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

TOLUWANI MATHEW ELEMOSHO

Analysing Model Attacks in Machine Learning Pipeline

Master's Thesis (30 ECTS)

Supervisor (s): Huber Flores, PhD
Abdul-Rasheed Ottun, MSc

Tartu 2024

Analysing Model Attacks in Machine Learning Pipeline

Abstract:

Machine learning models have evolved significantly and are integral to various sectors, including healthcare, finance, and transportation. However, their adoption has introduced vulnerabilities, particularly adversarial attacks that manipulate data to deceive these models. This thesis investigates the robustness of machine learning models against such attacks and explores the application of Explainable AI (XAI) techniques to enhance model transparency and security. Through a comprehensive literature review, this research identifies critical principles of trustworthy AI, including explainability, technical robustness, and human oversight.

The methodology systematically analyzes adversarial attacks, employing techniques like SHAP and LIME to evaluate model behaviour under different attack scenarios. The study also introduces a human-oversight dashboard designed to provide intuitive visualizations of these attacks, aiding in better understanding and mitigating vulnerabilities. Experimental results highlight the effectiveness of XAI in identifying and explaining adversarial manipulations, thereby improving the resilience of AI systems.

User studies reveal significant findings regarding the role of explanations in AI systems. Compared to no explanations, short explanations significantly enhance user engagement, preference for information, satisfaction, textual clarity, and trustworthiness. However, increasing the length of explanations from short to long yields minimal additional benefits. These results suggest concise explanations are highly effective in fostering user trust and engagement with AI systems.

This research contributes to the field by proposing robust defence mechanisms against adversarial attacks and emphasizing the role of human oversight in AI systems. It underscores the necessity for transparent, explainable, and resilient AI models to ensure their safe and ethical deployment in critical applications.

Keywords:

Machine Learning Pipeline, Adversarial Machine Learning, AI System Defense, Robustness in AI, Data Sanitisation, AI Transparency, Explainable AI (XAI), Interpretable AI, AI Explanation Methods, AI Decision-Making Explanation, Trustworthy AI, Ethical AI, User-Centered Design, Human-Computer Interaction (HCI)

CERCS:

P170 - Computer science, numerical analysis, system, control

Mudelrünnete analüüsimine masinõppe torujuhtmes

Lühikokkuvõte:

Masinõppemudelid on märkimisväärselt arenenud ja on lahutamatuks osaks erinevates sektorites, sealhulgas tervishoius, rahanduses ja transpordis. Kuid nende kasutuselevõtt on toonud kaasa haavatavusi, eriti ründavaid rünnakuid, mis manipuleerivad andmetega, et neid mudeleid petta. Käesolev lõputöö uurib masinõppemudelite vastupidavust selliste rünnakute vastu ja uurib Explainable AI (XAI) tehnikate rakendamist mudelite läbipaistvuse ja turvalisuse parandamiseks. Põhjaliku kirjanduse ülevaate kaudu tuvastab see uurimus usaldusväärse tehisintellekti kriitilised põhimõtted, sealhulgas seletatavuse, tehnilise vastupidavuse ja inimliku järelevalve.

Metoodika analüüsib süstemaatiliselt ründavaid rünnakuid, kasutades tehnikaid nagu SHAP ja LIME, et hinnata mudelite käitumist erinevates rünnakustsenaariumides. Uurimus tutvustab ka inimjärelevalve armatuurlauda, mis on loodud rünnakute intuitiivsete visualiseerimiste pakkumiseks, aidates paremini mõista ja vähendada haavatavusi. Eksperimentaalsed tulemused toovad esile XAI tõhususe ründavate manipulatsioonide tuvastamisel ja selgitamisel, parandades seeläbi AI süsteemide vastupidavust.

Kasutajauuringud toovad esile olulised leiud seoses selgituste rolliga AI süsteemides. Lühikesed selgitused suurendavad märkimisväärselt kasutajate kaasatust, eelistust teabe vastu, rahulolu, tekstilist selgust ja usaldusväärsust võrreldes selgituste puudumisega. Siiski, selgituste pikkuse suurendamine lühikestest pikkadeks annab minimaalselt täiendavat kasu. Need tulemused viitavad sellele, et lühikesed selgitused on väga tõhusad kasutajate usalduse ja kaasatuse suurendamisel AI süsteemidega.

See uurimus annab valdkonnale panuse, pakkudes välja tugevad kaitsemehhanismid ründavate rünnakute vastu ja rõhutades inimjärelevalve rolli AI süsteemides. See rõhutab vajadust läbipaistvate, seletatavate ja vastupidavate AI mudelite järele, et tagada nende ohutu ja eetiline kasutuselevõtt kriitilistes rakendustes.

Võtmesõnad:

Masinõppe torujuhe, võistlev masinõpe, tehisintellekti süsteemi kaitse, tehisintellekti töökindlus, andmete desinfitseerimine, tehisintellekti läbipaistvus, selgitatav tehisintellekt (XAI), tõlgendatav tehisintellekt, tehisintellekti selgitamise meetodid, tehisintellekti otsuste tegemise selgitus, usaldusväärne tehisintellekt, eetiline tehisintellekt, kasutajakeskne disain, inimese ja arvuti interaktsioon (HCI)

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

Acknowledgements

A very big thank you to my family, who have provided me with the support to write this work. I would like to thank my supervisor, Prof. Dr Huber Flores for giving me the opportunity to learn under his tutelage and for his outstanding contributions during every phase of the project. I would also like to thank my co-supervisor, Abdul-Rasheed Ottun, MSc for his support throughout the thesis.

My gratitude would not be complete without special mentions to Mayowa Olopade, Rasinthe Marasinghe, Adebola Gbiri, and Faustina Ekeh, for their relentless efforts during the experiments; their support cannot be overestimated in completing this work.

Tolu

Contents

List of Figures

List of Tables

1 Introduction

Machine learning models have experienced a remarkable evolution, progressing from simplistic rule-based systems for reasoning to sophisticated neural networks capable of discerning intricate patterns and associations from vast datasets with unprecedented accuracy and efficiency. This rapid evolution owes much to advances in chip and computation technology. The advancement has led to their widespread adoption in society across diverse application domains, including but not limited to education [2], criminal justice [3, 4], finance [3, 5], agriculture[5], transportation [6, 3], and healthcare [7, 3]. With the growing adoption of ML-based applications, legislative and regulatory initiatives, such as the European Union's Artificial Intelligence Act (EU AI ACT), have emerged to mandate the trustworthiness in developing, deploying, and using ML-based systems. The EU AI Act requires certifying and verifying the fundamental functionality of developed algorithms and applications before deploying them in key infrastructures. The EU regulations emphasize the importance of scalability, robustness, and transparency in the trustworthiness of the AI model[1].

AI models are a black box and often difficult to interpret, challenging to identify vulnerabilities and potential attack points. The need for more transparency in AI decision-making processes further complicates the development of effective defense strategies. For instance, adversaries can meticulously manipulate training inputs to undermine the learning capability of AI models, potentially leading to erroneous decisions with critical consequences for individuals and society at large. One example is in the healthcare industry, where AI models are used to analyze medical images for diagnosis. If an adversarial attack were to occur, a malicious actor could manipulate the image in a way that causes the AI model to misdiagnose a patient, potentially leading to harmful consequences. Human oversight plays a crucial role in ensuring the trustworthiness and effectiveness of AI systems. [8]. It involves monitoring and supervising AI models to understand their behavior, limitations, and decision-making processes. The goal is to make AI's decisions understandable and transparent to users. The integration of human input is essential for tailoring explanations to meet user needs, a process that involves anticipating and answering both explicit and implicit questions that users might have about AI decisions [9]. This approach is crucial in addressing the 'black box' nature of sophisticated AI algorithms, where the decision-making process is often opaque and complex for users to understand[2]. By incorporating human oversight, AI systems can benefit from context and ethical considerations, ensuring that decisions are made within acceptable moral and ethical boundaries[8].

In this research, we propose to leverage explainable AI (XAI) techniques to detect adversarial attacks targeting AI models and develop a user-friendly and intuitive dashboard that serves as a tool for visualizing and understanding these attacks. The primary objective is to equip researchers and developers with a tool that provides better insights into the security aspects of AI models. By doing so, they can better understand the complexities of adversarial tactics and the points of weakness within their systems. Ultimately, this will help enhance the resilience of AI models and, in turn, contribute to their trustworthiness.

This thesis is structured as follows: The Introduction in Chapter 1 lays out the general context and objectives. Chapter 2 conducts a comprehensive Literature Review, detailing the methodologies, databases consulted, keyword development, and the inclusion/exclusion criteria for the literature. This chapter also discusses the state of the art in explainability, AI robustness, and human oversight in AI systems. Chapter 3 describes the Methodology, starting with the data selection and pre-processing steps. It then outlines the selection criteria for the machine learning models used and explains the simulation of adversarial attacks, including the types of attacks and the metrics for measuring resilience. The chapter also integrates Explainable Artificial Intelligence (XAI), discusses the implementation of a Human-oversight dashboard within XAI, and conducts a user study on the impact of texts on the explainability of XAI. Chapter 4 presents the results, analyzing the outcomes of the applied methodologies and the performance of the proposed system under various conditions. Finally, the chapter focuses on the main findings and discussion. Following the main body, the thesis concludes with a References section, acknowledging all sourced materials, and an Appendix that includes a Glossary for terms used throughout the thesis.

2 Literature Review

The principles of AI trustworthiness comprise technical resilience, robustness, and security considerations during the development and deployment of AI. The risk exposures from the application of AI systems are from the vulnerability exploit associated with the system. The development of AI forms the basis of preventing and minimizing any risk or harm that can negatively impact humans or the immediate environment of AI. Adversaries' exploitations of AI vulnerabilities are so crafty that their orchestration and implementation are complicated to diagnose—these range from manipulating training data to inference attacks and unintended biases leading to discriminatory outcomes. Consequently, to effectively conduct diagnostics for threats, the internal workings of AI also need to be understood, and this relates to another trustworthy requirement: transparency, and explainability. A systematic literature review (SLR) is used in this study to look into different parts of explainable methods in machine learning. This study focuses on two fundamental principles of trustworthy AI—transparency and technical robustness—as outlined in the "Ethics Guidelines for Trustworthy AI" by the European High-Level Expert Group [1].

2.1 Databases Searched

A comprehensive and systematic search for literature was conducted using the SCOPUS database. SCOPUS is a popular bibliometric database chosen for its comprehensive coverage, popularity, and efficient retrieval capabilities through its advanced search[10]. To strategically perform an effective query of the database for collecting an extensive body of research works to be reviewed, keywords relating to the relevant AI trustworthiness properties were carefully extracted from conceptual definitions found in reliable frameworks and guidelines on trustworthy AI. This approach ensured the consideration of a wide range of valuable studies to be included in this study, thereby enhancing the thoroughness and relevance of the literature review.

2.2 Development of Keywords

The literature review process searches for terms associated with three of the principles of trustworthy AI: Transparency and Technical Robustness—according to the "Ethics Guidelines for Trustworthy AI" by the European High-Level Expert Group [1]. Leveraging the AI taxonomy suggested by AI Watch [11], the literature review methodology includes a strategic search for terms aligned with AI ethics combined with AI Transparency. Table 2 shows key phrases for the search database: "AI Ethics," "Trustworthy AI," "Ethics of AI," "Explainable Artificial Intelligence XAI," and related terms. We employed relevant terms from NIST's Taxonomy of Attacks, Defenses, and Consequences in Adversarial Machine Learning [12]. The taxonomy and its associated terminology comprehensively encompass the spectrum of adversarial attacks on machine learning systems, the defensive strategies against them, and the subsequent outcomes.

Furthermore, for the aspect of the study that examines human oversight in explainable AI, the keywords "machine learning," " visualisation, "explainable AI, " and related terms are selected to conduct a broad yet precise literature search, which is crucial for capturing studies that discuss human oversight in explainable AI systems.

Lastly, the keywords and search strings utilized for gathering the literature to be reviewed for this work were presented in Table 1 and Table 2 respectively.

Table 1. NIST Keywords and Their Synonyms

NIST Keywords	Synonymous Keywords
Adversarial Machine Learning	Adversarial machine learning
Attacks	Adversarial attack
Poisoning	Poisoning attack
ML models	Model poisoning
Evasion	Evasion attack
Defenses	AI system defense
Robustness Improvement	Robustness in AI
Data sanitization (reject on negative impact)	Data sanitization

2.3 Search Strategy

The search strategy utilized Boolean operators (AND, OR) to create complex search queries, enabling a focused yet comprehensive exploration of the literature. This method combined keywords and Boolean logic in the SCOPUS Library to retrieve relevant studies within the domains of AI transparency, technical robustness, and human oversight in explainable AI.

2.4 Inclusion and Exclusion Criteria

Explicit inclusion and exclusion criteria ensure that the systematic literature review is focused and comprehensive. The inclusion criteria are as follows: peer-reviewed journal articles or conference papers published from January 2014 onwards, available in English, and focusing on artificial intelligence as applied to transparency, technical robustness, or human oversight in explainable AI. The exclusion criteria are articles without the relevant keywords in the title, publications before 2014, non-peer-reviewed literature such as book reviews and editorials, papers that are not conference materials or academic journals, studies not written in English, and literature that does not directly address the three focused domains.

Table 2. Search strings

AI Transparency	**Technical Robustness**	**Human-oversight in XAI**
("AI Ethics" OR "Ethical AI" OR "Trustworthy AI" OR "AI Trustworthiness" OR "Ethical guidelines for AI" OR "AI moral principles") AND TITLE ("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") AND TITLE-ABS-KEY("XAI algorithms" OR "Explainability algorithms" OR "Interpretability models" OR "AI decision-making explanation" OR "SHAP" OR "Shapley Additive Explanations" OR "LIME" OR "Local Interpretable Model-agnostic Explanations" OR "Model-agnostic methods" OR "black-box models")	TITLE ("robustness in AI" OR "model poisoning" OR "Adversarial attack" OR "data sanitization" OR "AI system defense" OR "poisoning attack" OR "evasion attack" OR "Adversarial machine learning") AND ALL ("Explainable AI" OR "XAI" OR "Interpretable AI" OR "AI transparency" OR "Transparency in Machine Learning" OR "AI methods" OR "Understandable AI" OR "Accountable AI" OR "Explainable Artificial Intelligence")	TITLE ("visualization techniques" OR "visual analytics" OR "user-centric AI design" OR "human-computer interaction" OR "HCI in AI" OR "collaborative interfaces" OR "decision support systems" OR "human-in-the-loop" OR "interactive machine learning" OR "user-centered design" OR "interface" OR "human-centered AI" OR "user-centric explainable AI") AND 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")

2.5 Selection of Primary Studies

The process of selecting primary studies for this systematic literature review (SLR) was meticulously conducted across three distinct domains: AI transparency, technical robustness, and visualisation of explainable AI. Each domain followed a structured approach to filter the initial vast array of search results down to a set of studies most relevant for in-depth review. The overall aim was to ensure a comprehensive collection of high-quality, pertinent scholarly works. Here is a summary of the selection process for each domain:

AI Transparency

- Initial Query: The search commenced with 1,156 studies identified from an extensive initial query tailored to capture literature on AI transparency.
- Publication Year: After applying the publication year filter, 1,067 documents remained.
- Source Type: Refinement based on source type (conference proceedings and journals) led to 912 documents
- Publication Stage: Further filtering to only include documents in their final publication stage resulted in 889 documents.
- Open Access: Narrowing down to open access documents further reduced the count to 496.
- Content Type: The final filter, for documents that are both articles and conference papers, left 426 documents.

Technical Robustness

- Initial Query: The search commenced with 1,653 studies identified from an extensive initial query tailored to capture literature on AI transparency.
- Publication Year: After applying the publication year filter, 1,496 documents remained.
- Source Type: Refinement based on source type (conference proceedings and journals) led to 1,307 documents
- Publication Stage: Further filtering to only include documents in their final publication stage resulted in 1,287 documents.

- Open Access: Narrowing down to open access documents further reduced the count to 727.
- Content Type: The final filter, for documents that are both articles and conference papers, left 600 documents.

