling Gaps in Trustworthy Development of AI. *Science* **2021**, *374*, 1327–1329, doi:10.1126/science.abi7176.
33. Jain, S.; Luthra, M.; Sharma, S.; Fatima, M. Trustworthiness of Artificial Intelligence. In Proceedings of the 2020 6th International Conference on

Advanced Computing and Communication Systems (ICACCS); IEEE: Coimbatore, India, March 2020; pp. 907–912.

34. Tabassi, E. *AI Risk Management Framework: AI RMF (1.0)*; National Institute of Standards and Technology: Gaithersburg, MD, 2023; p. error: NIST AI 100-1;
35. Kiseleva, A.; Kotzinos, D.; De Hert, P. Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. *Front. Artif. Intell.* **2022**, *5*, 879603, doi:10.3389/frai.2022.879603.
36. Eckstein, J.; Moghadasi, N.; Körperich, H.; Weise Valdés, E.; Sciacca, V.; Paluszkiewicz, L.; Burchert, W.; Piran, M. A Machine Learning Challenge: Detection of Cardiac Amyloidosis Based on Bi-Atrial and Right Ventricular Strain and Cardiac Function. *Diagnostics* **2022**, *12*, 2693, doi:10.3390/diagnostics12112693.
37. Galaitsi, S.; Trump, B.D.; Keisler, J.M.; Linkov, I.; Kott, A. Cybertrust: From Explainable to Actionable and Interpretable AI (AI2). **2022**, doi:10.48550/ARXIV.2201.11117.
38. Mbiazi, D.; Bhange, M.; Babaei, M.; Sheth, I.; Kenfack, P.J. Survey on AI Ethics: A Socio-Technical Perspective. **2023**, doi:10.48550/ARXIV.2311.17228.
39. Bodria, F.; Giannotti, F.; Guidotti, R.; Naretto, F.; Pedreschi, D.; Rinzivillo, S. Benchmarking and Survey of Explanation Methods for Black Box Models. *Data Min Knowl Disc* **2023**, *37*, 1719–1778, doi:10.1007/s10618-023-00933-9.
40. Moghadasi, N.; Piran, M.; Valdez, R.S.; Baek, S.; Moghaddasi, N.; Polmateer, T.L.; Lambert, J.H. Process Quality Assurance of AI-Enabled Medical Diagnosis.; IEEE ISCV: Fez, Morocco, February 2024.
41. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160, doi:10.1109/ACCESS.2018.2870052.
42. Palacio, S.; Lucieri, A.; Munir, M.; Ahmed, S.; Hees, J.; Dengel, A. XAI Handbook: Towards a Unified Framework for Explainable AI. In Proceedings of the 2021 IEEE/CVF International Conference on Computer

Vision Workshops (ICCVW); IEEE: Montreal, BC, Canada, October 2021; pp. 3759–3768.

43. Manresa-Yee, C.; Roig-Maimó, M.F.; Ramis, S.; Mas-Sansó, R. Advances in XAI: Explanation Interfaces in Healthcare. In *Handbook of Artificial Intelligence in Healthcare*; Lim, C.-P., Chen, Y.-W., Vaidya, A., Mahorkar, C., Jain, L.C., Eds.; Intelligent Systems Reference Library; Springer International Publishing: Cham, 2022; Vol. 212, pp. 357–369 ISBN 978-3-030-83619-1.
44. Zolanvari, M.; Yang, Z.; Khan, K.; Jain, R.; Meskin, N. TRUST XAI: Model-Agnostic Explanations for AI With a Case Study on IIoT Security. *IEEE Internet Things J.* **2023**, *10*, 2967–2978, doi:10.1109/JIOT.2021.3122019.
45. Kapcia, M.; Eshkiki, H.; Duell, J.; Fan, X.; Zhou, S.; Mora, B. ExMed: An AI Tool for Experimenting Explainable AI Techniques on Medical Data Analytics. In Proceedings of the 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI); IEEE: Washington, DC, USA, November 2021; pp. 841–845.
46. Mehdiyev, N.; Majlatow, M.; Fettke, P. Interpretable and Explainable Machine Learning Methods for Predictive Process Monitoring: A Systematic Literature Review. **2023**, doi:10.48550/ARXIV.2312.17584.
47. Moghadasi, N. On a Framework for Enterprise Risk Management of AI for Healthcare Applications, University of Virginia, Department of Systems and Information Engineering, 2024.
48. Rawal, A.; McCoy, J.; Rawat, D.B.; Sadler, B.M.; Amant, R.St. Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives. *IEEE Trans. Artif. Intell.* **2022**, *3*, 852–866, doi:10.1109/TAI.2021.3133846.
49. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM: San Francisco California USA, August 13 2016; pp. 1135–1144.

