UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Software Engineering Curriculum

Marasinghe Mudiyanselage Rasinthe Marasinghe

Systematic Evaluation of Trustworthy AI Augmentation in Modern Applications

Master's Thesis (30 ECTS)

Supervisor: Huber Flores
Co-Supervisor: Abdul-Rasheed Ottun

Tartu 2024

Systematic Evaluation of Trustworthy AI Augmentation in Modern Applications

Abstract:
Artificial intelligence (AI) has become pervasive in various sectors, including healthcare, finance, education, and transportation, transforming how tasks are performed and decisions are made. However, the rapid integration of AI raised significant concerns about privacy, bias, security, and the opacity of AI systems, often referred to as "black boxes." These challenges highlight the critical gap in ensuring AI systems are both efficient and trustworthy. This research addresses this gap by focusing on the practical implementation of continuous human oversight in AI development. The study specifically evaluates an adaptive dashboard developed for the SPATIAL platform to enhance AI transparency and accountability. Through experiments with the Medical Analysis Module (MAM) that employs Explainable AI (XAI) techniques to provide role-specific explanations for stakeholders analyzing electrocardiogram (ECG) data, the research assesses the interpretability of AI-generated explanations and the system's performance under varying user loads. The findings demonstrate that tailored explanations significantly improve user understanding and trust while the system maintained robust performance, ensuring scalability and reliability. These insights provide valuable guidance for developing practical tools to enhance the monitoring and oversight of AI inferences, aligning with regulatory requirements for trustworthy AI.

Keywords: Artificial Intelligence, Trustworthy AI, Explainable AI, Measuring Trustworthiness, AI Applications, Explanatory Platform, Human Oversight

CERCS: P170 - Computer Science, Numerical Analysis, System, Control

Usaldusväärse AI süstemaatiline hindamine Laiendus kaasaegsetes rakendustes **Lühikokkuvõte:**

Tehisintellekt (AI) on levinud mitmetes sektorites, sealhulgas tervishoius, rahanduses, hariduses ja transpordis, muutes ülesannete täitmist ja otsuste tegemist. Kuid tehisintellekti kiire integreerimine on tekitanud märkimisväärset muret eraelu puutumatuse, eelarvamuste, turvalisuse ja tehisintellekti süsteemide, mida sageli nimetatakse „mustadeks kastideks“, läbipaistmatuse pärast. Need probleemid rõhutavad kriitilist puudujääki, mis on seotud tehisintellekti süsteemide tõhususe ja usaldusväärsuse tagamisega. Käesolevas uuringus käsitletakse seda puudujääki, keskendudes pideva inimliku järelevalve praktilisele rakendamisele tehisintellekti arendamisel. Uuringus hinnatakse konkreetselt SPATIAL platvormi jaoks välja töötatud adaptiivset armatuurlauda, et suurendada tehisintellekti läbipaistvust ja aruandekohustust. Uuringus hinnatakse eksperimentide kaudu meditsiinilise analüüsimooduliga (MAM), mis kasutab seletava tehisintellekti (XAI) meetodeid, et anda elektrokardiogrammi (EKG) andmeid analüüsivatele sidusrühmadele rollipõhiseid selgitusi, tehisintellekti genereeritud selgituste tõlgendatavust ja süsteemi toimimist erineva kasutajakoormuse korral. Tulemused näitavad, et kohandatud selgitused parandavad märkimisväärselt kasutajate arusaamist ja usaldust, samas kui süsteem säilitas tugeva jõudluse, tagades skaleeritavuse ja usaldusväärsuse. Need teadmised annavad väärtuslikke juhiseid praktiliste vahendite väljatöötamiseks, et tõhustada tehisintellekti järelduste jälgimist ja järelevalvet, mis on kooskõlas usaldusväärse tehisintellekti regulatiivsete nõuetega.

Võtmesõnad: Tehisintellekt, usaldusväärne tehisintellekt, seletatav tehisintellekt, usaldusväärsuse mõõtmine, tehisintellekti rakendused, seletav platvorm, inimlik järelevalve

CERCS: P170 - Arvutiteadus, numbriline analüüs, süsteem, juhtimine

Contents

List of Figures

List of Tables

1 Introduction

Artificial Intelligence (AI) has become ubiquitous in the society. It has transcended from just a mere conception to an advancing technology with profound capabilities to transform capabilities. The trend of AI is surging across different sectors from healthcare[2], finance[3], education[4], and transportation[5] This trend seems not to be slowing down as societies consistently experiencing more profound performance from AI on complex human tasks. Thus demonstrating its potential to impact everyone in society. From small wearable devices like rings and smart cards to personal ones (smartphones and smart speakers), household appliances, machinery, and vehicles, AI is embedded in nearly every facet of our lives. However, this surge of AI has become worrisome due to the risks associated with AI, ranging from privacy invasions to discrimination and biased decision-making that can be significantly consequential to society at large. Another concern is the opacity of AI's inference generation, especially for the models behind the AI systems that undertake complex and human-like tasks. These models are considered black boxes because it is unclear how the input of the AI system leads to the inference derived from the AI system [6]. In addition, AI systems can be easily compromised, making them vulnerable to a wide range of attacks that can jeopardise the objective of the system[1]. These challenges pose safety and ethical risks that raise concerns about the trustworthiness of AI systems. Thus, highlighting the need for AI systems that are not only efficient in performing the assigned tasks but also safe, ethical, and reliable.

Globally, stakeholders are making efforts to ensure that AI is trustworthy. legislation, regulations, and initiatives are being established to govern the design, development, and deployment of AI to guarantee safety and ensure that its functionalities comply with acceptable societal norms and values. AI legislation and regulations provide the legal frameworks and outline the operational guidelines and requirements for ensuring the development, deployment, and application of AI systems are ethical and responsible. The European Union (EU) AI ACT [7] is the foremost AI regulation to be codified in the world to regulate AI. It aims to protect the fundamental rights of humans without stifling AI innovation and application in European states. The ACT draws inspiration from the ethical principles and requirements for trustworthy AI of the EU High-Level Expert Group on an AI. It mandates that AI commissioning must comply with the seven key requirements for trustworthy AI, including human control and oversight, ethics, robustness, privacy, transparency, fairness, and societal well-being. Besides this ACT, other similar regulations are around the world are: US Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (AI)[8], and the Artificial Intelligence and Data Act (AIDA) of Canada [9]. Countries like China and the United Kingdom are also finalizing their regulations on AI. All AI legislation strives to promote ethical behavior, transparency, human supervision, and accountability in AI applications amongst AI practitioners. Despite the comprehensive stipulations in the regulation, practitioners lack practical guidance on implementation towards fulfilling some of the requirements of trustworthy AI like continuous oversight for monitoring and supervision of AI inference.

This research work seeks to address the gap in the practical implementation of trustworthy AI requirements on continuous human oversight during AI development by systematically evaluating an adaptive dashboard of an AI platform designed for accountable AI development in the context of EU Security and Privacy Accountable Technology Innovations, Algorithms, and machine Learning (SPATIAL) initiative [10] and the explanations that it presents. To achieve this aim, this work reviewed existing works in the literature relating to various trustworthy components of AI in different contexts, applications, and domains to understand existing efforts in complying with trustworthy AI requirements. Additionally, a Medical Analysis Module (MAM) that is a microservice in SPATIAL platform was used to appraise the explainability characteristic of trustworthiness. MAM uses Electrocardiograms (ECGs) of patients to generate descriptions and recommendations for stakeholders. The contributions from the results and findings hope to aid compliance by effectively demonstrating how the requirement can be translated into practical tools for end-users and regulatory bodies, particularly in the area of enhancing monitoring and oversight of AI inferences.

The following remaining part of this work is outlined as follows: Section 2 provides the background of the study, while the introduction of the SPATIAL architecture is presented in Section 3. The experiments carried out are discussed in Section 4, followed by the presentation of results in Section 5. In Section 6, we engage in the discussion, and Section 7 presents the conclusion.

2 Literature Review

Applications of AI are numerous and versatile. However, the black box-like nature of AI, the vulnerability of AI systems to indiscernible attacks and data-poisoning, perceived biases against underrepresented groups, and user privacy concerns, among many other things, create a perception of unreliability among users [1]. This brings about the need for research into AI and AI based applications, specifically the trustworthiness of such systems. As such, the research into the trustworthiness of AI-based applications is vast and diverse.

A few characteristics have been established as indicators of AI trustworthiness [1, 11] such as fairness, accuracy, transparency, robustness, privacy, accountability, explainability and interpretability [21]. Explainability and Interpretability, particularly in the context of AI trustworthiness, have two different definitions, but they are both tied to context. Interpretability focuses more on the underlying functioning that leads to decision-making - it explains how the decision was made, while explainability gives insights into the logic of the outcome for end users to understand - why the decision was made [12]. However, the explainability of AI can be challenging to compare across studies, owing to differences in how the model is built, the context in which it is used, and the number of parameters that contributed towards the decision [12].

Nonetheless, a mere search, "Trustworthy AI," on the SCOPUS database yielded 13,827 publications, while searching "Explainable AI" yielded 38,399 search results. This demonstrates the vastness of the amount of literature that exists on the topic. Although not all of this literature will prove relevant to this study, a clear search strategy must be established to sift out the applicable studies. Also, it is important to note that the two main topics in focus, the trustworthiness of modern AI-based applications and Explainable AI, are intertwined, as Explainable Artificial Intelligence (XAI) is a characteristic of Trustworthy AI. This stipulates the need for a systematic review of pre-existing work. Objectives of the literature review, literature selection process, criteria for inclusion and exclusion, and keyword selection will be detailed in subsequent sub-chapters. This study aims to build a system architecture that hosts services that evaluate the trustworthiness of AI-based applications with a distinctive focus on explainability and interpretability - where the caliber of the XAI model will be evaluated against stakeholder needs and its ability to detect anomalies in inputs. To explore pre-existing knowledge and apprehensions in the field, the Systematic Literature Review (SLR) method was applied, and key terms relevant to this research were used to filter relevant studies that explore the concepts of Trustworthy AI, explainability, and interpretability in AI systems.

2.1 Literature Review Methodology

Focusing on research questions, literature on the selected database SCOPUS was surveyed. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement for systematic reviews was followed when conducting the review [13] to ensure a systematic flow of information. By deep diving into the concept of XAI in particular, the degree to which XAI can detect changes in input in projects done in similar contexts and user expectations for XAI will be identified, setting a precedent to what's expected from the XAI component that's developed as part of this thesis.

2.1.1 Eligibility Criteria

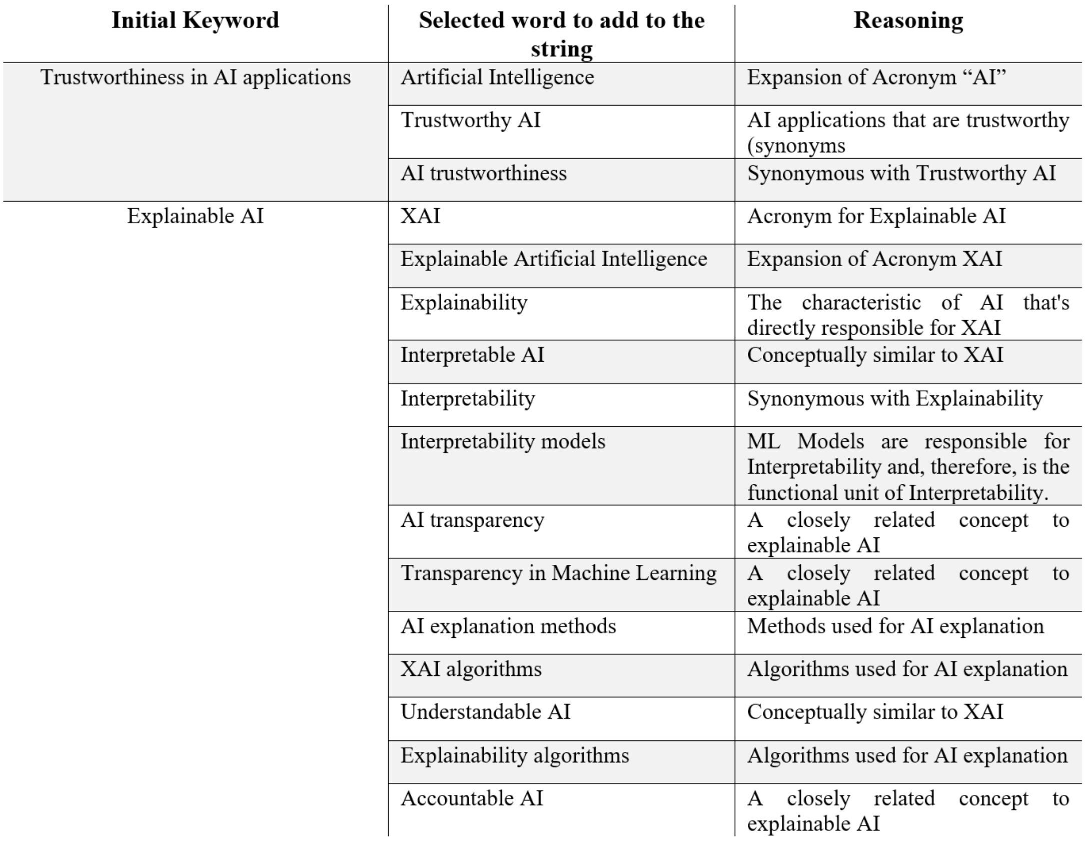
For the purpose of reviewing and scoping, research work was chosen to be included if the work contributed towards the Trustworthiness of AI-based applications and Explainable AI. For the purpose of this work, the definition of AI as presented by National Institute of Standards and Technology (NIST) is taken as the primary definition. Herein, the term AI is defined as "An Interdisciplinary field, usually regarded as a branch of computer science, dealing with models and systems for the performance of functions generally associated with human intelligence, such as reasoning and learning."[14]

2.1.2 Information Sources

The SCOPUS database was chosen as the primary database for SLR because it is considered an authentic database of scholarly publications curated by a board of subject matter experts[15]. It provides advanced analysis tools and indicators that facilitate customized searches with exclusion and/or inclusion criteria. Search strings that included appropriate keywords were used to call relevant research publications. To gather the keywords and for the purpose of obtaining reliable definitions that match closest to the context of this study, "The Language of Trustworthy AI: An In-Depth Glossary of Terms" by NIST [14] was used. NIST was considered a reliable source as its information is subjected to rigorous internal peer review to ensure the integrity of the scientific information. [28]

2.1.3 Search Strategy

The trustworthiness of AI can be characterized based on its explainability and interpretability, as it is crucial that the reasoning behind the AI model's decision-making is understood to determine whether or not its output can be trusted [12]. Therefore, it was determined that “Explainable AI” and “Trustworthiness in AI Applications” can be used as primary keywords. A trustworthy AI glossary published by NIST [14] was used to gather terms and abbreviations related to AI trustworthiness. From there, some terminology and abbreviations related to keywords were chosen and included in the search string. Terms were chosen based on the synonymity of meaning or the conceptual relevancy they had to the keyword in question. An overview of the terms chosen for each string is outlined in Figure 1. Search strings were designed with Boolean operators “AND” and “OR” to combine the terms and abbreviations gathered. Then, the strings were employed on SCOPUS to retain work with the keywords in the Title, Abstract, and Keyword sections of the research. The final version of the two strings used to filter through SCOPUS database is shown in Table 1.



Initial Keyword	Selected word to add to the string	Reasoning
Trustworthiness in AI applications	Artificial Intelligence	Expansion of Acronym “AI”
	Trustworthy AI	AI applications that are trustworthy (synonyms
	AI trustworthiness	Synonymous with Trustworthy AI
Explainable AI	XAI	Acronym for Explainable AI
	Explainable Artificial Intelligence	Expansion of Acronym XAI
	Explainability	The characteristic of AI that's directly responsible for XAI
	Interpretable AI	Conceptually similar to XAI
	Interpretability	Synonymous with Explainability
	Interpretability models	ML Models are responsible for Interpretability and, therefore, is the functional unit of Interpretability.
	AI transparency	A closely related concept to explainable AI
	Transparency in Machine Learning	A closely related concept to explainable AI
	AI explanation methods	Methods used for AI explanation
	XAI algorithms	Algorithms used for AI explanation
	Understandable AI	Conceptually similar to XAI
	Explainability algorithms	Algorithms used for AI explanation
	Accountable AI	A closely related concept to explainable AI

Figure 1. Terms and acronyms chosen to be included in respective search strings.

To ensure that the literature chosen for the systematic literature review has high relevance to the research topic, Inclusion and Exclusion criteria were established to filter the chosen literature further. The first inclusion criteria were that the title and the abstract had the established keywords, and more criteria were put in place to ensure the information’s relevance, specificity, and correctness.

Table 1. Search strings used to filter through SCOPUS database.

Search String for “Trustworthiness in AI Applications.”	**Search String for “Explainable AI.”**
TITLE-ABS-KEY (trustworthiness AND in AND ai OR artificial AND intelligence AND applications)	TITLE-ABS-KEY ("Explainable AI" OR "XAI" OR "Interpretable AI" OR "AI transparency" OR "Transparency in Machine Learning" OR "AI explanation methods" OR "Understandable AI" OR "Accountable AI" OR "Explainable Artificial Intelligence" OR "XAI algorithms" OR "explainability" OR "Explainability algorithms" OR "Interpretability models" OR "interpretability" OR "AI decision-making explanation)

2.2 Establishing the literature selection process

2.2.1 Literature Selection Strategy

Inclusion Criteria :

The SCOPUS database was initially searched to gather publications that had the chosen keywords in their title and abstracts. Literature was further retained in the search if they were published within the past decade (2014-2024), and further filtered to include texts that are in English, sourced from a journal or conference, and the document type being conference or article. Further, the literature was sorted based on whether the publication stage is the final and whether the work has been published with open access. These filters were applied sequentially to gather applicable literature.

Exclusion Criteria :

The exclusion criteria are more or less the inversion of the inclusion criteria. The publications were excluded if the title and the abstract did not include designated keywords. Next, items were excluded if the work was published before 2014, was in a language other than English, and was not sourced from a journal or conference. If the publication is in an intermediary stage/the final publication has not been made, the work was withdrawn from the selection. Next, if the work was not available on open-access, and finally, if the document/content type was not “Article” or “Conference”, the studies were filtered. It is important to note here, that if any one of the criteria mentioned above was met, the work was taken off from the aggregation of selected literature. It is not needed to meet multiple of these criteria for each study to be excluded.

2.3 Work Chosen for the Systematic Literature Review

2.3.1 Literature about “Explainable AI”

The initial filtering of literature with the keyword search string yielded a total of 80 research publications. Exclusion filters were applied, and Figure 2 provides an overview of the selection, and outlines how many publications remained after each filtration. In the end, 32 studies remained.

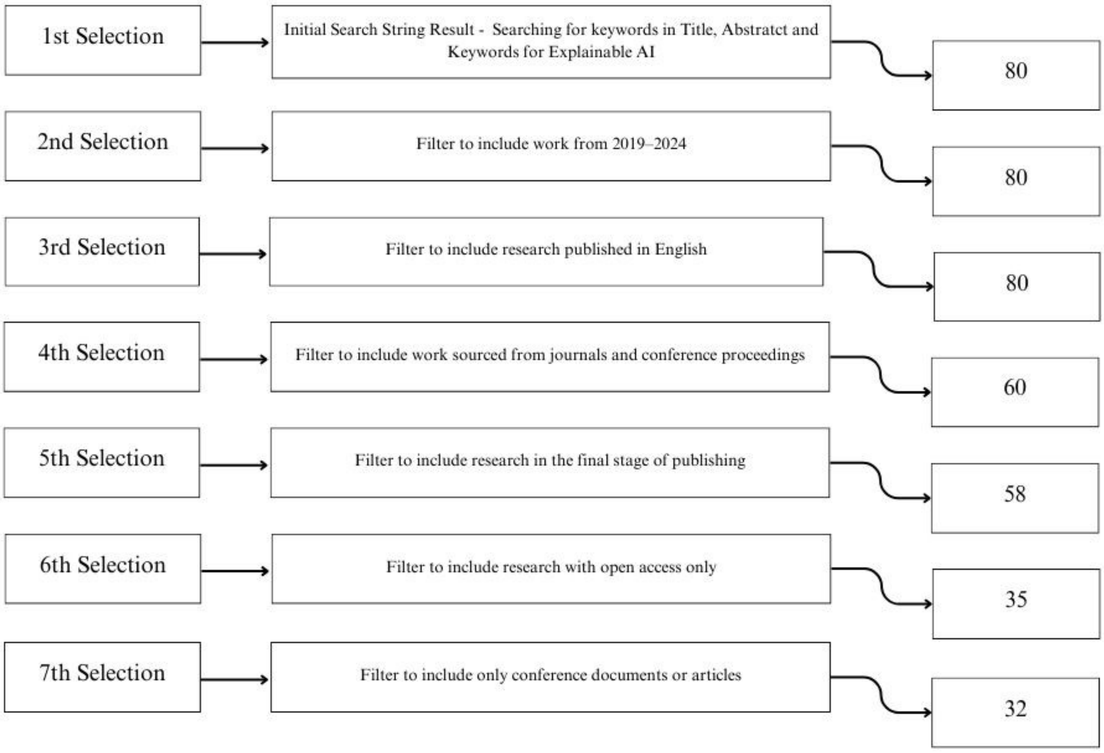

1st Selection
Initial Search String Result - Searching for keywords in Title, Abstratct and Keywords for Explainable AI
80
2nd Selection
Filter to include work from 2019–2024
80
3rd Selection
Filter to include research published in English
80
4th Selection
Filter to include work sourced from journals and conference proceedings
60
5th Selection
Filter to include research in the final stage of publishing
58
6th Selection
Filter to include research with open access only
35
7th Selection
Filter to include only conference documents or articles
32

Figure 2. Filtering research obtained for "Explainable AI" keyword using inclusion criteria.

2.3.2 Literature about “ Trustworthiness in AI Applications”

The initial filtering of literature with the keyword search string yielded only a total of 395 research publications. After the final filter adjustment only 110 publications remained.

Figure 3 outlines this process stepwise, highlighting the number of publications retained after each step.