Human-oversight in Explainable AI

- Initial Query: The search commenced with 11,263 studies identified from an extensive initial query tailored to capture literature on AI transparency.
- Publication Year: After applying the publication year filter, 10,125 documents remained.
- Source Type: Refinement based on source type (conference proceedings and journals) led to 8,729 documents
- Publication Stage: Further filtering to only include documents in their final publication stage resulted in 8,456 documents.
- Open Access: Narrowing down to open access documents further reduced the count to 4,811.
- Content Type: The final filter, for documents that are both articles and conference papers, left 3,984 documents.

Table 3 provides a detailed illustration of the study selection process. Initially, a search term was employed to sift through the 5010 studies across the three domains of AI transparency, technical robustness, and the human-oversight in explainable AI collected from the Scopus database. By using inclusion and exclusion criteria, 5010 studies were taken out of the pool for a number of reasons, such as language barriers, documents that were not relevant (like posters and reviews), and duplicates. This left a total of 5010 studies. These remaining studies underwent scrutiny by examining their titles and abstracts to determine their relevance to the review's scope. Studies that did not explicitly state their application domain or contribution in the title or abstract were retained for further evaluation. This scrutiny reduced the number to 131 studies.

2.6 State of the art

This section explores the concept of trustworthy AI, which includes principles ensuring ethical, fair, and accountable decision-making in AI systems [13], focusing on explainability and robustness. It starts by defining key terms related to the three core subdomains central to this research. First, it addresses the importance of transparency in machine

Table 3. Study Selection Process

Selection Phase	Criteria	AI Transparency	Technical Robustness	Human Oversight	Total Studies at Each Phase
Initial Search	"all"	1,156	1,653	11,263	14,072
Screened (Year Filter)	<10y	1,067	1,496	10,125	12,688
Eligibility (Source Type)	"j", "p"	912	1,307	8,729	10,948
Eligible (Publication Stage)	"final"	889	1,287	8,456	10,632
Included (Open Access)	"all"	496	727	4,811	6,034
Final Studies (Content Type)	"ar","cp"	426	600	3,984	5,010
Remaining After Title/Abs Screening	"t","abs"	31	28	72	131

learning (ML), emphasizing its critical role in enhancing model robustness [14]. Next, it examines AI model robustness, highlighting its significance and challenges. Finally, it discusses human oversight in explainable artificial intelligence (XAI), underscoring its role in making AI systems understandable and accountable. Each subdomain contributes to the broader understanding and implementation of reliable and trustworthy AI systems.

2.6.1 Trustworthy AI

Trustworthy AI encompasses several vital characteristics: explainability, fairness, accountability, and robustness. See 1 for complete taxonomy. Explainability allows AI systems to provide understandable reasons for their decisions and actions, crucial for user trust [15, 16, 17, 7]. Trustworthy AI ensures auditability and scrutinizes decisions for accuracy and fairness. Fairness addresses the need for unbiased decision-making, designing AI systems to mitigate biases based on age, gender, race, location, and other personal perceptions [18]. Accountability emphasizes the importance of tracing AI actions back to human operators or developers, ensuring responsibility. Robustness refers to AI's resilience against attacks and reliability under various conditions, including safeguarding against adversarial attacks [13]. These characteristics form the foundation of trustworthy AI, ensuring transparent, equitable, accountable, and durable AI technologies and fostering trust among users and stakeholders. Trustworthy AI promotes principles for ethical, transparent, and accountable frameworks within AI systems. Challenges persist, notably the vulnerability of AI models throughout their lifecycle, requiring robust methods to ensure trustworthiness. Specifically, in the context of this research, trustworthy AI aims to improve the safety and reliability of AI systems, particularly by addressing their susceptibility to adversarial attacks [13].

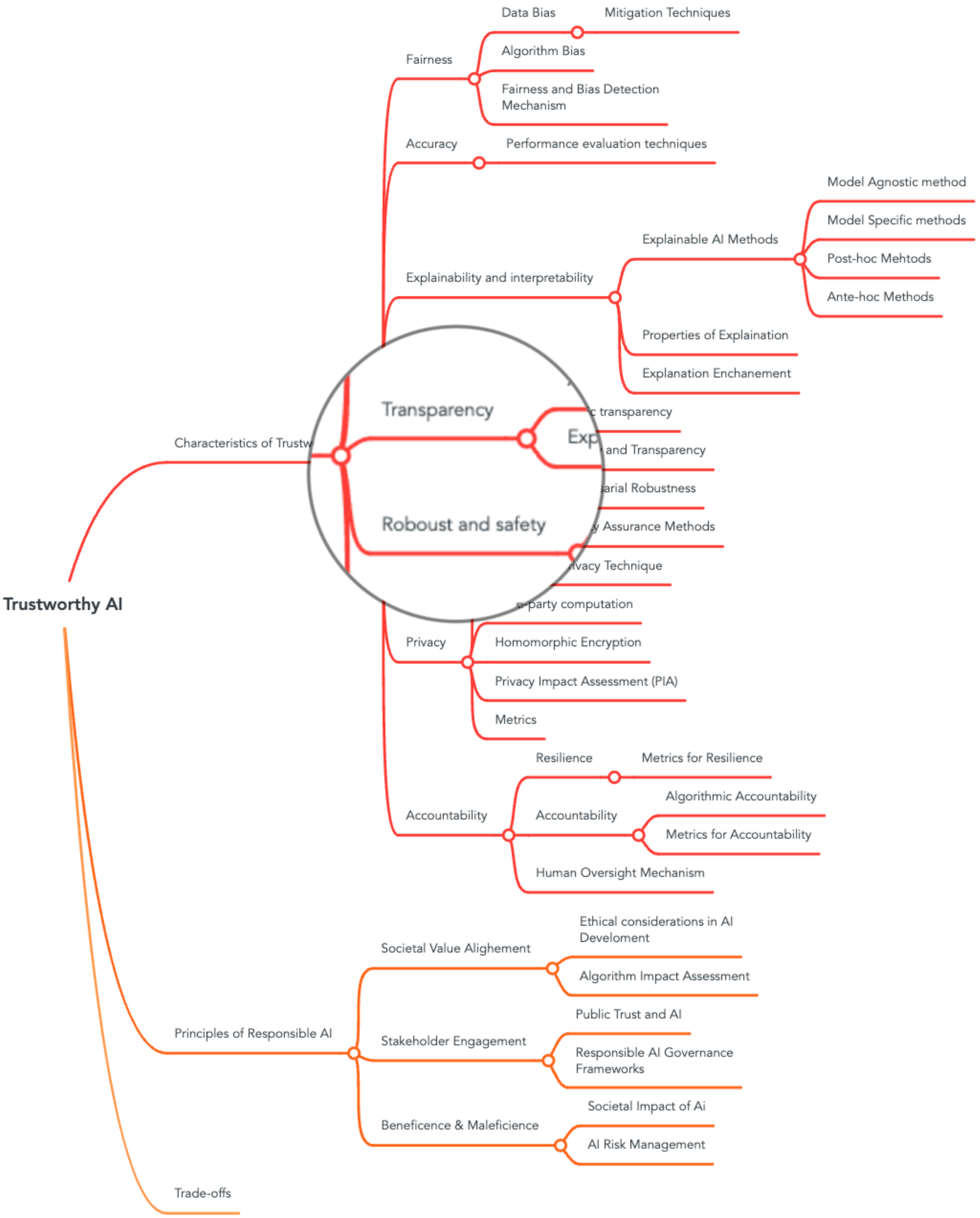


Figure 1. Trustworthy AI

2.6.2 Explainable AI

The push for explainable AI (XAI) arises from several critical factors, including legislative requirements such as the EU General Data Protection Regulation (GDPR). This regulation demands robustness in AI systems by mandating safeguards for automated decision-making processes, impacting any entity managing European data globally [19, 14]. Explainable AI helps understand how models respond to adversarial attacks, supporting the development of more resilient AI systems [20, 21]. Before creating an interpretable model, decision-makers and experts must consider how local interpretability can diagnose failures before retraining and assess the potential impacts on intended users [19]. Explainability in AI, primarily through Bayesian Neural Networks (BNNs), makes AI models more resistant to adversarial attacks [18] by providing a deeper understanding of how models make predictions, helping to identify and mitigate vulnerabilities more effectively. Bayesian methods for explainability, including Layer-Wise Relevance Propagation (LRP), produce more stable and interpretable explanations under adversarial conditions [22, 23, 18]. This suggests that explainable models can maintain their performance and reliability even when faced with attempts to deceive or confuse them.

Incorporating explainability features enhances AI models' robustness against attacks [16]. It provides insights into their decision-making process, enabling improvements in model design and training methodologies that enhance security and effectiveness[22]. Explainable AI improves the performance of AI systems by identifying key decision-making features, reducing data and computational power requirements, and enhancing model accuracy through these insights [5]. Explainable AI's ability to pinpoint instances of model uncertainty enables necessary adjustments, enhancing the system's robustness and fostering trust[5].

Local and Global Methods

Explainable AI methods can be categorized into local and global approaches. Global methods provide an overall understanding of model behaviour, while local methods focus on individual predictions. Visualization techniques and simpler models can approximate complex model decisions globally [24]. Local methods, such as LIME and SHAP, explain individual predictions by approximating complex models more simply [10, 25]. Counterfactual explanations suggest changes to alter outcomes for specific instances [3, 24].

Model-Specific and Agnostic Methods

Model-specific techniques explain particular models by analyzing their structure or data usage, working well for specific models but not others [24]. Model-agnostic methods like LIME and SHAP can be applied to any model to provide explanations without requiring knowledge of the model's internals [26].

Transparent, Opaque, and Post-Hoc Explainability Transparent models, such as Decision Trees and Bayesian Networks, offer clear visibility into their mechanisms, fostering user trust [20]. Opaque models, like Random Forests and Neural Networks, require additional methods to elucidate their decision-making processes [20]. Post-hoc explainability techniques, such as LIME and SHAP, explain non-interpretable models, ensuring broader applicability and understanding [26, 24]. explainable machine learning (ML) models can be broadly categorized into Transparent, opaque, and post-hoc explainability models [14]. Transparent models are like open books, offering clear visibility into their operational mechanisms and decision-making processes. These models, which include tree-based models like Decision Trees and Bayesian Networks, linear models such as Linear and Logistic Regression, and rule-based models, are inherently interpretable due to their straightforward and intuitive designs. This level of transparency is vital as it fosters trust among users by making the AI's decisions understandable [20]. In contrast, Opaque Models, which encompass techniques like Random Forests and Neural Networks, are characterized by their inherent complexity and lack of interpretability. These models often necessitate additional methods to elucidate their decision-making processes [20]. Models with post-hoc explainability are not intrinsically interpretable but can be made so with various techniques, such as model-agnostic methods like LIME or SHAP [26], which can be applied irrespective of the model's internal workings.

Model-specific techniques are designed to explain a particular type of model by examining its structure or data usage [24]. These methods work well for the specific models they target but are difficult to apply to other models. They are powerful because they can delve into model internals, like weights or structures. However, this specificity makes them hard to generalize [24]. Zacharias categorizes AI methods into intrinsic and extrinsic types [27]. Intrinsic methods are naturally understandable, designed to be clear and straightforward from the beginning without needing extra explanatory steps, making them accessible and easy to grasp, especially for beginners. Extrinsic methods, also known as post hoc explanations, are used to explain the decisions of more complex AI models after they have made a decision. Post-hoc explainability involves methods used after a model has been trained to help explain how it makes decisions. These methods are especially useful for complex models that are not easily understandable on their own [10]. These extrinsic methods can be divided into global and local approaches. Global methods aim to provide an overall understanding of how the model works. Nagahisarchoghaei describes global methods as explanations of how a machine learning model makes decisions overall, not just for one single prediction [10]. These methods look at all the data or features to explain the model's behavior in a general way. This is like looking at the forest instead of focusing on individual trees, giving a broad understanding of what the model pays attention to when making predictions [19]. For example, visualization techniques can show the overall effect of certain features on the model's predictions. This could be through graphs that reveal patterns or important features [10]. Another global method involves using simpler models to approximate and explain the complex model's decisions, offering a big-picture view of how the model works [24]. These global methods help build trust by showing that the model works in a way that makes sense across many different situations.

Nagahisarchoghaei and Adadi liken local methods to zooming in on one piece of a puzzle to see why it fits where it does [10, 3]. Nagahisarchoghaei highlights examples of local post-hoc methods, which include:

LIME, or Local Interpretable Model-agnostic Explanations, helps understand AI decisions by examining specific examples [28, 29]. Imagine LIME as a magnifying glass that zooms in on one decision made by AI. It modifies small parts of the data and observes how these changes affect the AI's decision. This helps identify what's important for the AI's decision in that specific case. By doing this, LIME creates a simpler model that explains why the AI made its decision, even if the original AI is very complex and hard to understand [4]. This method is akin to experimenting with a recipe by slightly altering ingredients to see which ones are crucial for the taste [19]. LIME focuses on local explanations, meaning it explains individual predictions by approximating the complex model in a simpler, more understandable way [19, 29].

Counterfactual explanations are local because they focus on specific instances or decisions made by a model [24]. They show how input needs to change to get a different result. For example, they can explain what would need to be different for a loan application to be approved instead of denied [3]. They work by looking at specific cases and suggesting changes that could alter the outcome [30].

SHAP, which stands for Shapley Additive exPlanations, explains decisions made by machine learning models by focusing on individual predictions [31]. It examines how much each feature in the data contributes to the final decision for a specific instance, making SHAP a local explanation method because it provides detailed insights for single predictions rather than explaining the model's behavior as a whole [32, 33]. This method is useful because it can give explanations for individual predictions, which is why it's considered a local explanation method [25].

2.6.3 Taxonomy of Robustness

Adversarial attacks manipulate data to deceive machine learning (ML) models, causing incorrect outcomes nearly undetectable to humans. These attacks target both the training phase (poisoning attacks) and the inference phase of ML models [34].

Apruzzese et al. [35] discuss different attacks on machine learning systems to detect network intrusions. One primary type is an evasion attack, where the attacker slightly changes harmful data so the system thinks it is safe. The primary objective of evasion attacks is to alter the malicious data to evade detection, leveraging the model's vulnerabilities without knowing the system's internals. Another type mentioned is inference attacks, which occur when the system is already working, and the attacker tries to make it fail by exploiting its decision-making process. The main goal of these attacks is to trick the system into making wrong decisions by taking advantage of how the system makes those decisions [36]. There is also a mention of poisoning attacks, which involve messing with the system's learning phase by adding or changing data and altering DNN models' behaviour according to attackers' intentions. The paper also talks about grey-box attacks, where the attacker has limited information about the system. There are also white-box attacks, where the attacker knows everything about the system and uses this knowledge to craft specific attacks. On the other hand, black-box attacks happen when the attacker does not know the system's details but can still find ways to trick it by guessing.

Naderi et al. [30] categorize different types of attacks based on how much the attacker knows about the model. For example, in a "black-box" attack, the attacker only knows the input and output of the model but not its inner workings. In contrast, a "white-box" attack means the attacker fully knows the model's architecture and parameters.

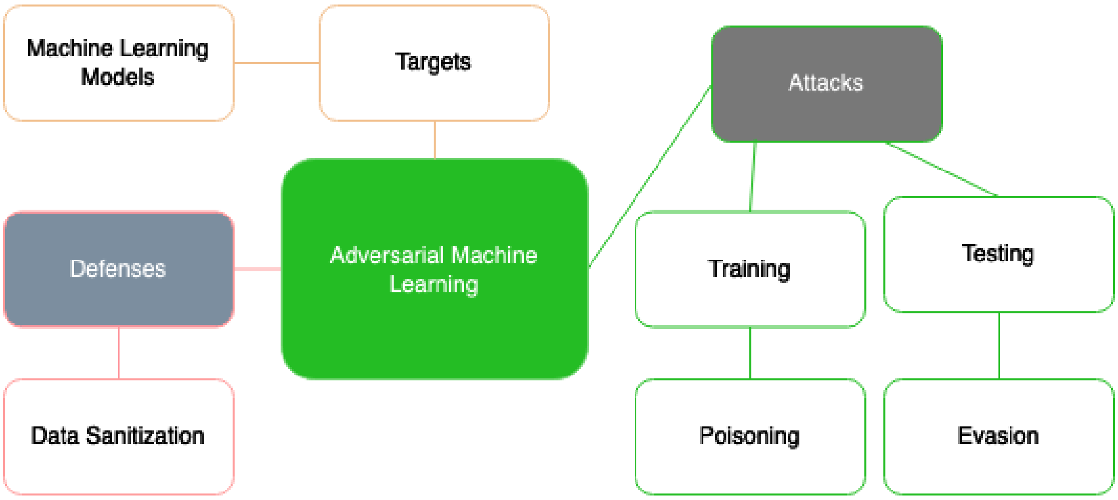


Figure 2. Example of the NIST Taxonomy organized in a Hierarchical manner

Targets In machine learning, a target, or vulnerability, refers to a weakness that attackers can exploit to cause incorrect predictions or decisions [34]. These vulnerabilities emerge from various sources, such as outsourcing the training process, using pre-trained models from third parties, or ineffective data validation on the network [6]. Palacio et al. [37] classify attack targets into stages: the Physical Domain of input sensors, the Digital Representation for pre-processing, and the Machine Learning Model. This research focuses on the machine learning model as the target for adversarial attacks.

