SPATIAL

GUIDING LIGHT IN THE AI LANDSCAPE

Illuminating the Path to a Trustworthy AI Ecosystem

INDEX

Overview

The SPATIAL project, a pioneering European initiative, has developed innovative solutions to enhance the trustworthiness and functionality of artificial intelligence (AI) systems. Over the past three years, this project has focused on four key use cases:

- Privacy-preserving AI on the Edge and Beyond: This use case emphasizes the importance of maintaining data privacy while deploying AI solutions at the edge and beyond.
- Improving Explainability, Resilience, and Performance of Cybersecurity Analysis of 5G/4G/IoT Networks: This area targets enhancing the security analysis of advanced networks, ensuring they are more explainable, resilient, and efficient.
- Resilient Cybersecurity Analytics: This use case focuses on developing robust and adaptive cybersecurity analysis tools that can withstand and respond to evolving threats.
- Accountable AI in Emergency eCall Systems: This use case is dedicated to ensuring that AI systems used in emergency call scenarios are accountable and reliable.

SPATIAL has also developed educational modules for the general public and more advanced users to support these advancements. These modules are designed to increase understanding and foster responsible usage of AI technologies.

Central to SPATIAL's efforts is the SPATIAL platform, designed to estimate the trustworthiness of AI by using a combination of metrics deployed as services. This platform qualitatively analyses and assesses AI models and SPATIAL use cases, aiming to provide higher levels of explainable AI solutions based on accountability, privacy, and resilience principles. Internally, the platform integrates several key components to achieve its design goals.

Additionally, the project introduces the COMPASS framework, which allows organisations to critically evaluate the technical innovation potentials and societal impacts of AI systems. This framework ensures responsible and trustworthy AI by providing a comprehensive roadmap for AI developers and auditors. It offers the flexibility needed for AI practitioners to tailor evaluation processes to different industries' specific needs and priorities, helping to identify essential skills and apply SPATIAL design principles effectively.

This document compiles the solutions and results achieved by the SPATIAL project within its three years of operation, showcasing the significant strides made in the realm of trustworthy AI.

Privacy-preserving AI on the Edge and Beyond Telefónica

Introduction

The first use case of the SPATIAL project leverages federated learning (FL) to build AI models using a distributed approach. The main benefit of using FL is that it removes the need to transfer training data to a central server.

Instead, FL clients train locally a version of the AI model using their local datasets. The models are then sent to a central server that aggregates all the client updates. FL reduces the communication overhead while increasing the privacy protection by default.

Nowadays, FL is seen as a promising computing paradigm that could allow training the AI models of the future by leveraging distributed computing capacities, such as those offered by existing communication networks.

Advantages

FLaaS demonstrates several key advantages with respect to classic centralized

ML training

Enhanced Privacy and Security: Data never leaves the local devices, minimising the risk of exposure and ensuring higher compliance with privacy laws.

Reduced Latency and Bandwidth Usage: By avoiding the need to transfer large datasets to a central server continuously, FL reduces the demand on network bandwidth and lowers latency, making it feasible for deployment in edge computing environments.

Scalability: FL can easily scale to accommodate many clients, each contributing to the global model without needing a powerful centralised infrastructure.

Robustness and Fault Tolerance: Decentralized training can improve the robustness of the learning process, as the system does not rely on a single point of failure. Even if some clients go offline, the training can continue with the remaining active clients.

FLaaS (Federated Learning as a Service) builds on these foundational principles to offer a comprehensive framework that simplifies the deployment and management of federated learning systems. FLaaS aims to provide a seamless experience for organisations looking to implement FL, offering tools and services that handle the complexities of client coordination, model aggregation, and communication protocols.

Implementation

In this context, Telefónica has developed FLaaS, the Federated Learning as a Service platform that attempts to facilitate practical FL on Android devices and can be extended to support new types of end devices. Similar to Machine Learning as a Service (MLaaS) platforms, FLaaS reduces the learning curve and configuration required to run FL projects.

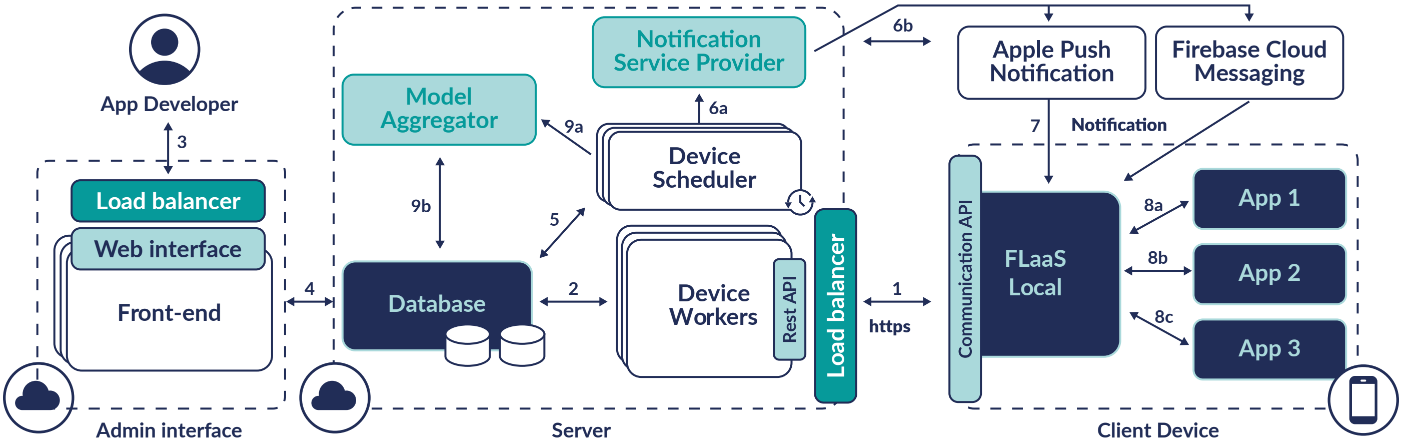


Figure 1

To achieve that, as depicted in the figure above, FLaaS offers a number of components and functionalities:

- Admin interface: This offers a user interface for the AI developer and system administrator to configure, run, and monitor the execution of FL projects.
- FLaaS server: The backend of the service performs the orchestration and aggregation required for the FL process.
- FLaaS local: Library available for Android devices that facilitates the training of models combining the data of multiple applications installed on the device using two different approaches: (i) combining their data (Joint Samples, JS); (ii) combining their models trained individually (Joint Models, JM)
- FLaaS communication service: This service enables a push-based communication channel between the server and the clients, which is required for FL tasks such as starting the training process.

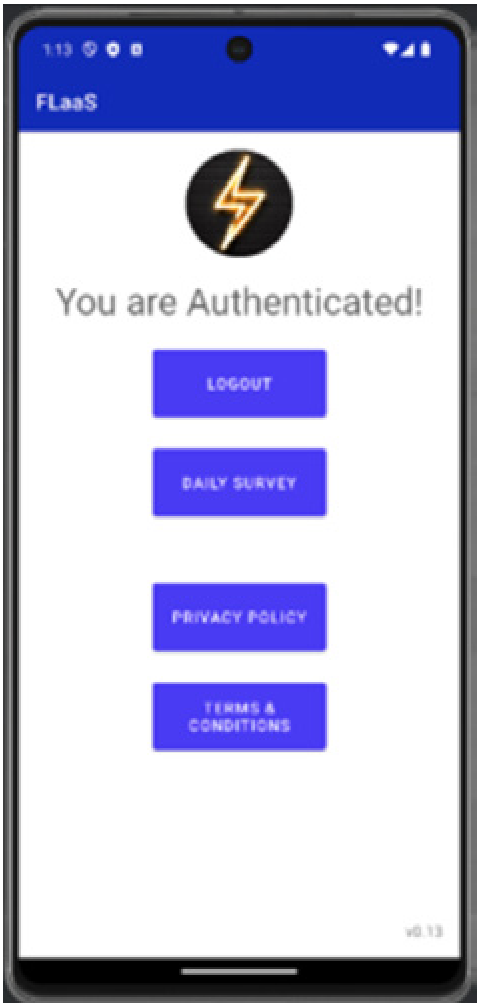
Despite not requiring the sharing of user data, FL does not offer complete privacy guarantees. Previous research has shown how successful privacy attacks can still be crafted against models even when trained using FL. To address this problem, thanks to the integration between FLaaS and the SPATIAL Differential Privacy (DP) component, available in the SPATIAL platform, we have incorporated a new privacy-related feature in FLaaS. It corresponds to the option of adding Central Differential Privacy (CDP) to the FL process. To include CDP, the admin has the option of configuring and parametrising the level of DP that is added to the aggregation process. By including CDP, we add an additional level of privacy protection to safeguard the privacy of the users participating in any FL project.