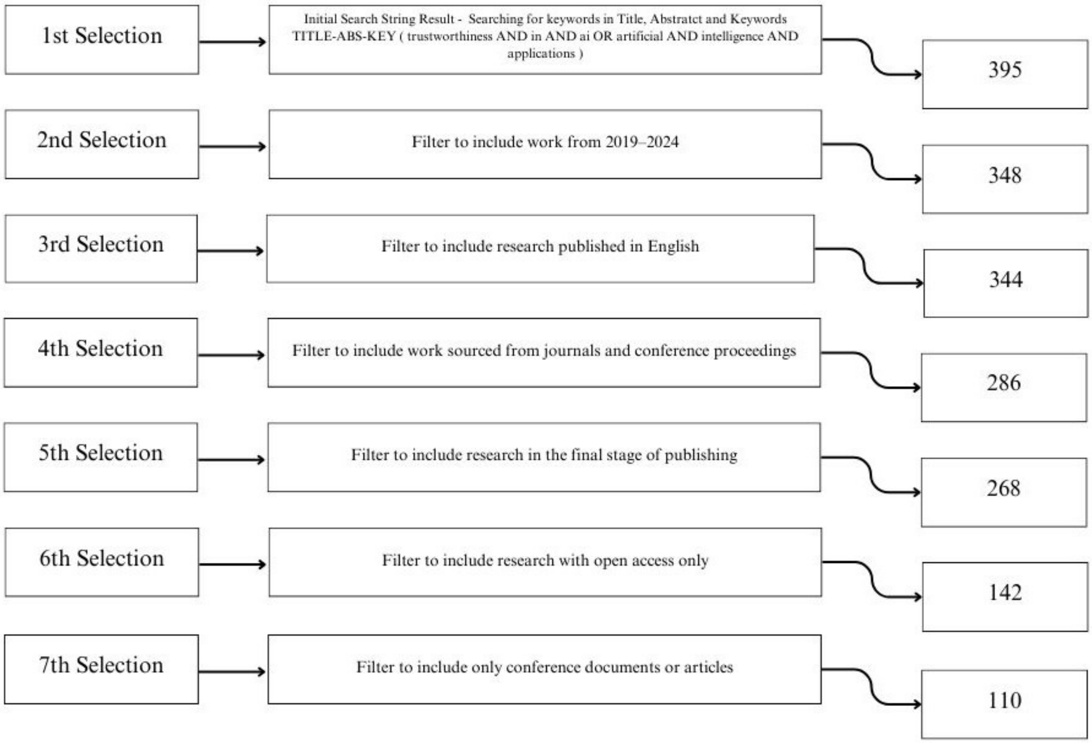

1st Selection
Initial Search String Result - Searching for keywords in Title, Abstratct and Keywords TITLE-ABS-KEY (trustworthiness AND in AND ai OR artificial AND intelligence AND applications)
395
2nd Selection
Filter to include work from 2019–2024
348
3rd Selection
Filter to include research published in English
344
4th Selection
Filter to include work sourced from journals and conference proceedings
286
5th Selection
Filter to include research in the final stage of publishing
268
6th Selection
Filter to include research with open access only
142
7th Selection
Filter to include only conference documents or articles
110

Figure 3. Filtering research based on inclusion criteria for "Trustworthiness in AI applications" keyword.

2.4 Artificial Intelligence and AI-based Applications

2.4.1 Uses of AI systems in the present day and Age

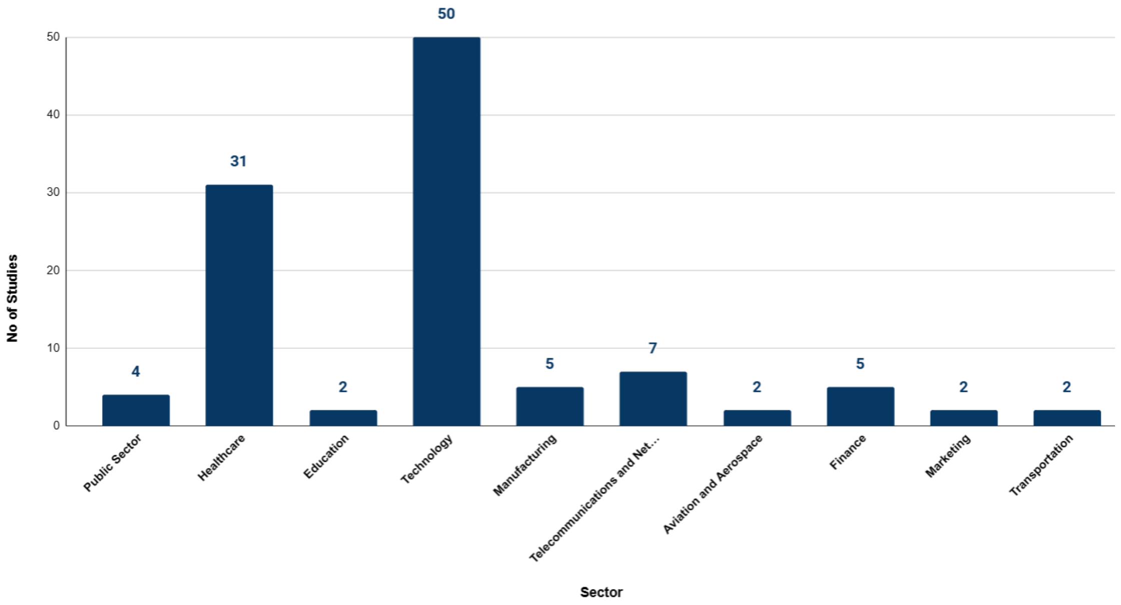


Figure 4. Number of studies based on the sector.

AI-based applications have spread in almost every industry, and many organizations, businesses, individuals, and governments in the world utilize them to advance in their respective use cases. An overall analysis of the chosen literature shows that AI-based applications are being used across different sectors, and Figure 4 provides an overview of the number of studies based on each sector. For instance, police can utilize it to enhance their investigation techniques, surveillance capabilities, crime prevention strategies, and order maintenance efforts [16]. In healthcare, the increased processing power has amplified the significance and applications of artificial intelligence, where AI is providing significant benefits to individuals and the healthcare system as a whole, including addressing patient inquiries, aiding in procedures, and driving advancements in pharmaceuticals [17]. In fact, Statista predicts that the AI healthcare market, with a value of 11 billion United States Dollars (USD) in 2021, will grow significantly to reach an impressive 187 billion USD by 2030 [17]. Financial institutions and the Fintech industry have adopted AI technology and implemented machine learning algorithms in trading, portfolio management, and investment advisory [18]. AI has gained footing in education, marketing, cybersecurity firms[19], in products and manufacturing [20]. It has also been integrated into day-to-day applications such as dating apps, chatbots[21], and smart home product management.

2.4.2 Construction of AI Models in a Nutshell and Decision-making Process of AI Models

AI models utilize Machine Learning (ML) techniques to provide outcomes and results, and Linear Regression (LR), logistic regression, Decision Tree (DT), Fuzzy rule-based systems (FRBS), Bayesian Networks (BNs), are some of the so-called machine learning models, used to generate these results. Ali et al., 2023 explain that decisions made by machine learning models, such as LR, logistic regression, DT, FRBS, or BNs, are often less accurate and unsuitable for large and complex data set analysis because of the non linearity of real-world data, thus rendering such models impracticable to be used in real-world applications - paving the way for the need to have more complex models such as Deep Neural Networks (DNNs). These models are complex, including many convolutional filters or kernels and neural units to compensate for the non-linearities of practical data. The immense potential of machine learning (ML) applications stems from the fact that their behavior is not defined by explicitly written rules but rather inferred from data. ML approaches are specifically designed to detect patterns in training data and integrate their statistical insights into a model. Deep neural networks, namely large machine learning models derived from extensive data, can do tasks that have hitherto eluded concise descriptions through human-imposed rules. Over time, AI-capable applications build and improve their models via a systematic workflow. AI's probabilistic nature is improved with each contribution. This procedure, shown in Figure 5, consists of several consecutive steps. First, data is gathered and processed using different methods, such as removing duplicates and missing information and applying data augmentation [22] techniques to improve the data quality. Next, the data is labeled, which is often done by human annotators, in order to prepare it for the AI system. After that, the training phase starts, during which a suitable algorithm (like a Random Forest or Support Vector Machine) is selected, and the training procedure (such as data parallelization or model partitioning [23]) is decided. Next, the model's performance is evaluated, usually using methods like cross-validation [24]. Ultimately, the model is implemented, and its effectiveness is assessed within the application.

Models in traditional architectures typically require retraining and redeployment upon the availability of new data. In contrast, model updates in more modern paradigms, such as federated learning, are controlled by a global aggregator that combines contributions from multiple clients. As a result, all contributors receive the revised model, providing a more dynamic and cooperative method of model improvement.

2.4.3 Large scale AI model- related issues

As mentioned earlier, complex models have significantly augmented accuracy of results and predictions. In fact, the study by Ali et al., 2023 [25] evidence that the more

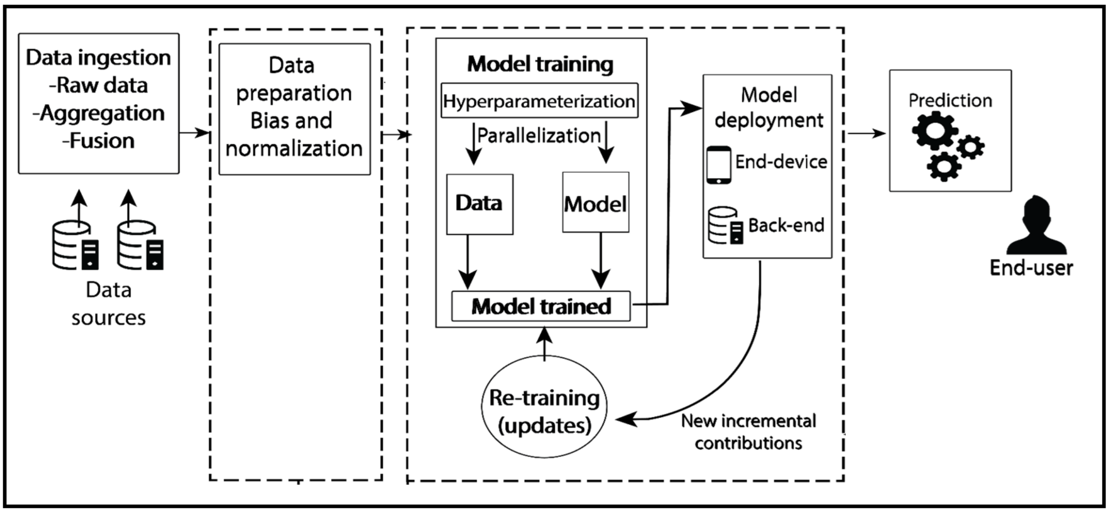
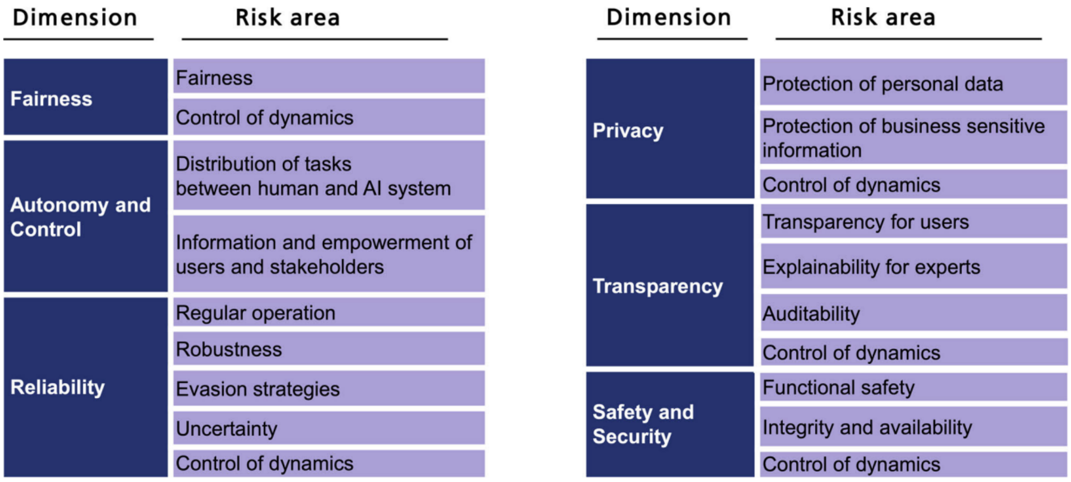


Figure 5. Standard machine learning pipeline.

profound the network is, the more accurate the generated results are. It is understood that given the above-mentioned applications of AI, the outcomes it presents need to be accurate, so complex models with infinite parameters and multiple layers are often used to obtain results that end users perceive as accurate. However, it has been identified that there's a tradeoff between the accuracy of a model and its understandability. In other words, the more accurate the model is, the more complex it is, thus making it harder for users to understand the outcome. Complex DNN models used at present often have millions of parameters that are considered when the model makes its decision. Between the number of layers and many variables a model considers when making a decision, it is tough to understand the model's reasoning, creating a "black-box" situation. This is the primary cause of doubt and distrust towards AI and AI-based applications in the present world. However, other factors promote adversity toward AI-based applications. Machine learning applications pose the risk of incorporating and replicating discriminating tendencies in actual data, which can have negative effects on individuals, organizations, businesses, and governments if such results are used in decision-making processes. A study by A. Schmitz et al.,2022, summarizes the Risks of AI as shown on Figure 6.

When presented with these threats, to ensure that the adversaries that AI brings to society do not outweigh the advantages, guidelines need to be put in place to ensure that AI and AI-based applications can be trusted. As is the case with traditional IT security and safety, preventing physical harm, property damage, or financial loss cannot be the only focus of these guidelines, as the intangible effects of AI on society and people also need to be considered.



Dimension	Risk area
Fairness	Fairness
	Control of dynamics
Autonomy and Control	Distribution of tasks between human and AI system
	Information and empowerment of users and stakeholders
Reliability	Regular operation
	Robustness
	Evasion strategies
	Uncertainty
	Control of dynamics

Dimension	Risk area
Privacy	Protection of personal data
	Protection of business sensitive information
	Control of dynamics
Transparency	Transparency for users
	Explainability for experts
	Auditability
	Control of dynamics
Safety and Security	Functional safety
	Integrity and availability
	Control of dynamics

Figure 6. AI risk scheme.

2.5 Trustworthy AI

The term trust in a technical context can be defined as the degree to which a user or a stakeholder has confidence that a product or system will behave as intended [14]. As outlined earlier, it is imperative that AI systems and models can be trusted. This brings forth the concept of trustworthy AI, Artificial Intelligence that can be trusted by humans. The concept of "AI trustworthiness" has been embraced as a comprehensive measure of excellence for AI applications, surpassing the traditional notions of IT security and safety.

2.5.1 Characteristics of Trustworthy AI

The successful implementation of ML technology in recent decades has primarily derived from the utilization of accuracy-based performance metrics. There has been a major focus on evaluating task performance using quantitative measures such as accuracy or loss, and the process of training AI models becomes manageable in terms of optimization. Meanwhile, even though the use of predictive accuracy indicates the excellence of an AI product, solely relying on accuracy as a benchmark for performance of a ML model revealed several new issues, as described under earlier topics. Therefore, additional metrics of trustworthiness for AI are considered more and more to ensure trustworthiness of AI based applications [1]. Certain attributes characterize trustworthiness in AI-based applications. Li et al., 2023 [1] among other studies [26] outline key elements of AI trustworthiness, encompassing robustness, explainability, generalization, reproducibility, transparency, fairness, privacy preservation, and accountability. A representation of

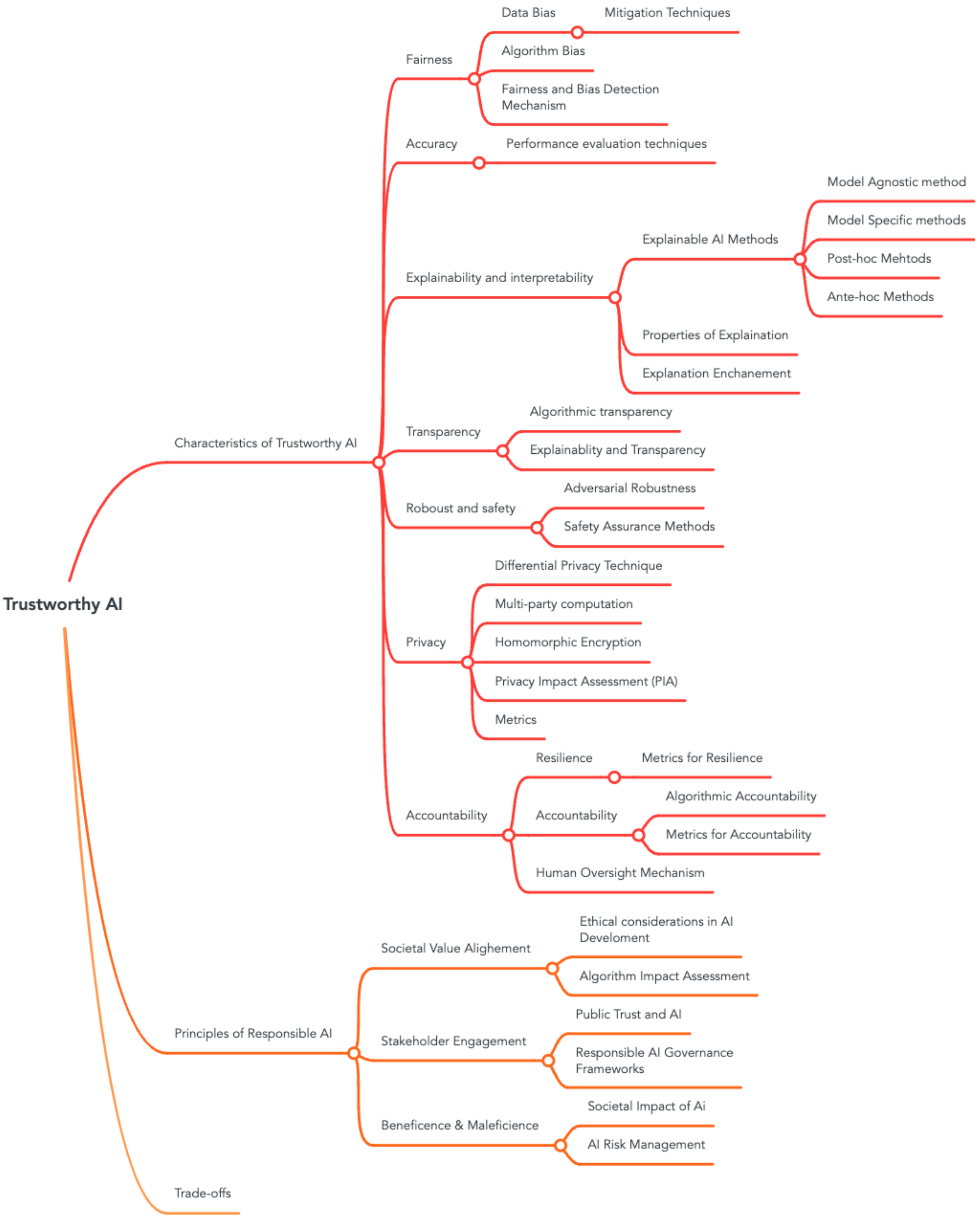

Trustworthy AI
Characteristics of Trustworthy AI
Fairness
Data Bias
Mitigation Techniques
Algorithm Bias
Fairness and Bias Detection Mechanism
Accuracy
Performance evaluation techniques
Explainability and interpretability
Explainable AI Methods
Model Agnostic method
Model Specific methods
Post-hoc Mehtods
Ante-hoc Methods
Properties of Explaination
Explanation Enchanement
Transparency
Algorithmic transparency
Explainablity and Transparency
Roboust and safety
Adversarial Robustness
Safety Assurance Methods
Privacy
Differential Privacy Technique
Multi-party computation
Homomorphic Encryption
Privacy Impact Assessment (PIA)
Metrics
Accountability
Resilience
Metrics for Resilience
Accountability
Algorithmic Accountability
Metrics for Accountability
Human Oversight Mechanism
Principles of Responsible AI
Societal Value Alighement
Ethical considerations in AI Delevolment
Algorithm Impact Assessment
Stakeholder Engagement
Public Trust and AI
Responsible AI Governance Frameworks
Beneficence & Maleficience
Societal Impact of Ai
AI Risk Management
Trade-offs

Figure 7. Trustworthy AI Overview.

characteristics of trustworthy AI and principles of Resposible AI has been outlined in Figure 7.

Robustness: Robustness, in broad terms, pertains to the capacity of an algorithm or system to effectively handle faults during execution, faulty inputs, or unfamiliar data. NIST Glossary for Trustworthy AI [14] defines Robustness as the ability of a machine learning model/algorithm to maintain correct and reliable performance under different conditions (e.g., unseen, noisy, or adversarially manipulated data). Studies indicate that the resilience of an artificial intelligence (AI) system can be evaluated based on three vulnerability factors: data, algorithms, and systems. To assess this metric, many conventional functional test strategies, such as monkey tests, as well as innovative testing approaches like mathematical proof of an AI model's adversarial robustness, can be employed [1].

Explainability and Interpretability: Explainability of an AI model is defined as the process of elucidating or revealing the decision-making mechanisms of a model. Interpretability also has an enmeshed definition with explainability. Interpretability is understanding the inner workings of the AI models, and understanding the model and its outcome by the information disclosed about the model itself. Interpretability helps developers, and explainability helps end users, to gain confidence in the decision-making process of AI and AI-based models. Interpretability focuses more on the underlying functioning that leads to decision-making - how the decision was made, while explainability gives insights into the logic of the outcome for end users to understand - why the decision was made [25].

Generalization: Generalisation refers to the ability of an AI model to extract information from a small amount of training data and use it to generate precise predictions about new, unknown data. Unseen or test inputs generally exhibit the same distribution as the training data. AI models can attain a satisfactory level of accuracy on training datasets, however, a notable disparity called the generalisation gap arises between their accuracy on training data and their accuracy on testing data. It is necessary to study statistical and deep learning methods in order to improve the generalisation of the model. Common techniques encompass cross-validation, regularisation, and data augmentation. In order to develop a contemporary AI model that relies on data, a substantial quantity of data is necessary during the training phase. As a result, producers and users incur significant expenses when they have to gather and annotate data in order to train the model for each specific activity. Conversely, AI models should possess the capability to make generalizations without the necessity of extensively gathering and annotating vast quantities of data across different domains, so a balance need to be obtained between reducing the generalization gap, and model training efforts and costs [1]. Generalisation is not commonly discussed in relation to trustworthy AI.

Reproducibility: Reproducibility, as defined by the NIST glossary, refers to the degree of agreement between measurement results of the same measurand when conducted under different measurement settings. Research indicates that achieving the lowest level of reproducibility necessitates replicating an experiment precisely using the exact implementation and data. On the other hand, a higher level of reproducibility can be attained by employing various implementations or data. In addition to the initial verification of research, a greater level of repeatability enhances comprehension of the research by differentiating the main aspects that impact effectiveness.