Developers can create machine learning models using three main methods: supervised, unsupervised, and reinforcement learning [7]. In supervised learning, the model trains with labelled data, learning to predict labels for new data. Unsupervised learning identifies patterns or clusters without labelled data. Reinforcement learning involves trial and error, where the model makes decisions and learns from the results to achieve objectives [7]. An example of an unsupervised machine learning system is an attacker's target intrusion detection system (IDS) that protect networks, particularly in complex environments like the Industrial Internet of Things (IIoT). This system get tricked by adversarial samples—altered data that appears normal to the IDS but contains malicious content. Attackers target components like Supervisory Control and Data Acquisition (SCADA) or database management systems, transforming average data into adversarial samples to bypass detection and compromise platform security [8].

Adversarial Attacks Adversarial attacks manipulate data to deceive targets i.e machine learning (ML) models, causing incorrect outcomes and being nearly undetectable to humans. These attacks target both the training phase (poisoning attacks) and the inference phase of ML models [34]. There are different types of attacks on machine learning systems used for detecting network intrusions [35]. One main type is called an evasion attack, where the attacker changes harmful data slightly so the system thinks it's safe. Another type mentioned is inference attacks, which occur when the system is already working and the attacker tries to make it fail by exploiting its decision-making process.The main goal of these attacks is to trick the system into making wrong decisions by taking advantage of how the system makes those decisions. There's also a mention of poisoning attacks, which involve messing with the system's learning phase by adding or changing data, altering DNN models' behavior according to attackers' intentions.

Attacks can be categorized based on how much the attacker knows about the model[30]. For example, in a "black-box" attack, the attacker only knows the input and output of the model but not its inner workings. In contrast, a "white-box" attack means the attacker has full knowledge of the model's architecture and parameters [8] talks about gray-box attacks, where the attacker has some limited information about the system but not everything.

Defenses: Robustness in AI refers to a system's resilience and reliability in maintaining functionality under various conditions, including adversarial attacks [19] It involves handling input perturbations and maintaining consistent outputs, ensuring the integrity of operations. Robustness is crucial, especially in systems using post-hoc explanation techniques like LIME and SHAP, which are vulnerable to manipulation. The deployment of robust AI systems requires developing novel techniques to withstand adversarial attacks, safeguarding the model's explanations and decisions[24]. Furthermore, AI Robustness prevents bias against underrepresented groups, and protects individual data, thereby enhancing user experience and societal trust.

Data sanitization involves cleaning or removing sensitive information from datasets to protect against unauthorized access or attacks. It helps prevent poisoning attacks, especially when using third-party datasets or tools. This process is crucial for machine learning models as it ensures data integrity during preprocessing without introducing biases or errors [35]. However, there are challenges faced during data sanitization. One major challenge is ensuring that the sanitization process does not inadvertently remove or alter data that could be crucial for analysis or operations. This requires a delicate balance between security and data integrity [35]. Another significant hurdle to data sanitization is the wide range of values that numeric features can have, which necessitates normalization to ensure that all features contribute equally to the analysis without any single feature dominating due to its scale [8].

Defensive distillation is a defence technique that trains a new model to avoid getting easily tricked by tiny changes in the data it sees. It does this by first making the model's predictions softer, which means it makes the model's output less extreme and more generalized, and these softened predictions serve as the new labels to train a second model, aiming to make this distilled model more robust against attacks [36]. However, it is essential to note that while defensive distillation can significantly enhance the model's resistance to specific attacks, it is only partially foolproof [36]. According to [38], researchers have developed two main defence strategies to protect machine learning models from attacks. The first strategy is adversarial training, where the model learns from original and modified data to recognize better. The second strategy involves using Generative Adversarial Networks (GANs), which create new data samples that closely resemble the original ones, helping the model to improve its detection capabilities[38]. However, implementing these strategies requires more computational resources and can be complex [38].

There are several ways to protect 3D point cloud models from attacks. One method changes the model's design to make it tougher against attacks by splitting the PointNet model and using a feature extractor and a discriminator. Another approach, DUP-Net, cleans the data from unwanted noise and makes the point cloud denser to bring it closer to its original form. IF-Defense uses a two-step process first to remove outliers and then adjust the remaining points to make the point cloud smooth and evenly spread out [30]. LPF-Defense trains models with a version of the data that only includes low-frequency information, which helps ignore the high-frequency changes made by attackers. Lastly, combining different strategies, like specific pooling operations, has made models more resistant to attacks. Modern defence strategies against attacks on machine learning models include adversarial training and Generative Adversarial Networks (GANs). Adversarial training involves mixing the original data with data that attackers have altered to make the system smarter in recognizing threats[38]. GANs, another advanced method, create new data samples that look very similar to the original ones, which helps in improving the system's ability to detect attacks, enhancing detection rates by up to 50%. However, these strategies require more computer power and can be complex[38].

Defence strategies against adversarial attacks on deep neural networks (DNNs) [39] can be categorized into preprocessing-based and adversarial-training-based methods [40]. Preprocessing-based defences aim to remove adversarial perturbations from the input images, employing JPEG Compression, Pixel Deflection, and Total Variance Minimization techniques. Another preprocessing approach involves adding randomness to the input to observe the variance in the outputs, exemplified by Random Smoothing, which uses Gaussian noise. On the other hand, adversarial training-based defences focus on retraining neural networks with adversarial samples to improve their robustness. Adversarial training-based defences include InceptionV3adv and InceptionResNetV2adv [41, 42], specifically designed to withstand adversarial attacks. Despite these efforts, adversarial attacks, especially those targeting the attention mechanisms of DNNs, have shown effectiveness even against these advanced defence strategies, indicating the ongoing challenge in securing DNNs against such threats.

Threats to Robustness One significant threat to robustness is adversarial attacks, where small, carefully designed changes to the input can trick the AI into making wrong decisions [19]. Techniques like LIME and SHAP are criticized for their vulnerability to adversarial attacks, allowing for arbitrary explanations[24]. Another threat is bias, where the AI might not treat all groups of users' data fairly because it learned from unfair data [19]. Another significant threat is data poisoning attacks, where attackers inject malicious data into the model's training set, leading to compromised performance [10]. Additionally, using black-box models, which are complex and not fully understood even by their creators, can also threaten robustness [10]. Lastly, the balance between accuracy, explainability, and tractability in AI models is delicate, and focusing too much on one aspect can compromise the others, affecting the overall robustness of the system [10].

Effectiveness and trade-offs of robustness-enhancing techniques Adversarial training and model retraining strategies enhance robustness but may decrease model accuracy due to overfitting adversarial examples [13]. Implementing these techniques can be computationally expensive and complex, potentially limiting their applicability in resource-constrained environments [30]. Additionally, there is a contradiction between enhancing model explanation and adversarial robustness, as advanced AI methods could potentially aid attackers in creating more effective adversarial examples [43].

2.6.4 Human Oversight and Explainable AI (XAI)

Promoting Explainable AI and Human-in-the-Loop Approaches The European Union promotes Explainable AI (XAI) to make AI decisions understandable to humans,

Table 4. Overview of Taxonomy in Machine Learning Security

Taxonomy	**Sub Points**	**Method / Attack / Defense**	**References**
Targets	Classification of Targets	Physical Domain of input sensors Digital Representation for pre-processing Machine Learning Model	[6, 10]
Methods of ML	Supervised learning		[7, 34]
	Unsupervised learning	-	[34]
	Reinforcement learning	-	[44]
	Examples of Attack on ML Model		[8]
Adversarial Attacks	Attacks Definition		[8]
	Types of Attacks	Evasion attack	[35, 38]
		Inference attacks	
		Poisoning attacks	[34, 38, 13]
	Classification of Attacks based on Knowledge	Black-box attack	[30, 45, 40, 43]
		White-box attack	
		Grey-box attack	
Defense	Defenses Definition		
	Defense Methods	Data sanitization	[36, 30, 38, 13]
		Defensive distillation	
		Adversarial Training	
		Generative Adversarial Networks (GANs)	
		Preprocessing-based defense methods	
		Adversarial-training-based defense methods	

enabling human-in-the-loop solutions that combine human and machine intelligence for sustainable and responsible outcomes [46]. Human-in-the-loop approaches enhance machine learning systems' functionality and reliability by facilitating human interaction. They aid dataset exploration, understanding, model validation, and trustworthiness through transparency and interpretable classifiers [47]. Human pattern recognition significantly improves data splits, underscoring the importance of human intuition and understanding in enhancing machine learning outcomes [47]. Decision systems are crucial to AI explainability as they help users understand how AI models make decisions, providing insights into the reasoning behind AI decisions and increasing transparency and trust in AI technologies [48].

Enhancing Trust and Understanding in AI Systems Decision Support Systems (DSS) can boost trust among healthcare providers by explaining the reasons behind decisions, especially when they aren't immediately clear. Understanding the sociocultural environment and addressing concerns about AI replacement is crucial for its acceptance. DSS can prevent mistakes by novices and support collaborative decision-making, enhancing patient care and healthcare efficiency [49, 18]. Involvement of human feedback in the development of AI explanations helps refine the explanatory process, making it more intuitive and satisfactory for users who may not have a technical background [9]. Non-experts often find it challenging to understand the decision-making process of complex "black box" algorithms due to their lack of transparency [50]. Research shows that a causality-based predictive model is interpretable without needing additional models, making it accessible to both experts and non-experts alike [50]. A user study with 18 non-expert participants revealed that the interface made it easy to grasp the model's workings, proving effective in demystifying complex algorithmic decisions for those without specialized knowledge of machine learning.

Tools, Customization, and Presentation in AI Fairness Current tools for assessing AI fairness, like AI Fairness 360 and FairVis, are primarily designed for data scientists, neglecting the needs of ordinary end-users. This lack of interfaces for non-technical users highlights a missed opportunity to include diverse perspectives, including cultural values like masculinity and uncertainty avoidance. Research suggests that creating interactive interfaces for end-users could democratize the fairness assessment process, allowing a wider audience to identify and mitigate biases in AI systems [51].

Customization and Effective Presentation in XAI Customization involves modifying something to meet specific requirements or preferences, such as adjusting a service or system to fit individual needs. In Explainable Artificial Intelligence (XAI), customization can involve adjusting AI systems to provide personalized explanations or adapt their functionality for different users or applications [4]. The effectiveness of AI explanations depends on their presentation [52, 53, 54]. Different formats like visualizations or textual representations affect users' backgrounds and expertise levels [55, 56, 57, 58]. Simpler formats like diagrams or natural language can demystify complex AI decisions, making them more accessible. The choice of presentation method should align with user preferences and understanding capacity. The complexity of explanations, whether through rule sets, decision trees, or graphical representations, must be tailored to match the user's comprehension level, boosting interpretability and fostering user trust. This customization is crucial for enhancing AI system interpretability.

User Interface (UI), User Experience (UX), and Interactive Techniques User Interface (UI) refers to the appearance and layout of a product, focusing on user interaction and quality. User Experience (UX), on the other hand, refers to the overall user experience, including ease and satisfaction, efficiency, dependability, and meeting user needs and expectations [59, 60]. Both aspects are crucial in creating a user-friendly and effective product [61].

Visual and Interactive Techniques Visual and interactive techniques in explanations leverage the human ability to recognize patterns and understand complex information quickly, making these explanations more accessible and effective [14, 50, 16]. Interactive explanations allow users to engage with the information, ask questions, and explore different outcomes, making the explanation process more effective by promoting a deeper understanding of the subject matter [14, 53].

3 Methodology

This chapter outlines a structured methodology for this study. The methodology is in four phases: the integration of explainable AI (XAI) techniques in a standard machine learning pipeline (see Figure 5(a)), simulation of adversarial attacks, the development of human oversight dashboards, and a user study to evaluate the effectiveness of AI explanation on the implemented system.

3.1 System Design

The system architecture, illustrated in Figure4, integrates various services centred around explainable AI (XAI). It includes a client-side and a backend server that handles data flow between the user and the machine learning pipeline. The sequence diagram in Fig3 illustrates the architecture's user authentication and request processing flow. In the process, the user logs in via the frontend application, which sends the login request to Okta for authentication. Okta processes the authentication and returns a success or error message to the front end. Upon successful authentication, the front end allows user access and sends a request to the API Gateway. The API Gateway validates the request and routes it to the appropriate microservice. The microservice processes the request, sending the response back through the API Gateway to the front end, which then delivers the response to the user. This design ensures secure, scalable, and maintainable user authentication and service request handling. This architecture enhances existing ML/FL systems by extending the conventional machine learning pipeline with steps to assess trustworthiness properties using XAI and specialised metrics (see Figure 9(b)). Integrating these metrics enhances AI model robustness, as relying solely on XAI (see Figure 9(b)) is insufficient for accurately identifying drifts and performance errors or their root causes [1]. The verification process involves developing a diagnostic profile to characterise and measure AI performance.

3.2 System implementation

3.2.1 Data selection and preprocessing

This research uses Montimage's metrics selection and raw network traffic datasets from various network experimental testbeds to construct a standard machine-learning pipeline 5a. This research uses Montimage's selection of metrics and raw network traffic datasets from various network experimental testbeds to create a machine-learning pipeline. It integrates eXplainable AI (XAI) methods like LIME and SHAP to improve the interpretability of AI models. The goal is to adjust trade-offs over time dynamically. The approach follows a correct-by-construction method, guided by stakeholder feedback. The development of a dashboard aims to provide precise insights into why specific