Our results have shown how the inclusion of CDP increases the resilience of the trained models against privacy attacks, reducing the capacity of a skilled attacker to retrieve sensitive information from them.

Potential Applications

FLaaS can be applied across various sectors where data privacy is paramount and distributed data sources are abundant:

- Healthcare: Training predictive models using patient data from multiple hospitals without compromising patient confidentiality.
- Finance: Developing fraud detection algorithms by combining insights from different financial institutions without sharing sensitive transaction data.
- Smart Cities: Enhancing urban planning and management by analyzing data from distributed IoT devices without aggregating all data centrally.



You are Authenticated!

Figure 2 Screenshot of the FLaaS demo app running on an Android device

Evaluation

We have evaluated FLaaS through rigorous testing in both controlled in-lab environments and real-world user studies to ensure its robustness, usability, and effectiveness.

In-Lab Testing

Our in-lab testing focused on validating the core functionalities of FLaaS, including model training, aggregation, and deployment. We conducted extensive performance benchmarking to assess the efficiency of the aggregation server and the scalability of the client SDK across various hardware configurations. Security assessments were performed to ensure that the privacy-preserving mechanisms, such as differential privacy and secure multiparty computation, functioned correctly and did not introduce significant performance bottlenecks.

User Studies

End-User Evaluation

To understand the impact of FLaaS on end-users, we conducted a series of field studies involving diverse participants using different types of devices. These studies focused on two main aspects:

- Device Performance: We monitored the effect of running FLaaS on device performance metrics such as CPU usage, memory consumption, battery life, and network bandwidth. Our goal was to ensure that FLaaS operates efficiently without significantly degrading the performance or usability of end-user devices.
- User Perception: Participants were asked to provide feedback on their experiences with FLaaS through structured interviews and questionnaires. This feedback helped us understand the perceived impact of FLaaS on their device performance and comfort level with the federated learn-

ing process. The results indicated a generally positive reception, with most users appreciating the privacy benefits and minimal impact on device performance.

AI Practitioner Survey

We also conducted a survey with 20 AI practitioners to evaluate the backend and administrative interface of FLaaS. This survey aimed to gather insights on the usability, functionality, and overall satisfaction with the system from those who would be managing and deploying FLaaS in real-world scenarios. Key aspects evaluated included:

- Ease of Use: Practitioners assessed the intuitive and user-friendly nature of the admin interface, focusing on the learning curve and ease of integration with existing workflows.
- Functionality: The survey evaluated the range of features available in the backend, such as monitoring tools, model management capabilities, and security configurations.
- Performance and Reliability: Practitioners provided feedback on the system's performance, reliability, and scalability when managing multiple federated learning clients and large datasets.

The survey results were overwhelmingly positive, with practitioners highlighting FLaaS's key strengths as seamless integration, comprehensive monitoring tools, and robust security features.

Sociotechnical Analysis

Finally, we performed a comprehensive sociotechnical analysis to evaluate the broader impact of FLaaS within organisational and social contexts. This analysis was performed using COMPASS, an algorithmic assessment framework developed within the SPATIAL project that considers various factors, including the Context, Accountability, Measures utilised and Privacy potentials.

The sociotechnical analysis confirmed that FLaaS not only meets technical requirements but also aligns well with ethical standards and organisational goals, highlighting its inherent focus on privacy protection and the need for continuous stakeholder engagement to maintain trust and acceptance.

Conclusion

In conclusion, our comprehensive evaluation of FLaaS through in-lab testing, user studies, AI practitioner surveys, and sociotechnical analysis has demonstrated its effectiveness and viability as a federated learning framework.

These evaluations have provided valuable insights that have guided the refinement of FLaaS, ensuring it is a robust, user-friendly, and privacy-preserving solution for modern AI model training.

Improving Explainability, Resilience, and Performance of Cybersecurity Analysis of 5G/4G/IoT Networks Montimage

Introduction

The rise in encrypted network traffic, driven by the use of HTTPS and Virtual Private Networks (VPN), poses challenges for traditional traffic analysis tools, as they struggle to detect malicious activity. This, along with the large amount of traffic from mobile and IoT devices, demands advanced Artificial Intelligence (AI)-based techniques. These techniques enable intelligent, adaptable, and autonomous security management, helping to address the increasing complexities and real-time anomaly detection of the 5G/IoT network. However, the adoption of AI methods in IoT and future mobile networks is still in its infancy, with three major issues in the 5G/IoT domain. The challenges in advancing AI research for 4G/5G/IoT networks include the lack of diverse real-world datasets due to privacy concerns among telecom operators. Additionally, current AI security solutions prioritize accuracy and performance metrics, often lacking explainability for decision-making, which is crucial for the reliability of 5G infrastructure. Furthermore, ML models are susceptible to adversarial attacks, presenting a significant challenge in ensuring robustness against such threats.

We develop three AI-based security applications that correspond to three main steps of Intrusion Detection and Response for real-time anomaly detection of the 5G/IoT network. Firstly, the Traffic Classification application characterizes network traffic to identify certain types of normal user activities, such as web browsing, chatting, or video streaming. Secondly, the Attack Detection application differentiates between malicious traffic and legitimate traffic to detect popular cyberattacks in 5G or IoT environments. Finally, the Root Cause Analysis (RCA) application employs a similarity-based machine learning approach to discover the root causes of problems in order to quickly identify appropriate solutions.

We aim to enhance the performance, explainability, and robustness of our AI-based security applications for network traffic analysis and anomaly detection in 5G/IoT networks, addressing the aforementioned challenges:

- **Deploying real 4G/5G/IoT testbeds**. It involves the deployment of real (private) 4G/5G/IoT networks as well as Security Analysis, providing detailed instructions for their setup.
- **XAI framework for resilient 5G/IoT traffic analytics.** We design and develop an open-source framework consisting of two main components: (1) Network Traffic Analysis and (2) XAI for Resiliency, that is integrated in the SPATIAL platform.

Extensive experimentation demonstrates the effectiveness of the AI models, evaluating accountability and resilience metrics using both public and private datasets through our testbeds, thereby validating the models' robustness and reliability. Additionally, our framework serves a broad range of user groups, including network administrators, security analysts, IT operations teams, cybersecurity researchers, enterprises, organizations, and academic institutions, utilizing the framework for experiments, validation, teaching, and research in network security and AI.