50. Tjoa, E.; Xie, F.; Shu, C.-H.; Xue, L.; Aghaeepour, N.; Guan, C. *Endorsed Attributions: eXplainable AI (XAI) with Voting Mechanism with Application in Healthcare*; 2023;
51. VanYe, C.M.; Li, B.E.; Koch, A.T.; Luu, M.N.; Adekunle, R.O.; Moghadasi, N.; Collier, Z.A.; Polmateer, T.L.; Barnes, D.; Slutzky, D.; et al. Trust and Security of Embedded Smart Devices in Advanced Logistics Systems. In Proceedings of the 2021 Systems and Information Engineering Design Symposium (SIEDS); IEEE: Charlottesville, VA, USA, April 30 2021; pp. 1–6.
52. Sage, A.P.; Lynch, C.L. Systems Integration and Architecting: An Overview of Principles, Practices, and Perspectives. *Syst. Engin.* **1998**, *1*, 176–227, doi:10.1002/(SICI)1520-6858(1998)1:3<176::AID-SYS3>3.0.CO;2-L.
53. Holt, J. *Systems Engineering Demystified*; Packt Publishing: S.l., 2021; ISBN 978-1-83898-580-6.
54. SEBoK Editorial Board *Guide to the Systems Engineering Body of Knowledge (SEBoK), Version 2.7*; 2022;
55. *Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities*; Walden, D.D., Roedler, G.J., Forsberg, K., Hamelin, R.D., Shortell, T.M., International Council on Systems Engineering, Eds.; 4th edition.; Wiley: Hoboken, New Jersey, 2015; ISBN 978-1-118-99941-7.
56. Almutairi, A.; Thorisson, H.; Wheeler, J.P.; Slutzky, D.L.; Lambert, J.H. Scenario-Based Preferences in Development of Advanced Mobile Grid Services and a Bidirectional Charger Network. *ASCE-ASME J. Risk Uncertainty Eng. Syst., Part A: Civ. Eng.* **2018**, *4*, 04018017, doi:10.1061/AJRUA6.0000962.
57. Moghadasi, N.; Collier, Z.A.; Koch, A.; Slutzky, D.L.; Polmateer, T.L.; Manasco, M.C.; Lambert, J.H. Trust and Security of Electric Vehicle-to-Grid Systems and Hardware Supply Chains. *Reliability Engineering & System Safety* **2022**, *225*, 108565, doi:10.1016/j.ress.2022.108565.
58. Keeney, R.L. COMMON MISTAKES IN MAKING VALUE TRADE-OFFS. *JSTOR* **2002**.

59. Keeney, R.L.; Raiffa, H.; Rajala, D.W. Decisions with Multiple Objectives: Preferences and Value Trade-Offs. *IEEE Trans. Syst., Man, Cybern.* **1979**, *9*, 403–403, doi:10.1109/TSMC.1979.4310245.
60. Stillwell, W.G.; Seaver, D.A.; Edwards, W. A Comparison of Weight Approximation Techniques in Multiattribute Utility Decision Making. *Organizational Behavior and Human Performance* **1981**, *28*, 62–77, doi:10.1016/0030-5073(81)90015-5.
61. Collier, Z.A.; Lambert, J.H. Evaluating Management Actions to Mitigate Disruptive Scenario Impacts in an E-Commerce Systems Integration Project. *IEEE Systems Journal* **2019**, *13*, 593–602, doi:10.1109/JSYST.2018.2812864.
62. Lazzerini, B.; Mkrtchyan, L. Analyzing Risk Impact Factors Using Extended Fuzzy Cognitive Maps. *IEEE Systems Journal* **2011**, *5*, 288–297, doi:10.1109/JSYST.2011.2134730.
63. Keeney, R.L. *Value-Focused Thinking: A Path to Creative Decisionmaking*; Harvard Univ. Press: Cambridge, Mass., 1992; ISBN 978-0-674-93198-5.
64. Belton, V.; Stewart, T.J. *Multiple Criteria Decision Analysis*; Springer US: Boston, MA, 2002; ISBN 978-1-4613-5582-3.
65. Clemen, R.T. *Making Hard Decisions with DecisionTools*; 2. rev. ed.; Duxbury Thomson Learning: Pacific Grove, Calif, 2001; ISBN 978-0-495-01508-6.
66. Karvetski, C.W.; Lambert, J.H. Evaluating Deep Uncertainties in Strategic Priority-Setting with an Application to Facility Energy Investments. *Syst. Engin.* **2012**, *15*, 483–493, doi:10.1002/sys.21215.
67. Loose, D.C. Systems and Information Engineering Ph.D. Dissertation, University of Virginia, Department of Systems and Information Engineering, 2023.
68. Moghadasi, N.; Piran, M.; Valdez, R.S.; Moghaddasi, N.; Loose, D.C.; Polmateer, T.L.; Lambert, J.H. Systems Modeling of Trust in AI-Enabled Medical Diagnosis.; Montreal, QC, Canada, April 2024.
69. Montemayor, C.; Halpern, J.; Fairweather, A. In Principle Obstacles for Empathic AI: Why We Can't Replace Human Empathy in Healthcare. *AI & Soc* **2022**, *37*, 1353–1359, doi:10.1007/s00146-021-01230-z.