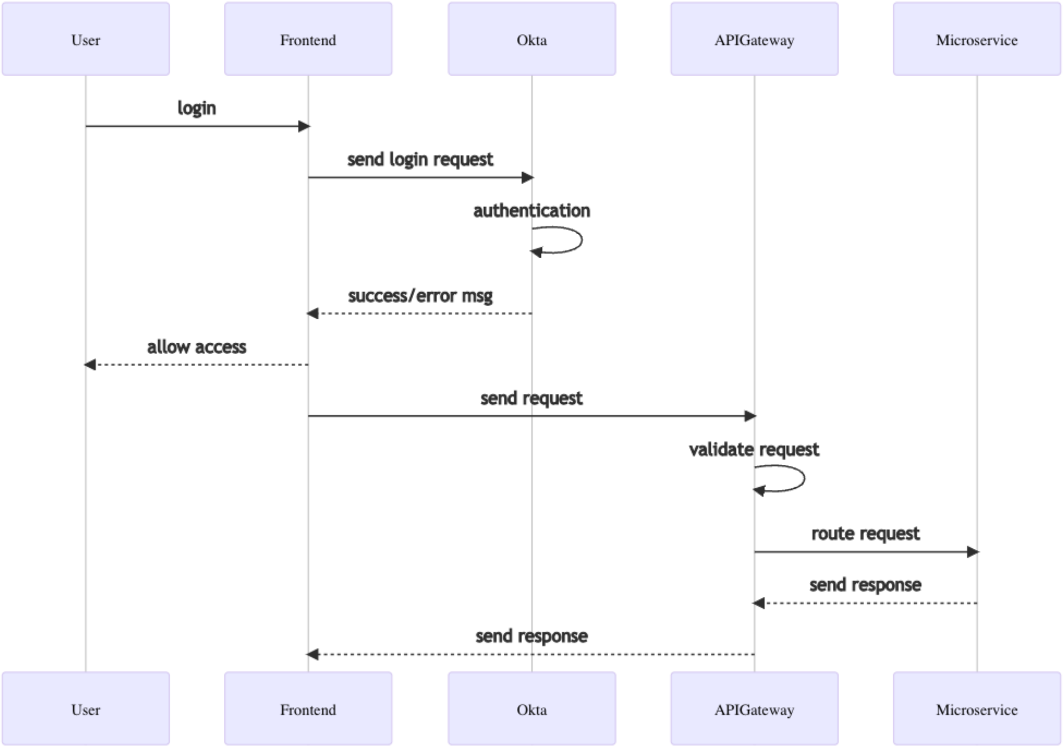


Figure 3. Sequence diagram of the SPATIAL architecture

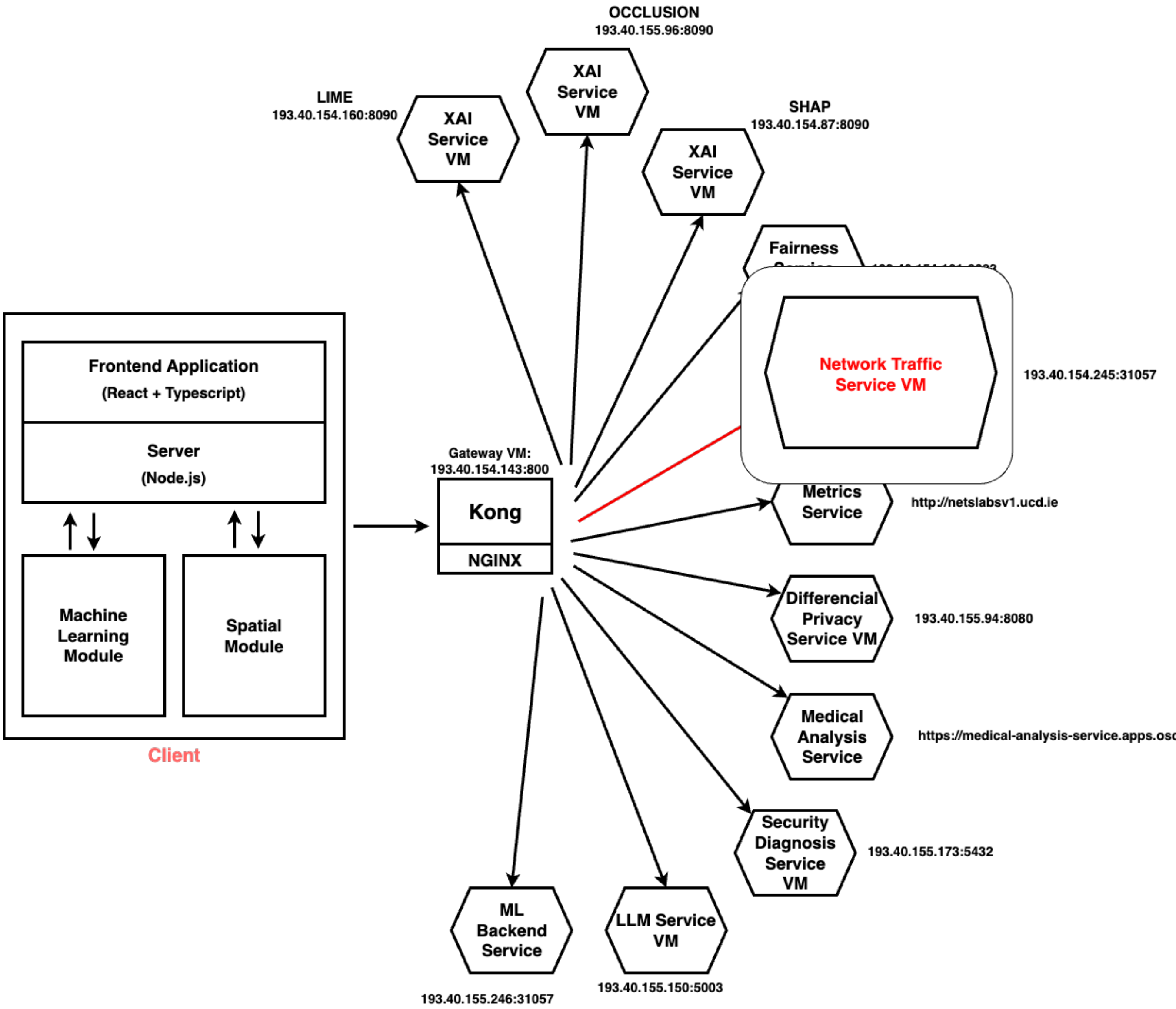


Figure 4. Overall underlying deployment of SPATIAL [1]

models are susceptible to certain threats, allowing stakeholders to understand adversarial tactics better and identify system vulnerabilities.

3.2.2 Machine learning pipeline

AI models systematically use the pipe to build and improve performance and security over time. Figure 5(a) deps of developing an AI model. The approach starts by converting an imbalanced dataset into a balanced one using raw network traffic data from several networks and a specialised experimental testbed provided by Montimage [62]. Data cleaning imputes missing data, eliminates duplication, and implements data enrichment techniques. Extracted features are converted into numerical vectors and standardised to prepare the complex network data for AI processing.

The dataset, comprising 382 distinct characteristics, is categorised into web, interactive, and video activities. Algorithms such as Neural Networks, LightGBM, and XGBoost train the dataset to extract and select features efficiently. These models handle vectorised data exceptionally well and generate precise classifications based on user activities, significantly enhancing the system's usefulness in diverse network contexts. The model's performance is evaluated using precision score, F1-score, and recall to improve resilience against adversarial attacks.

This implementation employs numerous Python libraries for machine learning, including NumPy [63], scikit-learn [64], TensorFlow [65], XGBoost [39], and LightGBM [66]. Tools like SHAP and LIME [26] implement explainable AI techniques, making the AI's decision-making process more transparent and understandable.

The platform's API gateway allows stakeholders to interact with these advanced machine-learning functionalities (see Figure 4). This gateway seamlessly integrates the client-side application and the backend services, providing an interface for request handling and data processing. Acting as a reverse proxy, the gateway enables secure data exchanges and enhances scalability by efficiently managing requests. Installing the Kong API Admin on a virtual machine optimises API management, allowing for service definition, routes, and security settings for external clients. Updating the Nginx [67] configuration ensures accurate routing through the Kong Gateway, effectively handling request paths. This setup supports high concurrency and low latency, enhancing scalability and providing efficient data exchange.

3.2.3 Human-Oversight Dashboard in XAI

Overview

This implementation includes an AI dashboard that enables users to understand and evaluate the trustworthy characteristics of AI models quantitatively. The dashboard facilitates modifications to either the AI model or the data upon which it is trained, resulting in a

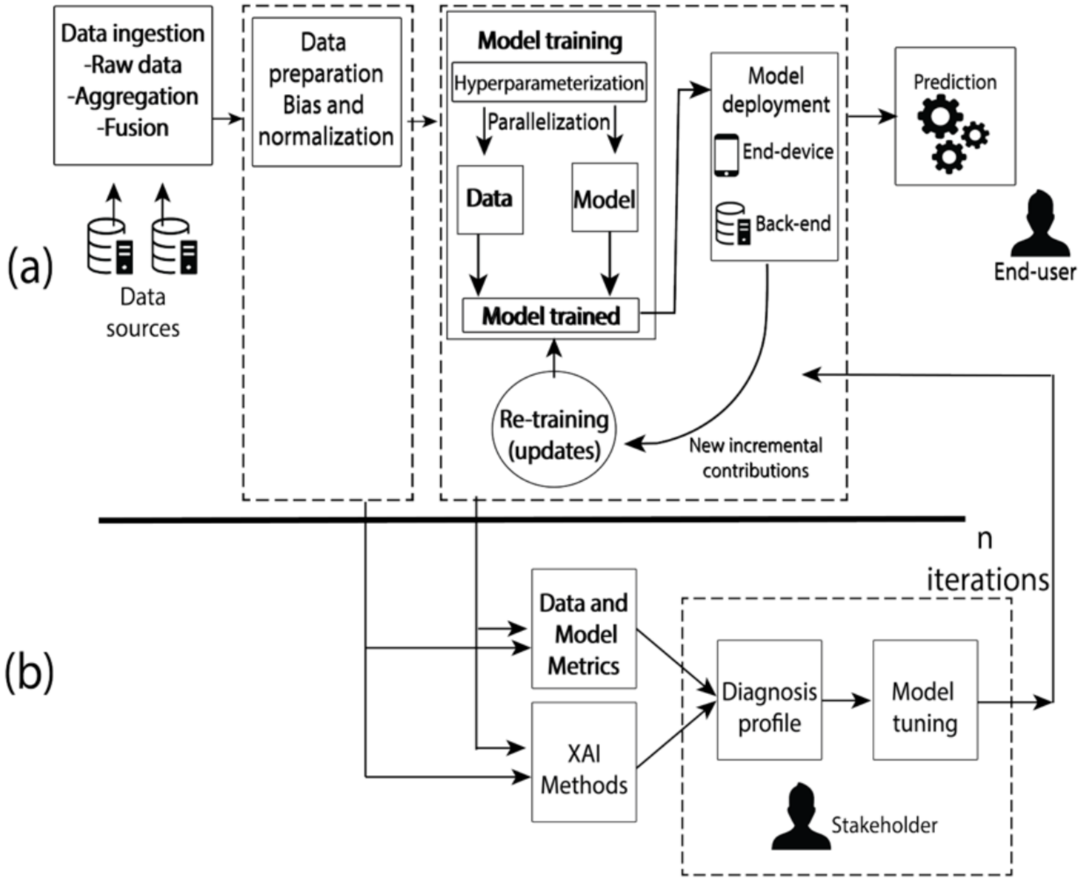


Figure 5. Standard pipeline to build AI models based on machine learning [1]

new version of the model or data and allows for iterative enhancements in subsequent versions. Additionally, it features a comparison tool that enables users to juxtapose different trustworthy properties from various versions of AI models or datasets. This tool is handy for assessing changes and improvements over time, providing a clear visual representation of how modifications have impacted the trustworthiness and performance of the AI system. This comprehensive approach enhances user interaction and bolsters the system's capability to deliver secure, reliable, and user-friendly functionality.

Front-End Development

The front end of the SPATIAL system is developed using React, a popular JavaScript library known for its efficiency and flexibility in building interactive user interfaces [68]. React's architecture allows for a modular approach, enabling developers to construct a dynamic and responsive user experience that integrates seamlessly with the various features of SPATIAL. To support React's development environment, Node.js is the underlying runtime, facilitating essential development tools like Babel and Webpack [69]. Babel helps transform modern JavaScript code into a backwards-compatible version for older browsers, while Webpack bundles various modules into static assets [70, 71].

For styling, the implementation leverages Bootstrap 5, a robust framework designed for developing responsive and mobile-first websites [72]. Bootstrap 5 provides a range of pre-designed components that can be easily customised and are responsive across devices. Alongside Bootstrap, Tailwind CSS is used for its utility-first approach to styling, allowing developers to apply styles directly in HTML, thereby enhancing the efficiency of creating custom designs and maintaining style consistency throughout the application [73].

Security and User Management

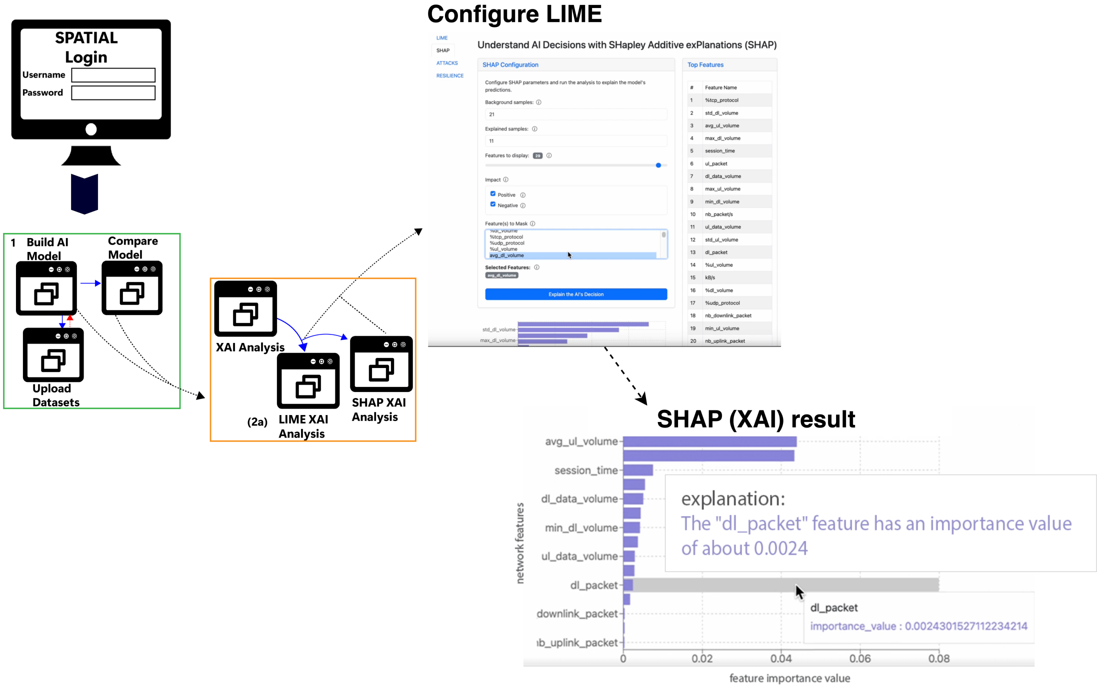
The integration of Okta addresses security and user management as critical components of the front end. Okta provides a comprehensive identity management solution, offering secure and robust authentication and authorisation services [74]. This integration ensures that only authenticated users can access certain features, safeguarding sensitive information and enhancing overall system security.

Key Features of the Dashboard

The end-to-end SPATIAL dashboard includes several vital interfaces:

- **Login Page**: Serves as the initial point of interaction for users, featuring username and password fields, a "Remember Me" checkbox for convenience on future visits, and a "Sign In" button alongside links for password retrieval and user support.

- **Build Model Page**: Allows users to set parameters for model creation, focusing on service types like "Network Traffic," where users can select datasets, define training and validation splits, and start the model-building process.
- **Model List**: Refer to Figures 9 and 8. This displays all AI models by ID and service type, with functionalities to view, download, or select models for further actions, including filtering options and access to associated datasets.
- **SPATIAL Dashboard**: Refer to Figures 9 and 8. This evaluates and compares AI model performance through model selection and configuration tools, featuring a confusion matrix for detailed performance analysis.
- **Adversarial Attack Interface**: Refer to Figures 9 and 8. This interface visualizes the impact of feature changes on AI behavior using graphs and provides LIME and SHAP values [75, 76] for deeper insights.
- **Resilience Metrics**: Refer to Figure 8. It evaluates model resilience against specific adversarial attacks, offering simulation tools that show performance metrics before and after attacks.
- **LIME Config and LIME Result Interfaces**: Provide detailed explanations of AI decisions using the LIME method, highlighting the influence of various features and aiding in transparency with visual outputs like pie charts. Refer to Figures 6 and 7
- **Compare XAI Result Section**: Allows users to contrast explanations from different XAI methods, ensuring clarity and reliability in interpreting AI decisions across models (See Fig9).



Understand AI Decisions with SHapley Additive exPlanations (SHAP)

SHAP Configuration

Configure SHAP parameters and run the analysis to explain the model's predictions.