Deploying real 5G/IoT testbeds

The 4G/5G testbed comprises an EPC-in-a-Box platform, representing a commercialized 4G/5G network core developed by Montimage and Cumucore. It allows for the quick creation of a complete 4G/5G network within minutes and serves as both a testing environment and a means to establish a small-scale mobile network for industrial use. The testbed consists of three main components: Radio Access Network (RAN), Evolved Packet Core (EPC) or 5G Core, and MMT. Once deployed, the testbed enables commercial devices to connect, granting access to Internet Protocol (IP) services in Public Data Networks and the internet. MMT-Probe analyses the real-time traffic between the RAN and EPC, ensuring compliance with security requirements, while MMT-Operator facilitates automated decision-making and responses to anomalies. A step-by-step guide for deploying a 5G testbed using open-source tools and technologies can be found in the repository **https://github.com/Montimage/cerberus-edge-configuration.**

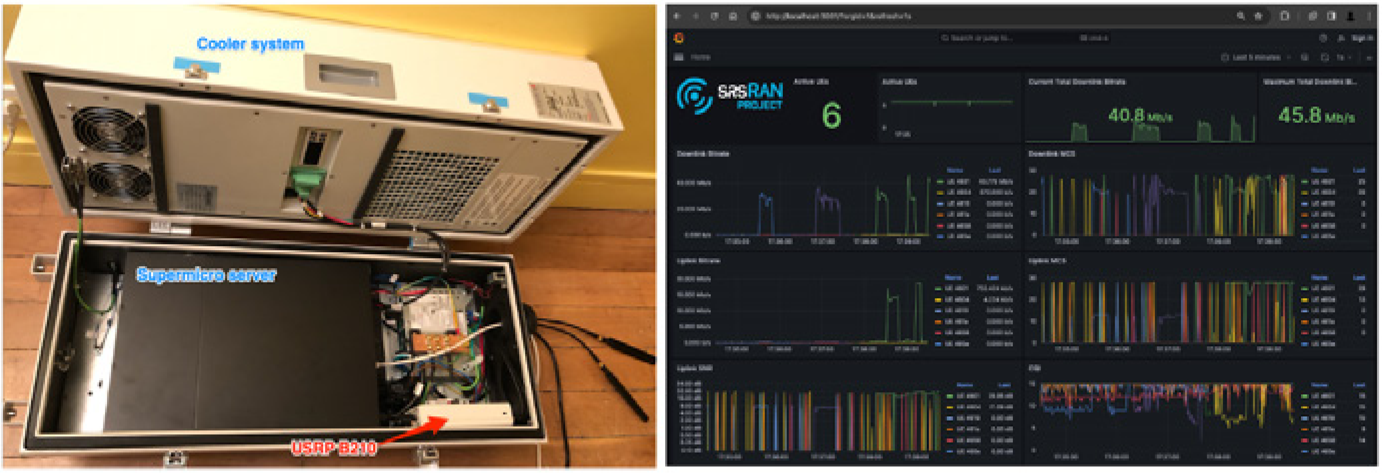


Figure 3 Inside of the supermicro server (left) and GUI of the 5G testbed (right)

The **IoT testbed** consists of various equipment, including Zolertia REMotes, a Raspberry Pi, and related accessories. These components form an IoT IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN). A border router mote acts as the gateway, collecting data from other motes and transmitting it to the server via a Universal Serial Bus (USB) connection on the Raspberry Pi. The MMT-Sniffer device captures network traffic and sends it to the Linux-based machine via USB, where MMT-IoT is deployed for traffic analysis and statistics extraction for the Root-cause Analysis module. The Raspberry Pi serves as the power source for the motes, hosts the server for handling sensed data, and receives the sniffed traffic for analysis by MMT-IoT.

XAI framework for resilient 5G/IoT traffic analytics

We design and implement the Montimage AI Platform (MAIP), an XAI framework with an intuitive and user-friendly interface for network traffic analysis and classification in 5G/IoT networks. It comprises two principal components: **Network Traffic Analysis and XAI for Resiliency.** While the **Network Traffic Analysis** component aims to fulfill the need for effective analysis and classification of encrypted traffic using advanced AI techniques, the **XAI for Resiliency** component aims to enhance the robustness of AI models built within the Network Traffic Analysis module, making them more resilient against different types of adversarial machine learning attacks.

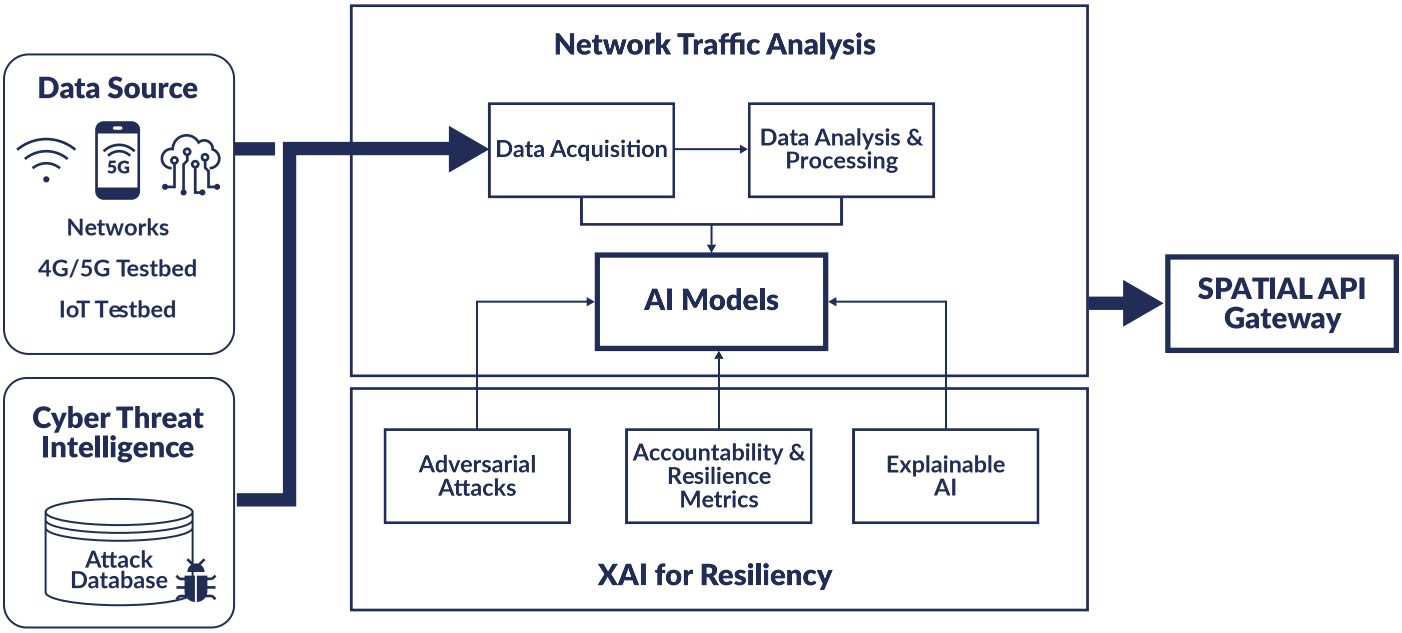


Figure 4

The **Network Traffic Analysis** component aims to meet the demand for effective analysis and classification of encrypted traffic using advanced AI techniques. The Data Acquisition module collects raw traffic data from networks or 4G/5G/IoT testbeds in either online or offline mode. Additionally, it can leverage Cyber Threat Intelligence (CTI) sources, e.g., deployed honeypots, to learn and continuously train our models using attack patterns and past anomaly information stored in the database, especially for anomaly detection applications. The Data Analysis and Processing phase employs the open source Montimage Monitoring Tools MMT-Probe to parse raw network traffic, extract network and application-based events (such as protocol field values and statistics), compute the features required for AI models and translate them into a numeric form. The modular architecture of the MMT-Probe allows for the addition of new protocols to parse. The extracted features comprise multiple parameters that are computable on raw traffic independently of whether the traffic is encrypted or not, including statistics involving byte and time information. For instance, for anomaly detection, we extract and employ 59 features, including basic features in packet headers and statistical features after performance traffic aggregation into flows. The AI models module is tasked with constructing various models to classify vectorized network traffic data for diverse objectives, such as classifying user activity, detecting anomalies in encrypted traffic, and conducting root cause analysis.