70. Abbas, S.W.; Hamid, M.; Alkanhel, R.; Abdallah, H.A. Official Statistics and Big Data Processing with Artificial Intelligence: Capacity Indicators for Public Sector Organizations. *Systems* **2023**, *11*, 424, doi:10.3390/systems11080424.
71. Dobson, G.P. Trauma of Major Surgery: A Global Problem That Is Not Going Away. *International Journal of Surgery* **2020**, *81*, 47–54, doi:10.1016/j.ijsu.2020.07.017.
72. Lu, Q.; Liu, K.; Zhang, W.; Li, T.; Shi, A.-H.; Ding, H.-F.; Yan, X.-P.; Zhang, X.-F.; Wu, R.-Q.; Lv, Y.; et al. End-to-End Vascular Anastomosis Using a Novel Magnetic Compression Device in Rabbits: A Preliminary Study. *Sci Rep* **2020**, *10*, 5981, doi:10.1038/s41598-020-62936-6.
73. Morris, P.J.; Knechtle, S.J. *Kidney Transplantation: Principles and Practice*; 7th edition.; Saunders, Elsevier: London, 2014; ISBN 978-1-4557-4096-3.
74. Al-Nahian, S.; Khan, O.S.; Alam, S.N.; Adhikary, A.B.; Aftabuddin, M. A Rapid and Reliable End-To-End Anastomosis Technique for Emergency Vascular Surgery--A Case Report. *Mymensingh Med J* **2016**, *25*, 158–160.
75. Boström, P.; Svensson, J.; Brorsson, C.; Rutegård, M. Early Postoperative Pain as a Marker of Anastomotic Leakage in Colorectal Cancer Surgery. *Int J Colorectal Dis* **2021**, *36*, 1955–1963, doi:10.1007/s00384-021-03984-w.
76. Hébert, J.; Eltonsy, S.; Gaudet, J.; Jose, C. Incidence and Risk Factors for Anastomotic Bleeding in Lower Gastrointestinal Surgery. *BMC Res Notes* **2019**, *12*, 378, doi:10.1186/s13104-019-4403-0.
77. Lebovitz, J.J.; Eller, J.L.; Sweeney, J.M.; Sasaki-Adams, D.; Darbar, A.; Palejwala, S.K.; Radon, A.-M.; Abdulrauf, S.I. Cerebral Revascularization for Giant Aneurysms of the Transitional Segment of the Internal Carotid Artery. In *Principles of Neurological Surgery*; Elsevier, 2012; pp. 291–308 ISBN 978-1-4377-0701-4.
78. Nakamura, T.; Takayama, Y.; Sato, T.; Watanabe, M. Risk Factors for Wound Infection After Laparoscopic Surgery for Colon Cancer. *Surgical Laparoscopy, Endoscopy & Percutaneous Techniques* **2020**, *30*, 45–48, doi:10.1097/SLE.0000000000000735.

79. *Critical Care Nephrology*; Ronco, C., Bellomo, R., Kellum, J.A., Ricci, Z., Eds.; Third edition.; Elsevier, Inc.: Philadelphia, PA, 2019;
80. Tuturov, A.O. The Role of Peripheral Nerve Surgery in a Tissue Reinnervation. *Chin Neurosurg Jl* **2019**, *5*, 5, doi:10.1186/s41016-019-0151-1.
81. Zhou, Y.; Wu, H. Comparison of End-to-Side versus Side-to-Side Anastomosis in Upper Limb Arteriovenous Fistula in Hemodialysis Patients: A Systematic Review and Meta-Analysis. *Front. Surg.* **2023**, *9*, 1079291, doi:10.3389/fsurg.2022.1079291.
82. Health Jade Team Anastomosis 2023.
83. Pagano, M. HEMI Fellow Sung Hoon Kang Receives Cohen Fund Grant for Work on a 3D-Printed Medical Device. *Hopkins Extreme Materials Institute* 2020.
84. Bao, Y.; Gong, W.; Yang, K. A Literature Review of Human–AI Synergy in Decision Making: From the Perspective of Affordance Actualization Theory. *Systems* **2023**, *11*, 442, doi:10.3390/systems11090442.
85. Chen, F.; Zhou, J.; Holzinger, A.; Fleischmann, K.R.; Stumpf, S. Artificial Intelligence Ethics and Trust: From Principles to Practice. *IEEE Intell. Syst.* **2023**, *38*, 5–8, doi:10.1109/MIS.2023.3324470.
86. Kim, J.; Scott, C.D. Robust Kernel Density Estimation. **2011**, doi:10.48550/ARXIV.1107.3133.
87. Węglarczyk, S. Kernel Density Estimation and Its Application. *ITM Web Conf.* **2018**, *23*, 00037, doi:10.1051/itmconf/20182300037.
88. Chen, Y.-C. A Tutorial on Kernel Density Estimation and Recent Advances. *Biostatistics & Epidemiology* **2017**, *1*, 161–187, doi:10.1080/24709360.2017.1396742.
89. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; First edition.; CRC Press: Boca Raton, FL, 2018; ISBN 978-1-351-45617-3.
90. Soh, Y.; Hae, Y.; Mehmood, A.; Hadi Ashraf, R.; Kim, I. Performance Evaluation of Various Functions for Kernel Density Estimation. *OJAppS* **2013**, *03*, 58–64, doi:10.4236/ojapps.2013.31B012.
91. Wilcox, R.R. *Introduction to Robust Estimation and Hypothesis Testing*; 5th ed.; Elsevier, Inc: Philadelphia, 2021; ISBN 978-0-12-820098-8.