Background samples: 21

Explained samples: 11

Features to display: 29

Impact

☑ Positive

☑ Negative

Feature(s) to Mask: %dl_volume, %tcp_protocol, %udp_protocol, %ul_volume, avg_dl_volume

Selected Features: avg_dl_volume

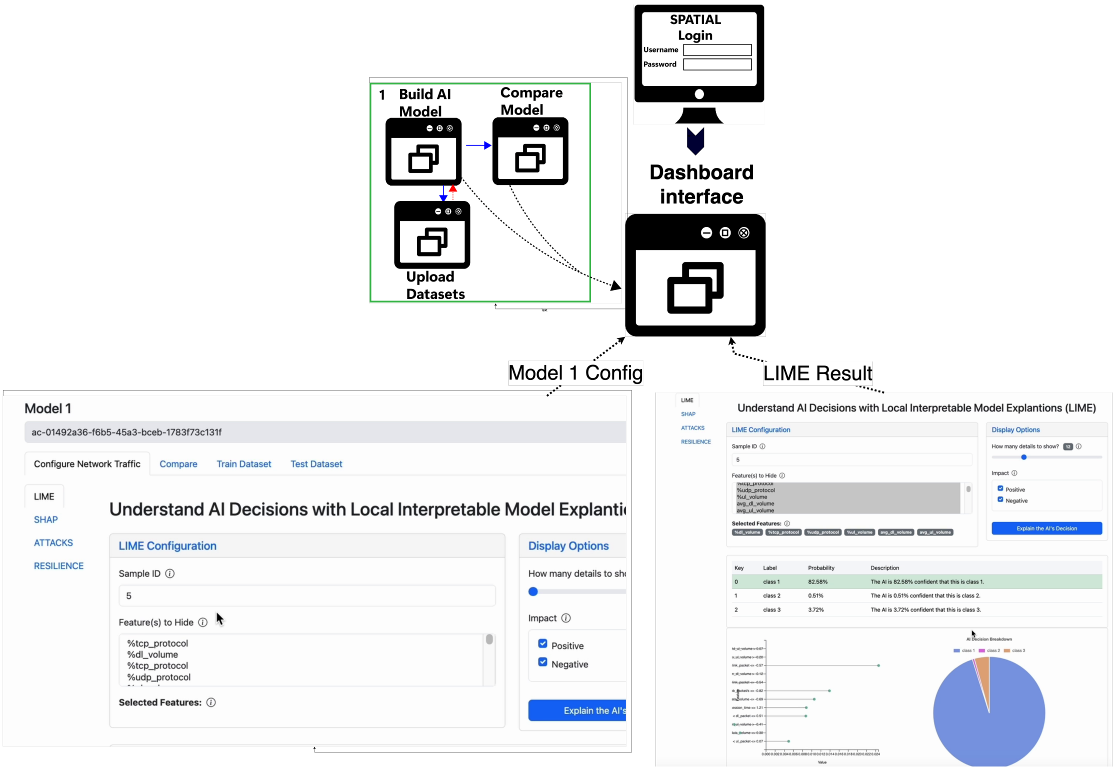
Top Features

#	Feature Name
1	%tcp_protocol
2	std_dl_volume
3	avg_ul_volume
4	max_dl_volume
5	session_time
6	ul_packet
7	dl_data_volume
8	max_ul_volume
9	min_dl_volume
10	nb_packet/s
11	ul_data_volume
12	std_ul_volume
13	dl_packet
14	%ul_volume
15	kB/s
16	%dl_volume
17	%udp_protocol
18	nb_downlink_packet
19	min_ul_volume
20	nb_uplink_packet

explanation:
The "dl_packet" feature has an importance value of about 0.0024

dl_packet
importance_value : 0.0024301527112234214

Figure 6. SHAP configuration and result



Model 1

ac-01492a36-f6b5-45a3-bceb-1783f73c131f

Understand AI Decisions with Local Interpretable Model Explanti

LIME Configuration

Sample ID ⓘ: 5

Feature(s) to Hide ⓘ:
%tcp_protocol
%dl_volume
%tcp_protocol
%udp_protocol

Selected Features: ⓘ

Display Options

How many details to sh

Impact ⓘ
☑ Positive
☑ Negative

Understand AI Decisions with Local Interpretable Model Explantions (LIME)

LIME Configuration

Sample ID ⓘ: 5

Feature(s) to Hide ⓘ:
%tcp_protocol
%udp_protocol
%ul_volume
avg_dl_volume
avg_ul_volume

Selected Features: ⓘ %dl_volume %tcp_protocol %udp_protocol %ul_volume avg_dl_volume avg_ul_volume

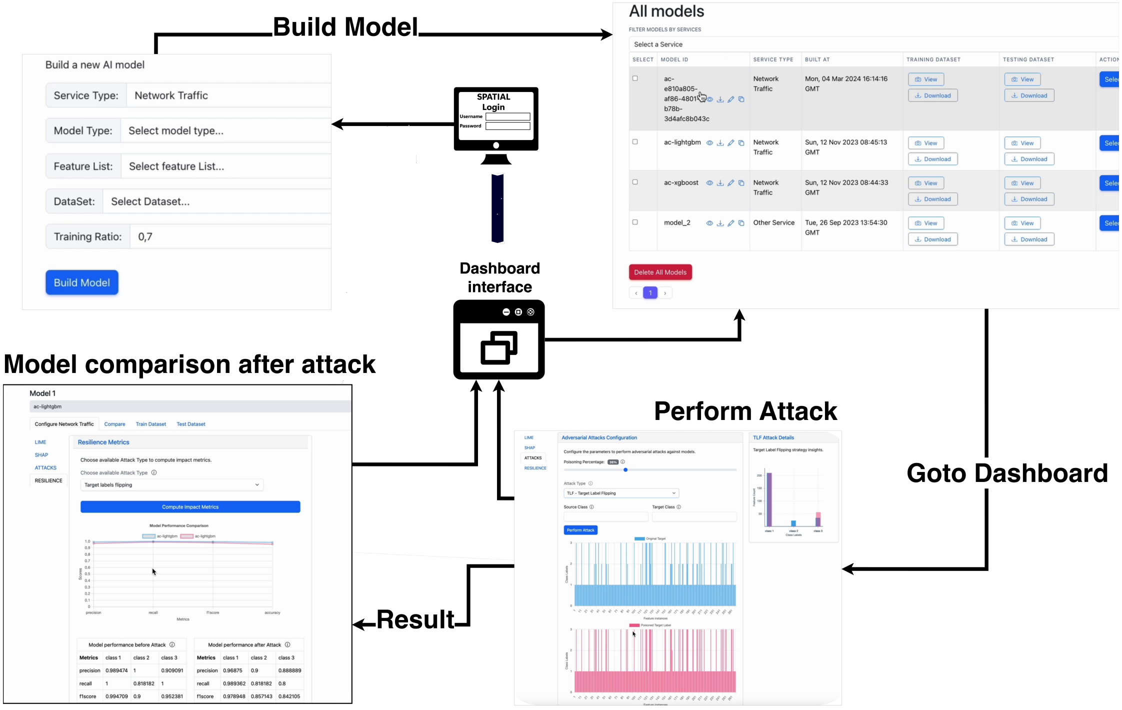
Display Options

How many details to show? 12 ⓘ

Impact ⓘ
☑ Positive
☑ Negative

Key	Label	Probability	Description
0	class 1	82.58%	The AI is 82.58% confident that this is class 1.
1	class 2	0.51%	The AI is 0.51% confident that this is class 2.
2	class 3	3.72%	The AI is 3.72% confident that this is class 3.

Figure 7. LIME configuration and result



Build a new AI model

Service Type: Network Traffic

Model Type: Select model type...

Feature List: Select feature List...

DataSet: Select Dataset...

Training Ratio: 0,7

All models

FILTER MODELS BY SERVICES

Select a Service

SELECT	MODEL ID	SERVICE TYPE	BUILT AT	TRAINING DATASET	TESTING DATASET	ACTION
□	ac-e810a805-af86-4801-b78b-3d4afc8b043c	Network Traffic	Mon, 04 Mar 2024 16:14:16 GMT	View Download	View Download	Selec
□	ac-lightgbm	Network Traffic	Sun, 12 Nov 2023 08:45:13 GMT	View Download	View Download	Selec
□	ac-xgboost	Network Traffic	Sun, 12 Nov 2023 08:44:33 GMT	View Download	View Download	Selec
□	model_2	Other Service	Tue, 26 Sep 2023 13:54:30 GMT	View Download	View Download	Selec

Model 1

ac-lightgbm

Choose available Attack Type to compute impact metrics.

Choose available Attack Type: Target labels flipping

Model performance before Attack			
Metrics	class 1	class 2	class 3
precision	0.989474	1	0.909091
recall	1	0.818182	1
f1score	0.994709	0.9	0.952381

Model performance after Attack			
Metrics	class 1	class 2	class 3
precision	0.96875	0.9	0.888889
recall	0.989362	0.818182	0.8
f1score	0.978948	0.857143	0.842105

Configure the parameters to perform adversarial attacks against models.

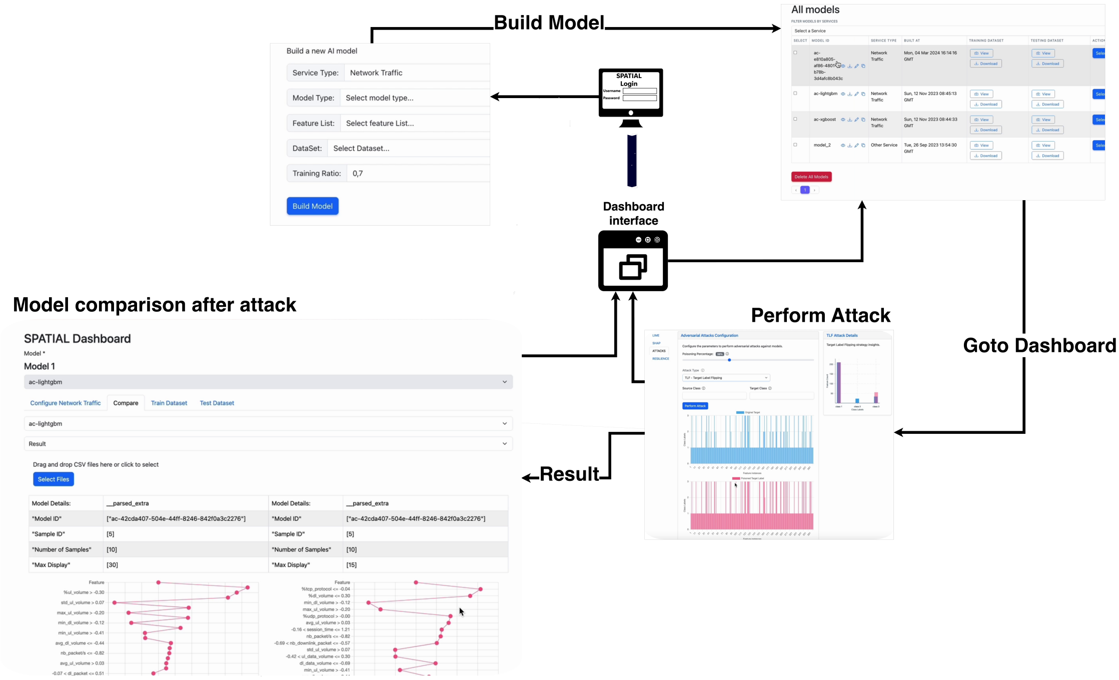
Poisoning Percentage: 36%

Attack Type: TLF - Target Label Flipping

Source Class:

Target Class:

Figure 8. Adversarial Attack Lifecycle using specialised metrics



Build a new AI model

Service Type: Network Traffic

Model Type: Select model type...

Feature List: Select feature List...

DataSet: Select Dataset...

Training Ratio: 0,7

All models

FILTER MODELS BY SERVICES

Select a Service

SELECT	MODEL ID	SERVICE TYPE	BUILT AT	TRAINING DATASET	TESTING DATASET	ACTION
□	ac-e810a805-af86-4801-b78b-3d4afc8b043c	Network Traffic	Mon, 04 Mar 2024 16:14:16 GMT	View Download	View Download	Selec
□	ac-lightgbm	Network Traffic	Sun, 12 Nov 2023 08:45:13 GMT	View Download	View Download	Selec
□	ac-xgboost	Network Traffic	Sun, 12 Nov 2023 08:44:33 GMT	View Download	View Download	Selec
□	model_2	Other Service	Tue, 26 Sep 2023 13:54:30 GMT	View Download	View Download	Selec

SPATIAL Dashboard

Model *

Model 1

ac-lightgbm

ac-lightgbm

Result

Drag and drop CSV files here or click to select

Model Details:	__parsed_extra	Model Details:	__parsed_extra
"Model ID"	["ac-42cda407-504e-44ff-8246-842f0a3c2276"]	"Model ID"	["ac-42cda407-504e-44ff-8246-842f0a3c2276"]
"Sample ID"	[5]	"Sample ID"	[5]
"Number of Samples"	[10]	"Number of Samples"	[10]
"Max Display"	[30]	"Max Display"	[15]

Figure 9. Adversarial Attack Lifecycle using XAI

Implementation Summary

The implementation features the SPATIAL AI dashboard, designed to enhance the trustworthiness of AI models through quantitative evaluation and iterative improvements. Users can modify AI models or their training data and utilise a comparison tool to assess attributes across versions. Developed using React for its efficient and interactive UI capabilities and supported by Node.js, the system uses Babel for browser compatibility and Webpack for module bundling. Interface styling is managed with Bootstrap 5 and Tailwind CSS, ensuring responsiveness and customisability. Okta integration addresses security and user management, which are critical components of the front end. Okta provides secure and robust authentication and authorisation services, ensuring that only authenticated users can access certain features. This approach safeguards sensitive information and enhances overall system security.

3.3 User Study Design

This research uses an experimental user study to explore how different levels of explanation impact user perceptions of AI decision-making.(See Appendix II) Three distinct conditions are compared:

In Group A, participants encounter AI decisions without any accompanying explanation, allowing an assessment of how AI outputs are interpreted without contextual data.

In Group B, participants receive short, concise explanations (see Figure 15), which communicate the rationale behind the AI's decisions, providing a basis for understanding the effects of minimal explanatory detail.

Finally, in Group C, participants receive long, detailed explanations (see Figure 16), delving deeply into the reasoning and data behind the AI's decisions.

This experiment explores how comprehensive information influences understanding and trust. By examining these three systems, this thesis aims to contribute valuable insights into how different levels of explanatory detail in AI systems can affect user outcomes, potentially guiding future designs of AI interfaces and decision-support systems.

3.3.1 Participants

Participants were carefully selected to ensure diverse representation, including technical and non-technical backgrounds, to capture a wide range of AI experiences. Participants were recruited through social media platforms, professional networks, and academic list serves. The target demographic included approximately 54 participants, with efforts to balance the group in terms of age, gender, and educational background. Participants were divided into two categories, A and B. Category A consisted of 32 participants, further divided into two subgroups: those who received no explanations and those who received short explanations. Category B included 18 participants from Category A's short explanation subgroup and 18 participants who received more detailed descriptions. A pre-screening process with a preliminary questionnaire ensured that all participants had a basic understanding of AI. Furthermore, a thorough consent process was conducted, where participants agreed to consent, clearly outlining the study's scope, the procedures for ensuring the anonymity of responses, and the protocols for handling personal data, safeguarding participant privacy, and maintaining ethical standards throughout the study.

3.3.2 Apparatus

Google Forms was chosen as the survey platform due to its accessibility and reliability, as well as its support for various question types and branching logic. Customising questions for each of the three distinct AI explanation systems under investigation is crucial. The survey was designed to include a demographic section that captures essential background information and specific questions that assess comprehension, satisfaction, trust, and engagement with AI decisions.

3.3.3 Data Analysis

This section details visualisations to gather, process, and interpret data to assess how different AI explanations affect user perception. A blend of quantitative and qualitative approaches is used for comprehensive analysis. The quantitative component employs the Mann-Whitney Test across groups exposed to no/short and short/detailed explanations, alongside regression analysis, to investigate the influence of these explanations on trust and engagement, factoring in demographic variables as potential confounders. Simultaneously, qualitative data from open-ended survey questions are analysed using thematic analysis to uncover themes that reveal each type of analysed explanation's emotional and perceptual effects. This dual-methodology approach allows for a holistic understanding by integrating quantitative metrics with qualitative insights, enhancing the findings' depth. This report integrates statistical outcomes with narrative insights, detailing the differences in explanation effectiveness and the reasons behind these differences, offering actionable insights for optimising AI explanations to suit diverse user groups.

3.3.4 Ethical Considerations

Ethical principles of confidentiality and transparency are adhered to in order to ensure the integrity of the research process and protect participant welfare. Data collected is anonymised, and personal identifiers are removed to prevent tracing back anonymised dual participants. Informed consent is obtained from all participants before they engage in the study. The consent form outlines the study's scope, participant involvement, and data confidentiality measures. This process ensures that ethical standards are upheld, maintaining the integrity and credibility of the research.