The **XAI for Resiliency** component aims to enhance the robustness of AI models built within the Network Traffic Analysis module, making them more resilient against various types of adversarial machine learning attacks. The Adversarial Attacks module focuses on injecting various evasion and poisoning adversarial attacks, such as label flipping attacks and Generative Adversarial Networks (GANs) attacks or integrating existing AI-based attack libraries for the robustness analysis of AI models. The Explainable AI module aims to produce post-hoc global and local explanations of predictions generated by our model. Specifically, we employ popular model-agnostic post-hoc XAI techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), to explain predictions of our models. Furthermore, we incorporate defense mechanisms, such as adversarial training and leveraging XAI techniques, to prevent attacks against the AI models. The relationship between XAI and adversarial attacks is intricate and multi-dimensional. On one hand, increasing the transparency and interpretability of a model can make it more vulnerable to adversarial attacks. By leveraging the explanations provided by XAI methods, attackers can identify model weaknesses and create more potent adversarial examples. On the other hand, if a model lacks transparency and interpretability, it becomes difficult not only to understand the reasoning behind its predictions but also to detect and address adversarial attacks. Hence, it is crucial to find a balance between XAI and adversarial attacks to develop secure and robust AI models. As we need to consider the trade-off between explainability, robustness and performance of our system, we measure quantifiable metrics for its accountability (e.g., through accuracy, currentness, and confidence metrics) and resilience (e.g., through impact, complexity metrics).

Our framework is implemented in Node.js, utilizing the MMT-Probe tool written in C for feature extraction. We leverage popular Python libraries for machine learning (e.g., numpy, scikit-learn, tensorflow, xgboost, lightgbm) and XAI (e.g., SHAP, LIME). These components are packaged in Docker containers, enabling users to effortlessly install and test them locally. Users have the flexibility to interact with the components either via the locally-hosted Swagger UI or directly through the SPATIAL platform’s API gateway. For an enhanced user experience, we’ve also developed a client-side built-in React, providing users with an intuitive and user- friendly interface to access all services. This includes extracting features, building or retraining the model, injecting adversarial attacks, producing explanations, and evaluating our model using different metrics, all provided by our two components.

Accountable AI in Emergency eCall System

Fraunhofer Fokus

Introduction

The SPATIAL Use Case 3 investigates the design, development, and integration of an accountable AI-based emergency detection into a next-generation emergency communication system. Specifically, the objective of the system developed in this use case is to automatically detect emergencies by analysing data gathered by IoT sensors with state-of-the-art AI technologies. Subsequently, after detecting an emergency, a corresponding emergency call (eCall) to a trained medical professional will be automatically initiated so that the call taker can initiate medical help immediately.

During the emergency call, both the patient and medical expert answering the call have access to all relevant data encompassing the sensor data, the decision of the employed AI models, and explanations describing why the AI model recognized an emergency in the available data. Therefore, the medical expert receiving the eCall can get a precise overview of the current situation, better instruct available first responders, and initiate the necessary medical countermeasures more effectively. As a result, the explainability and accountability features provided by the system developed in Use Case 3 can potentially improve and accelerate the entire emergency chain.

EMYNOS Next-Generation Emergency Communication System

As a foundation for the realisation of the emergency communication, we will build on the results of the EU-funded H2020 EMYNOS (nExt generation eMergencY commuNicatiOnS) project. The EMYNOS framework is a prototype of a so-called Next Generation 112 (NG112) emergency communication system that allows to realisation of VoIP emergency services. Compared to traditional phone-based emergency communication systems, EMYNOS offers several advantages. For example, EMYNOS allows the transmission of rich-media information during the call, such as video or geolocation. Furthermore, the IP-based infrastructure also enables to sharing of additional information, e.g. sensor data, directly with the emergency call centre during an emergency call. Therefore, the EMYNOS NG112 emergency communication system provides an excellent basis for the accountable AI-based emergency detection envisioned in Use Case 3.

Use Case 3 Architecture

The architecture of the Use Case 3 system, along with all involved actors and their interactions, are illustrated in Figure 5. In the light green and red boxes, we can see the patient (caller side) and the medical professional answering the eCall (Public Safety Answering Point - PSAP), respectively. Both use an adapted version of the OpenSource VoIP client Linphone to connect to the EMYNOS network and conduct VoIP-based eCalls. In addition, we can see that relevant cardiovascular data is collected on the patient side by IoT sensors. This data is later used for AI-based detection of emergencies.

Furthermore, Figure 5 shows that the EMYNOS framework shown in blue has been integrated into the SPATIAL ecosystem and extended by two additional services—the Medical Analysis Module (MAM) and the Enhanced Interpretability Module (EIM). These services process the collected data and extend the functionality of the EMYNOS framework to perform accountable AI-based emergency detection and generation of corresponding explanations, which help users understand the decision-making of the employed models. We will briefly introduce these SPATIAL services and their functionality in the following.

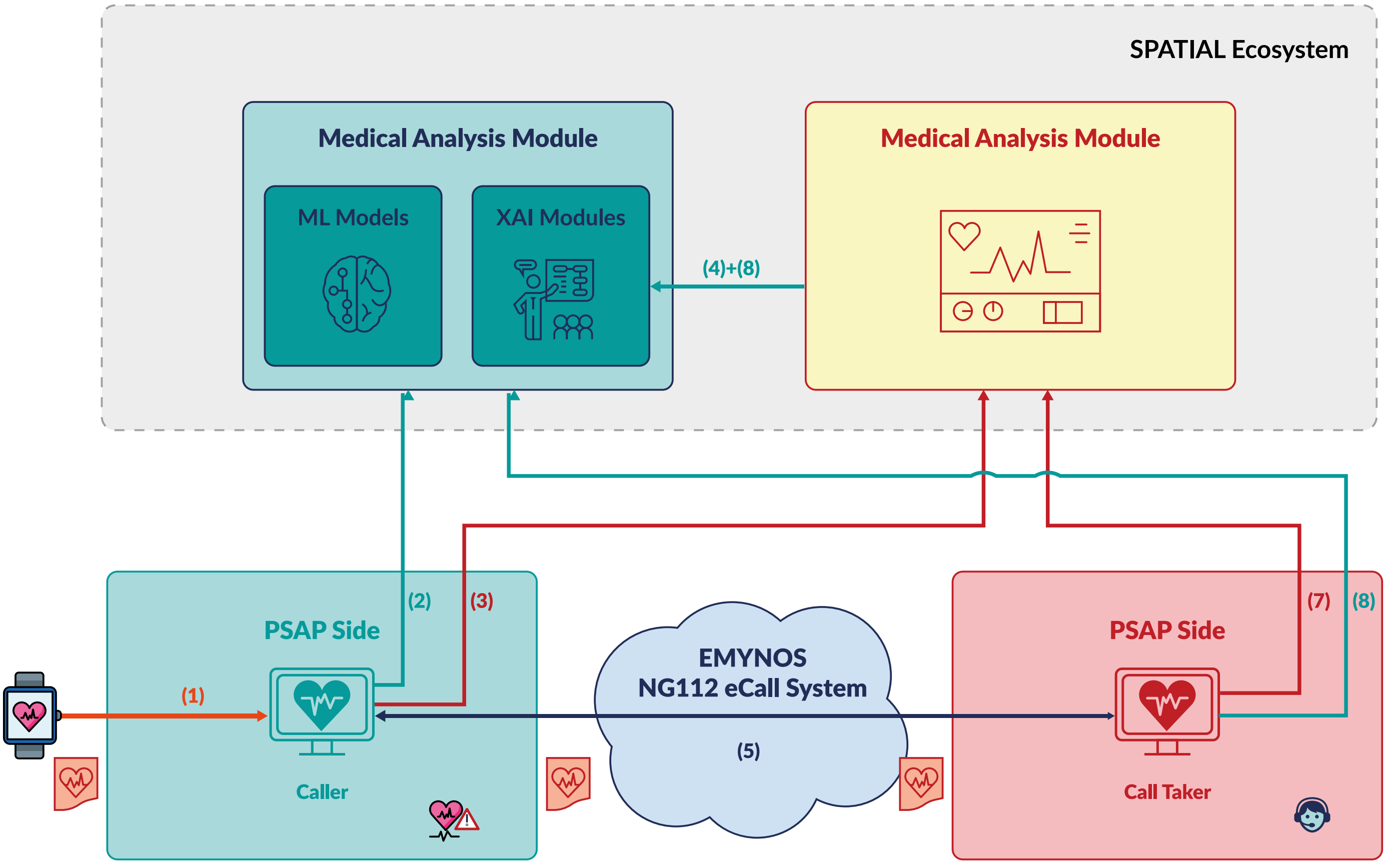


Figure 5 The system's architecture developed in Use Case 3 summarises all relevant components, involved actors, and their interactions.