Transparency : NIST defines transparency as a property of a system that ensures appropriate information about the system is made available to relevant stakeholders. Appropriate information for system and model transparency can include aspects such as features, performance, limitations, components, procedures, measures, design goals, design choices and assumptions, data sources, and labeling protocols [14]. To better understand this concept, [1] cites an interesting example that highlights the transparency of an AI model with regard to a biometric identification system. Users typically express concern regarding the purpose and utilization of their biometric information. Business operators prioritize precision and resilience against attacks in order to effectively manage threats, while the government sectors are primarily focused on ensuring that the AI system adheres to established norms, laws, and regulations. The availability of this information stipulates transparency and helps establish public trust in AI systems. However, inappropriate disclosure of some aspects of a system can violate security, privacy or confidentiality requirements.

Fairness: There are so-called biases in prediction models. A bias is a systematic error, and unwanted bias places privileged groups at systematic advantage and unprivileged groups at systematic disadvantage, leading to "unfair results. The Metric of Fairness quantifies unwanted bias in training data or models. As mentioned earlier in the chapter, AI systems assist processes such as recruitment, financial risk evaluation, and facial recognition, and having unnecessary biases can lead to adverse social consequences. Marginalized groups may face consistent disadvantages in hiring processes or be disproportionately affected by criminal risk profiling. This can undermine the confidence society as a whole have in AI, but also hinders the progress and use of AI technology for the benefit of society. Therefore, it is important to measure the fairness of AI models to ensure users can trust its outcomes.

Accountability: Accountability for systems, is a measurement that ensures that actions of an entity can be traced uniquely to the entity. Accountability is a fundamental aspect of the AI development lifecycle, as it necessitates that the individuals involved in the development, execution, and maintenance of the AI system be responsible for explaining how their actions are consistent with intuitive human values [1].

Privacy : Privacy, as defined by the NIST glossary, refers to the state of being free from intrusion into one's private life or affairs. This intrusion occurs when there is an unjust or illegal collection and use of data about an individual. AI-powered applications may have access to a diverse array of data, such as an individual's name, age, gender, facial image, fingerprints, and other related information. Anonymity, confidentiality, and a strong dedication to privacy protection are considered crucial aspects in assessing the trustworthiness of an AI system. Government entities are developing an increasing amount of policies to oversee the privacy of data. The General Data Protection Regulation (GDPR) is a comprehensive regulatory framework that compels organisations to use appropriate steps to safeguard user privacy [1]. Differential privacy is a widely used approach for assessing privacy. It is a technique used to quantify the amount of information that may be inferred about an individual from the result of a calculation. The process relies on the randomized introduction of "noise." Noise refers to the introduction of random changes to data inside a dataset, with the purpose of making it more difficult to identify specific individuals through direct or indirect identifiers [14].

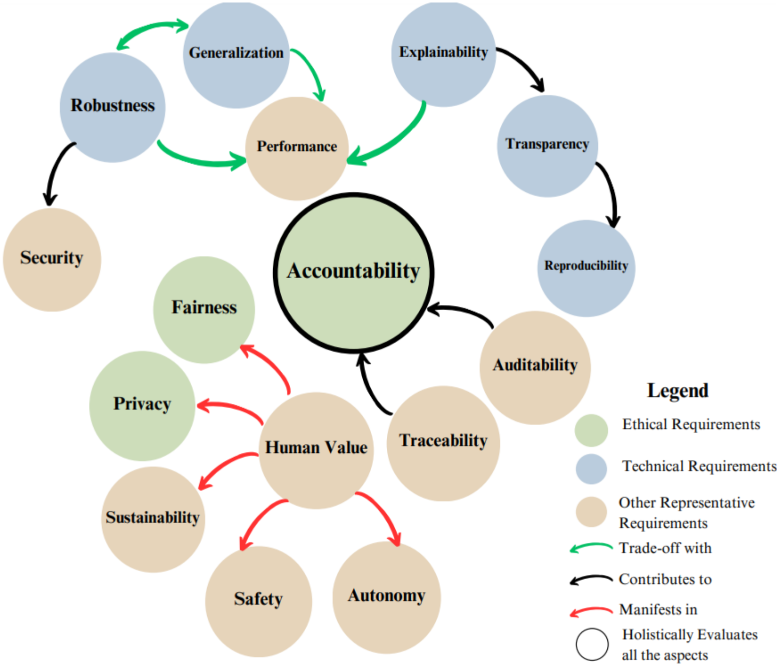


Figure 8. Relationships between different aspects of AI trustworthiness. [1]

The above mentioned characteristics are also connected to each other and serves as metrics to a wide range of other representative requirements. The connectivity between these characteristics are outlined in Figure 8. Additionally, it has been noted that to implement trustworthy AI effectively, it is necessary to comprehend its connection to the measures taken throughout the development process, despite the fact that it is commonly stated as a system need. [27].

2.5.2 Human-in-the-loop supervision of AI-based applications

It is evident that, Human social morality and ethical principles should be followed by artificially intelligent systems intended for decision-making and other purposes, to ensure trustworthiness. It is a crucial need to prevent personal prejudice and discrimination and to promote the reliability and trustworthiness of AI systems [28]. The best, most accurate outcome may not always be the correct course of action, as the sustainability, morality, and ethical standing of decisions also may need to be considered depending on the context. Pal, 2020 presents an example for this: For instance, an AI system designed to aid farmers should not solely focus on optimising crop production, but should also consider the ecological consequences of excessive nitrogen fertiliser usage [29]. Both human and artificial intelligence has advantages of its own [29]. It should be mentioned that when AI helps people make decisions, those decisions may perform differently than those made by the AI alone or by humans acting without AI help. In the researched scenario of a Swedish labor market decision-support system, for instance, caseworkers are told to follow the automated advice mainly but can override it in specific situations to reach the best decision [30].

In fact, trustworthiness is promoted by combining the two to ensure that humans are in the loop and monitoring the system. Person-machine cooperation and integration are expected to become mainstream trends of intelligent society in the future, and a man-machine-object ternary system with a person in the loop will be created. Humans and machines typically interact extensively in this system through monitoring, collecting, transmitting, analyzing, fusing, and using information [28]. The system design mode of "human in the loop" should be used especially in judicial, medical, military, and other applications that involve significant human interests to guarantee the maintenance of the final decision-making right by human beings [28]. This makes it necessary to create a strong and effective evaluation and monitoring system to realise whole-process supervision on artificial intelligence algorithm design, product development, data collecting, and product application as well as to promptly identify and address new issues and challenges in the creation of man-machine inclusive society [28]. Also, it is important to note the accountability aspect of human oversight. Specifically, it means that the system should be designed to enable human decision-makers to answer for the choices they make with AI assistance. For instance, caseworkers must be able to somehow supervise the AI and come to a judgement, considering not only the AI's output but also other important factors, and if they are officially in charge of making choices there should be a system in place that ensures that the accountability of such decisions can be traced back to the decision maker [30].

2.5.3 Regulations that ensure AI trustworthiness

As important as it is to have characteristics that define trustworthiness, it is imperative to have legal regulations to ensure that these characteristics are sought after. At the moment, more than 700 AI policy proposals from 60 nations and territories, together with over 170 new AI-related legislation projects exist, and the characteristics necessary to verify and validate the appropriate development and application of AI models are specified by AI regulations. However, it is evident that there are varying perspectives worldwide regarding the appropriate use of AI. So many guidelines exist because of the variety of national and territorial AI strategies that influence official guidelines and regulations on AI. Although criteria vary widely, common components are included in many AI guidelines, particularly in Western programs. Specifically, both European and US American guidelines agree that privacy, fairness, and transparency are important aspects of AI, in addition to the already recognized aims of reliability, safety, and security. These dimensions are considered necessary for ensuring the quality of AI. [31] The EU AI Act, proposed by the European Commission on April 21, 2021, and passed on 13 March 2024, pioneers in this context as the first comprehensive regulation of AI by a significant regulatory authority. Furthermore, the EU-AI Act strongly emphasizes human autonomy and oversight as crucial criteria for ensuring trustworthy AI [7].

Another pertinent guideline in the European legal ecosystem with regard to AI would be the General Data Protection Regulations (GDPR). It establishes the protocols for handling personal data within the European Union (EU), emphasizing the importance of transparency, impartiality, security, privacy, trust, and explanation in the development of AI-based solutions and software. Other existing regulations include, for instance, the Health Insurance Portability and Accountability Act (HIPAA), the Cybersecurity Law of China, the California Consumer Privacy Act (CCPA) and the Data Governance Act [32]. Considering the above mentioned aspects, modern applications need to include tools or methods that let people understand the inference capabilities of artificial intelligence. However, this necessitates a comprehensive examination of the AI model's construction, which warrants its own challenges.

2.5.4 How do Existing AI-based Applications support trustworthiness?

Existing regulations for AI and the ever-growing popularity of AI have motivated researchers and developers to consider the trustworthiness aspect of AI. Studies emphasize that the European Commission's trustworthy AI principles, in particular, which stress the importance of data protection, security, and responsible governance, serve as the foundation for AI's universal acceptance [33]. It has been studied that there's an apparent link between social perceptions of AI's impact and factors such as trustworthiness, associated hazards, and usage/acceptance [33]. According to the survey I conducted on the literature acquired for the "trustworthiness in AI applications" keyword string, embedding trustworthiness in applications is an increasing trend in modern AI-powered applications. A recapitulation of the survey is outlined in Table 2. This shows conclusively that existing AI-based applications support trustworthiness characteristics. Therefore, a subsequent need arises, for a platform that can monitor the inference capabilities of such AI systems, so that human operators can visualize and evaluate the trustworthiness of the application. So as the natural next step, the abundance and workings of platforms that can monitor the inference capability of the underlying AI model of such applications were examined.

2.5.5 Existing-Platforms that monitor and gauge AI inference capabilities

The selected literature library of 110 studies contained one study that dedicated efforts toward assuring the trustworthiness of high-risk AI systems, to comply with the EU AI Act. To be deployed, high-risk classified AI systems need to meet seven human-centric and trustworthy standards. The study evaluated them throughout the AI life cycle. Researchers Stettinger et al, 2023 employed risk assessment methodologies such as Operational Design Domain (ODD) and Behavior Competency (BC) , to evaluate residual hazards, in other words, to evaluate the risks of the AI systems - inversely proposing a brief idea about the trustworthiness of the same. The methodology of this study followed a continuous application of the ODD and BC throughout the AI life cycle, focusing on the trustworthiness assurance framework and process for AIS certification, and proposes a trustworthiness assurance assessment for High-Risk AI applications [26]. This is the closest approach I found from the selected literature, focusing on creating a "trustworthiness evaluation" for AI applications, but the evaluation of trustworthiness is indirect and not extended toward trustworthy AI characteristics in this context. This highlights the existing gap in research for a platform that monitors and gauges AI inference capabilities and AI trustworthiness.

Table 2. Research papers discussing trustworthy properties of AI models.

AI Model	Trustworthy Properties						
	Accuracy	**Explainability/ Interpretability**	**Fairness/ Bias**	**Privacy**	**Security**	**Robustness**	**Transparency**
ML / DL	[34, 35, 36]	[37, 38, 18] [39, 40, 41]	[36, 35, 42] [39, 43]	[36, 42, 43] [44, 45, 46]	[36, 42, 43] [44, 45]	[36, 47, 43, 44]	[39, 41, 48]
CNN	[49]	[50, 49, 51] [52, 53, 54] [55, 56, 57] [58, 30, 59] [60, 61]	[62, 50, 49]			[61, 63, 5]	[56, 30, 20]
DNN	[64, 29]	[65, 66, 67] [68, 69, 29]	[65]	[65, 70]	[71, 72]	[71]	[29]
BNN		[73]					
Naive Bayes classifier	[74]						
Multi-sensor Data Fusion model	[75]						
PGAR model		[4]	[4]		[4]	[4]	[4]
Vision-Language		[76]	[76]			[76]	
Random Forest	[3]	[3]	[3]				
GAM model		[77]					
GAN model		[78]					

2.5.6 Modern Architectures

Building a platform that monitors trustworthy AI metrics involves employing an architecture that suits the underlying functionality. Therefore, an analysis needed to be done on the underlying system architectures of modern applications to determine the most suitable architecture for the SPATIAL platform. There has been an increase in the attention given to design and development considerations. In the initial stages of development, within a simple client-server framework, end devices functioning as clients transmit requests to the server. Upon reaching the server, the request undergoes processing and a response is then transmitted back to the client (as shown in Figure 9).

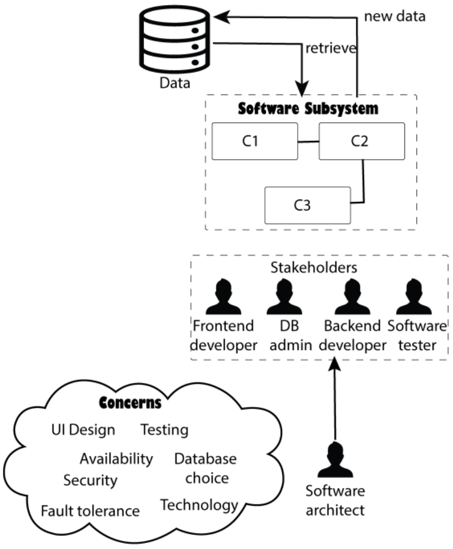


Figure 9. Basic client-server architecture.

Subsequently, more advanced frameworks are developed to gather data in a centralized manner (at the server) from users interacting with applications. The data is subsequently utilized to train machine learning models in order to enhance certain functionality over time (Figure 10. Advancements have enabled these systems to collect data from clients in a distributed manner, allowing for the utilization of more resilient datasets for model training.

At the moment, the global model receives training using client-provided data in a manner that respects their privacy, such as through Federated Learning (FL). Once the model is developed, it is then distributed to all the end devices. Figure 11 expands upon the ML architecture introduced in [79] to illustrate the most latest advances in distributed training.

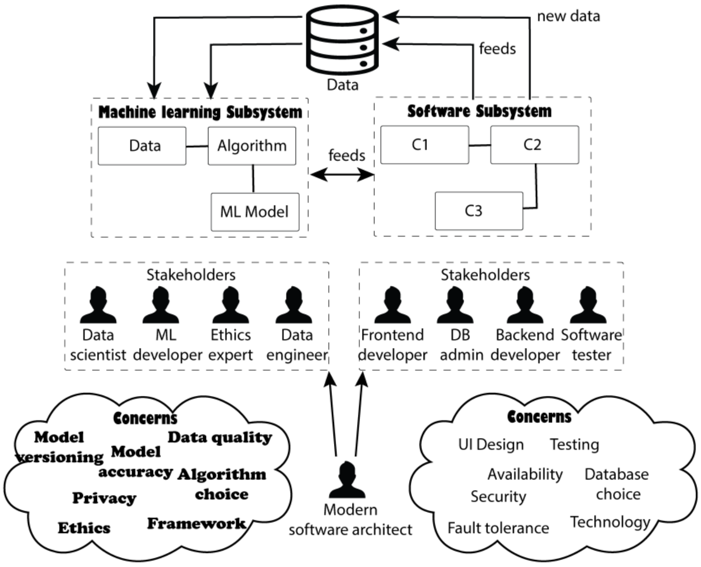


Figure 10. Machine learning architecture.

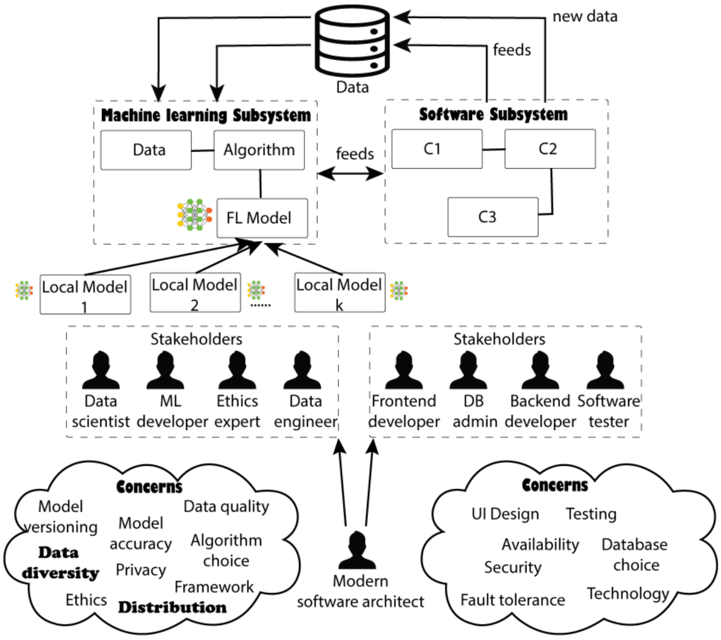


Figure 11. Federated learning architecture.

2.6 Explainable AI

2.6.1 What is XAI and Why it is needed

This chapter deep dives into explainability and sets precedence for the explainability microservice of the SPATIAL architecture. As such, this section describes explainability and explainable AI (XAI) and establishes the need for this measurement. Furthermore, XAI taxonomy and different XAI methods are also outlined.

AI and AI-based models are currently being utilized widely, across many fields and applications. However, as mentioned previously, the results or outputs of such AI models are sometimes difficult to comprehend, and challenging to trust, as users don't have an idea as to how this outcome was generated. This is known as the black box nature of AI, and is currently challenging the trustworthiness of AI. The concept of XAI was brought about to specify a system's capacity to explain the actions of AI-based applications. From the user's perspective, it sheds some light on the reasoning behind an AI model's decision-making [12] and helps shed some light on the "black-box" like nature that's typical in complex AI models. The terms interpretability and explainability are used here as intertwined concepts that complement each other. Interpretability helps developers, and explainability helps end users gain confidence in the decision-making process of AI and AI-based models. Interpretability focuses more on the underlying functioning that leads to decision-making - how the decision was made, while explainability gives insights into the logic of the outcome for end users to understand - why the decision was made. The amount of literature and research on AI explainability has steadily risen over time, indicating a growing interest in XAI and its associated applications. Furthermore, it is evident that various sectors have been actively pursuing the attainment of explainability in AI-based applications. These trends are clearly apparent in the literature chosen for this review. The distribution of literature over the recent years is shown in Figure 12 as evidence of these trends. Table 3 showcases XAI publications categorized by sector.

In recent years, many different approaches have arisen with the goal of providing explanations for AI algorithm behavior and therefore increasing its explainability. Because several of these techniques exhibit fundamentally distinct properties, we will provide a quick overview of the taxonomy of explainable AI approaches in the next section. Following that, we will briefly introduce some relevant state-of-the-art XAI methods that can support the explainability and thus accountability of AI-based systems and networks.

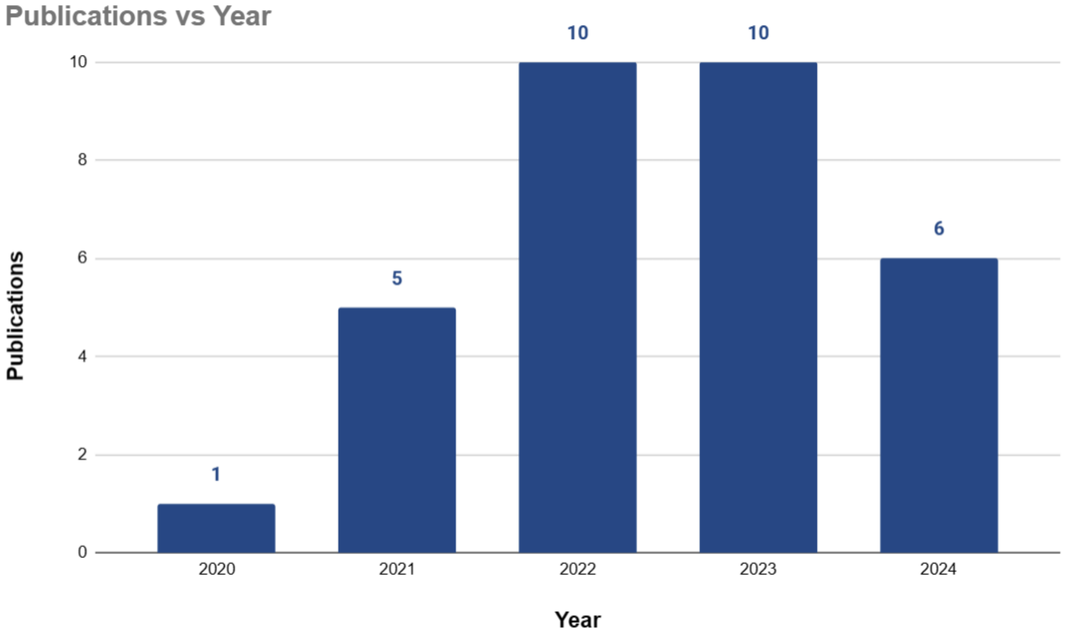


Figure 12. XAI publications over the recent years.

Table 3. XAI publications based on the sector.

Sector	**#**	**XAI method(s)**	**AI Model(s)**
Healthcare	[80]	LIME/GradCAM/ SHAP/DeepLIFT	DL
	[81]	SHAP/ LRP/ GradCAM	ML
	[82]	CAM	NN
	[83]	Causal Interpretability	ML
	[84]	LRP/ BERT-LRP	ML
	[85]	LIME/ SHAP	DNN
	[86]	LLM	Rule-based model
	[87]	SHAP/ BRCG/ Protodash/ LIME/ CEM	ML-based Wireless Network IDS

Continued on next page

Table 3 – continued from previous page

Sector	#	XAI method(s)	AI Model(s)
	[88]	SHAP	ML Algorithms
	[89]	SHAP/ LIME/ Heat map	CNN
	[90]	LIME	DNN
Education	[91]	LIME	DNN
Technology	[92]	Post-Hoc/ Model-specific explanations	ML
	[93]	SHAP	DNN
	[94]	Concept SA/ LRP /Propagation rules TCAV/ LIME	DNN
	[73]	LRP/ SHAP	BNN
	[86]	Model-agnostic/ Model-specific explanations	ML
	[95]	FED-XAI	Decision Trees Rule-based Models
	[96]	GradCAM/ LRP/ SmoothGrad/ LIME/ Integrated Gradients (IG)	CNN
	[97]	LIME SHAP	ML
	[98]	LORE/ SHAP/ LIME/ Anchors/ GradCAM/ CEM	DL
Telecommunication	[99]	DeepSHAP	DL
Psychology	[100]	Grad-CAM LRP	DL
Finance	[101]	TreeSHAP	NN
Manufacturing	[102]	Global explanations	RL

2.6.2 XAI Taxonomy

The XAI methodologies can be categorised into numerous groups according to different criteria [103, 104]. Below, we will explore the most prevalent taxonomies based on Explainable Artificial Intelligence (XAI). Methods of Explainable Artificial Intelligence (XAI) that belong to these categories are not necessarily limited to each specific group. Some methods can belong to multiple categories within a taxonomy.