92. Hüllermeier, E.; Waegeman, W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Mach Learn* **2021**, *110*, 457–506, doi:10.1007/s10994-021-05946-3.
93. Menapace, L.; Colson, G.; Raffaelli, R. Risk Aversion, Subjective Beliefs, and Farmer Risk Management Strategies. *American J Agri Economics* **2013**, *95*, 384–389, doi:10.1093/ajae/aas107.
94. Bhatt, U.S.; Newman, D.E.; Carreras, B.A.; Dobson, I. Understanding the Effect of Risk Aversion on Risk. In Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences; IEEE: Big Island, HI, USA, 2005; pp. 64b–64b.
95. Khan, M.A.; Arshad, H.; Khan, W.Z.; Alhaisoni, M.; Tariq, U.; Hussein, H.S.; Alshazly, H.; Osman, L.; Elashry, A. HGRBOL2: Human Gait Recognition for Biometric Application Using Bayesian Optimization and Extreme Learning Machine. *Future Generation Computer Systems* **2023**, *143*, 337–348, doi:10.1016/j.future.2023.02.005.
96. Chakraborty, T.; Seifert, C.; Wirth, C. Explainable Bayesian Optimization. **2024**, doi:10.48550/ARXIV.2401.13334.
97. Daniel James, L. *Practical Bayesian Optimization*; University of Alberta, 2008;
98. Jasper, S.; et al. Scalable Bayesian Optimization Using Deep Neural Networks. *International conference on machine learning* **2015**, *37*, 2171–2180.
99. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; The MIT Press, 2005; ISBN 978-0-262-25683-4.
100. Krasser, M. *Bayesian Optimization*; 2018;
101. Yang, J. How Exactly Does Bayesian Optimization Work? 2019.
102. Borjali, A.; Chen, A.F.; Muratoglu, O.K.; Morid, M.A.; Varadarajan, K.M. Deep Learning in Orthopedics: How Do We Build Trust in the Machine? *Healthcare Transformation* **2020**, heat.2019.0006, doi:10.1089/heat.2019.0006.
103. Roshan, K.; Zafar, A. Utilizing XAI Technique to Improve Autoencoder Based Model for Computer Network Anomaly Detection with Shapley Additive Explanation(SHAP). **2021**, doi:10.48550/ARXIV.2112.08442.
104. Shams Khoozani, Z.; Sabri, A.Q.M.; Seng, W.C.; Seera, M.; Eg, K.Y. Navigating the Landscape of Concept-Supported XAI: Challenges,

Innovations, and Future Directions. *Multimed Tools Appl* **2024**, doi:10.1007/s11042-023-17666-y.

105. Hung, L.-P.; Xu, C.-H.; Wang, C.-S.; Chen, C.-L. Applying the Shapley Value Method to Predict Mortality in Liver Cancer Based on Explainable AI. In *Smart Grid and Internet of Things*; Deng, D.-J., Chao, H.-C., Chen, J.-C., Eds.; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering; Springer Nature Switzerland: Cham, 2023; Vol. 497, pp. 133–143 ISBN 978-3-031-31274-8.
106. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. Explainable AI for Trees: From Local Explanations to Global Understanding. **2019**, doi:10.48550/ARXIV.1905.04610.
107. Patro, S.P.; Padhy, N. A Secure Remote Health Monitoring for Heart Disease Prediction Using Machine Learning and Deep Learning Techniques in Explainable Artificial Intelligence Framework. In Proceedings of the The 10th International Electronic Conference on Sensors and Applications; MDPI, November 15 2023; p. 78.
108. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. **2017**, doi:10.48550/ARXIV.1705.07874.
109. Biecek, P.; Burzykowski, T. *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*; Chapman & Hall/CRC data science series; First edition.; CRC Press: Boca Raton London New York, 2021; ISBN 978-0-367-13559-1.
110. Ribeiro, J.; Silva, R.; Cardoso, L.; Alves, R. Does Dataset Complexity Matters for Model Explainers? In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data); IEEE: Orlando, FL, USA, December 15 2021; pp. 5257–5265.
111. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*; Second edition.; Christoph Molnar: Munich, Germany, 2022; ISBN 9798411463330.

112. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. **2013**, doi:10.48550/ARXIV.1309.6392.
113. Liu, H. *Generalized Additive Model*; University of Minnesota Duluth, Duluth, MN, 2008;
114. Guidotti, R. Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking. *Data Min Knowl Disc* **2022**, doi:10.1007/s10618-022-00831-6.
115. Vij, A.; Nanjundan, P. Comparing Strategies for Post-Hoc Explanations in Machine Learning Models. In *Mobile Computing and Sustainable Informatics*; Shakya, S., Bestak, R., Palanisamy, R., Kamel, K.A., Eds.; Lecture Notes on Data Engineering and Communications Technologies; Springer Singapore: Singapore, 2022; Vol. 68, pp. 585–592 ISBN 9789811618659.
116. Kawakura, S.; Hirafuji, M.; Ninomiya, S.; Shibasaki, R. Adaptations of Explainable Artificial Intelligence (XAI) to Agricultural Data Models with ELI5, PDPbox, and Skater Using Diverse Agricultural Worker Data. *EJAI* **2022**, *1*, 27–34, doi:10.24018/ejai.2022.1.3.14.
117. Baughman, R.P.; Culver, D.A.; Judson, M.A. A Concise Review of Pulmonary Sarcoidosis. *Am J Respir Crit Care Med* **2011**, *183*, 573–581, doi:10.1164/rccm.201006-0865CI.
118. Deubelbeiss, U.; Gemperli, A.; Schindler, C.; Baty, F.; Brutsche, M.H. Prevalence of Sarcoidosis in Switzerland Is Associated with Environmental Factors. *European Respiratory Journal* **2010**, *35*, 1088–1097, doi:10.1183/09031936.00197808.
119. Lehtonen, J.; Uusitalo, V.; Pöyhönen, P.; Mäyränpää, M.I.; Kupari, M. Cardiac Sarcoidosis: Phenotypes, Diagnosis, Treatment, and Prognosis. *European Heart Journal* **2023**, *44*, 1495–1510, doi:10.1093/eurheartj/ehad067.
120. Williams, O.; Fatima, S. Granuloma. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2024.
121. Tana, C.; Mantini, C.; Donatiello, I.; Mucci, L.; Tana, M.; Ricci, F.; Cipollone, F.; Giamberardino, M.A. Clinical Features and Diagnosis of Cardiac Sarcoidosis. *JCM* **2021**, *10*, 1941, doi:10.3390/jcm10091941.

122. Christopher Frey, H.; Patil, S.R. Identification and Review of Sensitivity Analysis Methods. *Risk Analysis* **2002**, *22*, 553–578, doi:10.1111/0272-4332.00039.
123. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-Agnostic Interpretability of Machine Learning. **2016**, doi:10.48550/ARXIV.1606.05386.
124. Didugu, C. Anchors: A Simple Introduction. *Medium* 2023.
125. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. *AAAI* **2018**, *32*, doi:10.1609/aaai.v32i1.11491.
126. Blackman, R. *Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI*; Harvard Business Review Press: Boston, Massachusetts, 2022; ISBN 978-1-64782-281-1.
127. Macri, R.; Roberts, S.L. The Use of Artificial Intelligence in Clinical Care: A Values-Based Guide for Shared Decision Making. *Current Oncology* **2023**, *30*, 2178–2186, doi:10.3390/curroncol30020168.
128. Al Kuwaiti, A.; Nazer, K.; Al-Reedy, A.; Al-Shehri, S.; Al-Muhanna, A.; Subbarayalu, A.V.; Al Muhanna, D.; Al-Muhanna, F.A. A Review of the Role of Artificial Intelligence in Healthcare. *JPM* **2023**, *13*, 951, doi:10.3390/jpm13060951.
129. Naik, N.; Hameed, B.M.Z.; Shetty, D.K.; Swain, D.; Shah, M.; Paul, R.; Aggarwal, K.; Ibrahim, S.; Patil, V.; Smriti, K.; et al. Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Front. Surg.* **2022**, *9*, 862322, doi:10.3389/fsurg.2022.862322.
130. Crockett, K.; Colyer, E.; Latham, A. The Ethical Landscape of Data and Artificial Intelligence: Citizen Perspectives. In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI); IEEE: Orlando, FL, USA, December 5 2021; pp. 1–9.
131. Ashiku, L.; Threlkeld, R.; Canfield, C.; Dagli, C. Identifying AI Opportunities in Donor Kidney Acceptance: Incremental Hierarchical Systems Engineering Approach. In Proceedings of the 2022 IEEE International Systems Conference (SysCon); IEEE: Montreal, QC, Canada, April 25 2022; pp. 1–8.