Medical Analysis Module

The Medical Analysis Module (MAM) is a REST-based service that offers powerful tools for fast and accurate medical diagnoses while enhancing accountability and transparency through the use of explainable AI (XAI) methods. It employs advanced algorithms and machine learning techniques to visualise and analyse medical data, gain insights that can aid in diagnosing various health conditions, provide accurate and reliable medical diagnoses, and explain the decision-making of the underlying ML models. This enables patients to receive faster and more reliable diagnoses while medical experts can make more informed decisions. Furthermore, the MAM provides developers with functionalities to administer ML models operating on multivariate time series data (i.e. uploading, managing, deletion) and evaluate them regarding various performance indicators (e.g. accuracy, recall, precision).

Finally, the MAM also provides tools to generate local XAI explanations for the hosted ML models. The Medical Analysis Module is designed in an adaptive and modularised manner, allowing it to be easily extended to multiple medical applications. However, within the scope of SPATIAL, the provided service will be limited to the analysis of myocardial infarctions in ECG data, as we will discuss below.

The MI detection capability of the developed AI models hosted at the MAM. In the context of SPATIAL, the detection of myocardial infarctions (MIs), commonly known as heart attacks, is currently being studied as an exemplary emergency scenario in Use Case 3. For this purpose, 12-lead ECG sensor data is utilised, which describes the heart's electrical activity and allows the detection of various cardiovascular pathologies in a patient. In addition to professional clinical devices, such ECG readings can also be reliably collected by IoT sensors such as smart watches or chest bands and integrated into the Use Case 3 system.

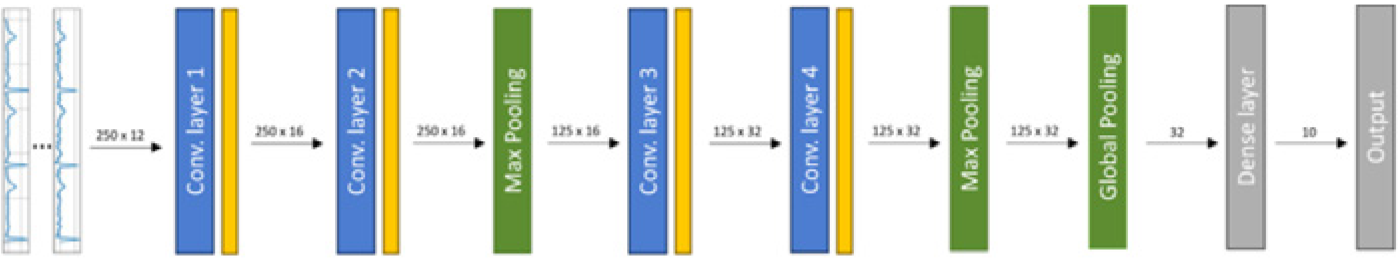


Figure 6 Architecture of the developed 1D CNN showing convolutional (blue), dropout (yellow), pooling (green) and dense (grey) layers. In Use Case 3, this model is employed for the AI-based MI detection in provided 12-lead ECG sensor data. Further details regarding the CNN model are published in[1]

To analyse the ECGs and reliably detect indications for MIs, ML models were utilised in Use Case 3 and made available at the MAM. A CNN model was developed, which performs one-dimensional convolutions on the available multivariate time series ECG data. The architecture of the developed model is visualised in Figure 6. It includes four convolutional layers, two max-pooling layers, a global average pooling layer followed by a dense layer, and a probabilistic output provided by the sigmoid function. The ReLU activation is applied to all layers except the output layer. Dropout layers are added after each convolutional layer to serve as a regularisation technique, helping to mitigate overfitting. The developed CNN reliably detects MIs in the PTB-XL[2] benchmarking dataset with an accuracy of 96.21%, precision of 91.55%, recall of 93.13%, and AUROC of 98.91%

The resulting CNN model is hosted at the MAM and utilised to perform the MI detection in the eCall scenario developed in Use Case 3. More details on the architecture of the CNN model, applied training processes, further performance indicators, and additional analyses were described by Knof et al.[3].

[1] H. Knof, P. Bagave, M. Boerger, N. Tcholtchev, and A. Y. Ding, "Exploring CNN and XAI-based Approaches for Accountable MI Detection in the Context of IoT-enabled Emergency Communication Systems," in Proceedings of the 13th International Conference on the Internet of Things, in IoT '23. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 50–57. doi: 10.1145/3627050.3627057.

[2] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," Sci. Data, vol. 7, no. 1, p. 154, Dec. 2020, doi: 10.1038/s41597-020-0495-6.

XAI explanations to increase the accountability of the system. In addition to the reliable detection of MIs, the accountability of the AI system also plays a decisive role in Use Case 3. The aim is to shed light on the decision-making of the employed AI models and explain why the models have detected indications for an MI in specific ECGs. To this end, various XAI methods are applied in Use Case 3 and provided as tools at the MAM. Currently, the MAM supports generating local explanations using LRP[4] and SHAP[5]. The employed XAI methods aim to identify the data points most relevant to MI detection (highlighted in dark red) in the analysed ECGs. This information is then visualised as a heatmap over the ECG signal so that users can directly identify the anomalous ECG segments. Figure 7 exemplifies two investigated visualisation forms for generated LRP explanations for a specific ECG. Figure 7 (a) visualises the individual data features relevant for MI detection, while in Figure 7 (b), the importance is aggregated over time intervals.

These local explanations enable medical professionals to directly identify meaningful and abnormal segments in the ECG signal, thereby understanding and verifying the model's decision towards MI detection. Furthermore, these explanations allow end users to understand the models' decision-making and enable developers to shed light on their inner workings, allowing them to improve them in a targeted manner.

[3] H. Knof, P. Bagave, M. Boerger, N. Tcholtchev, and A. Y. Ding, "Exploring CNN and XAI-based Approaches for Accountable MI Detection in the Context of IoT-enabled Emergency Communication Systems," in Proceedings of the 13th International Conference on the Internet of Things, in IoT '23. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 50–57. doi: 10.1145/3627050.3627057.

[4] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-Wise Relevance Propagation: An Overview," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., in Lecture Notes in Computer Science. , Cham: Springer International Publishing, 2019, pp. 193–209. doi: 10.1007/978-3-030-28954-6_10.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dezember 2017, pp. 4768–4777.