Model-agnostic vs Model-specific methods: Model-agnostic methods for Explainable Artificial Intelligence (XAI) refer to techniques that are not limited by the fundamental components of an AI algorithm while producing predictions. They assist in deciphering the decision-making process of black box models and offer developers ample freedom to implement them across different machine learning models. However, model agnostic explanation approaches often have limited access to the machine learning models they are working with, as they treat them as black-box functions and only have access to the model's output. These methods can be applied extensively and are highly flexible because they do not rely on any knowledge about the internal workings of the model, such as the topology, learnt parameters (weights, biases), and activation levels, especially in the case of neural networks. In contrast, model-specific approaches are tailored for individual models and utilise fundamental components of a machine learning model to interpret the results. Model-specific approaches are better suited for identifying precise details of machine learning models, but they lack flexibility.

Local vs Global methods: XAI methods can be categorised into two primary groups depending on the extent of the function employed by the interpreter to generate an explanation. There are both local and global explanations. Local explainers are specifically designed to analyse and interpret a specific part of the model function that influences the conclusion for a particular data point. When creating an explanation, the ML function explores the immediate proximity of the data point. In contrast, global approaches consider the ML function as a whole entity when producing explanations for the inference. These approaches are often slow but reliable, while local methods are rapid but occasionally unpredictable.

Pre-model, In-model, and Post-model explainers: XAI approaches can be categorised into three primary groups based on the stage at which they are implemented in the development process: pre-model, in-model, and post-model. Pre-model approaches are mostly employed during the dataset preparation phase in the model development pipeline. These methods are valuable for doing data analysis, performing feature engineering, and quickly understanding any underlying patterns observed in the data. Explainable Artificial Intelligence (XAI) methodologies are integrated within the Machine Learning (ML) algorithms. This encompasses all the transparent models, such as linear regression or decision trees. Furthermore, model explanations are produced by making alterations to the current ML model architectures using inherently transparent models. Post-hoc/post-model explanations are utilised subsequent to the training of a machine learning model. This allows us to ascertain the knowledge acquired by the model during the training phase.

Surrogate vs Visualization: XAI approaches are categorised into two primary groups based on the specific aspect that is explained during the process. Surrogate model-based explainers produce explanations by using an approximate model of the black-box model, which is taught to imitate the behaviour of the actual model. Surrogate models are generally intrinsically interpretable. Alternatively, individuals can employ visualisation techniques such as heatmaps and graphs on the original black-box model to investigate its internal mechanisms without the need for a representation. These XAI approaches are classified as belonging to the visualisation category.

2.6.3 XAI Methods

LIME: Local Interpretable Model-agnostic Explanations (LIME), as described by Ribeiro et al. (2016) [105] is a commonly utilised method for evaluating the results of black-box models across various domains and applications. LIME, as its name implies, provides a localised explanation by focusing on a specific collection of data to approximate explanations for model predictions. This technique is feasible because any intricate model exhibits linear behaviour at a local level. However, LIME has lately been well-known for its fast performance (compared to other global explanation techniques) and ease. It is able to understand outputs from any type of black-box model, regardless of the model it is wrapped around (model-agnostic).

SHAP: Shapley Additive Explanations (SHAP) is a model-agnostic technique in Explainable Artificial Intelligence (XAI) that quantifies the contribution of each feature value towards a certain prediction. When it comes to explaining individual predictions, the method employed is based on the concept of Shapley values. The Shapley values are a widely used technique in cooperative game theory that addresses the issue of equitable distribution of rewards among participants in a group. Given that players may contribute to winning in varying degrees, the allocation of rewards should correspondingly reflect their individual contributions. This notion is utilised to elucidate local AI forecasts and discern the varying contributions of features to the ultimate prediction. In this case, Shapley values are employed to compute the impact of each feature on the prediction by evaluating its incremental contribution for every conceivable combination of characteristics [106].

LRP: Layer-wise Relevance Propagation (LRP) is an explanation approach that relies on propagation. It necessitates access to the underlying components of the model, such as its topology, weights, and activations. However, this supplementary knowledge regarding the model enables LRP to streamline and thereby enhance its ability to solve the explanation problem more effectively. LRP, or Layer-wise Relevance Propagation, differs from model agnostic methods by leveraging the structure of a deep neural network to explain its predictions. It achieves this by redistributing the relevance factors, layer by layer, starting from the model's output and propagating them onto the input variables, such as pixels. Each redistribution can be viewed as the resolution of a simple explanatory problem, because it only involves two adjacent layers [107].

3 Methodology

This section describes the systematic approach employed to design, implement, and deploy the SPATIAL platform, which utilizes a microservice-based architecture to enhance the trustworthiness of AI applications through thorough analysis and monitoring.

3.1 System Design

The architecture of the SPATIAL platform (See Figure 13), is characterized by a microservice pattern, enabling the development of advanced AI applications. This approach involved creating independent, containerized services, each meticulously designed for specific SPATIAL functionalities. Every microservice encapsulates its unique set of functions, databases, and libraries, promoting a modular and scalable system. The Application Programming Interface (API) Gateway, serves as a centralized access point for clients and enabling efficient communication and integration across diverse services. It streamlines request processing and facilitates the discovery and utilization of microservices. Overall, the SPATIAL architecture embodies modularity, scalability, and resilience, empowering stakeholders with enhanced explanatory capabilities, quantifiable data, and efficient system management tools.

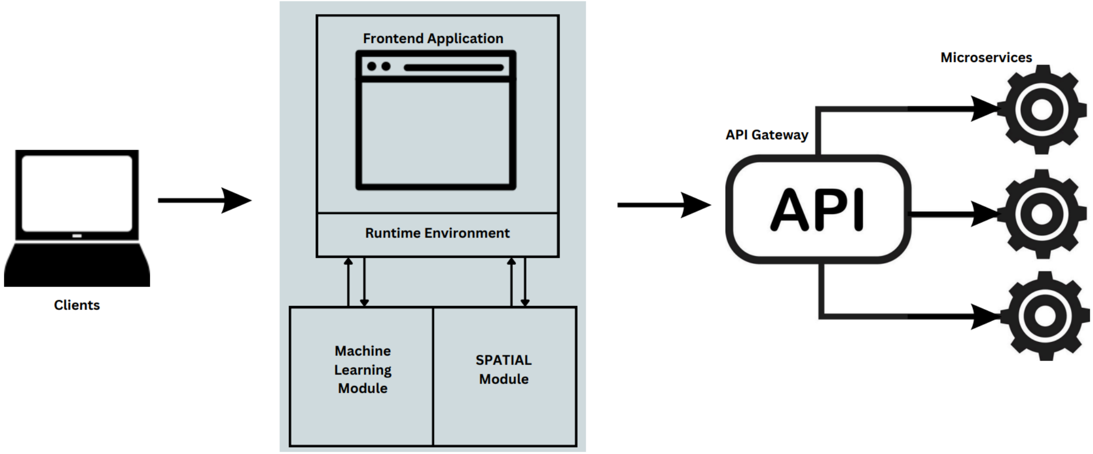


Figure 13. The SPATIAL Architecture.

Figure 14 illustrates the sequence of interactions between key objects in the SPATIAL architecture during a user login and request handling process.

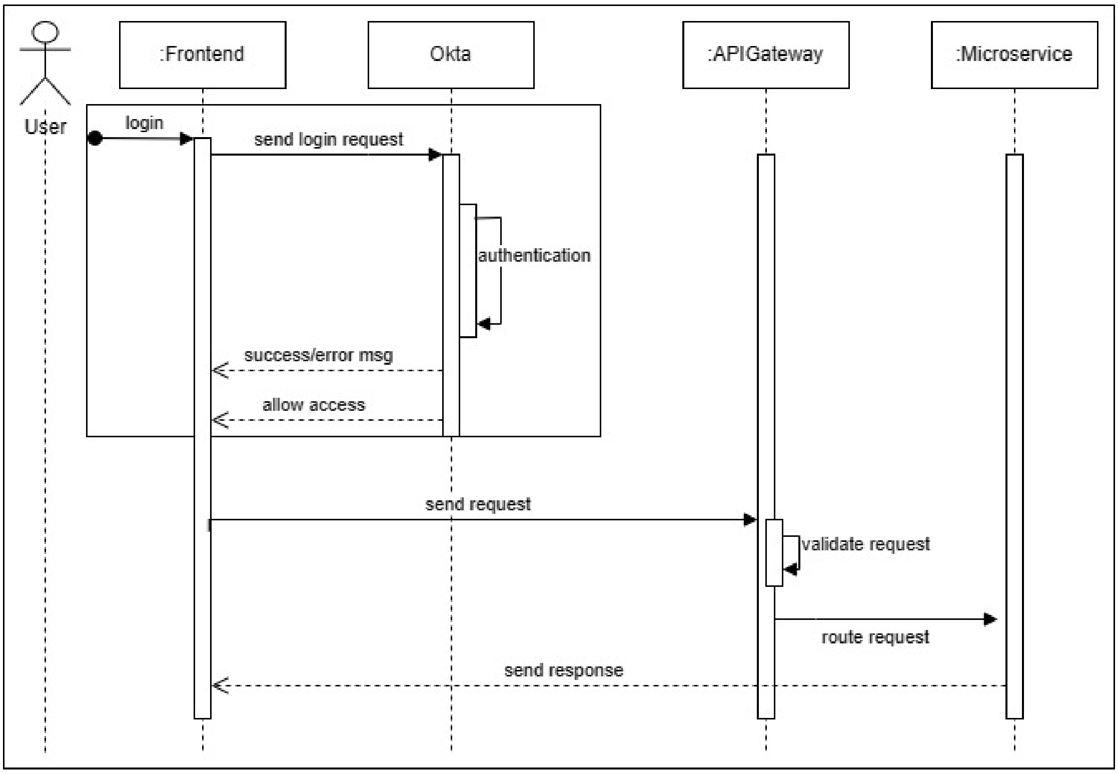


Figure 14. Sequence diagram of the SPATIAL architecture.

3.1.1 Augmented Machine Learning Pipeline

Our SPATIAL architecture augments latest ML/FL system architectures by building upon the standard machine learning pipeline that is available to construct the models. Figure 15(a) shows the additional steps introduced in the pipeline, such that trustworthiness properties can be verified using XAI and additional specialized metrics (see Figure 15(b)). This verification consists in building a diagnosis profile about the application that is analysed, such that it is possible to characterize and quantify the performance of AI models. The use of a combination of metrics is meant to improve the robustness when analysing AI models.

Indeed, XAI alone can be used to detect possible issues in models, but the provided insights are insufficient to accurately identify drifts and errors in performance nor their root causes. In addition to this, as trustworthy properties depict trade-offs, the main reason for extending the existing pipeline is to provide a way of tuning these trade-offs over time, following a tuning based on a human-in-the-loop (stakeholders) feedback and to avoid post de-facto verification, in which tuning is difficult to be implemented once the final model is deployed. Thus, our SPATIAL architecture fosters a correct-by-construction approach as models are updated over time. In parallel to this, Figure 16

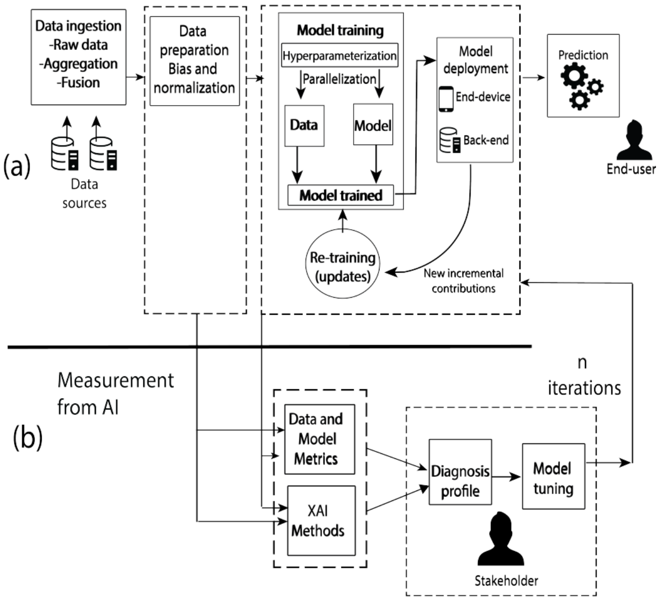


Figure 15. Augmented pipeline to analyze trustworthy trade-offs.

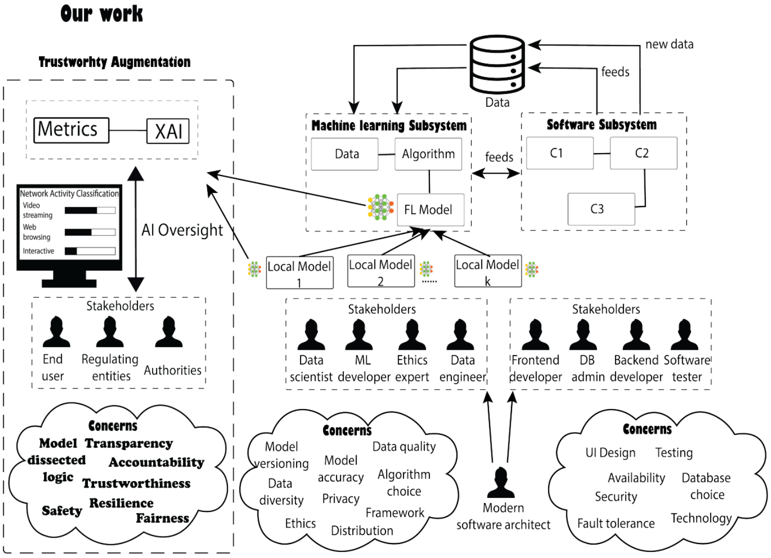


Figure 16. Conceptual modern system architecture equipped with methods to monitor trustworthiness.

also shows how modern system architectures are augmented when considering this new functionality. It is possible to observe the new concerns when developing not just applications that rely on AI but also ensure that AI is trustworthy.

3.1.2 AI Sensors in SPATIAL

Figure 17 illustrates virtual sensors that continuously monitor and analyze a trustworthy property over a period of time. Every trustworthy property is connected to an artificial intelligence sensor. The sensors are integrated into the target application, and their reliability is measured. These sensors, integrated as APIs within applications, allow for the measurement of AI compliance with existing criteria, providing information on how well the model adheres to required specifications. An API instrumentation offers a significant benefit by allowing the intensive computation required for the functionality of an AI sensor to be outsourced or offloaded to the server.

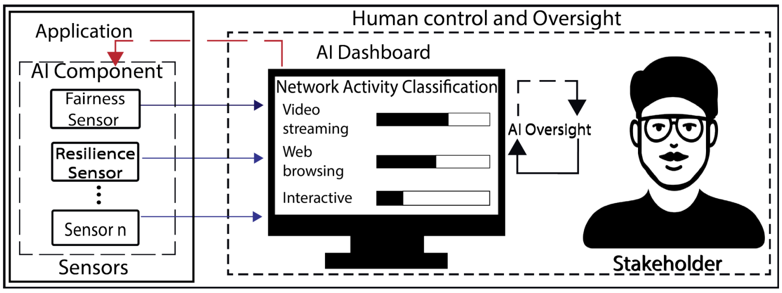


Figure 17. SPATIAL concept overview.

3.1.3 AI Dashboard in SPATIAL

SPATIAL's AI dashboard provide a user interface that allows humans to monitor the analysis of AI models, both for the data used and the AI model itself once it has been trained. AI dashboards clearly display all the measurable data collected by the AI sensors to users. This enables human specialists to collaborate in overseeing the development of the model and tuning the AI system to solve trade-offs in trustworthiness properties, while adhering to legal requirements.

3.2 System Implementation

This section details the practical steps taken to implement the SPATIAL platform, with a particular emphasis on the deployment of its microservices and the integrations of microservices into the API Gateway.

3.2.1 Deployment of the SPATIAL Platform

The deployment of the SPATIAL platform involved a systematic approach to configuring and launching its various microservices to ensure the overall functionality and performance of the system. Figure 18 shows the deployment of our augmented software architecture. We next provide a detail description of each component implementation.

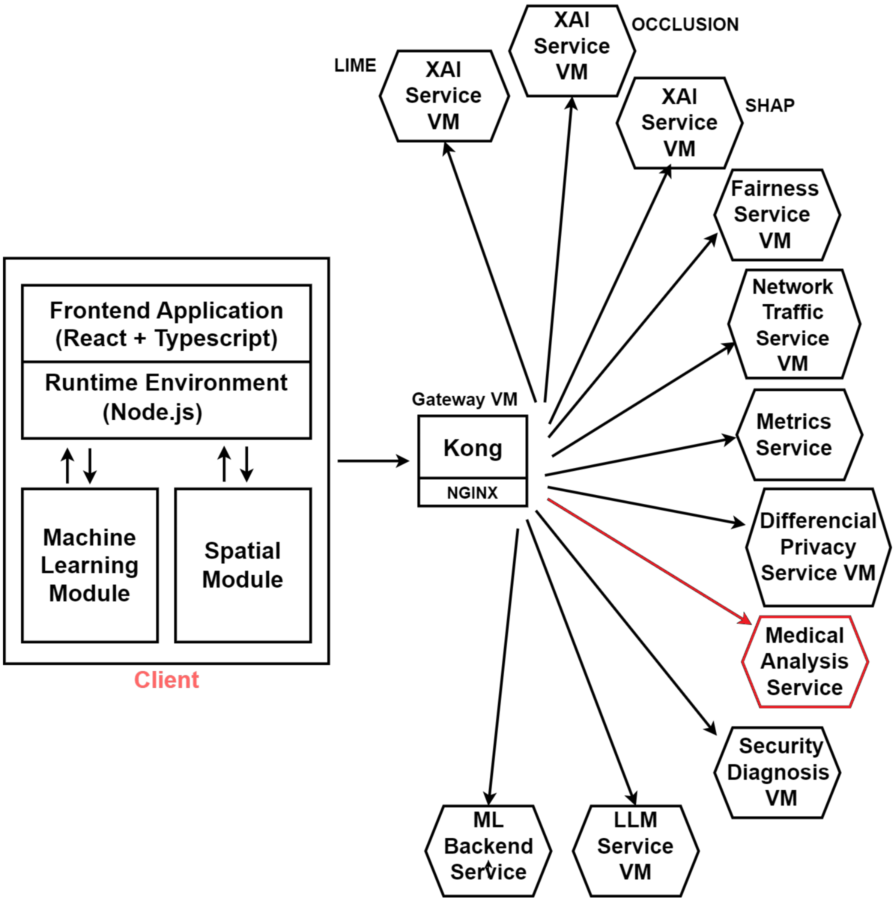


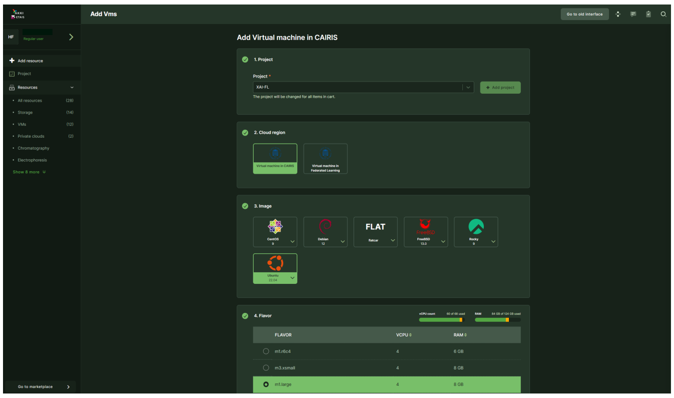
Figure 18. SPATIAL system deployment.

Back-end implementation: SPATIAL employs a micro-service pattern to assess the trustworthiness of AI by utilizing a combination of metrics and services. The core concept is that each micro-service focuses on characterizing a specific, trustworthy property, e.g., micro-service for explainability, micro-service for fairness. Micro-service patterns enable easy replacement of metrics for quantifying trustworthiness. This is beneficial since there is currently a discrepancy between legal and technical trustworthiness. Thus, metrics that align better with legal requirements can be easily modified in SPATIAL. Node.js serves as a foundational runtime environment in our architecture, preceding the API Gateway. It is used to create server-side applications that can handle several requests at the same time by using an asynchronous, event-driven programming approach. Our API gateway [108] is built using the open-source Kong technology, which allows for seamless connection with OpenAPI and provides extensive configuration options for continuous integration. The API Gateway coordinates communication, guaranteeing that each micro-service obtains the required input, handles it, and provides the accurate response. We used NGINX [109] to define Upstreams and API addresses in the configuration file to target particular URL paths to route to the corresponding micro-services. The metrics and services that assess trustworthiness are containerized as micro-services using Docker, following a request/response model. In the SPATIAL framework, a virtual machine (VM) is initially set up, onto which Docker [110] images containing all necessary dependencies and configurations are loaded. The use of Docker containers streamlines the deployment process by standardizing and isolating the environment, which facilitates consistent deployment experiences across various setups.

Front-end implementation: The frontend of SPATIAL is developed using React, offering users an easy interface for seamless integration with SPATIAL's functionalities. Node.js is the essential runtime environment for React's development tools, such as Babel and Webpack. The Bootstrap 5 framework is used for creating responsive designs, while Tailwind CSS is used for customizing styles, resulting in visually attractive user interface components. In addition, the SPATIAL client incorporates Okta for identity management, guaranteeing safe and strong authentication and authorization capabilities for access control. We employ D3.js, Chart.js, and Papaparse to handle datasets and create interactive charts. Papaparse is specifically used for processing CSV data.