132. Aggarwal, N.; Matheny, M.E.; Shachar, C.; Wang, S.X.Y.; Thadaney-Israni, S. Artificial Intelligence in Healthcare. In *The Oxford Handbook of AI Governance*; Bullock, J.B., Chen, Y.-C., Himmelreich, J., Hudson, V.M., Korinek, A., Young, M.M., Zhang, B., Eds.; Oxford University Press, 2022 ISBN 978-0-19-757932-9.
133. Mishra, P. *Practical Explainable aI Using Python: Artificial Intelligence Model Explanations Using Python-Based Libraries, Extensions, and Frameworks*; Apress: New York, 2022; ISBN 978-1-4842-7157-5.
134. Agarwal, R.; Samant, P.; Bansal, A.; Agarwal, R. Artificial Intelligence for Iris-Based Diagnosis in Healthcare. In *Handbook of Metrology and Applications*; Aswal, D.K., Yadav, S., Takatsuji, T., Rachakonda, P., Kumar, H., Eds.; Springer Nature Singapore: Singapore, 2023; pp. 1–31 ISBN 978-981-19155-0-5.
135. Kasthuri, N.; Meeradevi, T. AI-Driven Healthcare Analysis. In *Smart Systems for Industrial Applications*; Venkatesh, C., Rengarajan, N., Ponmurugan, P., Balamurugan, S., Eds.; Wiley, 2022; pp. 269–285 ISBN 978-1-119-76200-3.
136. Khedkar, S.; Subramanian, V.; Shinde, G.; Gandhi, P. Explainable AI in Healthcare. *SSRN Journal* **2019**, doi:10.2139/ssrn.3367686.
137. Hashimoto, D.A.; Rosman, G.; Rus, D.; Meireles, O.R. Artificial Intelligence in Surgery: Promises and Perils. *Annals of Surgery* **2018**, *268*, 70–76, doi:10.1097/SLA.0000000000002693.
138. Newman, J. *A Taxonomy of Trustworthiness for Artificial Intelligence*; Center for Long-Term Cybersecurity UC Berkeley, 2023;
139. Couzin-Frankel, J. Medicine Contends with How to Use Artificial Intelligence. *Science* **2019**, *364*, 1119–1120, doi:10.1126/science.364.6446.1119.
140. Dobrowolska, B.; Jędrzejkiewicz, B.; Pilewska-Kozak, A.; Zarzycka, D.; Ślusarska, B.; Deluga, A.; Kościołek, A.; Palese, A. Age Discrimination in Healthcare Institutions Perceived by Seniors and Students. *Nurs Ethics* **2019**, *26*, 443–459, doi:10.1177/0969733017718392.
141. Chu, L.C.; Anandkumar, A.; Shin, H.C.; Fishman, E.K. The Potential Dangers of Artificial Intelligence for Radiology and Radiologists. *Journal of*

the American College of Radiology **2020**, *17*, 1309–1311, doi:10.1016/j.jacr.2020.04.010.

142. Feehan, M.; Owen, L.A.; McKinnon, I.M.; DeAngelis, M.M. Artificial Intelligence, Heuristic Biases, and the Optimization of Health Outcomes: Cautionary Optimism. *JCM* **2021**, *10*, 5284, doi:10.3390/jcm10225284.
143. Patel, K.; Mistry, C.; Mehta, D.; Thakker, U.; Tanwar, S.; Gupta, R.; Kumar, N. A Survey on Artificial Intelligence Techniques for Chronic Diseases: Open Issues and Challenges. *Artif Intell Rev* **2022**, *55*, 3747–3800, doi:10.1007/s10462-021-10084-2.
144. Comaniciu, D. Artificial Intelligence for Healthcare. In Proceedings of the Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; ACM: Virtual Event CA USA, August 23 2020; pp. 3603–3603.
145. Kaur, S.; Singla, J.; Nkenyereye, L.; Jha, S.; Prashar, D.; Joshi, G.P.; El-Sappagh, S.; Islam, Md.S.; Islam, S.M.R. Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives. *IEEE Access* **2020**, *8*, 228049–228069, doi:10.1109/ACCESS.2020.3042273.
146. Xie, Y.; Gunasekeran, D.V.; Balaskas, K.; Keane, P.A.; Sim, D.A.; Bachmann, L.M.; Macrae, C.; Ting, D.S.W. Health Economic and Safety Considerations for Artificial Intelligence Applications in Diabetic Retinopathy Screening. *Trans. Vis. Sci. Tech.* **2020**, *9*, 22, doi:10.1167/tvst.9.2.22.
147. Kohli, P.S.; Arora, S. Application of Machine Learning in Disease Prediction. In Proceedings of the 2018 4th International Conference on Computing Communication and Automation (ICCCA); IEEE: Greater Noida, India, December 2018; pp. 1–4.
148. Kalaiselvan, V.; Sharma, A.; Gupta, S.K. "Feasibility Test and Application of AI in Healthcare"—with Special Emphasis in Clinical, Pharmacovigilance, and Regulatory Practices. *Health Technol.* **2021**, *11*, 1–15, doi:10.1007/s12553-020-00495-6.
149. Xiao, L.; Chen, Y.; Xing, Y.; Mou, L.; Zhang, L.; Li, W.; Xie, S.; Sun, M. The Analysis and AI Prospect Based on the Clinical Screening Results of Chronic Diseases. In *Proceedings of the 11th International Conference on Computer*