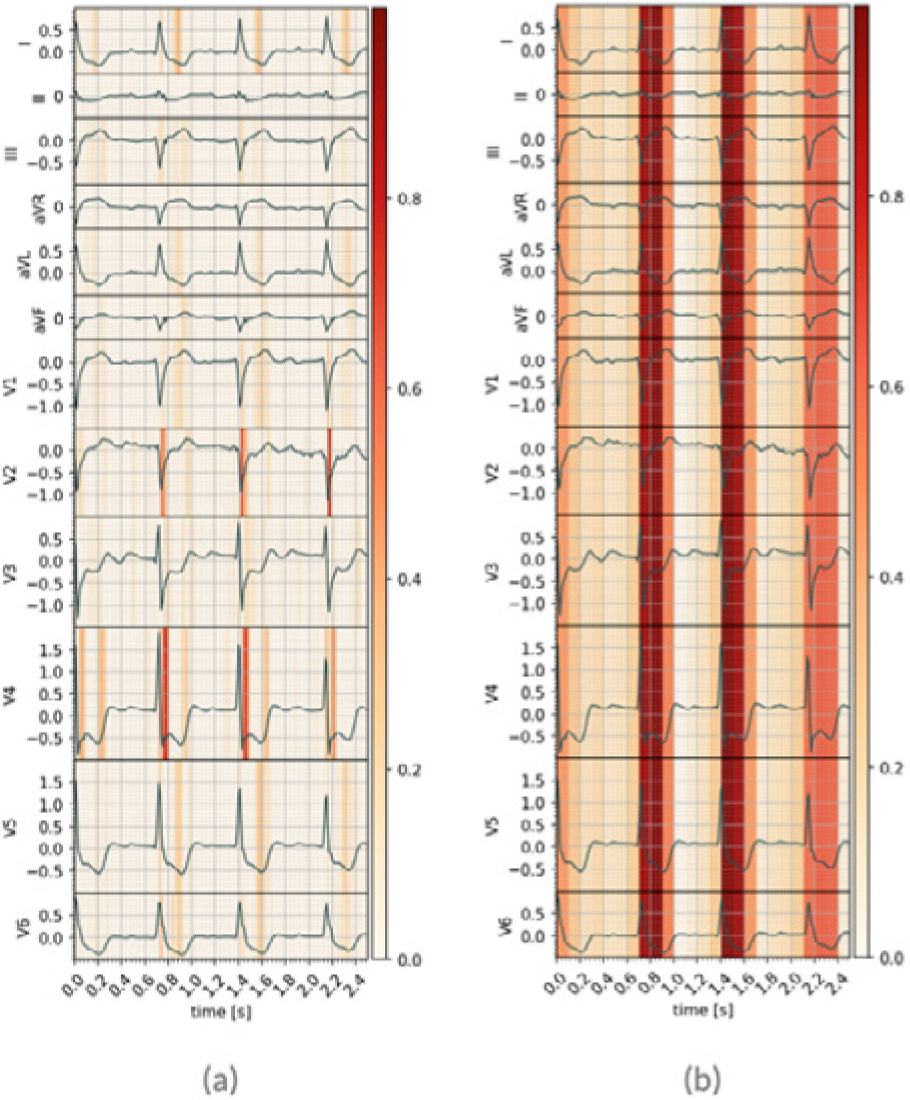


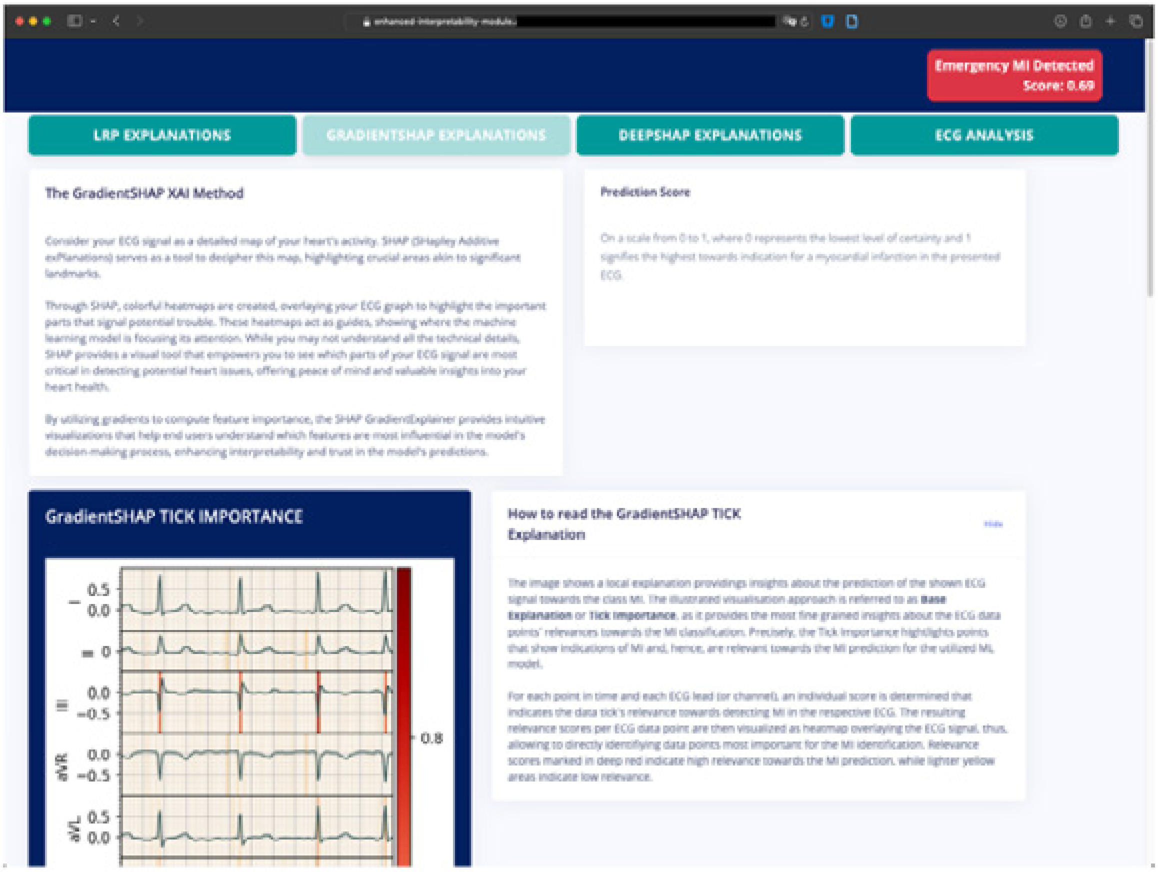
Figure 7 Sample local explanations using the XAI method LRP [3]. In Figure (a), the individual relevant data points are identified for each lead, resulting in a fine-grained explanation of the significant points. In Figure (b), this information determines the relevant time segments across the leads, enabling direct visual identification of prominent time segments.

Enhanced Interpretability Module

The second SPATIAL service utilised in Use Case 3 is the Enhanced Interpretability Module (EIM). The idea behind the EIM is motivated by the fact that explanations provided by state-of-the-art XAI methods are heavily influenced by the needs of data scientists. However, different stakeholders use these explanations after deployment in the application domain. As interpretability is subjective, and is influenced by the prior knowledge of the users, the interpretability of such explanations for stakeholders with different domain knowledge becomes challenging. Therefore, the Enhanced Interpretability Module proposes an interactive interface with flexible explanations so that the user can interact with multiple levels of explanations and achieve the needed interpretability from the SPATIAL platform.

In the scope of SPATIAL, the EIM is realised as a Web-App and integrated into the Use Case 3 eCall scenario. The EIM utilises the functionality of the MAM to enable users to easily apply and generate explanations from the available XAI methods for the hosted MI detection models.

Subsequently, the EIM enhances the XAI explanations with additional meta information and descriptions on reading and understanding the provided explanations. As a result, the interpretability of the explanations is increased for users. A screenshot with a sample explanation of the latest version of the EIM is shown in Figure 8.



The GradientSHAP XAI Method

Consider your ECG signal as a detailed map of your heart's activity. SHAP (SHapley Additive exPlanations) serves as a tool to decipher this map, highlighting crucial areas akin to significant landmarks.

Through SHAP, colorful heatmaps are created, overlaying your ECG graph to highlight the important parts that signal potential trouble. These heatmaps act as guides, showing where the machine learning model is focusing its attention. While you may not understand all the technical details, SHAP provides a visual tool that empowers you to see which parts of your ECG signal are most critical in detecting potential heart issues, offering peace of mind and valuable insights into your heart health.

By utilizing gradients to compute feature importance, the SHAP GradientExplainer provides intuitive visualizations that help end users understand which features are most influential in the model's decision-making process, enhancing interpretability and trust in the model's predictions.