3.2.2 Micro-service Deployment

To deploy SPATIAL microservices, the initial step involves creating Virtual Machines (VMs). This process is seamlessly facilitated by the Virtual Desktop/Cloud Service provided by the High Performance Computing (HPC) Center [111]. Affiliated with the University of Tartu, the HPC Center excels in maintaining and advancing scientific computing infrastructure. Their Virtual Desktop/Cloud Service, powered by robust AMD EPYC servers, plays a crucial role in enabling the smooth and efficient creation of VMs. HPC Center is also affiliated with other cloud vendors in cooperation with LUMI partners. For streamlined access to our Cloud services and VM deployment, minu.etais.ee serves as the primary portal. This user-centric interface simplifies the process of requesting, managing, and optimizing virtual hardware resources. As depicted in Figure 19, ordering a virtual machine requires a VM name and selection of a VM image. Selecting the initial VM resource profile is flavour, it will set the initial resource profile for a VM - how much RAM, virtual CPU cores (vCPU), and storage it will have. Selecting VM flavor will also update "System volume size" with the option to override it manually (to a higher custom value). The size of "Data Volume" can be customized and incremented in 1GB steps. "System volume" must be at least 10GB, whereas "System volume" and "Data volume" must be equal to or less than VPC's total Storage. By default, no incoming connections will be allowed for a VM. Predefined Security Groups (firewall rules) must be linked to a VM in order to open up access (like ssh, Hypertext Transfer Protocol (HTTP), etc.). A summary is depicted in Figure 20.



Add Virtual machine in CAIRIS

1. Project

Project *: XAI-FL

The project will be changed for all items in cart.

2. Cloud region

● Virtual machine in CAIRIS

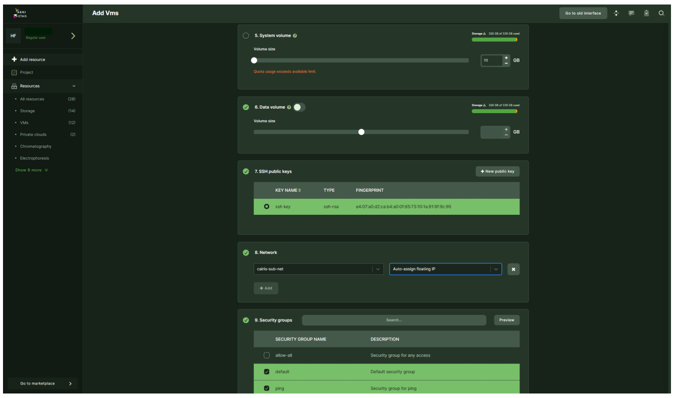
○ Virtual machine in Federated Learning

3. Image

4. Flavor

	FLAVOR	VCPU	RAM
○	m1.r6c4	4	6 GB
○	m3.xsmall	4	8 GB
●	m1.large	4	8 GB

Figure 19. Example for ordering VMs and corresponding resource profile at HPC center of University of Tartu.



5. System volume

Volume size: 10 GB

Quota usage exceeds available limit.

6. Data volume

Volume size: GB

7. SSH public keys

	KEY NAME	TYPE	FINGERPRINT
●	ssh key	ssh-rsa	e4:07:a0:d2:ca:b4:a0:0f:65:73:10:1a:91:9f:9c:95

8. Network

cairis-sub-net | Auto-assign floating IP

9. Security groups

	SECURITY GROUP NAME	DESCRIPTION
□	allow-all	Security group for any access
☑	default	Default security group
☑	ping	Security group for ping

Figure 20. Example listing further configuration details for the ordered VMs.

To access VMs seamlessly, we employ the x2go client [112] , providing a user-friendly interface for efficient remote connections. Below are the steps we follow to configure the VM for optimal use.

1. Connect to the VM:

```
ssh ubuntu@{external IP of the VM}
```

To securely access the VM.

2. Update the packages:

```
sudo apt update
```

To ensure the system is up to date.

3. Install XFCE Desktop:

```
sudo apt install xfce4
```

To enhance the user interface by installing XFCE desktop environment.

4. Install x2go Server:

```
sudo apt install x2goserver
```

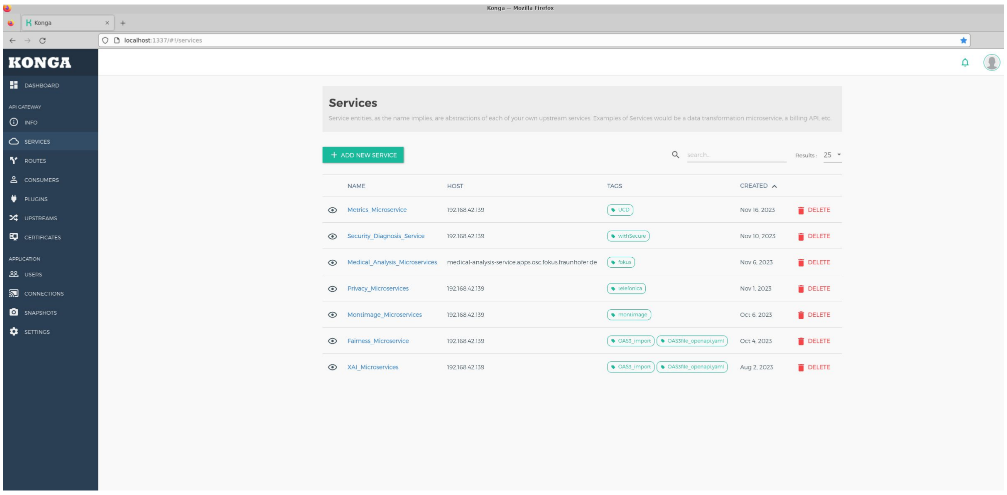
To enable seamless remote connection by installing the x2go server.

Deploying the microservices involved leveraging Docker [110] for efficiency and consistency. After confirming Docker on the VM, we adopted a streamlined approach. This included creating a Docker image encapsulating the microservice's dependencies and configurations, followed by deployment through a Docker container. This Dockerized deployment simplified the process and provided a standardized, isolated environment, ensuring a seamless deployment experience across different instances.

3.2.3 Integration of Services into the API Gateway

The integration of components in the context of Kong Gateway [108] involves the seamless coordination and configuration of services, routes, upstream applications, and NGINX [109] within the overall architecture. This integration process ensures efficient traffic management, routing of requests, and reliable distribution of workloads across backend services. The integration starts in the main admin interface of Kong (see Figure 21).

Service Definition: In Kong Gateway [108], the service was defined by providing a name and connection details for the upstream application. The connection details, including protocol, host, port, and path, were specified either as a single string in the Uniform Resource Locator (URL) field or through individual values. This was achieved through



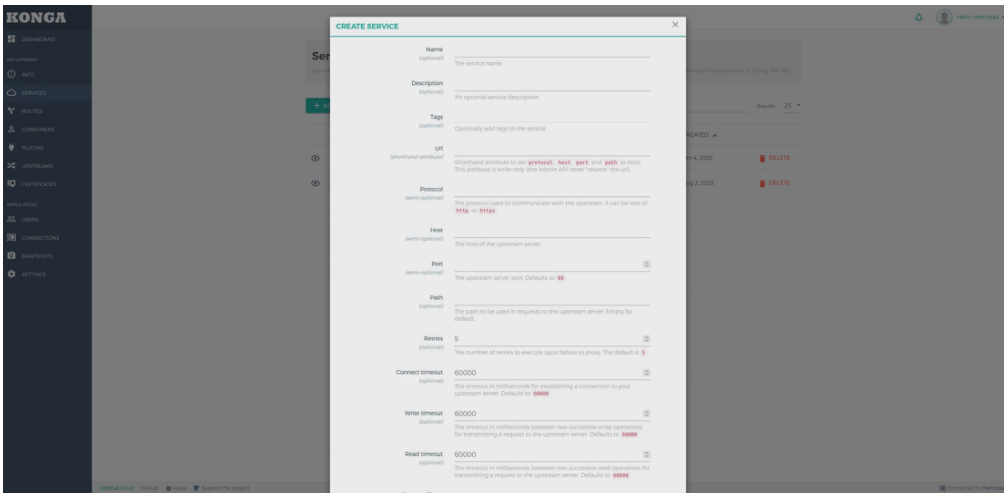
Services

Service entities, as the name implies, are abstractions of each of your own upstream services. Examples of Services would be a data transformation microservice, a billing API, etc.

NAME	HOST	TAGS	CREATED	
Metrics_Microservice	192.168.42.139	UCD	Nov 16, 2023	DELETE
Security_Diagnosis_Service	192.168.42.139	withSecure	Nov 10, 2023	DELETE
Medical_Analysis_Microservices	medical-analysis-service.apps.osc.fokus.fraunhofer.de	fokus	Nov 6, 2023	DELETE
Privacy_Microservices	192.168.42.139	telefonica	Nov 1, 2023	DELETE
Montimage_Microservices	192.168.42.139	montimage	Oct 6, 2023	DELETE
Fairness_Microservice	192.168.42.139	OAS3_import OAS3file_openapi.yaml	Oct 4, 2023	DELETE
XAI_Microservices	192.168.42.139	OAS3_import OAS3file_openapi.yaml	Aug 2, 2023	DELETE

Figure 21. The admin interface of the Kong API gateway.

Kong Gateway's administration interface or API (see Figure 22), where administrators or developers interacted with the Kong API to create and configure services.



CREATE SERVICE

Name (optional): The service name.

Description (optional): An optional service description.

Tags (optional): Optionally add tags to the service

Url (shorthand-attribute): Shorthand attribute to set `protocol`, `host`, `port` and `path` at once. This attribute is write-only (the Admin API never "returns" the url).

Protocol (semi-optional): The protocol used to communicate with the upstream. It can be one of `http` or `https`.

Host (semi-optional): The host of the upstream server.

Port (semi-optional): The upstream server port. Defaults to `80`.

Path (optional): The path to be used in requests to the upstream server. Empty by default.

Retries (optional): 5 — The number of retries to execute upon failure to proxy. The default is `5`.

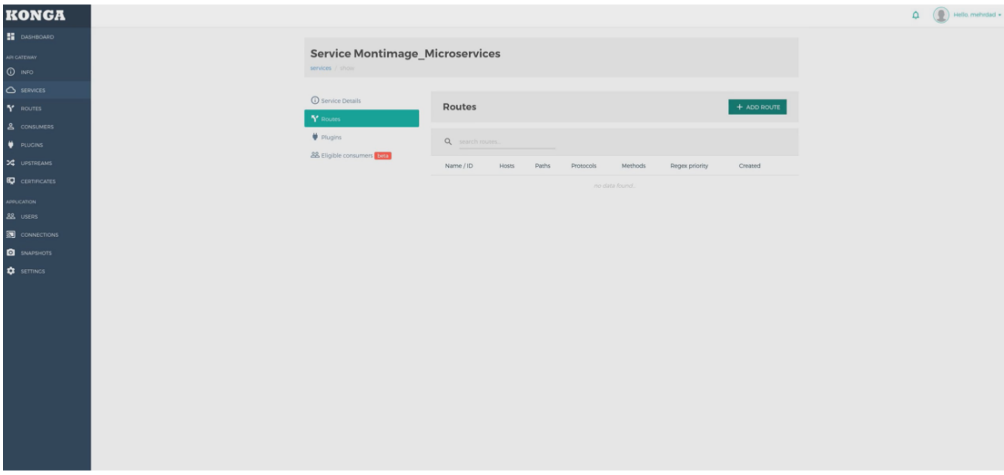
Connect timeout (optional): 60000 — The timeout in milliseconds for establishing a connection to your upstream server. Defaults to `60000`

Write timeout (optional): 60000 — The timeout in milliseconds between two successive write operations for transmitting a request to the upstream server. Defaults to `60000`

Read timeout (optional): 60000 — The timeout in milliseconds between two successive read operations for transmitting a request to the upstream server. Defaults to `60000`

Figure 22. Steps to define a new SPATIAL service in the Kong API gateway.

Route Configuration: Routes were configured by associating them with the previously defined services in Kong Gateway [108]. Basic route parameters such as name, paths, and references to existing services were set up during the service definition process. Advanced configurations, including protocols, hosts, methods, headers, and tags, were adjusted based on the specific requirements for routing requests (see Figure 23 and Figure 24).



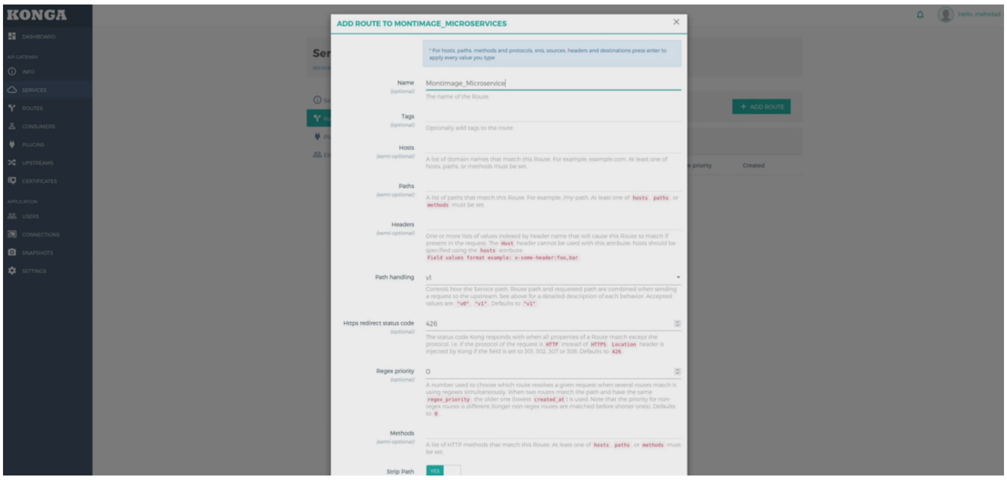
Service Montimage_Microservices

Routes

Name / ID	Hosts	Paths	Protocols	Methods	Regex priority	Created

no data found...

Figure 23. Interface to configure for an exemplary SPATIAL service.



ADD ROUTE TO MONTIMAGE_MICROSERVICES

* For hosts, paths, methods and protocols, snis, sources, headers and destinations press enter to apply every value you type

Name (optional): Montimage_Microservice
The name of the Route.

Tags (optional):
Optionally add tags to the route

Hosts (semi-optional):
A list of domain names that match this Route. For example: example.com. At least one of hosts, paths, or methods must be set.

Paths (semi-optional):
A list of paths that match this Route. For example: /my-path. At least one of `hosts`, `paths`, or `methods` must be set.

Headers (semi-optional):
One or more lists of values indexed by header name that will cause this Route to match if present in the request. The `Host` header cannot be used with this attribute: hosts should be specified using the `hosts` attribute.
`Field values format example: x-some-header:foo,bar`

Path handling: v1
Controls how the Service path, Route path and requested path are combined when sending a request to the upstream. See above for a detailed description of each behavior. Accepted values are: `"v0"`, `"v1"`. Defaults to `"v1"`.

Https redirect status code (optional): 426
The status code Kong responds with when all properties of a Route match except the protocol, i.e. if the protocol of the request is `HTTP` instead of `HTTPS`. `Location` header is injected by Kong if the field is set to 301, 302, 307 or 308. Defaults to `426`.

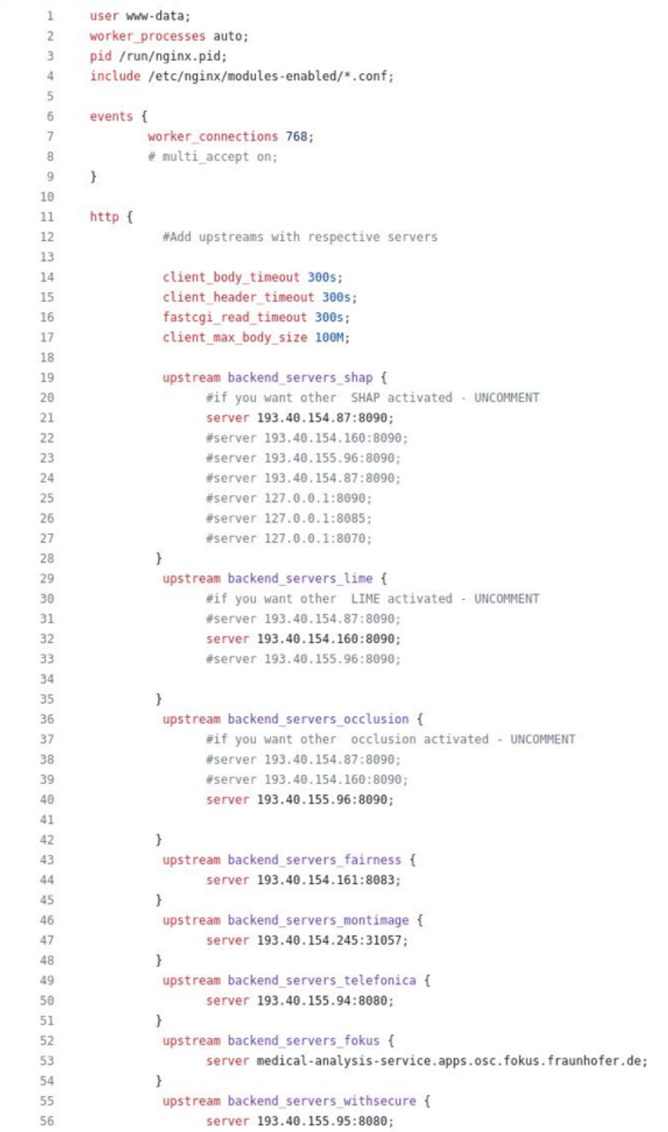
Regex priority (optional): 0
A number used to choose which route resolves a given request when several routes match it using regexes simultaneously. When two routes match the path and have the same `regex_priority`, the older one (lowest `created_at`) is used. Note that the priority for non-regex routes is different (longer non-regex routes are matched before shorter ones). Defaults to `0`.

Methods (semi-optional):
A list of HTTP methods that match this Route. At least one of `hosts`, `paths`, or `methods` must be set.

Strip Path: YES

Figure 24. Details for configuring a route for an exemplary SPATIAL service.

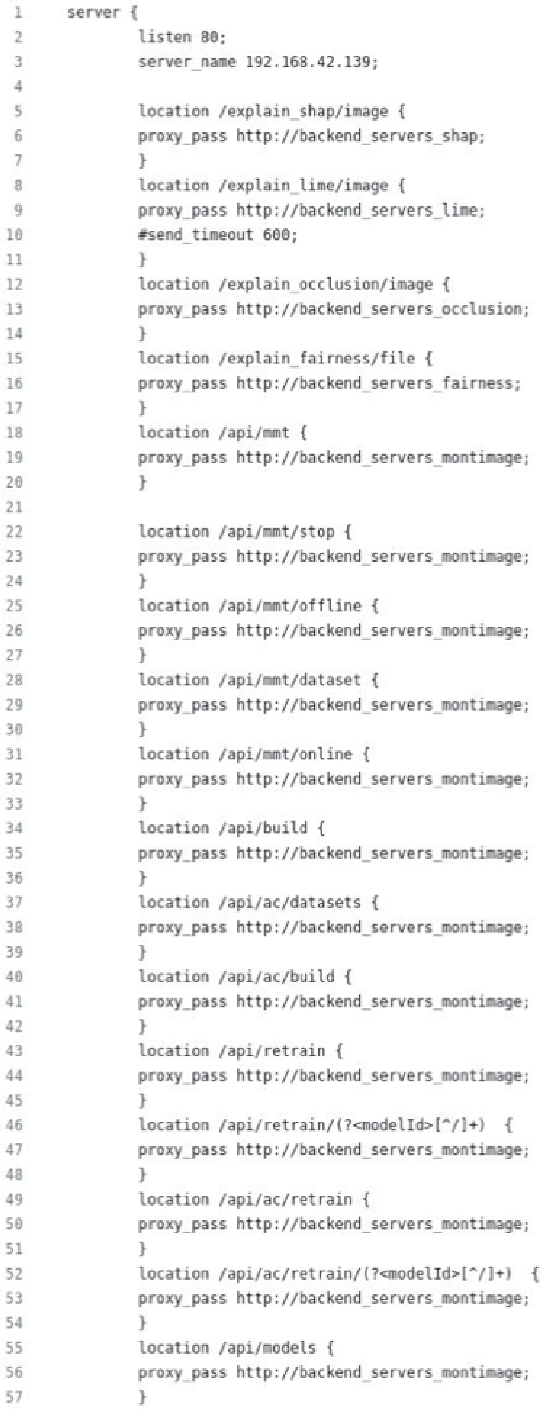
NGINX Upstream Configuration: Within the NGINX [109] configuration file, an upstream block was defined. This involved editing the NGINX configuration file, typically located in a directory like '/etc/nginx/conf.d/' or a similar location. Addresses of backend services were specified within the upstream block. These addresses represent where each service is running, and they could be cloud deployments or virtual machines. Figure 25 illustrates the NGINX configuration file for the current statue of the SPATIAL platform.



```
user www-data;
worker_processes auto;
pid /run/nginx.pid;
include /etc/nginx/modules-enabled/*.conf;

events {
        worker_connections 768;
        # multi_accept on;
}

http {
          #Add upstreams with respective servers

          client_body_timeout 300s;
          client_header_timeout 300s;
          fastcgi_read_timeout 300s;
          client_max_body_size 100M;

          upstream backend_servers_shap {
                #if you want other  SHAP activated - UNCOMMENT
                server 193.40.154.87:8090;
                #server 193.40.154.160:8090;
                #server 193.40.155.96:8090;
                #server 193.40.154.87:8090;
                #server 127.0.0.1:8090;
                #server 127.0.0.1:8085;
                #server 127.0.0.1:8070;
         }
          upstream backend_servers_lime {
                #if you want other  LIME activated - UNCOMMENT
                #server 193.40.154.87:8090;
                server 193.40.154.160:8090;
                #server 193.40.155.96:8090;

         }
          upstream backend_servers_occlusion {
                #if you want other  occlusion activated - UNCOMMENT
                #server 193.40.154.87:8090;
                #server 193.40.154.160:8090;
                server 193.40.155.96:8090;

         }
          upstream backend_servers_fairness {
                server 193.40.154.161:8083;
         }
          upstream backend_servers_montimage {
                server 193.40.154.245:31057;
         }
          upstream backend_servers_telefonica {
                server 193.40.155.94:8080;
         }
          upstream backend_servers_fokus {
                server medical-analysis-service.apps.osc.fokus.fraunhofer.de;
         }
          upstream backend_servers_withsecure {
                server 193.40.155.95:8080;
```

Figure 25. NGINX configuration file for the SPATIAL platform.