Engineering and Networks; Liu, Q., Liu, X., Chen, B., Zhang, Y., Peng, J., Eds.; Lecture Notes in Electrical Engineering; Springer Nature Singapore: Singapore, 2022; Vol. 808, pp. 553–562 ISBN 9789811665530.

150. Lee, E.E.; Torous, J.; De Choudhury, M.; Depp, C.A.; Graham, S.A.; Kim, H.-C.; Paulus, M.P.; Krystal, J.H.; Jeste, D.V. Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **2021**, *6*, 856–864, doi:10.1016/j.bpsc.2021.02.001.
151. Quinn, T.P.; Senadeera, M.; Jacobs, S.; Coghlan, S.; Le, V. Trust and Medical AI: The Challenges We Face and the Expertise Needed to Overcome Them. *Journal of the American Medical Informatics Association* **2021**, *28*, 890–894, doi:10.1093/jamia/ocaa268.
152. Moghadasi, N.; Valdez, R.S.; Piran, M.; Moghaddasi, N.; Loose, D.C.; Polmateer, T.L.; Lambert, J.H. Artificial Intelligence in Healthcare: A Systems Approach to Risk Analysis. *MDPI Systems Journal*.
153. Stop Talking about Tomorrow's AI Doomsday When AI Poses Risks Today. *Nature* **2023**, *618*, 885–886, doi:10.1038/d41586-023-02094-7.
154. Chimatapu, R.; Hagras, H.; Starkey, A.; Owusu, G. Explainable AI and Fuzzy Logic Systems. In *Theory and Practice of Natural Computing*; Fagan, D., Martín-Vide, C., O'Neill, M., Vega-Rodríguez, M.A., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2018; Vol. 11324, pp. 3–20 ISBN 978-3-030-04069-7.
155. Arya, V.; Bellamy, R.K.E.; Chen, P.-Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilović, A.; et al. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. **2019**, doi:10.48550/ARXIV.1909.03012.
156. Bhattacharya, M.; Penica, M.; O'Connell, E.; Southern, M.; Hayes, M. Human-in-Loop: A Review of Smart Manufacturing Deployments. *Systems* **2023**, *11*, 35, doi:10.3390/systems11010035.
157. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Algorithmic Bias In Health Care: A Path Forward 2019.

158. Richardson, J.P.; Smith, C.; Curtis, S.; Watson, S.; Zhu, X.; Barry, B.; Sharp, R.R. Patient Apprehensions about the Use of Artificial Intelligence in Healthcare. *npj Digit. Med.* **2021**, *4*, 140, doi:10.1038/s41746-021-00509-1.
159. Wang, B.; Asan, O.; Mansouri, M. Patients' Perceptions of Integrating AI into Healthcare: Systems Thinking Approach. In Proceedings of the 2022 IEEE International Symposium on Systems Engineering (ISSE); IEEE: Vienna, Austria, October 24 2022; pp. 1–6.
160. Moghadasi, N.; Piran, M.; Baek, S.; Valdez, R.S.; Porter, M.D.; Johnson, D.; Lambert, J.H. Systems Analysis of Bias and Risk in AI Enabled Medical Diagnosis.; IEEE, September 15 2023.
161. Anthony (Tony)Cox, L. What's Wrong with Risk Matrices? *Risk Analysis* **2008**, *28*, 497–512, doi:10.1111/j.1539-6924.2008.01030.x.
162. Cox, L.A. (Tony); Babayev, D.; Huber, W. Some Limitations of Qualitative Risk Rating Systems. *Risk Analysis* **2005**, *25*, 651–662, doi:10.1111/j.1539-6924.2005.00615.x.
163. Krisper, M. Problems with Risk Matrices Using Ordinal Scales. **2021**, doi:10.48550/ARXIV.2103.05440.
164. Rozell, D.J. A Cautionary Note on Qualitative Risk Ranking of Homeland Security Threats. *The Journal of the NPS Center for Homeland Defense and Security* 2015.
165. Murdock, H.; de Bruijn, K.; Gersonius, B. Assessment of Critical Infrastructure Resilience to Flooding Using a Response Curve Approach. *Sustainability* **2018**, *10*, 3470, doi:10.3390/su10103470.
166. Koren, D.; Kilar, V.; Rus, K. Proposal for Holistic Assessment of Urban System Resilience to Natural Disasters. *IOP Conf. Ser.: Mater. Sci. Eng.* **2017**, *245*, 062011, doi:10.1088/1757-899X/245/6/062011.
167. Sarker, P.; Lester, H.D. Post-Disaster Recovery Associations of Power Systems Dependent Critical Infrastructures. *Infrastructures* **2019**, *4*, 30, doi:10.3390/infrastructures4020030.
168. Poulin, C.; Kane, M.B. Infrastructure Resilience Curves: Performance Measures and Summary Metrics. *Reliability Engineering & System Safety* **2021**, *216*, 107926, doi:10.1016/j.ress.2021.107926.