3.3.5 Limitations

This study has limitations, including generalizability, interactivity, and bias. The findings may not apply to all types of AI models or user demographics, as the visualisations are specific to the SHAP algorithm, and participants are visualisations of a specific demographic. Future research should consider a broader range of AI explanation methods and a more diverse participant pool to enhance generalizability. Additionally, participant responses may be influenced by subjective interpretations or biases towards technology, which could skew perceptions.

4 Results

4.1 Experiment Use Case

We discuss two primary types of attacks on network traffic classification systems: poisoning attacks during the training phase and evasion attacks during the testing or inference phase. We evaluate the resilience of an explanatory platform by analyzing its behaviour during these attack scenarios. To illustrate, we implemented a poisoning attack on a network activity classification dataset by altering the original dataset used to train the model. In this study, we executed three types of poisoning attacks: random label flipping, where we randomly changed labels to degrade the model's utility; targeted label poisoning, which involved selectively flipping labels to bias the dataset and shift the model's decision boundary; and GAN-based attacks aimed at diminishing the model's utility, creating a biased dataset targeting specific labels, and examining the model's internal reactions without disrupting its performance.

4.1.1 Analysis of Adversarial Attack Simulations

To evaluate the model's resilience to these attacks, we analyzed SHAP (Shapley Additive exPlanations) values for three scenarios: unpoisoned, randomly poisoned, and GAN-based poisoned models. We examined the impact of a new GAN-generated poisoning attack compared to conventional random and targeted poisoning attacks on a neural network-based activity classification model. Our study found that the GAN-generated attack causes less disruption because the new examples added are similar to the original dataset, thereby maintaining model utility. As a result, we observed more minor differences in accuracy between the original and the poisoned models, which translates to a lesser overall impact and makes detection more challenging. In contrast, the random and targeted poisoning attacks cause more significant deviations from the original model, leading to substantial disruptions and higher complexity metrics. This analysis highlights the value of using Explainable AI (XAI) metrics to detect and understand changes induced by poisoning attacks, thereby aiding in assessing model resilience.

This finding is crucial for stakeholders when deploying models, as it helps in planning resource allocation and defense strategies. Overall, this thesis offers a detailed view of the vulnerabilities of machine learning models to different types of adversarial attacks and highlights the importance of strong evaluation metrics to protect against these threats.

4.2 Results of User Study

4.2.1 Visualisation Without and With Text Explanation

As shown in Table 5, the statistical analysis for Group A, which compared no explanation to short explanations, demonstrated significant differences in several key areas. The

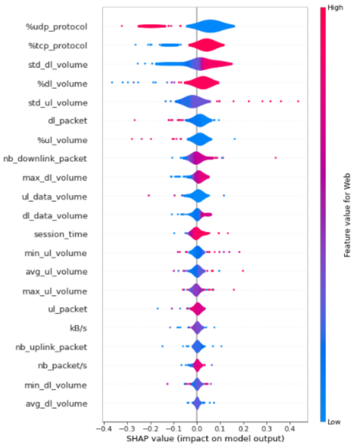


Figure 10. No poisoning

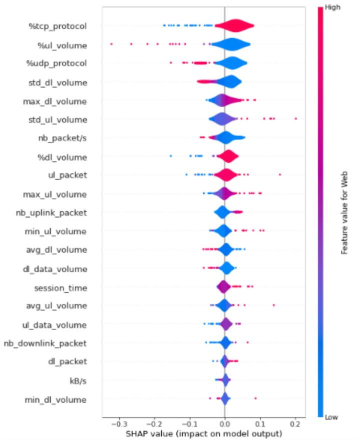


Figure 11. 50% random flipping

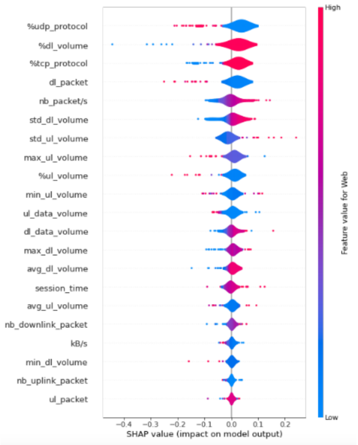


Figure 12. 50% GAN poisoning

Mann-Whitney U test applied to assess the clearness of the decision versus textual clarity resulted in a statistic of 101.0 and a p-value of 0.049, indicating a statistically significant difference at the 5% level. This finding suggests that participants' perceptions of decision clearness and textual clarity varied significantly. Additionally, the trustworthiness of the AI decision revealed substantial differences, with a statistic of 68.0 and a p-value of 0.0021, underscoring a notable variation in how trustworthy participants found the AI decisions. Engagement levels also showed significant differences, evidenced by a Mann-Whitney U statistic of 86.0 and a p-value of 0.0134.

4.2.2 Visualisation With Short and Long Explanation

The analysis for Group B (See Table 5), which compared short to long explanations, revealed no statistically significant differences across various metrics. Textual clarity, as assessed through how clearly participants understood the text accompanying the chart, showed no significant variation, with a Mann-Whitney U statistic of 203.0 and a p-value of 0.1824. Similarly, satisfaction levels with the explanations provided (U = 193.5, p = 0.3103), participants' trust in the AI decision (U = 147.0, p = 0.8423), engagement with the explanation format (U = 205.0, p = 0.1563), and ease of digesting the information (U = 185.5, p = 0.1773) did not differ significantly between the two groups.

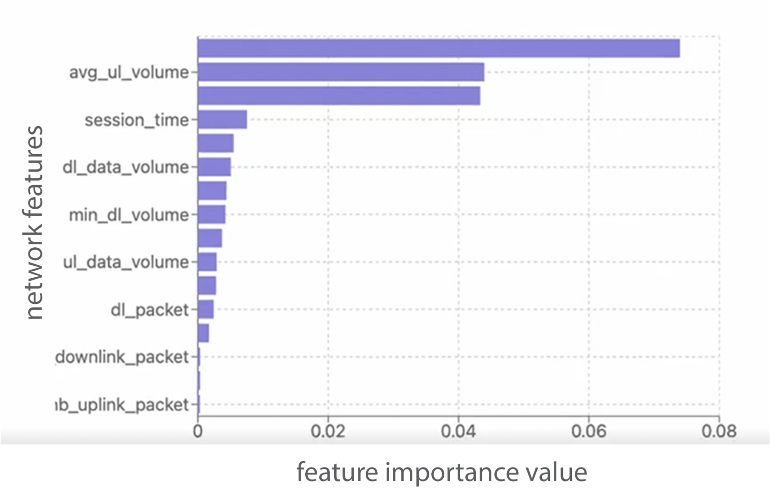
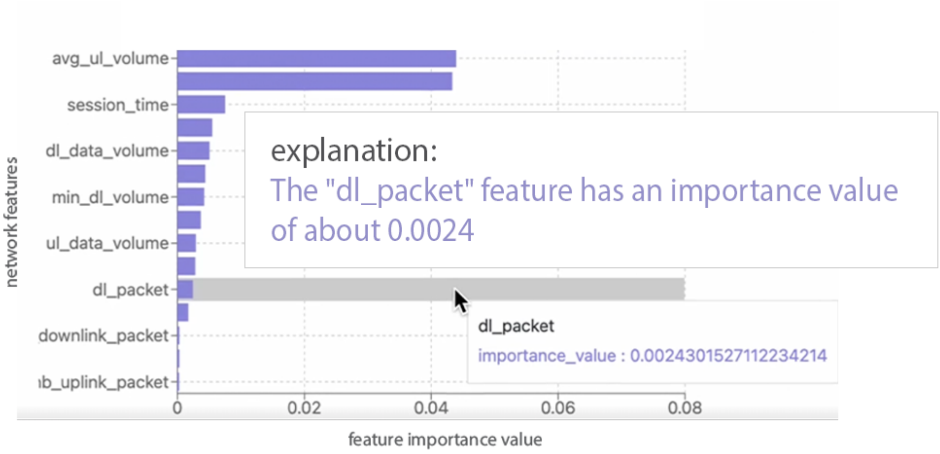


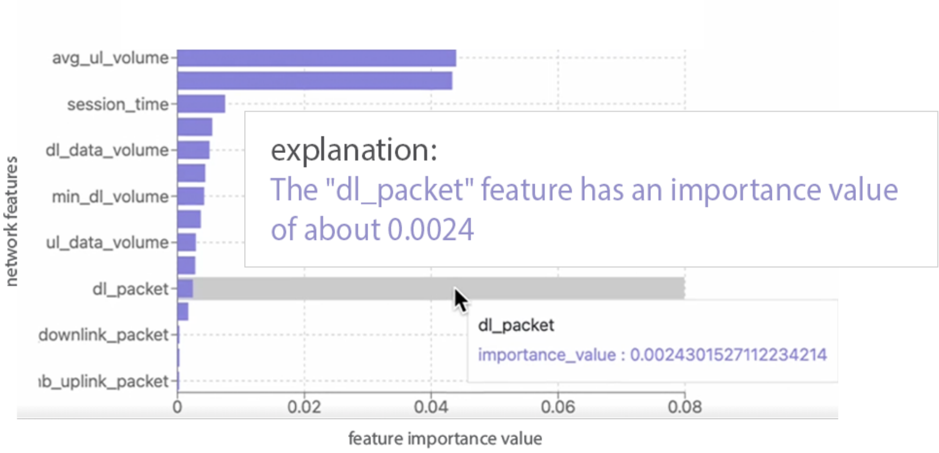
Figure 13. Visualisation with no Explanation



explanation:
The "dl_packet" feature has an importance value of about 0.0024

dl_packet
importance_value : 0.0024301527112234214

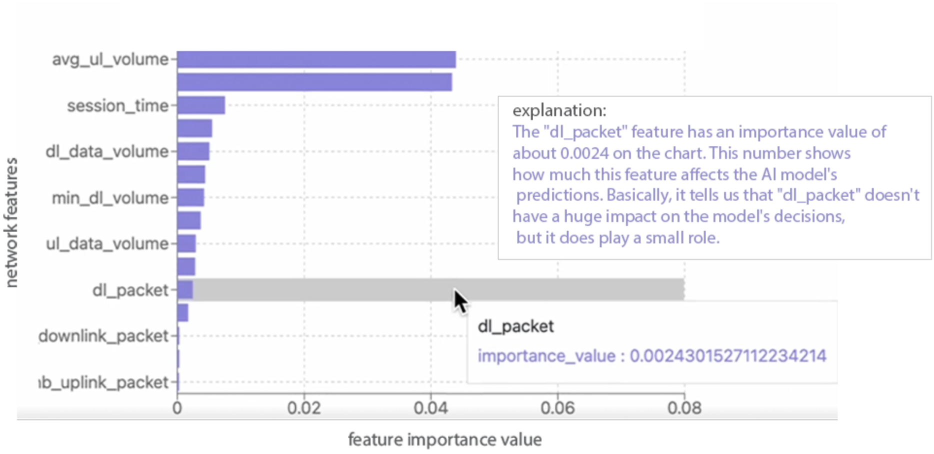
Figure 14. Visualisation with short Explanation



explanation:
The "dl_packet" feature has an importance value of about 0.0024

dl_packet
importance_value : 0.0024301527112234214

Figure 15. Visualization with short Explanation



explanation:
The "dl_packet" feature has an importance value of about 0.0024 on the chart. This number shows how much this feature affects the AI model's predictions. Basically, it tells us that "dl_packet" doesn't have a huge impact on the model's decisions, but it does play a small role.

dl_packet
importance_value : 0.0024301527112234214

Figure 16. Visualization with long Explanation

Table 5. Statistical Analysis Results for User Study

A (None vs Short Explanation)		
Metric	**U Statistic**	**p-Value**
Engagement	86.0	0.0134
Preference for Information	92.0	0.0225
Satisfaction	78.5	0.007
Textual Clarity	101.0	0.049
Trustworthiness	68.0	0.0021
B (Short vs Long Explanation)		
Metric	**U Statistic**	**p-Value**
Ease of Information Digestion	185.5	0.1773
Engagement	205.0	0.1563
Satisfaction	193.5	0.3103
Textual Clarity	203.0	0.1824
Trustworthiness of Decision	147.0	0.8423

5 Main Findings

This chapter presents the primary findings from an experimental user study designed to explore how different levels of explanation impact user perceptions of AI decision-making. The study compared three distinct conditions: no explanation (Group A), short explanations (Group B), and lengthy explanations (Group C). The results are based on statistical analyses, focusing on various metrics to evaluate the effectiveness and impact of the explanations provided.

Table 6. Statistical Analysis Results for User Study

A (None vs Short Explanation)			
Metric	**U Statistic**	**p-Value**	**Result**
Engagement	86.0	0.0134	Significant
Preference for Information	92.0	0.0225	Significant
Satisfaction	78.5	0.007	Significant
Textual Clarity	101.0	0.049	Significant
Trustworthiness	68.0	0.0021	Significant
B (Short vs Long Explanation)			
Metric	**U Statistic**	**p-Value**	**Result**
Ease of Information Digestion	185.5	0.1773	Not Significant
Engagement	205.0	0.1563	Not Significant
Satisfaction	193.5	0.3103	Not Significant
Textual Clarity	203.0	0.1824	Not Significant
Trustworthiness of Decision	147.0	0.8423	Not Significant

These findings indicate that even a short explanation significantly improves user engagement, preference for information, satisfaction, textual clarity, and trustworthiness compared to not explaining at all. This underscores the value of providing concise and clear explanations to enhance user experience.

In contrast, none of the metrics showed significant differences when comparing short explanations to long explanations (B: Short vs. Long Explanation). These suggest that increasing the length of the explanation does not significantly improve user outcomes over short explanations. Users do not necessarily benefit from more detailed explanations; they value clarity and relevance.

6 Discussion

The findings highlight the importance of explaining AI decisions. Compared to no explanations, short explanations significantly enhance user engagement, preference for information, satisfaction, textual clarity, and trustworthiness. However, increasing the length and detail of the explanations (short vs. long) shows no improvement in these metrics. These results suggest minimal explanations can effectively improve user perceptions of AI systems, and adding more detail may yield little benefit. Overly detailed explanations overwhelm users or make the information more challenging to digest, negating the benefits of providing more comprehensive insights. Users do not necessarily benefit from more detailed explanations; they value clarity and relevance over quantity of information.

The practical implications of these findings are significant. The study will prioritize creating explanations that are easy to understand and directly relevant to the user's context. This approach enhances the user experience and promotes greater acceptance and trust in AI technologies. The implementation can better support decision-making processes and improve overall satisfaction by focusing on the most critical information for users and presenting it straightforwardly.

Moreover, these insights can guide the development of best practices for AI interface design. Such practices may include standardized methods for generating explanations that strike the right balance between informative and accessible. As AI continues integrating into various sectors, these principles will ensure that the technology serves its intended purpose effectively and ethically. By fostering an environment where users feel informed and confident in the AI's capabilities, we can drive broader adoption and more effective utilization of AI systems.

6.1 Future Work

Future research should explore several areas to build on these findings. One key area is understanding the elements within short explanations that enhance user engagement and trust most effectively. Identifying these elements can help create even more refined and effective explanatory frameworks. Additionally, studies should examine the impact of explanations across different user demographics and contexts to determine if certain groups benefit more from detailed explanations than others. This could involve sector-specific investigations, such as in healthcare, finance, or education, where the nature of AI decisions and user needs may vary significantly.

Furthermore, research should also explore the role of interactive explanations, where users can choose the level of detail they want to receive. This could help balance the need for brevity and comprehensiveness, allowing users to tailor explanations to their preferences and requirements.

6.2 Conclusion

The study demonstrates that short, clear explanations significantly enhance user perceptions of AI systems compared to no explanations, without additional benefit from longer, more detailed explanations. These findings emphasize the importance of clarity and relevance in AI-generated explanations. By prioritizing user-friendly and contextually relevant information, AI developers can improve user engagement, satisfaction, and trust, fostering broader acceptance and more effective use of AI technologies. Future research should continue to refine these strategies, exploring specific elements that enhance explanatory effectiveness and tailoring explanations to diverse user needs.