API Address Definition: API addresses were defined within the NGINX [109] configuration file (see Figure 26), often using location blocks. Location blocks were configured to establish rules for routing requests to specific API endpoints based on criteria such as Uniform Resource Locator (URL) paths. These configurations determined how incoming requests targeting particular URLs or Uniform Resource Identifier (URI) paths were routed to the corresponding backend services.



```
server {
        listen 80;
        server_name 192.168.42.139;

        location /explain_shap/image {
        proxy_pass http://backend_servers_shap;
        }
        location /explain_lime/image {
        proxy_pass http://backend_servers_lime;
        #send_timeout 600;
        }
        location /explain_occlusion/image {
        proxy_pass http://backend_servers_occlusion;
        }
        location /explain_fairness/file {
        proxy_pass http://backend_servers_fairness;
        }
        location /api/mmt {
        proxy_pass http://backend_servers_montimage;
        }

        location /api/mmt/stop {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/mmt/offline {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/mmt/dataset {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/mmt/online {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/build {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/ac/datasets {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/ac/build {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/retrain {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/retrain/(?<modelId>[^/]+)  {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/ac/retrain {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/ac/retrain/(?<modelId>[^/]+)  {
        proxy_pass http://backend_servers_montimage;
        }
        location /api/models {
        proxy_pass http://backend_servers_montimage;
        }
```

Figure 26. SPATIAL service IP address specified in the NGINX configuration file.

4 Experiments

This section describes the experiments designed to evaluate the Medical Analysis Module (MAM) integrated within the SPATIAL Platform. The initial experiment investigates the effectiveness of role-specific explanations integrated within the SPATIAL Platform to understand user needs and improve the interpretability of AI-generated explanations. Finally, the capacity-load performance of the module was assessed by analyzing the behaviour of the system under varying user loads to determine the scalability and reliability of the module within the SPATIAL platform.

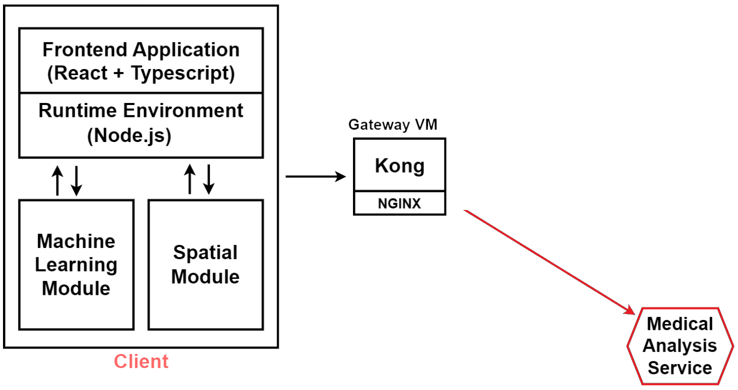


Figure 27. Medical Analysis Module.

4.1 Stakeholder-Adaptive Explanation

The medical analysis module (MAM) caters to medical experts and AI developers, two stakeholders considered within the SPATIAL project. The module analyzes electrocardiogram (ECG) signals from an e-calling medical system using XAI methods to enable the stakeholders to comprehend and interpret the underlying data effectively. The system leverages a fall detection model to trigger emergency medical assistance for aged patients when falling is detected in their daily activity. The role-specific explanations using LIME and SHAP explainable AI methods are mainly to aid medical experts who are monitoring emergency calls. This experiment evaluates the interpretability and relevance of AI-generated explanations for different user profiles using an interactive interface to improve the explanation to meet stakeholder's needs. The experiment was conducted by the Fraunhofer Institute for Open Communication Systems (FOKUS) as they maintain propriety access to the e-calling application. However, the translation of the findings to the development of an adaptive and interpretable dashboard and other design experimentation is the major experimental contribution of this work.

4.1.1 Experimental Setup

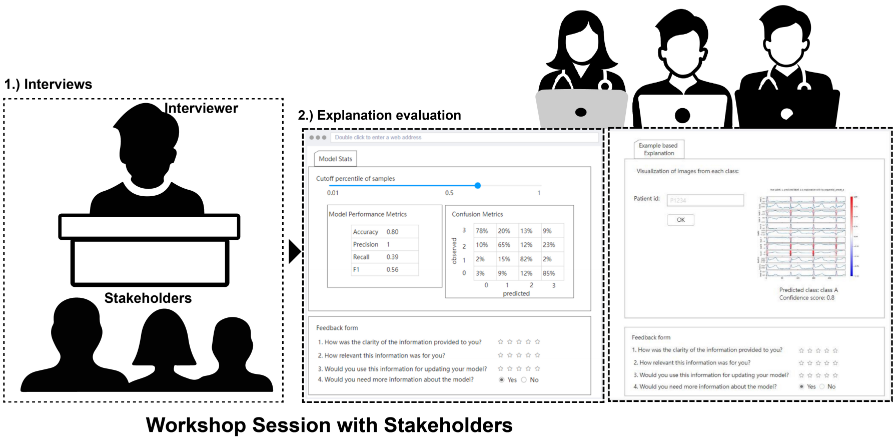
Experimental design and methodology: The investigators carefully considered the appropriate experimental method that could adequately achieve the objective of the investigation. Given that context, the perception of users of the MAM, and the adaptive explanations are relevant to the evaluation, a qualitative study was selected. Specifically, a participatory study was conducted. This study actively engaged the stakeholders in the design and assessment of an interactive interface for the emergency e-calling system. Designed mockups were presented to stakeholders at several workshops or mockup sessions and interviews to understand their needs and requirements.

E-Calling Medical Application: This is a mobile application, part of an e-calling system that utilizes accelerometer sensor data to detect the falling of an elderly person. As the falling event is detected by the underlying AI model in the application, the application triggers an emergency call to request medical assistance.

Apparatus: During the prototyping, Mockplus, a wireframe tool, was utilized for creating a static prototype of the interactive interface. The mockup included interactive elements such as buttons, text inputs, and tabs for information presentation. Likewise, the Python-based Explainer Dashboard library to develop a fully functional web app for deriving and presenting explanations generated about the ECG images to stakeholders.

Stakeholders: Two stakeholders were involved in the experiment. The AI developers are responsible for developing AI models and generating explanations.The second stakeholder, medical experts, makes medical decisions based on AI predictions and explanations.

Procedure: The experiment had three phases: (1) No Explanation, where users received only generic statistics about model performance; (2) Global Explanations, where data visualizations using dimensionality reduction techniques were provided; and (3) Local Explanations, where users interacted with specific data instances for detailed explanations. Participants, selected from AI developers and medical experts, engaged with these different levels of explanations, performing tasks and providing feedback on the clarity, relevance, and influence of the explanations. Quantitative metrics like interaction time and consistency of explanations were recorded, alongside qualitative feedback collected through surveys and interviews. Follow-up interviews offered deeper insights into user experiences and needs. This is illustrated in figure 28. The iterative design process incorporated user feedback to refine the interface, aiming to provide clear, relevant, and influential explanations, ultimately enhancing user confidence and decision-making accuracy.


1.) Interviews
Interviewer
Stakeholders
2.) Explanation evaluation
Workshop Session with Stakeholders

Double click to enter a web address

Model Stats

Cutoff percentile of samples: 0.01 — 0.5 — 1

Model Performance Metrics

Accuracy	0.80
Precision	1
Recall	0.39
F1	0.56

Confusion Metrics

observed \ predicted	0	1	2	3
3	78%	20%	13%	9%
2	10%	65%	12%	23%
1	2%	15%	82%	2%
0	3%	9%	12%	85%

Feedback form

1. How was the clarity of the information provided to you? ☆ ☆ ☆ ☆ ☆
2. How relevant this information was for you? ☆ ☆ ☆ ☆ ☆
3. Would you use this information for updating your model? ☆ ☆ ☆ ☆ ☆
4. Would you need more information about the model? ● Yes ○ No

Example based Explanation

Visualization of images from each class:

Patient id: P1234

OK

Predicted class: class A
Confidence score: 0.8

Feedback form

1. How was the clarity of the information provided to you? ☆ ☆ ☆ ☆ ☆
2. How relevant this information was for you? ☆ ☆ ☆ ☆ ☆
3. Would you use this information for updating your model? ☆ ☆ ☆ ☆ ☆
4. Would you need more information about the model? ● Yes ○ No

Figure 28. Interview Procedure and feedback

4.2 Capacity-load Performance Experiment

The development of the stakeholder-adaptive explanation laid the foundation for enhancing the transparency and trustworthiness of AI systems. Building on this, the performance evaluation of the medical analysis module (MAM) was conducted to ensure that the SPATIAL platform could maintain its robust and scalable performance while providing explanations under varying loads. The evaluation focuses on understanding how the system behaves under different user loads, which is critical for ensuring reliability and responsiveness in a production environment. The load tests allow the MAM to be analyzed in more detail and application-specific ways, allowing room to identify improvement areas that are only relevant to this service. The load tests are performed by sending requests against the MAM via the API Gateway link. This provides an estimate for utilizing MAM in operational settings. As a result, the system under test per isolated test consists of the MAM to be tested and the API Gateway. Figure 29 illustrates the abstract test setup for the MAM.

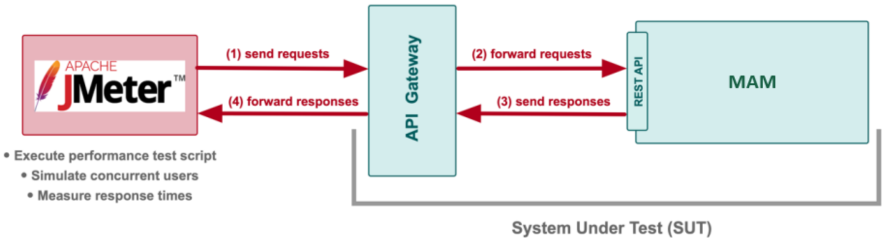


Figure 29. Abstract setup for MAM service integrated into the SPATIAL platform

4.2.1 Experimental Setup

Experimental design and methodology: The load testing was conducted using Apache JMeter[113], on a local machine setup. JMeter was particularly set up to simulate different levels of user load on the MAM system by generating and sending HTTP requests to the designated REST API endpoints. The configuration enabled scalable, repeatable tests across varying loads up to a maximum of 100 concurrent users.

The JMeter configuration for the capacity testing experiment was setup to replicate realistic user interactions with the MAM's REST APIs. It included Thread Groups to simulate concurrent users, each set up to send requests to predefined API endpoints. HTTP Request Samplers were created using specified HTTP methods, URLs, and body data to mimic regular user requests accurately. Listeners were used to tracking key metrics such as response times, throughput, and success/error rates in real-time, providing instant feedback on the system's performance under stress. However, for the purposes of this experiment, our primary focus remained on the response time metric. Timers were used to space out the requests, ensuring a realistic traffic flow while not overwhelming the server. (See Figure 30 for JMeter configuration)

REST APIs tested: Several REST API endpoints within the MAM were tested, including:

```
POST /emergency_detection/mi_detection/predict
```

This endpoint is used for the automatic emergency detection. It expects a 12-lead ECG signal encoded in WFDB as JSON input and returns 1) the probability that the ECG contains indications for an acute MI, 2) the predicted class, and 3) a 70oolean flag describing whether the situation is expected to be an emergency. For the emergency detection, a pre-defined default model is used.

```
POST /medical_analysis/ecg_analysis/visualize_ecg
```

This endpoint can be utilised to visualise a 12-lead ECG signal. Similar to other endpoints of the MAM, it takes a 12-lead ECG signal encoded in WFDB as JSON input. Afterwards, a PNG image of the plotted signal will be provided.

```
POST /model/{model_id}/explain/{xai_method}
```

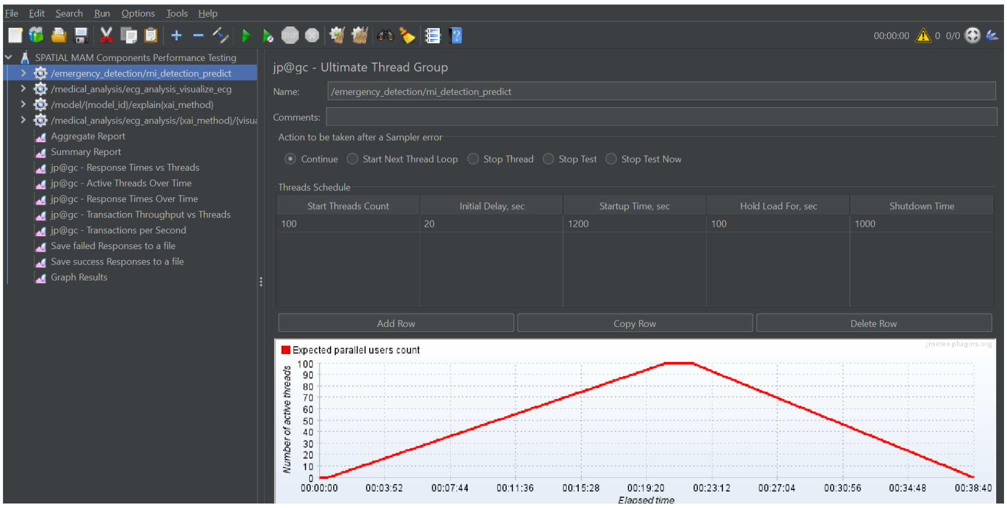
In order to generate numeric explanations for the decision-making of hosted ML models, this endpoint can be used. First of all, the endpoint requires 12-lead ECG signal encoded in WFDB as JSON input. Furthermore, the user needs to specify the ID of the model to use as well as the XAI method to apply in the request path parameters.

```
POST /medical_analysis/ecg_analysis/{xai_method}/{
    visualization_approach}
```

This endpoint can generate visual explanations for ML models performing ECG analysis. The user needs to specify which XAI method to be selected. Moreover, the user can choose between three visualisation methods by selecting the path parameter visualization Approach accordingly. The following values are currently supported:

1) ‘tick importance’, which highlights the relevant time ticks in the ECG signal,
2) ‘time importance’, which marks relevant aggregated time segments, and
3) ‘lead importance’, which highlights the most important relevant ECG lead.

Finally, the endpoint also requires a 12-lead ECG signal encoded in WFDB as JSON input. The endpoint will then return a PNG Image visualising the relevances as a heatmap overlaying the original ECG signal.


File
Edit
Search
Run
Options
Tools
Help
00:00:00
0
0/0
SPATIAL MAM Components Performance Testing
/emergency_detection/mi_detection_predict
/medical_analysis/ecg_analysis_visualize_ecg
/model/{model_id}/explain{xai_method}
/medical_analysis/ecg_analysis/{xai_method}/{visua
Aggregate Report
Summary Report
jp@gc - Response Times vs Threads
jp@gc - Active Threads Over Time
jp@gc - Response Times Over Time
jp@gc - Transaction Throughput vs Threads
jp@gc - Transactions per Second
Save failed Responses to a file
Save success Responses to a file
Graph Results
Add Row
Copy Row
Delete Row
Expected parallel users count
jmeter-plugins.org
Number of active threads
100
90
80
70
60
50
40
30
20
10
0
00:00:00
00:03:52
00:07:44
00:11:36
00:15:28
00:19:20
00:23:12
00:27:04
00:30:56
00:34:48
00:38:40
Elapsed time

jp@gc - Ultimate Thread Group

Name: /emergency_detection/mi_detection_predict

Comments:

Action to be taken after a Sampler error

● Continue ○ Start Next Thread Loop ○ Stop Thread ○ Stop Test ○ Stop Test Now

Threads Schedule

Start Threads Count	Initial Delay, sec	Startup Time, sec	Hold Load For, sec	Shutdown Time
100	20	1200	100	1000

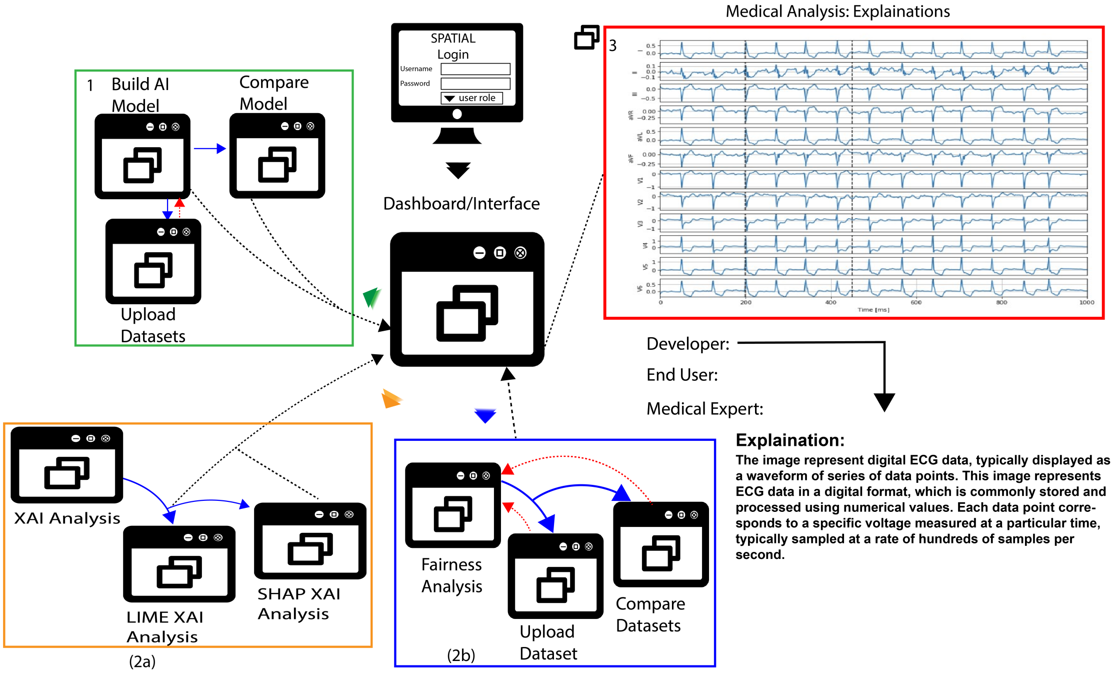
Figure 30. JMeter configuration for the load testing.

5 Results

5.1 Stakeholder-Adaptive Explanation

The experiment on the explanations derived from the Medical Analysis Module demonstrated that the interactive interface effectively provided clear, relevant, and influential explanations tailored to the needs of AI developers and medical experts. Feedback examination revealed pertinent design considerations that were incorporated into the development of the platform's MAM explanation dashboard .

Figure 31 illustrates the SPATIAL platform's dashboard for medical analysis, focusing on ECG data. It outlines the process from dataset upload, AI model building, and comparison to XAI analysis using LIME and SHAP and fairness assessment. The central dashboard integrates these functions, providing a unified interface for different user roles, including developers, end-users, and medical experts. The medical analysis section displays ECG data with detailed explanations to enhance interpretability. This setup ensures that users can effectively understand and utilize complex medical data, improving the overall usability and transparency of the AI system.



Explaination:

The image represent digital ECG data, typically displayed as a waveform of series of data points. This image represents ECG data in a digital format, which is commonly stored and processed using numerical values. Each data point corresponds to a specific voltage measured at a particular time, typically sampled at a rate of hundreds of samples per second.

Figure 31. Explanations based on user role.

Implicationn of result: Stakeholders tend to find the explanations on the dashboard to be generally clear and relevant, which positively influences their confidence and decision-making. Interaction time varied with explanation complexity, indicating the cognitive effort required. Quantitative metrics showed consistency across different XAI methods, while qualitative feedback highlighted areas for further improvement. Overall, the iterative design and user feedback significantly enhanced the usability and interpretability of the emergency e-calling system's interactive interface. (See Appendix I for other explanation figures)

5.2 Performance Testing Results

We assessed the performance of the MAM by testing it with up to 100 concurrent demanding users. As previously mentioned, the performance was evaluated based on the response time. Since the MAM operates on an HTTP-based REST API, this statistic primarily provides external users with relevant information about the anticipated performance of the component in a real-world operational environment.

The load test was executed for a duration of 7.9 hours. The load on the service was gradually increased by executing and managing concurrent threads for different endpoints. A user load of up to 100 concurrent users was simulated. In addition, a grand total of 18267 samples were utilized for the MAM service in the studies, out of which 17475 were successful. The average response time for the analyzed API endpoint is displayed in Figure 32. In addition, Figure 32 depicts the 95 percent confidence interval of the observed measurements. Furthermore, it is essential to mention that only the succeeded requests are considered.

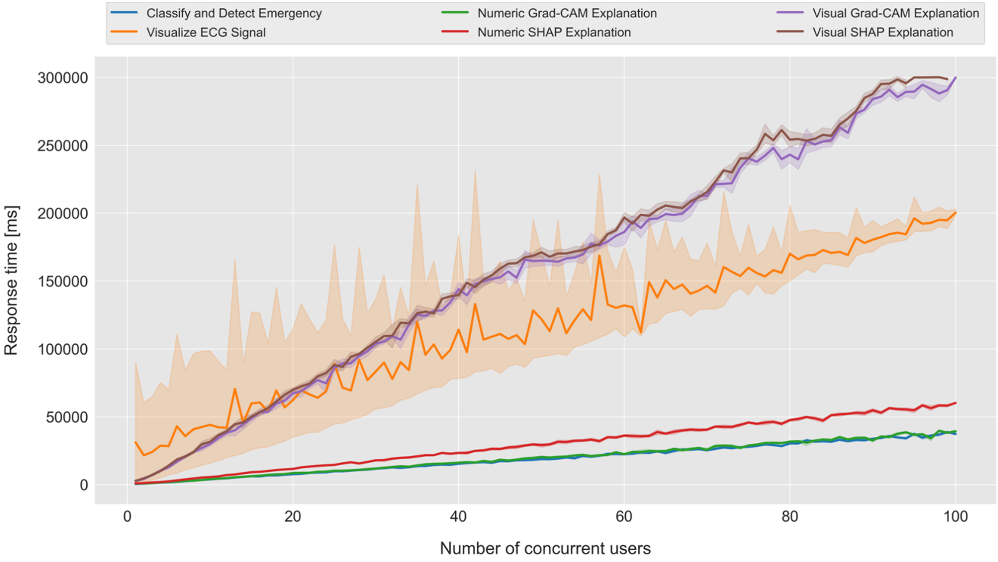


Figure 32. JMeter load testing results.

As we can see, the response time of the service endpoints increases significantly with increasing user load. Although processing requests for high user loads takes more time, the MAM can still reliably answer requests for up to 100 parallel users. As shown in Figure 32, the endpoints that only calculate and return numerical values are significantly more performant than those that plot and return ECG signals or XAI explanations. Precisely, the endpoints "Classify and Detect Emergency", "Numeric Grad-Cam Explanation", and "Numeric SHAP Explanation" are among those that return numerical values. All other endpoints generate and respond with PNG images.

In addition, we can observe that the functions "Classify and Detect Emergency" and "Numeric Grad-Cam Explanation" are the most efficient. These manifest a minimum response time of 512ms and 651ms, and a maximum response time of 45206ms and 46752ms, respectively. Interestingly, the calculation of the Gradient-weighted Class Activation Mapping (GradCAM) values performs similarly to the pure inference of the hosted TensorFlow models, indicating the Grad-CAM method's computational efficiency. In contrast, the calculation of the SHAP values for the "Numeric SHAP Explanation" function is significantly more computationally intensive, resulting in a higher response time. Here, the minimum response time was measured at 851ms, and the maximum value was at 67990ms for generating explanations for ECG classifications.