Prediction Score

On a scale from 0 to 1, where 0 represents the lowest level of certainty and 1 signifies the highest towards indication for a myocardial infarction in the presented ECG.

How to read the GradientSHAP TICK Explanation

The image shows a local explanation providings insights about the prediction of the shown ECG signal towards the class MI. The illustrated visualisation approach is referred to as **Base Explanation** or **Tick Importance**, as it provides the most fine grained insights about the ECG data points' relevances towards the MI classification. Precisely, the Tick Importance highlights points that show indications of MI and, hence, are relevant towards the MI prediction for the utilized ML model.

For each point in time and each ECG lead (or channel), an individual score is determined that indicates the data tick's relevance towards detecting MI in the respective ECG. The resulting relevance scores per ECG data point are then visualized as heatmap overlaying the ECG signal, thus, allowing to directly identifying data points most important for the MI identification. Relevance scores marked in deep red indicate high relevance towards the MI prediction, while lighter yellow areas indicate low relevance.

Figure 8 Screenshot of the Enhanced Interpretability Module, a subcomponent that provides XAI explanations for the ECG analysis and enhances them with additional meta-information and descriptions to increase users' interpretability.

Resilient Cybersecurity Analytics

Withsecure

The rise of artificial intelligence (AI) over the last few years is undeniable. It is as fast as it is surprising, playing a dominant role in many areas of technology and our lives. This development of new technology brings new opportunities and raises many questions about the trust we place in it. Trustworthy AI has been created to ensure that the features we build using AI are ethical, explainable, reliable, transparent, and robust. The last concept is extremely important in the area of cybersecurity, although its importance also extends to other domains. AI systems must be designed to be robust and resilient to various types of attacks and failures. This includes ensuring that AI systems can function even when faced with malicious attacks, system failures, or unexpected inputs. Robustness requires that AI systems are tested thoroughly and can operate in different environments and scenarios. In other words, AI systems must be reliable, safe, and able to handle unexpected situations.

We can distinguish two most common types of attacks on models: model evasion and data poisoning. **Data poisoning** attack begins when adversaries, known as "threat actors", somehow gain access to the training dataset and can contaminate the data by modifying entries or introducing tampered data into the training dataset. This attack is very effective but requires the threat actor to gain access to the training data, which makes this attack vector quite difficult. **Model evasion** attacks are among the most common types of inference attacks that target the prediction-making process of machine learning (ML) systems. Evasion attacks target the inference phase of the machine learning model lifecycle and compromise the integrity of the machine learning model's predictions. They exploit misclassifications using well-crafted malicious inputs, so-called 'adversarial examples', to confuse machine learning models into making an incorrect prediction. As shown in a seminal work from Goodfellow, Shlens, and Szegedy, after adding a minimum amount of noise to the testing data, which is imperceptible to a human, the classification result has been successfully compromised from a "panda" to a "gibbon" with surprisingly high confidence (99.3%). Consider a situation where an attacker takes control of electronic signs, especially stop and speed limit signs; the consequences of such attacks can be devastating. Those attack patterns don't require extra noise, but small changes to a stop sign may be enough for the network to recognise it as a "50 km/h" speed sign. The above example highlights the vulnerability of AI models to security attacks, as even minor modifications to the input data can lead to significant errors in the model's predictions. The consequence of such attacks is the erosion of trust in AI systems and potential negative impacts on various applications relying on these models.

In our work, we concentrated on model evasion because it is the easiest and cheapest attack for the attacker and also the most realistic scenario. This case is realistic because we assume that the attacker does not have access to the model, and does not know the weights or model architecture, but can query it multiple times, see the model output, and tune their attack based on that.

A key piece of technology built as part of the Spatial project is the security diagnosis library, which aims to identify security vulnerabilities and describe the security state of ML systems. This library focuses on assessing and improving resilience against evasion attacks, which are the most common adversarial ML attacks. Many methods and tools have been developed to generate adversarial examples for performing evasion attacks. These can be used for empirical vulnerability assessment: they are executed against an ML model to infer how vulnerable it is against this threat. However, existing methods are typically restricted to a few domains and types of input, e.g., image, sound or text, and to a few types of ML models, such as neural networks. To apply to many other ML applications, the security diagnosis library implements generic approaches to evasion attacks and the generation of adversarial examples.

We have introduced several important and unique features to the security diagnosis library, enabling effective and comprehensive testing of the model's resilience:

- **Applicability to a wide range of ML models.** Our solutions allow us to run security tests against various types of models, e.g. SVM, Naïve Bayes, Random Forrest, Gradient-boosted decision trees, but also Deep Neural Networks.
- **Applicability to various frameworks.** The security diagnosis library enables the testing of different ML frameworks, such as Scikit-learn, TensorFlow, PyTorch, XGBoost, ONNX, etc.
- **Various types of input data.** Model input features can have different modalities compatible with the analysed model, such as tabular data in CSV format or image files.
- **Wide range of evasion attacks.** In the latest version, security diagnosis library enables testing models using Simba, Zoe HSJ and NES attacks, each with its own set of configuration parameters that can be defined by the user/attacker.
- **Attack hyperparameter optimisation capability**. Our solution offers a method for users to search for the optimal parameters for a specified attack. This labour-intensive process can be set up to run in the background, and upon identifying the ideal combination, the attack is automatically executed using the discovered parameters.
- **Iterative measurement of attack progress. This functionality allows one to observe attack effectiveness during execution and collect relevant metrics,** which can be stored and later used for in-depth analysis and visualisation.
- **Storing best adversarial sample.** During the attack, an interactive search is conducted for the optimal sample to introduce the minimal perturbations of features necessary to achieve the attack's goal. As some algorithms eventually begin to yield less effective samples, this functionality allows for selecting the best sample rather than merely the most recent one.
- **Modifying attack success criterion.** This functionality enables the execution of successful attacks on binary classification models that utilise a custom classification threshold to determine the target class. This is especially beneficial in fields such as cybersecurity, where the output probability must be exceedingly high for a sample to be deemed malicious for certain machine learning models.
- **Definition of feature space constraints.** It allows the definition of various sampling strategies that govern the modification of input features to launch an attack successfully. These strategies can be relative or absolute regarding feature values and may include value clipping, rounding, monotonic constraints, etc. Additionally, the user can decide if they should be applied globally or separately for each feature.
- **Custom metrics for model security.** We have developed metrics for complexity, detectability, and

global feature distortions. These metrics facilitate the assessment of an attack's effectiveness and the model's security, and notably, they enable the comparison of different models' security against one another.

- **Security assessment report.** A proper assessment of the model's resilience should be grounded in a comprehensive and intelligible report. Users have the option to specify whether to preserve various file artifacts generated during attacks and whether to save the analysis results in HTML format.

As the use of ML models in contemporary IT systems grows, so does the concern for their security. Ensuring their security is crucial, as initial research using our library indicates that even well-functioning models can become vulnerable to attacks when the right tools are employed. The security diagnosis library aims to equip AI developers with the means to secure their creations and to balance the trade-off between performance and security finely.

The **COMPASS** framework.
Erasmus University Rotterdam

The **COMPASS** framework enables organizations to critically evaluate both technical innovation potentials and societal impacts of AI systems to ensure a responsible and trustworthy AI system. COMPASS provides a roadmap for AI developers and auditors to navigate the complex landscape of AI system and its impact on society. This framework offers flexibility for AI practitioners to tailor an evaluation process to the industries` specific needs and priorities through a self-assessment process. It helps identify where to leverage necessary skills by making use of the SPATIAL design principles.

By incorporating all components of COMPASS into the life cycle of AI systems, trust, fairness, and societal benefits can be fostered in the era of AI.