169. Mousavi, B.A.; Heavey, C.; Azzouz, R.; Ehm, H.; Millauer, C.; Knobloch, R. Use of Model-Based System Engineering Methodology and Tools for Disruption Analysis of Supply Chains: A Case in Semiconductor Manufacturing. *Journal of Industrial Information Integration* **2022**, *28*, 100335, doi:10.1016/j.jii.2022.100335.
170. Nowak, C.P.; Konietschke, F. Simultaneous Inference for Kendall's Tau. *Journal of Multivariate Analysis* **2021**, *185*, 104767, doi:10.1016/j.jmva.2021.104767.
171. Muñoz-Pichardo, J.M.; Lozano-Aguilera, E.D.; Pascual-Acosta, A.; Muñoz-Reyes, A.M. Multiple Ordinal Correlation Based on Kendall's Tau Measure: A Proposal. *Mathematics* **2021**, *9*, 1616, doi:10.3390/math9141616.
172. Flight, R.M.; Bhatt, P.S.; Moseley, H.N. *Information-Content-Informed Kendall-Tau Correlation: Utilizing Missing Values*; Bioinformatics, 2022;
173. World Population Review US Immigration by Country 2022 2022.
174. Urbański, M. Comparing Push and Pull Factors Affecting Migration. *Economies* **2022**, *10*, 21, doi:10.3390/economies10010021.
175. Brandirectory HEALTHCARE 2022 RANKING. *HEALTHCARE 2022 RANKING* 2022.
176. Leswing, K.; Cortes, G. Apple Grew More Slowly than Google, Amazon, Microsoft and Meta, and Has so Far Dodged Major Layoffs 2023.
177. Allen, A. How Pfizer Won the Pandemic, Reaping Outsize Profit and Influence 2022.
178. Croux, C.; Dehon, C. Influence Functions of the Spearman and Kendall Correlation Measures. *Stat Methods Appl* **2010**, *19*, 497–515, doi:10.1007/s10260-010-0142-z.
179. Moghadasi, N.; Lambert, J.H. On Evaluating System Resilience by the Degree of Order Disruption. *INCOSE International Symp* **2023**, *33*, 739–751, doi:10.1002/iis2.13049.
180. Hooker, C. Introduction to Philosophy of Complex Systems: A. In *Philosophy of Complex Systems*; Elsevier, 2011; pp. 3–90 ISBN 978-0-444-52076-0.
181. Peterson, T. Leveraging INCOSE Resources to Identify Systems Engineering Best Practices.; 2010.

182. Duenser, A.; Douglas, D.M. Whom to Trust, How and Why: Untangling Artificial Intelligence Ethics Principles, Trustworthiness, and Trust. *IEEE Intell. Syst.* **2023**, *38*, 19–26, doi:10.1109/MIS.2023.3322586.
183. Schmid, A.; Wiesche, M. The Importance of an Ethical Framework for Trust Calibration in AI. *IEEE Intell. Syst.* **2023**, *38*, 27–34, doi:10.1109/MIS.2023.3320443.
184. Ueda, D.; Kakinuma, T.; Fujita, S.; Kamagata, K.; Fushimi, Y.; Ito, R.; Matsui, Y.; Nozaki, T.; Nakaura, T.; Fujima, N.; et al. Fairness of Artificial Intelligence in Healthcare: Review and Recommendations. *Jpn J Radiol* **2023**, doi:10.1007/s11604-023-01474-3.
185. Shams, M.; Choudhari, J.; Reyes, K.; Prentzas, S.; Gapizov, A.; Shehryar, A.; Affaf, M.; Grezenko, H.; Gasim, R.W.; Mohsin, S.N.; et al. The Quantum-Medical Nexus: Understanding the Impact of Quantum Technologies on Healthcare. *Cureus* **2023**, doi:10.7759/cureus.48077.