The functions "Visual SHAP Explanation" and "Visual Grad-CAM Explanation" perform the same calculations as those just discussed for calculating the numerical relevance based on the XAI methods SHAP and Grad-CAM. However, these functions also plot these values and overlay them on the original ECG as a heat map. As we can observe from Figure 32, the computational and transmission overhead for creating and sending the images is significant. A minimum response time of 2557ms and a maximum response time of 300867ms were measured for "Visual Grad-CAM Explanation". In the case of "Visual SHAP Explanation", values of 2657ms and 300745ms were observed. This indicates that plotting is significantly more computationally expensive. In this context, the "Visualise ECG Signal" function provides an estimate of the pure effort required for plotting, where a minimum response time of 1669ms and a maximum value of 242872ms is manifested. The increased overhead for creating the images and the increased amount of data required for transferring the images suggest using the numeric functions and moving the plotting to the clients for time-critical applications.

6 Discussion

6.1 Stakeholder-Adaptive Explanation

The experiment conducted on stakeholder-adaptive explanations within the Medical Analysis Module (MAM) demonstrated the effectiveness of an interactive interface in addressing the specific requirements of AI developers and medical experts. It enhanced clarity, relevance, and influence by tailoring explanations based on user roles and preferences. Incorporating iterative feedback from stakeholders proved invaluable in refining the explanation dashboard, ensuring that it met the evolving needs of end-users. The enhanced interpretability of AI-driven medical analysis not only improves user confidence and accuracy in decision-making but also facilitates transparency, which is essential for trust-building in sensitive healthcare applications.

The iterative design process employed in the experiment highlights the necessity of stakeholder collaboration in creating effective AI interfaces. The participatory study, which included workshops, mockup sessions, and interviews, provided valuable insights into requirements and preferences of the users. This feedback was helpful in improving the explanation dashboard, making it more relevant and user-friendly. The distinction between global and local explanations addressed the stakeholders' various cognitive needs, resulting in a deeper understanding of the AI models' predictions.

Furthermore, the results revealed that explanations not only improved the trustworthiness of AI systems but also facilitated better decision-making by providing clear and actionable insights. This relates with existing literature, emphasizing the importance of explainability in building trust in AI applications [93].

6.2 Performance Testing

The performance evaluation of the Medical Analysis Module (MAM) within the SPATIAL platform provides critical insights into its operational efficiency and scalability under varying loads. The results showed that the platform could handle a significant load without laying down performance, ensuring reliable and efficient operation even in high-demand situations. This is particularly important in real-world applications where AI systems need to process enormous amounts of data and provide prompt responses.

The experimental configuration simulated up to 100 concurrent users interacting with MAM's REST API endpoints. Endpoints tested included emergency detection, ECG visualization, and the generation of numerical and visual explanations using SHAP and Grad-CAM techniques. The findings revealed significant differences in performance depending on the type of request. Endpoints that return numerical results, such as "Classify and Detect Emergency" and "Numeric Grad-CAM Explanation," had much shorter response times than those that generated and returned visual explanations, such as "Visual SHAP Explanation" and "Visual Grad-CAM Explanation." This insight is crucial because it highlights the importance of balancing computational efficiency and the depth of analysis provided. For time-sensitive applications, offloading plotting tasks to client-side process could reduce response time issues, ensuring that the MAM remains responsive even under high load conditions.

7 Conclusion

In this paper, we set out to systematically evaluate the augmentation of trustworthy AI in modern applications through the development of the SPATIAL architecture. This study aimed to fill the void in the practical implementation of trustworthy AI requirements regarding continuous human oversight throughout AI development, by methodically assessing an adaptive dashboard of an AI platform intended for trustworthy AI development within the framework of the EU SPATIAL initiative. SPATIAL, we conclude, diagnoses the functionality of artificial intelligence by integrating various techniques that characterize and quantify the inference process of AI. By utilizing an interactive interface to assess the interpretability and relevance of AI-generated explanations for various user profiles, this study enhanced the explanations to better cater to the stakeholders' requirements. With regard to the reliable explainability metric, we improved the user experience by tailoring explanations for stakeholders concerning Medical Analysis Module (MAM) to their respective user roles. The stakeholder-adaptive explanation was considered satisfactory due to the integration of user feedback into interface refinement. The SPATIAL architecture was able to deliver explanations that were unambiguous, pertinent, and persuasive, ultimately resulting in increased user confidence and improved decision-making accuracy for the MAM.

Moreover, we effectively showcased the platform's efficacy in delivering explanations tailored to stakeholders' needs while ensuring the dependable operation of AI systems. The results and findings are expected to provide valuable insights for compliance, specifically in the realm of enhancing monitoring and oversight of AI inferences. They will effectively demonstrate how the requirement can be operationalized into practical tools that can be utilized by end-users and regulatory authorities.

References

[1] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Comput. Surv.*, 55(9), jan 2023.

[2] J. Zhang and Z.-M. Zhang. Ethics and governance of trustworthy medical artificial intelligence. *BMC Medical Informatics and Decision Making*, 23(1), 2023. Publisher: BioMed Central Ltd.

[3] P. Giudici, M. Centurelli, and S. Turchetta. Artificial Intelligence risk measurement. *Expert Systems with Applications*, 235, 2024. Publisher: Elsevier Ltd.

[4] D. Baneres, A.E. Guerrero-Roldán, M.E. Rodríguez-González, and A. Karadeniz. A predictive analytics infrastructure to support a trustworthy early warning system. *Applied Sciences (Switzerland)*, 11(13), 2021. Publisher: MDPI AG.

[5] Y. Wang, H. Zen, M.F.M. Sabri, X. Wang, and L.C. Kho. Towards Strengthening the Resilience of IoV Networks—A Trust Management Perspective. *Future Internet*, 14(7), 2022. Publisher: MDPI.

[6] Marco Almada. Governing the black box of artificial intelligence. November 7 2023.

[7] Eu artificial intelligence act. `https://artificialintelligenceact.eu/`, 2024. Accessed: May 10, 2024.

[8] The White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, 2023.

[9] Innovation, Science and Economic Development Canada. Artificial intelligence and data act (aida), 2022.

[10] Detection mechanisms to identify data biases and exploratory studies about different data quality trade-offs for ai-based systems, 2020. Accessed: date-of-access.

[11] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: A review. *ACM Comput. Surv.*, 55(2), jan 2022.

[12] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023.

[13] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The prisma 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, 88:105906, 2021.

[14] Daniel Atherton, Reva Schwartz, Peter Fontana, and Patrick Hall. The language of trustworthy ai: An in-depth glossary of terms. OTHER, 2023. Accessed May 10, 2024.

[15] Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies, 2020.

[16] J.E. Paulsen. Ai, trustworthiness, and the digital dirty Harry problem. *Nordic Journal of Studies in Policing*, 8(2), 2020. Publisher: Universitetsforlaget AS.

[17] U. Upadhyay, A. Gradisek, U. Iqbal, E. Dhar, Y.-C. Li, and S. Syed-Abdul. Call for the responsible artificial intelligence in the healthcare. *BMJ Health and Care Informatics*, 30(1), 2023. Publisher: BMJ Publishing Group.

[18] M. Rizinski, H. Peshov, K. Mishev, L.T. Chitkushev, I. Vodenska, and D. Trajanov. Ethically Responsible Machine Learning in Fintech. *IEEE Access*, 10:97531–97554, 2022. Publisher: Institute of Electrical and Electronics Engineers Inc.

[19] G. Rjoub, J. Bentahar, O. Abdel Wahab, R. Mizouni, A. Song, R. Cohen, H. Otrok, and A. Mourad. A Survey on Explainable Artificial Intelligence for Cybersecurity. *IEEE Transactions on Network and Service Management*, 20(4):5115–5140, 2023. Publisher: Institute of Electrical and Electronics Engineers Inc.

[20] R. Seidel, K. Schmidt, N. Thielen, and J. Franke. Trustworthiness of machine learning models in manufacturing applications using the example of electronics manufacturing processes. In Valente A., Carpanzano E., and Boer C., editors, *Procedia CIRP*, volume 107, pages 487–492. Elsevier B.V., 2022. Journal Abbreviation: Procedia CIRP.

[21] J. Matulis and R. McCoy. Relief in Sight? Chatbots, In-baskets, and the Overwhelmed Primary Care Clinician. *Journal of General Internal Medicine*, 38(12):2808–2815, 2023. Publisher: Springer.

[22] W. Fan. Data quality: From theory to practice. *ACM SIGMOD Record*, 44(3):7–18, 2015.

[23] Z. Jia et al. Exploring hidden dimensions in parallelizing convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 2279–2288, 2018.

[24] A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27:1413–1432, 2017.

[25] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99:101805, 2023.

[26] G. Stettinger, P. Weissensteiner, and S. Khastgir. Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. *IEEE Access*, 12:22718–22745, 2024. Publisher: Institute of Electrical and Electronics Engineers Inc.

[27] E. Hohma and C. Lütge. From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems. *AI (Switzerland)*, 4(4):904–925, 2023. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[28] L. Xiuquan and J. Shaoru. On the Needs of Artificial Intelligence Technical Regulation in the Man-machine Symbiosis Society. In *IFAC-PapersOnLine*, volume 53, pages 491–494. Elsevier B.V., 2020. Issue: 5 Journal Abbreviation: IFAC-PapersOnLine.

[29] N.R. Pal. In Search of Trustworthy and Transparent Intelligent Systems With Human-Like Cognitive and Reasoning Capabilities. *Frontiers in Robotics and AI*, 7, 2020. Publisher: Frontiers Media S.A.

[30] A. Berman, K. de Fine Licht, and V. Carlsson. Trustworthy AI in the public sector: An empirical analysis of a Swedish labor market decision-support system. *Technology in Society*, 76, 2024. Publisher: Elsevier Ltd.

[31] A. Schmitz, M. Akila, D. Hecker, M. Poretschkin, and S. Wrobel. Vertrauenswürdige KI - Warum und Wie? Ein Ansatz zur systematischen Qualitätssicherung beim Einsatz von ML Komponenten. *At-Automatisierungstechnik*, 70(9):793–804, 2022. Publisher: De Gruyter Oldenbourg.

[32] S.Z. El Mestari, G. Lenzini, and H. Demirci. Preserving data privacy in machine learning systems. *Computers and Security*, 137, 2024. Publisher: Elsevier Ltd.

[33] M. Gerlich. Perceptions and Acceptance of Artificial Intelligence: A Multi-Dimensional Study. *Social Sciences*, 12(9), 2023. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[34] R.L. Stroup, K.R. Niewoehner, R.D. Apaza, D. Mielke, and N. Maurer. Application of AI in the NAS - The Rationale for AI-Enhanced Airspace Management. In *AIAA IEEE Dig Avionics Syst Conf Proc*, volume 2019-September. Institute of Electrical and Electronics Engineers Inc., 2019. Journal Abbreviation: AIAA IEEE Dig Avionics Syst Conf Proc.

[35] M.A. Onari, I. Grau, M.S. Nobile, and Y. Zhang. Trustworthy Artificial Intelligence in Medical Applications: A Mini Survey. In *CIBCB - IEEE Conf. Comput. Intell. Bioinform. Comput. Biol.* Institute of Electrical and Electronics Engineers Inc., 2023. Journal Abbreviation: CIBCB - IEEE Conf. Comput. Intell. Bioinform. Comput. Biol.

[36] L.F. De Arizón, E.R. Viera, M. Pilco, A. Perera, G. De Maeztu, A. Nicolau, M. Furlano, and R. Torra. Artificial intelligence: a new field of knowledge for nephrologists. *Clinical Kidney Journal*, 16(12):2314–2326, 2023. Publisher: Oxford University Press.

[37] M.R. Pinsky, A. Bedoya, A. Bihorac, L. Celi, M. Churpek, N.J. Economou-Zavlanos, P. Elbers, S. Saria, V. Liu, P.G. Lyons, B. Shickel, P. Toral, D. Tscholl, and G. Clermont. Use of artificial intelligence in critical care: opportunities and obstacles. *Critical Care*, 28(1), 2024. Publisher: BioMed Central Ltd.

[38] Z. Li, W. Fang, C. Zhu, Z. Gao, and W. Zhang. AI-Enabled Trust in Distributed Networks. *IEEE Access*, 11:88116–88134, 2023. Publisher: Institute of Electrical and Electronics Engineers Inc.

[39] A. Ferrario, M. Loi, and E. Viganò. In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philosophy and Technology*, 33(3):523–539, 2020. Publisher: Springer.

[40] F.R. Da Silva Oliveira and F.B. De Lima Neto. Method to Produce More Reasonable Candidate Solutions With Explanations in Intelligent Decision Support Systems. *IEEE Access*, 11:20861–20876, 2023. Publisher: Institute of Electrical and Electronics Engineers Inc.

[41] C. Chandler, P.W. Foltz, and B. Elvevåg. Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness. *Schizophrenia Bulletin*, 46(1):11–14, 2020. Publisher: Oxford University Press.

[42] W.T. Hutiri and A. Yi Ding. Towards Trustworthy Edge Intelligence: Insights from Voice-Activated Services. In Ardagna C.A., Bian H., Chang C.K., Chang R.N., Damiani E., Dustdar S., Marco J., Singh M., Teniente E., Ward R., Wang Z., Xhafa F., and Zhang J., editors, *Proc. - IEEE Int. Conf. Serv. Comput., SCC*, pages 239–248. Institute of Electrical and Electronics Engineers Inc., 2022. Journal Abbreviation: Proc. - IEEE Int. Conf. Serv. Comput., SCC.

[43] M.A. Butt, A. Qayyum, H. Ali, A. Al-Fuqaha, and J. Qadir. Towards secure private and trustworthy human-centric embedded machine learning: An emotion-aware facial recognition case study. *Computers and Security*, 125, 2023. Publisher: Elsevier Ltd.

[44] D. Yu, Z. Xie, Y. Yuan, S. Chen, J. Qiao, Y. Wang, Y. Yu, Y. Zou, and X. Zhang. Trustworthy decentralized collaborative learning for edge intelligence: A survey. *High-Confidence Computing*, 3(3), 2023. Publisher: Shandong University.

[45] M.A. Mohammed, M. Boujelben, and M. Abid. A Novel Approach for Fraud Detection in Blockchain-Based Healthcare Networks Using Machine Learning. *Future Internet*, 15(8), 2023. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[46] N.T. Cam and V.T. Kiet. FlwrBC: Incentive Mechanism Design for Federated Learning by Using Blockchain. *IEEE Access*, 11:107855–107866, 2023. Publisher: Institute of Electrical and Electronics Engineers Inc.

[47] L. Fidon, M. Aertsen, F. Kofler, A. Bink, A.L. David, T. Deprest, D. Emam, F. Guffens, A. Jakab, G. Kasprian, P. Kienast, A. Melbourne, B. Menze, N. Mufti, I. Pogledic, D. Prayer, M. Stuempflen, E. Van Elslander, S. Ourselin, J. Deprest, and T. Vercauteren. A Dempster-Shafer Approach to Trustworthy AI With Application to Fetal Brain MRI Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3784–3795, 2024. Publisher: IEEE Computer Society.

[48] L. Wu, J. Xu, S. Thakkar, M. Gray, Y. Qu, D. Li, and W. Tong. A framework enabling LLMs into regulatory environment for transparency and trustworthiness and its application to drug labeling document. *Regulatory Toxicology and Pharmacology*, 149, 2024. Publisher: Academic Press Inc.

[49] P. Giudici and E. Raffinetti. SAFE Artificial Intelligence in finance. *Finance Research Letters*, 56, 2023. Publisher: Elsevier Ltd.

[50] M. Socha, W. Prażuch, A. Suwalska, P. Foszner, J. Tobiasz, J. Jaroszewicz, K. Gruszczynska, M. Sliwinska, M. Nowak, B. Gizycka, G. Zapolska, T. Popiela,

G. Przybylski, P. Fiedor, M. Pawlowska, R. Flisiak, K. Simon, J. Walecki, A. Cieszanowski, E. Szurowska, M. Marczyk, and J. Polanska. Pathological changes or technical artefacts? The problem of the heterogenous databases in COVID-19 CXR image analysis. *Computer Methods and Programs in Biomedicine*, 240, 2023. Publisher: Elsevier Ireland Ltd.

[51] M.H. Wang, K.K.-L. Chong, Z. Lin, X. Yu, and Y. Pan. An Explainable Artificial Intelligence-Based Robustness Optimization Approach for Age-Related Macular Degeneration Detection Based on Medical IOT Systems. *Electronics (Switzerland)*, 12(12), 2023. Publisher: MDPI.

[52] F. Schwendicke, W. Samek, and J. Krois. Artificial Intelligence in Dentistry: Chances and Challenges. *Journal of Dental Research*, 99(7):769–774, 2020. Publisher: SAGE Publications Inc.

[53] T. Shyamalee, D. Meedeniya, G. Lim, and M. Karunarathne. Automated Tool Support for Glaucoma Identification With Explainability Using Fundus Images. *IEEE Access*, 12:17290–17307, 2024. Publisher: Institute of Electrical and Electronics Engineers Inc.

[54] J. Feng, D. Wang, and Z. Gu. Bidirectional Flow Decision Tree for Reliable Remote Sensing Image Scene Classification. *Remote Sensing*, 14(16), 2022. Publisher: MDPI.

[55] E. Shoemaker, H. Malik, H. Narman, and J. Chaudri. Explaining the Unseen: Leveraging XAI to Enhance the Trustworthiness of Black-Box Models in Performance Testing. In Shakshuki E., editor, *Procedia Comput. Sci.*, volume 224, pages 83–90. Elsevier B.V., 2023. Journal Abbreviation: Procedia Comput. Sci.

[56] C. Dindorf, W. Teufl, B. Taetz, G. Bleser, and M. Fröhlich. Interpretability of input representations for gait classification in patients after total hip arthroplasty. *Sensors (Switzerland)*, 20(16):1–14, 2020. Publisher: MDPI AG.

[57] E.S. Maddy and S.A. Boukabara. MIIDAPS-AI: An Explainable Machine-Learning Algorithm for Infrared and Microwave Remote Sensing and Data Assimilation Preprocessing - Application to LEO and GEO Sensors. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8566–8576, 2021. Publisher: Institute of Electrical and Electronics Engineers Inc.

[58] K. Ma, S. He, G. Sinha, A. Ebadi, A. Florea, S. Tremblay, A. Wong, and P. Xi. Towards Building a Trustworthy Deep Learning Framework for Medical Image Analysis. *Sensors*, 23(19), 2023. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[59] T. Qamar and N.Z. Bawany. Understanding the black-box: towards interpretable and reliable deep learning models. *PeerJ Computer Science*, 9, 2023. Publisher: PeerJ Inc.

[60] K.E. Brown and D.A. Talbert. Using Explainable AI to Measure Feature Contribution to Uncertainty. In Bartak R., Franklin M., and Keshtkar F., editors, *Proc. Int. Fla. Artif. Intell. Res. Soc. Conf., FLAIRS*, volume 35. Florida Online Journals, University of Florida, 2022. Journal Abbreviation: Proc. Int. Fla. Artif. Intell. Res. Soc. Conf., FLAIRS.

[61] M. Kwiatkowska and X. Zhang. When to Trust AI: Advances and Challenges for Certification of Neural Networks. In Ganzha M., Maciaszek L., Maciaszek L., Paprzycki M., Slezak D., Slezak D., and Slezak D., editors, *Proc. Conf. Comput. Sci. Intell. Syst., FedCSIS*, pages 25–37. Institute of Electrical and Electronics Engineers Inc., 2023. Journal Abbreviation: Proc. Conf. Comput. Sci. Intell. Syst., FedCSIS.

[62] D. Dealcala, I. Serna, A. Morales, J. Fierrez, and J. Ortega-Garcia. Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma. In Shahriar H., Teranishi Y., Cuzzocrea A., Sharmin M., Towey D., Majumder AKM.J.A., Kashiwazaki H., Yang J.-J., Takemoto M., Sakib N., Banno R., and Ahamed S.I., editors, *Proc Int Comput Software Appl Conf*, volume 2023-June, pages 1167–1172. IEEE Computer Society, 2023. Journal Abbreviation: Proc Int Comput Software Appl Conf.

[63] V. Moskalenko and V. Kharchenko. Resilience-aware MLOps for AI-based medical diagnostic system. *Frontiers in Public Health*, 12, 2024. Publisher: Frontiers Media SA.

[64] Y. Choi, M.W. Wahi-Anwar, and M.S. Brown. SimpleMind: An open-source software environment that adds thinking to deep neural networks. *PLoS ONE*, 18(4 April), 2023. Publisher: Public Library of Science.

[65] D. Franco, L. Oneto, N. Navarin, and D. Anguita. Toward learning trustworthily from data combining privacy, fairness, and explainability: An application to face recognition. *Entropy*, 23(8), 2021. Publisher: MDPI AG.

[66] N. Khan, S. Coleri, A. Abdallah, A. Celik, and A.M. Eltawil. Explainable and Robust Artificial Intelligence for Trustworthy Resource Management in 6G Networks. *IEEE Communications Magazine*, 62(4):50–56, 2024. Publisher: Institute of Electrical and Electronics Engineers Inc.

[67] I. Guarino, G. Aceto, D. Ciuonzo, A. Montieri, V. Persico, and A. Pescape. Explainable Deep-Learning Approaches for Packet-Level Traffic Prediction of

Collaboration and Communication Mobile Apps. *IEEE Open Journal of the Communications Society*, 5:1299–1324, 2024. Publisher: Institute of Electrical and Electronics Engineers Inc.

[68] N. Herwig and P. Borghesani. Explaining deep neural networks processing raw diagnostic signals. *Mechanical Systems and Signal Processing*, 200, 2023. Publisher: Academic Press.

[69] R. Luo, J. Xing, L. Chen, Z. Pan, X. Cai, Z. Li, J. Wang, and A. Ford. Glassboxing deep learning to enhance aircraft detection from sar imagery. *Remote Sensing*, 13(18), 2021. Publisher: MDPI.

[70] Y. Wang, R. Xiao, X. Wang, and A. Liu. Constructing autonomous, personalized, and private working management of smart home products based on deep reinforcement learning. In Liu A. and Kara S., editors, *Procedia CIRP*, volume 119, pages 72–77. Elsevier B.V., 2023. Journal Abbreviation: Procedia CIRP.