COMPASS stands for:

- **CONTEXT:** Defining and determining the context of AI system, including who is involved in its lifecycle, and who is affected by the system.
- **OPENNESS AND TRANSPARENCY:** Ensuring AI systems and algorithms are transparent and understandable for all stakeholders, considering access and documentation.
- **MEASURES:** Iteratively developing mechanisms for evaluating AI systems so that they operate fairly and reliably.
- **PRIVACY POTENTIALS:** Safeguarding user privacy and data protection throughout the AI life -cycle.
- **ACCOUNTABILITY:** Highlighting trustworthiness and holding AI systems and developers accountable for their actions.
- **SECURITY and SAFETY:** Implementing security measures and safety precautions throughout the AI life cycle to minimize the potential attacks.
- **SUSTAINABILITY:** Integrating mechanisms to maintain the reliability and performance of AI systems while providing sustainable and environmentally friendly AI solutions.

The adoption of **COMPASS** is intended as a self-reflexive evaluation process to develop reliable, bias reduced processes throughout the AI lifecycle. Strengths and weakness in AI development and deployment become immediately visible, to define areas for improvement or development. COMPASS allows an organisation to:

- identify weaknesses at every stage of the life cycle,
- determine potential reasons for these weaknesses (e.g. biased training data, inadequate tracking systems),
- implement best practices through guidelines and resources.

COMPASS provides an actionable evaluation process that guides users towards effective trustworthy AI development processes.

For more information, please click **here**.

The Creation Process of the Educational Modules

Minnalearn

In October 2023, we proudly launched two educational modules, **“Trustworthy AI”** and **“Advanced Trustworthy AI,”** which aim to equip learners with essential skills in understanding and managing artificial intelligence systems responsibly. Here’s a detailed look at the creation and promotion of these modules.

From the outset, our primary goal was to ensure the modules were highly relevant to our target audience. To achieve this, we incorporated interviews and testing into as many phases of the project as possible.

The creation began with two concept sprints, each lasting approximately 1.5 months.

First Sprint

University Workshops and Stakeholder Interviews: We initiated workshops and interviews to gather diverse perspectives.
Concept Drafting and Market Research: These insights fed into drafting and refining the initial concepts.
Concept Packaging and Testing: The drafted concepts were then packaged and tested for feedback.

Second Sprint

Further Interviews and Feedback: This phase involved additional interviews with universities, businesses, and other stakeholders.
Finalizing the Concept: We integrated the feedback to finalize the module concepts.

Developing the Course Material

After finalization, the actual content creation process started.

- Academic Source Material: Universities involved in the project gathered and reviewed academic sources to ensure the content was accurate and relevant.
- Instructional Design: Experts and instructional designers outlined the course structure, defined learning objectives, and devised assessment strategies.
- Copywriting: Copywriters translated academic material into engaging, clear content suitable for online learning.
- Visual Content Creation: Artists and graphic designers developed visual aids to enhance the learning experience.
- Assembling the Course: The course material was integrated into an online learning platform, incorporating text, quizzes, and visual elements.

Testing the Education Module

Once assembled, we invited graduates of the Building AI online course to beta-test the modules. We received nearly 310 applications, and the testers provided extensive feedback on various aspects of the course. The content, clarity, usability, technical aspects, and educational effectiveness were thoroughly tested.

A dedicated team of content experts and instructional designers reviewed the feedback and made necessary revisions to improve the course.

Course Details and Learning Objectives

Trustworthy AI

Available at this link **(https://courses.minnalearn.com/en/courses/trustworthy-ai/)**

Learning Objectives

Understand AI's impact on societal dynamics, environmental factors, market structures, and democratic processes.

Assess potential biases within AI systems and recognize these biases in decision-making processes. Evaluate the necessity of explainability in AI applications and employ approaches to enhance transparency.

Understand security and privacy risks associated with AI systems and address these risks effectively, along with familiarizing oneself with relevant legal obligations.

Trustworthy AI

Available at this link **(https://courses.minnalearn.com/en/courses/trustworthy-ai/)**

Learning Objectives

- Apply metrics and toolkits to quantify bias within AI systems and employ mitigation techniques.
- Utilize industry-standard libraries to implement explainable AI (XAI) methods.
- Analyze security and privacy risks specific to AI and develop strategies to mitigate these risks, ensuring responsible AI deployment.

Promoting the Course

We promoted the courses extensively within and beyond our network:

- University Outreach: Targeting students and staff of participating universities and those who had taken the Elements of AI course.
- Email Campaigns: Engaging our innovation networks through targeted emails.
- Newsletters and Social Media Campaigns: Dedicated newsletters and active promotion on social media.

Current Reach and Future Goals

By the end of November, the “Trustworthy AI” course had 180 students enrolled, and the “Advanced Trustworthy AI” course had 31 students. Enrollments span across North America, Europe, Asia, and Oceania.

Our aim is to enable the worldwide adoption of these crucial skills. Therefore, the courses are open to anyone willing to learn. We continue promoting them to universities, companies, and individual learners, striving for wide-scale adoption and impact.

We aim to foster a future where AI is developed and deployed responsibly, ethically, and effectively by equipping learners globally with these critical skills.

SPATIAL Platform

University of Tartu

The SPATIAL platform is a platform for estimating the trustworthiness of AI based on a combination of metrics deployed in the form of services to qualitatively analyze and assess AI models and SPATIAL use cases. The platform aims to provide a higher level of explainable AI solutions based on accountability, privacy, and resilience. Internally, it comprises several key components for achieving its design goals.

The platform consists of various components such as the quality metrics components, the explainability components, and the interactive interface components. The quality metrics components are cascaded to various services that evaluate the AI model accountability (using accuracy, effectiveness, confidence, compacity, and consistency), resilience (using impact metric), and privacy (using differential privacy). The explainability components reveal the internal workings of the AI models through various methods, making the decision-making process understandable to different stakeholders in a concise and easy-to-understand way. The interactive Interface component is an adaptive user interactive that enables stakeholders to engage with the platform. It adapts information (result and analysis) provision to the user's profile. Insights from quality metric analysis, explanations, and models within the platform are transformed into comprehensive reports. These components enable SPATIAL to provide a powerful and versatile platform that empowers users to assess and understand the trustworthiness of AI models confidently.

Structurally, the SPATIAL platform is designed based on microservice architecture, allowing for each component to have a collection of independent services running individual processes to achieve the design goals of the platform. Additionally, this design choice gives room for modularity, scalability, and efficient validation of various properties contributing to the overall trustworthiness of the AI model before deployment. For instance, the quality metric component can be further cascaded into dedicated microservices for accountability metrics, resilience metrics, and privacy metrics. Each microservice can then perform an in-depth analysis of its specific aspect, contributing to a comprehensive assessment of the AI model.

Functionalities

The SPATIAL platform is designed and developed to simplify the process of building and refining AI models. Its model-building functionalities allow for the construction and building of AI models within the platform. One of its standout features is the Explainable AI (XAI) service, which uses various methods to make the AI's decision-making process transparent. This transparency allows users to understand how AI models arrive at their conclusions, enhancing trust in the technology. By providing clear and adaptive explanations through the integrated LLM service, developers can better explain results to different users, boosting their confidence in the AI's methods.

SPATIAL also includes quality metric services that enable model refinement to meet regulatory standards. The platform's fairness service examines training data to identify and address potential biases, ensuring the AI models produce fair and unbiased results. Additionally, SPATIAL offers a privacy service to ensure that models comply with data privacy regulations, protecting sensitive information and promoting responsible data handling practices. The accountability service within SPATIAL evaluates key aspects of AI models such as accuracy, transparency, and traceability. This comprehensive assessment fosters trust and allows users to understand the decision-making processes of their AI models. Together, these services make SPATIAL a reliable and ethical choice for developing AI solutions.

DISCLAIMER

The information, documentation and figures available in this publication are provided by the SPATIAL project's consortium under EC grant agreement 101021808 and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

This document does not reflect the opinion of any SPATIAL partner or any organisations with which the experts are affiliated. The SPATIAL project and its consortium partners are not liable for any consequence stemming from the reuse of this publication.

Funded by
the European Union

SPATIAL