References

[1] (2020) Detection mechanisms to identify data biases and exploratory studies about different data quality trade-offs for ai-based systems. Accessed: date-of-access. [Online]. Available: https://spatial-h2020.eu/d3-1-detection-mechanisms-to-identify-data-biases-and-exploratory-studies-about-different-data-qua

[2] E. Melo, I. Silva, D. Costa, C. Viegas, and T. Barros, "On the Use of eXplainable Artificial Intelligence to Evaluate School Dropout," *Educ. Sci.*, vol. 12, no. 12, 2022, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85144686899&doi=10.3390%2feducsci12120845&partnerID=40&md5=b09ee2ee9871b1dd48ca5fa8eaf82ac4

[3] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018, publisher: Institute of Electrical and Electronics Engineers Inc. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053352477&doi=10.1109%2fACCESS.2018.2870052&partnerID=40&md5=cb385461376b3fd0420c138ded6d133f

[4] B. Mohammed, "A Review on Explainable Artificial Intelligence Methods, Applications, and Challenges," *Indones. J. Electr. Eng. Informatics*, vol. 11, no. 4, pp. 1007–1024, 2023, publisher: Institute of Advanced Engineering and Science. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85183189918&doi=10.52549%2fijeei.v11i4.5151&partnerID=40&md5=2f016dfe7bf6939a45cdd3a081206efe

[5] I. Kok, F. Okay, O. Muyanli, and S. Ozdemir, "Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14 764–14 779, 2023, publisher: Institute of Electrical and Electronics Engineers Inc. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85162887757&doi=10.1109%2fJIOT.2023.3287678&partnerID=40&md5=772fe50b94fd51f5f40fbb7b41f3e23b

[6] J. Valtl and V. Issakov, "Universal Adversarial Attacks on the Raw Data From a Frequency Modulated Continuous Wave Radar," *IEEE Access*, vol. 10, pp. 114 092–114 102, 2022, publisher: Institute of Electrical and Electronics Engineers Inc. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141608438&doi=10.1109%2fACCESS.2022.3218349&partnerID=40&md5=6d0f68d6e42959690e8c96dc2ce192a8

[7] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases

and beyond," *Inf. Fusion*, vol. 77, pp. 29–52, 2022, publisher: Elsevier B.V. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127068702&doi=10.1016%2fj.inffus.2021.07.016&partnerID=40&md5=30808f642dec8ce43d426fc67aa356b6

[8] M. Zolanvari, Z. Yang, K. Khan, R. Jain, and N. Meskin, "TRUST XAI: Model-Agnostic Explanations for AI With a Case Study on IIoT Security," *IEEE Internet Things J.*, vol. 10, no. 4, pp. 2967–2978, 2023, publisher: Institute of Electrical and Electronics Engineers Inc. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124068081&doi=10.1109%2fJIOT.2021.3122019&partnerID=40&md5=ec47ff7dc2d5ecb6071f1ca03a9822ce

[9] F. Sovrano and F. Vitali, "Generating User-Centred Explanations via Illocutionary Question Answering: From Philosophy to Interfaces," *ACM Trans. Interact. Intelligent Syst.*, vol. 12, no. 4, 2022, publisher: Association for Computing Machinery. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85122367649&doi=10.1145%2f3519265&partnerID=40&md5=7ed1d7c5b2d606f5fe455f3db41df6c3

[10] M. Nagahisarchoghaei, N. Nur, L. Cummins, N. Nur, M. Karimi, S. Nandanwar, S. Bhattacharyya, and S. Rahimi, "An Empirical Survey on Explainable AI Technologies: Recent Trends, Use-Cases, and Categories from Technical and Application Perspectives," *Electronics (Switzerland)*, vol. 12, no. 5, 2023, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85149666852&doi=10.3390%2felectronics12051092&partnerID=40&md5=e9ffc21d2e6e4c94c9668ce705e20c96

[11] S. Samoili, M. L. Cobo, E. Gómez, G. De Prato, F. Martínez-Plumed, and B. Delipetrev, "Ai watch. defining artificial intelligence. towards an operational definition and taxonomy of artificial intelligence," 2020.

[12] National Institute of Standards and Technology (NIST), "A taxonomy and terminology of adversarial machine learning," NIST, Tech. Rep. 8269, 2019. [Online]. Available: https://csrc.nist.gov/pubs/ir/8269/ipd

[13] P. Biczyk and Wawrowski, "Detecting and Isolating Adversarial Attacks Using Characteristics of the Surrogate Model Framework," *Appl. Sci.*, vol. 13, no. 17, 2023, publisher: Multidisciplinary Digital Publishing Institute (MDPI). [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85170375917&doi=10.3390%2fapp13179698&partnerID=40&md5=2517ffd20dcce2c781eebd8761a59913

[14] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, “Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications,” *Inf. Fusion*, vol. 81, pp. 59–83, 2022, publisher: Elsevier B.V. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85120935859&doi=10.1016%2fj.inffus.2021.11.003&partnerID=40&md5=a02d66f00840b8163e0d22bbed530af5

[15] S. Sachan, J.-B. Yang, D.-L. Xu, D. Benavides, and Y. Li, “An explainable AI decision-support-system to automate loan underwriting,” *Expert Sys Appl*, vol. 144, 2020, publisher: Elsevier Ltd. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85075989564&doi=10.1016%2fj.eswa.2019.113100&partnerID=40&md5=5ec482b90ba8e4eee4ae693fa8d34dd6

[16] N. Scharowski, S. Perrig, M. Svab, K. Opwis, and F. Brühlmann, “Exploring the effects of human-centered AI explanations on trust and reliance,” *Frontier. Comput. Sci.*, vol. 5, 2023, publisher: Frontiers Media SA. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85175111420&doi=10.3389%2ffcomp.2023.1151150&partnerID=40&md5=c12c6a113ec39a5d9072c56be7854119

[17] C. Cai, J. Jongejan, and J. Holbrook, “The effects of example-based explanations in a machine learning interface,” in *Int Conf Intell User Interfaces Proc IUI*, vol. Part F147615. Association for Computing Machinery, 2019, pp. 258–262, journal Abbreviation: Int Conf Intell User Interfaces Proc IUI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065577583&doi=10.1145%2f3301275.3302289&partnerID=40&md5=d598433d5a952339ba1fb903dd733220

[18] A. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. Becker, and C. Mooney, “Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review,” *Appl. Sci.*, vol. 11, no. 11, 2021, publisher: MDPI AG. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107556681&doi=10.3390%2fapp11115088&partnerID=40&md5=32428f478e5780cc8ba68bd71a0055e8

[19] M. 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,” *Brief. Bioinform.*, vol. 24, no. 5, 2023, publisher: Oxford University Press. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85172424276&doi=10.1093%2fbib%2fbbad236&partnerID=40&md5=b47d9a8b5752511ab34aef8735a57ea3

[20] P. Hamm, M. Klesel, P. Coberger, and H. Wittmann, “Explanation matters: An experimental study on explainable AI,” *Electron. Mark.*, vol. 33, no. 1,

2023, publisher: Springer Science and Business Media Deutschland GmbH. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159139811&doi=10.1007%2fs12525-023-00640-9&partnerID=40&md5=6948a019167b01fb1a8d2125d9837e14

[21] W. Tan, J. Renkhoff, A. Velasquez, Z. Wang, L. Li, J. Wang, S. Niu, F. Yang, Y. Liu, and H. Song, "NoiseCAM: Explainable AI for the Boundary Between Noise and Adversarial Attacks," in *IEEE Int Conf Fuzzy Syst.* Institute of Electrical and Electronics Engineers Inc., 2023, journal Abbreviation: IEEE Int Conf Fuzzy Syst. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85178506370&doi=10.1109%2fFUZZ52849.2023.10309766&partnerID=40&md5=8acae8336d5cc410ecc022387793fd12

[22] G. Carbone, L. Bortolussi, and G. Sanguinetti, "Resilience of Bayesian Layer-Wise Explanations under Adversarial Attacks," in *Proc Int Jt Conf Neural Networks*, vol. 2022-July. Institute of Electrical and Electronics Engineers Inc., 2022, journal Abbreviation: Proc Int Jt Conf Neural Networks. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140725807&doi=10.1109%2fIJCNN55064.2022.9892788&partnerID=40&md5=fa51bea416880126b6dd5db303f1b770

[23] B. Kwon, M.-J. Choi, J. Kim, E. Choi, Y. Kim, S. Kwon, J. Sun, and J. Choo, "RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records," *IEEE Trans Visual Comput Graphics*, vol. 25, no. 1, pp. 299–309, 2019, publisher: IEEE Computer Society. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052616093&doi=10.1109%2fTVCG.2018.2865027&partnerID=40&md5=53aa10f57ad3d41f2147ef7d9845dbf6

[24] M. Sahakyan, Z. Aung, and T. Rahwan, "Explainable Artificial Intelligence for Tabular Data: A Survey," *IEEE Access*, vol. 9, pp. 135 392–135 422, 2021, publisher: Institute of Electrical and Electronics Engineers Inc. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117052096&doi=10.1109%2fACCESS.2021.3116481&partnerID=40&md5=5fbdf1396d0454a8008fbf66d108d609

[25] S. Chakraborty, S. Mittermaier, and C. Carbonelli, "Understanding the Behavior of Gas Sensors Using Explainable AI †," *Eng. Proc.*, vol. 27, no. 1, 2022, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85145377253&doi=10.3390%2fecsa-9-13350&partnerID=40&md5=99cc8082b74a9372bd29bec764e1aca3

[26] J. Aechtner, L. Cabrera, D. Katwal, P. Onghena, D. Valenzuela, and A. Wilbik, “Comparing User Perception of Explanations Developed with XAI Methods,” in *IEEE Int Conf Fuzzy Syst*, vol. 2022-July. Institute of Electrical and Electronics Engineers Inc., 2022, journal Abbreviation: IEEE Int Conf Fuzzy Syst. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138821769&doi=10.1109%2fFUZZ-IEEE55066.2022.9882743&partnerID=40&md5=a184bdf930cd911a8175c612d2680cf3

[27] J. Zacharias, M. von Zahn, J. Chen, and O. Hinz, “Designing a feature selection method based on explainable artificial intelligence,” *Electron. Mark.*, vol. 32, no. 4, pp. 2159–2184, 2022, publisher: Springer Science and Business Media Deutschland GmbH. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142888081&doi=10.1007%2fs12525-022-00608-1&partnerID=40&md5=e9bbf03e2a79c76946bb1b0125350c38

[28] M. Bhandari, T. Shahi, and A. Neupane, “Evaluating Retinal Disease Diagnosis with an Interpretable Lightweight CNN Model Resistant to Adversarial Attacks,” *J. Imaging*, vol. 9, no. 10, 2023, publisher: Multidisciplinary Digital Publishing Institute (MDPI). [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85175253859&doi=10.3390%2fjimaging9100219&partnerID=40&md5=e044a81370ea0ee35fb37a0990e546da

[29] H. Mehta and K. Passi, “Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI),” *Algorithms*, vol. 15, no. 8, 2022, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137146369&doi=10.3390%2fa15080291&partnerID=40&md5=f703b2d38a651c891f9ed142d954769f

[30] H. Naderi and I. Bajic, “Adversarial Attacks and Defenses on 3D Point Cloud Classification: A Survey,” *IEEE Access*, vol. 11, pp. 144 274–144 295, 2023, publisher: Institute of Electrical and Electronics Engineers Inc. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85181784265&doi=10.1109%2fACCESS.2023.3345000&partnerID=40&md5=31637ed5403923334f135b5ddf9c533d

[31] V. Khanna, K. Chadaga, N. Sampathila, R. Chadaga, S. Prabhu, S. K S, A. Jagdale, and D. Bhat, “A decision support system for osteoporosis risk prediction using machine learning and explainable artificial intelligence,” *Heliyon*, vol. 9, no. 12, 2023, publisher: Elsevier Ltd. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85179106514&doi=10.1016%2fj.heliyon.2023.e22456&partnerID=40&md5=bbf02f026d4caaa90027ef3d38ade3e4

[32] W. Kulasooriya, R. Ranasinghe, U. Perera, P. Thisovithan, I. Ekanayake, and D. Meddage, “Modeling strength characteristics of basalt fiber reinforced concrete using multiple explainable machine learning with a graphical user interface,” *Sci. Rep.*, vol. 13, no. 1, 2023, publisher: Nature Research. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85167772136&doi=10.1038%2fs41598-023-40513-x&partnerID=40&md5=bc3a74f9b14f2238fb085bb5792abf4b

[33] A. Laios, E. Kalampokis, R. Johnson, A. Thangavelu, C. Tarabanis, D. Nugent, and D. De Jong, “Explainable Artificial Intelligence for Prediction of Complete Surgical Cytoreduction in Advanced-Stage Epithelial Ovarian Cancer,” *J. Pers. Med.*, vol. 12, no. 4, 2022, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129109272&doi=10.3390%2fjpm12040607&partnerID=40&md5=58a47640c1e7ea9446a59083741baa82

[34] M. Altoub, F. AlQurashi, T. Yigitcanlar, J. Corchado, and R. Mehmood, “An Ontological Knowledge Base of Poisoning Attacks on Deep Neural Networks,” *Appl. Sci.*, vol. 12, no. 21, 2022, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141856900&doi=10.3390%2fapp122111053&partnerID=40&md5=dfcb3f0f12b23d27b4007de296ffd48a

[35] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, and M. Colajanni, “Modeling Realistic Adversarial Attacks against Network Intrusion Detection Systems,” *Digit. Threats: Res. Pract.*, vol. 3, no. 3, 2022, publisher: Association for Computing Machinery. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142517093&doi=10.1145%2f3469659&partnerID=40&md5=58e8add6ab9cadad065f689e8e89fa79

[36] H. Tang, F. Catak, M. Kuzlu, E. Catak, and Y. Zhao, “Defending AI-Based Automatic Modulation Recognition Models Against Adversarial Attacks,” *IEEE Access*, vol. 11, pp. 76 629–76 637, 2023, publisher: Institute of Electrical and Electronics Engineers Inc. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85165271753&doi=10.1109%2fACCESS.2023.3296805&partnerID=40&md5=6eb877cf24df9938b20a7ea90a3e1a7b

[37] S. Palacio, A. Lucieri, M. Munir, S. Ahmed, J. Hees, and A. Dengel, “XAI Handbook: Towards a Unified Framework for Explainable AI,” in *Proc IEEE Int Conf Comput Vision*, vol. 2021-October. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 3759–3768, journal Abbreviation: Proc IEEE Int Conf Comput Vision. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123055408&doi=10.1109%2fICCVW54120.2021.00420&partnerID=40&md5=723755686ce299652266f15e5ea7a679

[38] R. Taheri, R. Javidan, M. Shojafar, P. Vinod, and M. Conti, "Can machine learning model with static features be fooled: an adversarial machine learning approach," *Cluster Comput.*, vol. 23, no. 4, pp. 3233–3253, 2020, publisher: Springer. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85082771494&doi=10.1007%2fs10586-020-03083-5&partnerID=40&md5=8bae283ce4bb30d7cd8dde98555e4948

[39] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[40] S. Chen, Z. He, C. Sun, J. Yang, and X. Huang, "Universal Adversarial Attack on Attention and the Resulting Dataset DAmageNet," *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 4, pp. 2188–2197, 2022, publisher: IEEE Computer Society. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125883864&doi=10.1109%2fTPAMI.2020.3033291&partnerID=40&md5=5b6a1c4cbc2f88d7feb5776f4e985788

[41] A. Minagi, H. Hirano, and K. Takemoto, "Natural Images Allow Universal Adversarial Attacks on Medical Image Classification Using Deep Neural Networks with Transfer Learning," *J. Imaging*, vol. 8, no. 2, 2022, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124215538&doi=10.3390%2fjimaging8020038&partnerID=40&md5=eb8501b7843ba32523ec28a5f2ff8048

[42] K. Koga and K. Takemoto, "Simple Black-Box Universal Adversarial Attacks on Deep Neural Networks for Medical Image Classification," *Algorithms*, vol. 15, no. 5, 2022, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129593850&doi=10.3390%2fa15050144&partnerID=40&md5=0b5845c23e15222e6f6bd4460c68ca7a

[43] Y. Chai, R. Liang, S. Samtani, H. Zhu, M. Wang, Y. Liu, and Y. Jiang, "Additive Feature Attribution Explainable Methods to Craft Adversarial Attacks for Text Classification and Text Regression," *IEEE Trans Knowl Data Eng*, vol. 35, no. 12, pp. 12 400–12 414, 2023, publisher: IEEE Computer Society. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159647344&doi=10.1109%2fTKDE.2023.3270581&partnerID=40&md5=d4c5580679e983d4e9fbf278355b59f7