[71] A. Nechi, A. Mahmoudi, C. Herold, D. Widmer, T. Kurner, M. Berekovic, and S. Mulhem. Practical Trustworthiness Model for DNN in Dedicated 6G Application. In *Int. Conf. Wirel. Mob. Comput. Netw. Commun.*, volume 2023-June, pages 312–317. IEEE Computer Society, 2023. Journal Abbreviation: Int. Conf. Wirel. Mob. Comput. Netw. Commun.

[72] D.P. Swamy, C. Berghoff, V. Danos, F. Langer, T. Markert, G. Schneider, A. von Twickel, and F. Woitschek. Towards Audit Requirements for AI-Based Systems in Mobility Applications. In Mori P., Lenzini G., and Furnell S., editors, *Int. Conf. Inf. Syst. Secur. Priv.*, pages 339–348. Science and Technology Publications, Lda, 2023. Journal Abbreviation: Int. Conf. Inf. Syst. Secur. Priv.

[73] M.C.A. Clare, M. Sonnewald, R. Lguensat, J. Deshayes, and V. Balaji. Explainable Artificial Intelligence for Bayesian Neural Networks: Toward Trustworthy Predictions of Ocean Dynamics. *Journal of Advances in Modeling Earth Systems*, 14(11), 2022. Publisher: John Wiley and Sons Inc.

[74] K. Rrmoku, B. Selimi, and L. Ahmedi. Application of Trust in Recommender Systems—Utilizing Naive Bayes Classifier. *Computation*, 10(1), 2022. Publisher: MDPI.

[75] X. Tian, J. Shao, and M. Yang. Macroeconomic Early Warning Method Based on Support Vector Machine under Multi-Sensor Data Fusion Technology. *Mobile Information Systems*, 2022, 2022. Publisher: Hindawi Limited.

[76] M. Vatsa, A. Jain, and R. Singh. Adventures of Trustworthy Vision-Language Models: A Survey. In Wooldridge M., Dy J., and Natarajan S., editors, *Proc.*

AAAI Conf. Artif. Intell., volume 38, pages 22650–22658. Association for the Advancement of Artificial Intelligence, 2024. Issue: 20 Journal Abbreviation: Proc. AAAI Conf. Artif. Intell.

[77] A. Apostolidis, N. Bouriquet, and K.P. Stamoulis. AI-Based Exhaust Gas Temperature Prediction for Trustworthy Safety-Critical Applications. *Aerospace*, 9(11), 2022. Publisher: MDPI.

[78] M. Cheng, Y. Li, S. Nazarian, and P. Bogdan. From rumor to genetic mutation detection with explanations: a GAN approach. *Scientific Reports*, 11(1), 2021. Publisher: Nature Research.

[79] H. Muccini et al. Software architecture for ml-based systems: what exists and what lies ahead. In *2021 IEEE/ACM AI WAIN*, pages 121–128, 2021.

[80] S. Bharati, M.R.H. Mondal, and P. Podder. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? *IEEE Transactions on Artificial Intelligence*, 5(4):1429–1442, 2024. Publisher: Institute of Electrical and Electronics Engineers Inc.

[81] A. Lombardi, D. Diacono, N. Amoroso, P. Biecek, A. Monaco, L. Bellantuono, E. Pantaleo, G. Logroscino, R. De Blasi, S. Tangaro, and R. Bellotti. A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease. *Brain Informatics*, 9(1), 2022. Publisher: Springer Science and Business Media Deutschland GmbH.

[82] E. Tjoa and C. Guan. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2021. Publisher: Institute of Electrical and Electronics Engineers Inc.

[83] P. Bagave, M. Westberg, R. Dobbe, M. Janssen, and A.Y. Ding. Accountable AI for Healthcare IoT Systems. In *Proc. - IEEE Int. Conf. Trust, Priv. Secur. Intell. Syst., Appl., TPS-ISA*, pages 20–28. Institute of Electrical and Electronics Engineers Inc., 2022. Journal Abbreviation: Proc. - IEEE Int. Conf. Trust, Priv. Secur. Intell. Syst., Appl., TPS-ISA.

[84] M.R. Karim, T. Islam, M. Shajalal, O. Beyan, C. Lange, M. Cochez, D. Rebholz-Schuhmann, and S. Decker. Explainable AI for Bioinformatics: Methods, Tools and Applications. *Briefings in Bioinformatics*, 24(5), 2023. Publisher: Oxford University Press.

[85] E. Neri, G. Aghakhanyan, M. Zerunian, N. Gandolfo, R. Grassi, V. Miele, A. Giovagnoni, and A. Laghi. Explainable AI in radiology: a white paper of the Italian Society of Medical and Interventional Radiology. *Radiologia Medica*, 128(6):755–764, 2023. Publisher: Springer-Verlag Italia s.r.l.

[86] C. Meske, E. Bunde, J. Schneider, and M. Gersch. Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 39(1):53–63, 2022. Publisher: Taylor and Francis Ltd.

[87] Y. Okada, Y. Ning, and M.E.H. Ong. Explainable artificial intelligence in emergency medicine: an overview. *Clinical and Experimental Emergency Medicine*, 10(4):354–362, 2023. Publisher: Korean Society of Emergency Medicine.

[88] F.G.Z. Kharrat, C. Gagne, A. Lesage, G. Gariépy, J.-F. Pelletier, C. Brousseau-Paradis, L. Rochette, E. Pelletier, P. Lévesque, M. Mohammed, and J. Wang. Explainable artificial intelligence models for predicting risk of suicide using health administrative data in Quebec. *PLoS ONE*, 19(4 April), 2024. Publisher: Public Library of Science.

[89] J. Cao, T. Zhou, S. Zhi, S. Lam, G. Ren, Y. Zhang, Y. Wang, Y. Dong, and J. Cai. Fuzzy inference system with interpretable fuzzy rules: Advancing explainable artificial intelligence for disease diagnosis—A comprehensive review. *Information Sciences*, 662, 2024. Publisher: Elsevier Inc.

[90] M.M. Hassan, S.A. AlQahtani, M.S. AlRakhami, and A.Z. Elhendi. Transparent and Accurate COVID-19 Diagnosis: Integrating Explainable AI with Advanced Deep Learning in CT Imaging. *CMES - Computer Modeling in Engineering and Sciences*, 139(3):3101–3123, 2024. Publisher: Tech Science Press.

[91] D. Hooshyar, R. Azevedo, and Y. Yang. Augmenting Deep Neural Networks with Symbolic Educational Knowledge: Towards Trustworthy and Interpretable AI for Education. *Machine Learning and Knowledge Extraction*, 6(1):593–618, 2024. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[92] U.-E. Habiba, J. Bogner, and S. Wagner. Can Requirements Engineering Support Explainable Artificial Intelligence? Towards a User-Centric Approach for Explainability Requirements. In Knauss E., Mussbacher G., Arora C., Bano M., and Schneider J.-G., editors, *Proc. Int. Conf. Requir. Eng.*, pages 162–165. IEEE Computer Society, 2022. Journal Abbreviation: Proc. Int. Conf. Requir. Eng.

[93] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities

and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. Publisher: Elsevier B.V.

[94] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 2023. Publisher: Elsevier B.V.

[95] A. Renda, P. Ducange, F. Marcelloni, D. Sabella, M.C. Filippou, G. Nardini, G. Stea, A. Virdis, D. Micheli, D. Rapone, and L.G. Baltar. Federated Learning of Explainable AI Models in 6G Systems: Towards Secure and Automated Vehicle Networking. *Information (Switzerland)*, 13(8), 2022. Publisher: MDPI.

[96] A. Arias-Duart, F. Pares, D. Garcia-Gasulla, and V. Gimenez-Abalos. Focus! Rating XAI Methods and Finding Biases. In *IEEE Int Conf Fuzzy Syst*, volume 2022-July. Institute of Electrical and Electronics Engineers Inc., 2022. Journal Abbreviation: IEEE Int Conf Fuzzy Syst.

[97] Z. Zhang, H.A. Hamadi, E. Damiani, C.Y. Yeun, and F. Taher. Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access*, 10:93104–93139, 2022. Publisher: Institute of Electrical and Electronics Engineers Inc.

[98] N. Capuano, G. Fenza, V. Loia, and C. Stanzione. Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access*, 10:93575–93600, 2022. Publisher: Institute of Electrical and Electronics Engineers Inc.

[99] A. Nascita, A. Montieri, G. Aceto, D. Ciuonzo, V. Persico, and A. Pescape. Improving Performance, Reliability, and Feasibility in Multimodal Multitask Traffic Classification with XAI. *IEEE Transactions on Network and Service Management*, 20(2):1267–1289, 2023. Publisher: Institute of Electrical and Electronics Engineers Inc.

[100] M. Fraile, J. Lindblad, C. Fawcett, N. Sladoje, and G. Castellano. Automatic analysis of infant engagement during play: An end-to-end learning and Explainable AI pilot experiment. In *ICMI Companion - Companion Publ. Int. Conf. Multimodal Interact.*, pages 403–407. Association for Computing Machinery, Inc, 2021. Journal Abbreviation: ICMI Companion - Companion Publ. Int. Conf. Multimodal Interact.

[101] F. Sovrano and F. Vitali. An objective metric for Explainable AI: How and why to estimate the degree of explainability. *Knowledge-Based Systems*, 278, 2023. Publisher: Elsevier B.V.

[102] S. Schwaiger, M. Aburaia, A. Aburaia, and W. Woeber. EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR ROBOT ARM CONTROL. In Katalinic B. and Getreidemarkt 9 Wien IFLT-IMS TU Wien, DAAAM International, editors, *Annals DAAAM Proc. Int. DAAAM Symp.*, volume 32, pages 640–647. DAAAM International Vienna, 2021. Issue: 1 Journal Abbreviation: Annals DAAAM Proc. Int. DAAAM Symp.

[103] Alejandro Barredo Arrieta et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[104] A. Singh, S. Sengupta, and V. Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

[105] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[106] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, Red Hook, NY, USA, December 2017.

[107] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[108] Kong. Kong gateway. `https://github.com/Kong/kong`, 2023. Accessed: 2023-12-01.

[109] NGINX. Nginx - advanced load balancer, web server, & reverse proxy, December 2023. Accessed December 1, 2023.

[110] Docker. Docker: Accelerated container application development, May 2022. Accessed: 2022-05-10.

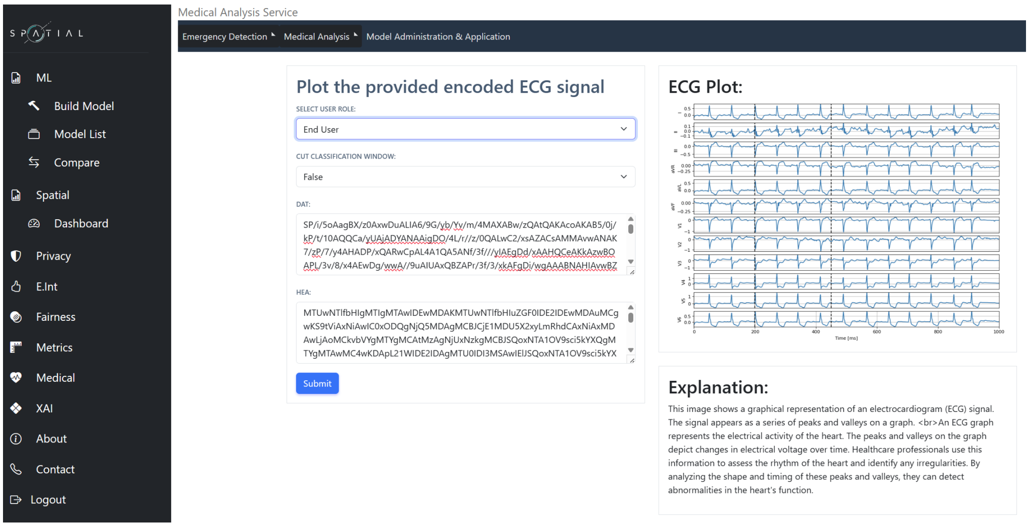
[111] University of tartu high performance computing center. `https://hpc.ut.ee/`. Accessed: May 10, 2024.

[112] X2go, 2023. Accessed December 1, 2023.

[113] Apache Software Foundation. Apache jmeter.

Appendix

I. Generated explanations from MAM based on the user role



Plot the provided encoded ECG signal

SELECT USER ROLE: End User

CUT CLASSIFICATION WINDOW: False

DAT:

SP/i/5oAagBX/z0AxwDuALIA6/9G/yb/Yv/m/4MAXABw/zQAtQAKAcoAKAB5/0j/
kP/t/10AQQCa/yUAjADYANAAigDO/4L/r//z/0QALwC2/xsAZACsAMMAvwANAK
7/zP/7/y4AHADP/xQARwCpAL4A1QA5ANf/3f///yIAEgDd/xAAHQCeAKkAzwBO
APL/3v/8/x4AEwDg/wwA//9uAIUAxQBZAPr/3f/3/xkAFgDi/wgAAABNAHIAvwBZ

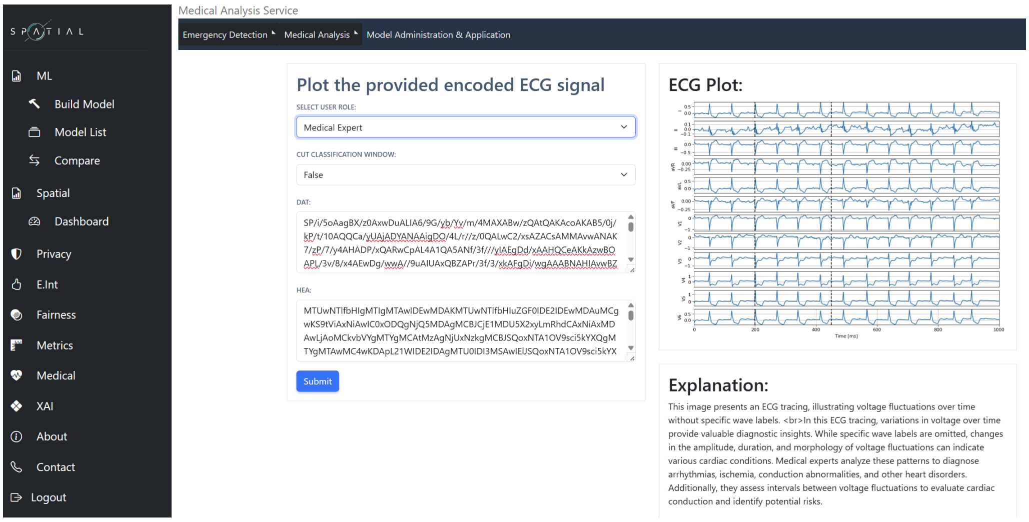
HEA:

MTUwNTlfbHIgMTIgMTAwIDEwMDAKMTUwNTlfbHIuZGF0IDE2IDEwMDAuMCg
wKS9tViAxNiAwIC0xODQgNjQ5MDAgMCBJCjE1MDU5X2xyLmRhdCAxNiAxMD
AwLjAoMCkvbVYgMTYgMCAtMzAgNjUxNzkgMCBJSQoxNTA1OV9sci5kYXQgM
TYgMTAwMC4wKDApL21WIDE2IDAgMTU0IDI3MSAwIElJSQoxNTA1OV9sci5kYX

Explanation:

This image shows a graphical representation of an electrocardiogram (ECG) signal. The signal appears as a series of peaks and valleys on a graph.
An ECG graph represents the electrical activity of the heart. The peaks and valleys on the graph depict changes in electrical voltage over time. Healthcare professionals use this information to assess the rhythm of the heart and identify any irregularities. By analyzing the shape and timing of these peaks and valleys, they can detect abnormalities in the heart's function.

Figure 33. Explanation for the end user



Plot the provided encoded ECG signal

SELECT USER ROLE: Medical Expert

CUT CLASSIFICATION WINDOW: False

DAT:

SP/i/5oAagBX/z0AxwDuALIA6/9G/yb/Yv/m/4MAXABw/zQAtQAKAcoAKAB5/0j/
kP/t/10AQQCa/yUAjADYANAAigDO/4L/r//z/0QALwC2/xsAZACsAMMAvwANAK
7/zP/7/y4AHADP/xQARwCpAL4A1QA5ANf/3f///yIAEgDd/xAAHQCeAKkAzwBO
APL/3v/8/x4AEwDg/wwA//9uAIUAxQBZAPr/3f/3/xkAFgDi/wgAAABNAHIAvwBZ

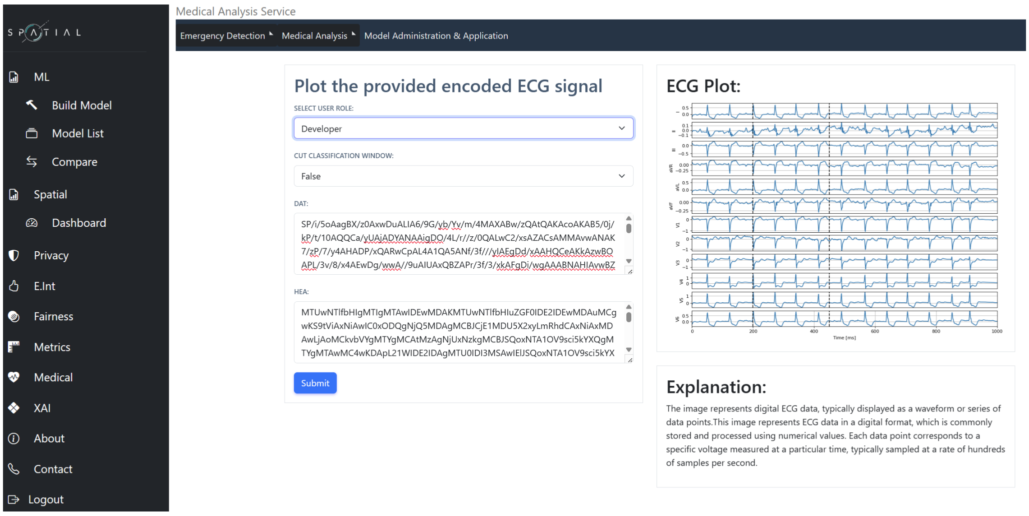
HEA:

MTUwNTlfbHIgMTIgMTAwIDEwMDAKMTUwNTlfbHIuZGF0IDE2IDEwMDAuMCg
wKS9tViAxNiAwIC0xODQgNjQ5MDAgMCBJCjE1MDU5X2xyLmRhdCAxNiAxMD
AwLjAoMCkvbVYgMTYgMCAtMzAgNjUxNzkgMCBJSQoxNTA1OV9sci5kYXQgM
TYgMTAwMC4wKDApL21WIDE2IDAgMTU0IDI3MSAwIElJSQoxNTA1OV9sci5kYX

Explanation:

This image presents an ECG tracing, illustrating voltage fluctuations over time without specific wave labels.
In this ECG tracing, variations in voltage over time provide valuable diagnostic insights. While specific wave labels are omitted, changes in the amplitude, duration, and morphology of voltage fluctuations can indicate various cardiac conditions. Medical experts analyze these patterns to diagnose arrhythmias, ischemia, conduction abnormalities, and other heart disorders. Additionally, they assess intervals between voltage fluctuations to evaluate cardiac conduction and identify potential risks.

Figure 34. Explanation for the medical expert



Medical Analysis Service

Plot the provided encoded ECG signal

SELECT USER ROLE: Developer

CUT CLASSIFICATION WINDOW: False

DAT:

SP/i/5oAagBX/z0AxwDuALIA6/9G/yb/Yv/m/4MAXABw/zQAtQAKAcoAKAB5/0j/kP/t/10AQQCa/yUAjADYANAAigDO/4L/r//z/0QALwC2/xsAZACsAMMAvwANAK7/zP/7/y4AHADP/xQARwCpAL4A1QA5ANf/3f///yIAEgDd/xAAHQCeAKkAzwBOAPL/3v/8/x4AEwDg/wwA//9uAIUAxQBZAPr/3f/3/xkAFgDi/wgAAABNAHIAvwBZ

HEA:

MTUwNTlfbHIgMTIgMTAwIDEwMDAKMTUwNTlfbHIuZGF0IDE2IDEwMDAuMCgwKS9tViAxNiAwIC0xODQgNjQ5MDAgMCBJCjE1MDU5X2xyLmRhdCAxNiAxMDAwLjAoMCkvbVYgMTYgMCAtMzAgNjUxNzkgMCBJSQoxNTA1OV9sci5kYXQgMTYgMTAwMC4wKDApL21WIDE2IDAgMTU0IDI3MSAwIElJSQoxNTA1OV9sci5kYX

ECG Plot:

Explanation:

The image represents digital ECG data, typically displayed as a waveform or series of data points.This image represents ECG data in a digital format, which is commonly stored and processed using numerical values. Each data point corresponds to a specific voltage measured at a particular time, typically sampled at a rate of hundreds of samples per second.

Figure 35. Explanation for the developer

Acronyms

AI Artificial Intelligence. 8

AIDA Artificial Intelligence and Data Act. 8

API Application Programming Interface. 36

BC Behavior Competency. 26

BNs Bayesian Networks. 17

CCPA California Consumer Privacy Act. 25

DNNs Deep Neural Networks. 17

DT Decision Tree. 17

ECGs Electrocardiograms. 9

EU European Union. 8

FL Federated Learning. 28

FOKUS Fraunhofer Institute for Open Communication Systems. 48

FRBS Fuzzy rule-based systems. 17

GDPR General Data Protection Regulation. 23

GradCAM Gradient-weighted Class Activation Mapping. 55

HIPAA Health Insurance Portability and Accountability Act. 25

HPC High Performance Computing. 41

HTTP Hypertext Transfer Protocol. 42

LIME Local Interpretable Model-agnostic Explanations. 34

LR Linear Regression. 17

LRP Layer-wise Relevance Propagation. 34

MAM Medical Analysis Module. 9

ML Machine Learning. 17

NIST National Institute of Standards and Technology. 11

ODD Operational Design Domain. 26

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses. 11

SHAP Shapley Additive Explanations. 34

SLR Systematic Literature Review. 10

SPATIAL Security and Privacy Accountable Technology Innovations, Algorithms, and machine Learning. 9

URI Uniform Resource Identifier. 46

URL Uniform Resource Locator. 46

USD United States Dollars. 16

VMs Virtual Machines. 41

XAI Explainable Artificial Intelligence. 10

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Marasinghe Mudiyanselage Rasinthe Marasinghe**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Systematic Evaluation of Trustworthy AI Augmentation in Modern Applications,
   *(*title of thesis)

   supervised by Huber Flores and Abdul-Rasheed Ottun.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Marasinghe Mudiyanselage Rasinthe Marasinghe
15/05/2024