[44] C. Zheng, C. Zhen, H. Xie, and S. Yang, "Towards Secure Multi-Agent Deep Reinforcement Learning: Adversarial Attacks and Countermeasures," in *IEEE Conf. Dependable Secur. Comput., DSC SECSOC Workshop, PASS4IoT Workshop SICSA Int. Paper/Poster Compet. Cybersecur.* Institute of Electrical and

Electronics Engineers Inc., 2022, journal Abbreviation: IEEE Conf. Dependable Secur. Comput., DSC SECSOC Workshop, PASS4IoT Workshop SICSA Int. Paper/Poster Compet. Cybersecur. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141038022&doi=10.1109%2fDSC54232.2022.9888828&partnerID=40&md5=fb02f7a65f61cbee3b93abb40ccdd94c

[45] A. Singh, L. Awasthi, M. Shorfuzzaman, A. Alsufyani, and M. Uddin, "Chained Dual-Generative Adversarial Network: A Generalized Defense Against Adversarial Attacks," *Comput. Mater. Continua*, vol. 74, no. 2, pp. 2541–2555, 2023, publisher: Tech Science Press. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141892073&doi=10.32604%2fcmc.2023.032795&partnerID=40&md5=6a2e5b75d821c3f55be7a76ef1997b4b

[46] C. Turner, O. Okorie, and J. Oyekan, "XAI Sustainable Human in the Loop Maintenance," in *IFAC-PapersOnLine*, Barbieri B., Romero D., Emmanouilidis C., Parlikad A., and Sepideh S., Eds., vol. 55. Elsevier B.V., 2022, pp. 67–72, issue: 19 Journal Abbreviation: IFAC-PapersOnLine. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142235223&doi=10.1016%2fj.ifacol.2022.09.185&partnerID=40&md5=7a4fdda4cba42bc3de2092128d617092

[47] V. Estivill-Castro, E. Gilmore, and R. Hexel, "Constructing Explainable Classifiers from the Start—Enabling Human-in-the Loop Machine Learning," *Information*, vol. 13, no. 10, 2022, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140467816&doi=10.3390%2finfo13100464&partnerID=40&md5=22c05ee2e98766db109295f11c54f70a

[48] F. McCabe and C. Baber, "Feature Perception in Broadband Sonar Analysis - Using the Repertory Grid to Elicit Interface Designs to Support Human-Autonomy Teaming," in *ICMI - Proc. Int. Conf. Multimodal Interact.* Association for Computing Machinery, Inc, 2021, pp. 487–493, journal Abbreviation: ICMI - Proc. Int. Conf. Multimodal Interact. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119015591&doi=10.1145%2f3462244.3479918&partnerID=40&md5=a6451dcb92a848e534d116c5b91be665

[49] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, and D. Pedreschi, "FairLens: Auditing black-box clinical decision support systems," *Inf. Process. Manage.*, vol. 58, no. 5, 2021, publisher: Elsevier Ltd. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108332873&doi=10.1016%2fj.ipm.2021.102657&partnerID=40&md5=5baef6cbf23a8b6c54abb95fffcbc14d

[50] M. Hoque and K. Mueller, "Outcome-Explorer: A Causality Guided Interactive Visual Interface for Interpretable Algorithmic Decision Making," *IEEE Trans Visual Comput Graphics*, vol. 28, no. 12, pp. 4728–4740, 2022, publisher:

IEEE Computer Society. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85112636053&doi=10.1109%2fTVCG.2021.3102051&partnerID=40&md5=01b1a2f55d9c2b9ca06166b267bf1c22

[51] Y. Nakao, S. Stumpf, S. Ahmed, A. Naseer, and L. Strappelli, "Toward Involving End-users in Interactive Human-in-the-loop AI Fairness," *ACM Trans. Interact. Intelligent Syst.*, vol. 12, no. 3, 2022, publisher: Association for Computing Machinery. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137831302&doi=10.1145%2f3514258&partnerID=40&md5=238453e3daa70459bc1777010bf09a2d

[52] A. Páez, "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)," *Minds Mach*, vol. 29, no. 3, pp. 441–459, 2019, publisher: Springer Netherlands. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066806049&doi=10.1007%2fs11023-019-09502-w&partnerID=40&md5=8539ebc1e0627f47a7fbf793bfb24fa6

[53] H. Mucha, S. Robert, R. Breitschwerdt, and M. Fellmann, "Interfaces for Explanations in Human-AI Interaction: Proposing a Design Evaluation Approach," in *Conf Hum Fact Comput Syst Proc*. Association for Computing Machinery, 2021, journal Abbreviation: Conf Hum Fact Comput Syst Proc. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105812326&doi=10.1145%2f3411763.3451759&partnerID=40&md5=3a7dd1b28b449b7674b7f1bf8134766d

[54] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 3, pp. 740–770, 2021, publisher: MDPI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142788243&doi=10.3390%2fmake3030037&partnerID=40&md5=6deb5297b93cc307fc213c563d902a38

[55] J. Wang, L. Wang, Y. Zheng, C.-C. Yeh, S. Jain, and W. Zhang, "Learning-From-Disagreement: A Model Comparison and Visual Analytics Framework," *IEEE Trans Visual Comput Graphics*, vol. 29, no. 9, pp. 3809–3825, 2023, publisher: IEEE Computer Society. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129348278&doi=10.1109%2fTVCG.2022.3172107&partnerID=40&md5=a7f98ea4fccd11f753e4f04677c1655b

[56] N. Andrienko, G. Andrienko, L. Adilova, and S. Wrobel, "Visual Analytics for Human-Centered Machine Learning," *IEEE Comput Graphics Appl*, vol. 42, no. 1, pp. 123–133, 2022, publisher: IEEE Computer Society. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.

0-85123905854&doi=10.1109%2fMCG.2021.3130314&partnerID=40&md5=144409f52551710076354df70147c83a

[57] Y. Jia, J. McDermid, N. Hughes, M. Sujan, T. Lawton, and I. Habli, "The Need for the Human-Centred Explanation for ML-based Clinical Decision Support Systems," in *Proc. - IEEE Int. Conf. Healthc. Informatics, ICHI.* Institute of Electrical and Electronics Engineers Inc., 2023, pp. 446–452, journal Abbreviation: Proc. - IEEE Int. Conf. Healthc. Informatics, ICHI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85181579288&doi=10.1109%2fICHI57859.2023.00064&partnerID=40&md5=c41b859f4152299ab04eb83065c8eb27

[58] M.-A. Zöller, W. Titov, T. Schlegel, and M. Huber, "XAutoML: A Visual Analytics Tool for Understanding and Validating Automated Machine Learning," *ACM Trans. Interact. Intelligent Syst.*, vol. 13, no. 4, 2023, publisher: Association for Computing Machinery. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85181445629&doi=10.1145%2f3625240&partnerID=40&md5=faaf94ea0978fba6b1095d8ed68641b9

[59] F. Sovrano and F. Vitali, "From Philosophy to Interfaces: An Explanatory Method and a Tool Inspired by Achinstein's Theory of Explanation," in *Int Conf Intell User Interfaces Proc IUI.* Association for Computing Machinery, 2021, pp. 81–91, journal Abbreviation: Int Conf Intell User Interfaces Proc IUI. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104510754&doi=10.1145%2f3397481.3450655&partnerID=40&md5=370118a7584878524e33cc3f53f4b382

[60] T. Schoonderwoerd, W. Jorritsma, M. Neerincx, and K. van den Bosch, "Human-centered XAI: Developing design patterns for explanations of clinical decision support systems," *Int J Hum Comput Stud*, vol. 154, 2021, publisher: Academic Press. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108794824&doi=10.1016%2fj.ijhcs.2021.102684&partnerID=40&md5=7957c1e57e652d7896599445148e1e73

[61] R. Galanti, M. de Leoni, M. Monaro, N. Navarin, A. Marazzi, B. Di Stasi, and S. Maldera, "An explainable decision support system for predictive process analytics," *Eng Appl Artif Intell*, vol. 120, 2023, publisher: Elsevier Ltd. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85147194002&doi=10.1016%2fj.engappai.2023.105904&partnerID=40&md5=08e44b3980adc2c97299efcf3b8b8f9c

[62] Montimage, "Montimage," https://www.montimage.com/, 2024, accessed: Apr. 29, 2024.

[63] “Numpy,” https://numpy.org/, 2024, accessed: Apr. 29, 2024.

[64] “scikit-learn: Machine learning in python,” https://scikit-learn.org/, 2024, accessed: Apr. 29, 2024.

[65] “Tensorflow,” https://tensorflow.org/, 2024, accessed: Apr. 29, 2024.

[66] “Lightgbm: Light gradient boosting machine,” https://lightgbm.readthedocs.io/, 2024, accessed: Apr. 29, 2024.

[67] nginx, “nginx,” https://www.nginx.com/, 2024, accessed: Apr. 29, 2024.

[68] “React - a javascript library for building user interfaces,” accessed: 2023-05-14. [Online]. Available: https://reactjs.org/

[69] “Node.js,” https://nodejs.org/, 2024, accessed: Apr. 29, 2024.

[70] “Babel - the compiler for writing next generation javascript,” accessed: 2023-05-14. [Online]. Available: https://babeljs.io/

[71] “Webpack - a static module bundler for modern javascript applications,” accessed: 2023-05-14. [Online]. Available: https://webpack.js.org/

[72] “Bootstrap 5 - the most popular html, css, and js library in the world,” accessed: 2023-05-14. [Online]. Available: https://getbootstrap.com/

[73] “Tailwind css - rapidly build modern websites without ever leaving your html,” accessed: 2023-05-14. [Online]. Available: https://tailwindcss.com/

[74] “Okta,” https://www.okta.com/, 2024, accessed: Apr. 29, 2024.

[75] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should i trust you?” explaining the predictions of any classifier,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1135–1144, 2016. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[76] “Shap - explainable ai by the use of shapley values,” accessed: 2023-05-14. [Online]. Available: https://shap.readthedocs.io/

I Glossary

AI	Artificial Intelligence
API	Application Programming Interface
BNN	Bayesian Neural Network
DSS	Decision Support System
EU	European Union
FL	Federated Learning
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HCI	Human-Computer Interaction
IEEE	Institute of Electrical and Electronics Engineers
IIoT	Internet of Things
IDS	Intrusion Detection Server
JPEG	Joint Photographic Experts Group
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-Wise Relevance Propagation
ML	Machine Learning
NIST	National Institute of Standards and Technology
PICO	Population, Intervention, Comparison, Outcome
SCADA	Supervisory Control and Data Acquisition
SCOPUS	A comprehensive bibliometric database
SHAP	Shapley Additive Explanations
SLR	Systematic Literature Review
VM	Virtual Machine
XAI	Explainable Artificial Intelligence

II Questionnaire

Here are the extracted questions from the three Google Forms, separated accordingly:

No Explanation on AI Explanations

Demographic Questions:

1. What is your age?
   - Under 18
   - 18-24
   - 25-34
   - 35-44
   - 45-54
   - 55-64
   - 65 or older
2. What is your gender?
   - Male
   - Female
   - Prefer not to say
3. Education Level
   - Some high school
   - High school graduate
   - Self-employed
   - Unemployed
   - Retired
   - Prefer not to say
4. In which field do you work or study?
   - Technology/Engineering
   - Business/Finance
   - Education

- Health Care
- Arts and Entertainment
- Science and Research

Evaluating AI Explanation:

1. Identifying Key Features: Which feature appears to contribute most to the prediction according to the chart?

   - dl_data_volume
   - ul_data_volume
   - Dl_packet
   - Avg_ul_volume
   - downlink_packet

2. Visual Clarity: How easy was it to discern the importance of each feature from the chart?

   - Very difficult
   - 1
   - 2
   - 3
   - 4
   - Very easy

3. Relative Importance: How does the importance of the feature labeled 'dl_packet' compare to 'session_time'?

   - Much less important
   - 1
   - 2
   - 3
   - 4
   - Much more important

4. Need for Additional Information: How much did you need additional textual information to understand the chart?

   - Did not need any additional information

- 1
- 2
- 3
- 4
- Needed a lot more information

5. Ease of Understanding: How easy was it to understand the chart?
   - Very difficult
   - 1
   - 2
   - 3
   - 4
   - Very easy
6. Overall Comprehension: With the text, how well did you understand the reasons behind the ranking of each feature?
   - Did not understand at all
   - 1
   - 2
   - 3
   - 4
   - Understood perfectly
7. Relative Importance: How does the importance of the feature labeled ’dl_packet’ compare to ’session_time’?
   - Much less important
   - 1
   - 2
   - 3
   - 4
   - Much more important
8. Feature Ranking: Please rank the features from the most important to the least important based on their visual length in the chart.

Short Explanation of AI Prediction

Demographic Questions:

1. What is your age?
   - Under 18
   - 18-24
   - 25-34
   - 35-44
   - 45-54
   - 55-64
   - 65 or older
2. What is your gender?
   - Male
   - Female
   - Prefer not to say
3. Education Level
   - Some high school
   - High school graduate
   - Self-employed
   - Unemployed
   - Retired
   - Prefer not to say
4. In which field do you work or study?
   - Technology/Engineering
   - Business/Finance
   - Education
   - Health Care
   - Arts and Entertainment
   - Science and Research

Evaluating AI Explanation:

1. Textual Clarity: How clear was the text accompanying the chart?
   - Very unclear
   - 1
   - 2
   - 3
   - 4
   - Very clear

2. Contribution of Text to Understanding: How much did the text help you understand the chart?
   - Did not help at all
   - 1
   - 2
   - 3
   - 4
   - Helped a great deal

3. Integration of Text and Visuals: How well did the text integrate with the visual elements of the chart?
   - Very poorly integrated
   - 1
   - 2
   - 3
   - 4
   - Very well integrated

4. Overall Comprehension: How well did you understand the information presented in the chart with text?
   - Did not understand at all
   - 1
   - 2

    - 3
    - 4
    - Understood perfectly

5. Satisfaction: How satisfied were you with the explanation provided by the text in the chart?

    - Not satisfied at all
    - 1
    - 2
    - 3
    - 4
    - Extremely satisfied

6. Feature Ranking: Please rank the features from the most important to the least important based on their visual length in the chart.

Long Explanation of AI Prediction

Demographic Questions:

1. What is your age?

    - Under 18
    - 18-24
    - 25-34
    - 35-44
    - 45-54
    - 55-64
    - 65 or older

2. What is your gender?

    - Male
    - Female
    - Prefer not to say

3. Education Level

- Some high school
- High school graduate
- Self-employed
- Unemployed
- Retired
- Prefer not to say

4. In which field do you work or study?

   - Technology/Engineering
   - Business/Finance
   - Education
   - Health Care
   - Arts and Entertainment
   - Science and Research

Evaluating AI Explanation:

1. Textual Clarity: How clear was the text accompanying the chart?

   - Very unclear
   - 1
   - 2
   - 3
   - 4
   - Very clear

2. Contribution of Text to Understanding: How much did the text help you understand the chart?

   - Did not help at all
   - 1
   - 2
   - 3
   - 4
   - Helped a great deal

3. Integration of Text and Visuals: How well did the text integrate with the visual elements of the chart?
    - Very poorly integrated
    - 1
    - 2
    - 3
    - 4
    - Very well integrated
4. Overall Comprehension: How well did you understand the information presented in the chart with text?
    - Did not understand at all
    - 1
    - 2
    - 3
    - 4
    - Understood perfectly
5. Satisfaction: How satisfied were you with the explanation provided by the text in the chart?
    - Not satisfied at all
    - 1
    - 2
    - 3
    - 4
    - Extremely satisfied
6. Overall Comprehension: With the text, how well did you understand the reasons behind the ranking of each feature?
    - Did not understand at all
    - 1
    - 2
    - 3

- 4
- Understood perfectly

7. Satisfaction with Textual Information: How satisfied were you with the amount and quality of the information provided in the text?

   - Not satisfied at all
   - 1
   - 2
   - 3
   - 4
   - Extremely satisfied

8. Feature Ranking: Please rank the features from the most important to the least important based on their visual length in the chart.

III Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Toluwani Mathew Elemosho**,
(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,

   Analysing Model Attacks in Machine Learning Pipeline,

   (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.

Toluwani Mathew Elemosho
15/05/2024