TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Janno Jaal 212061IVCM

THREAT MODELLING FOR AI/ML-BASED HEALTHCARE SYSTEMS

Master's Thesis

Supervisor: Hayretdin Bahşi
PhD

Tallinn 2024

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Janno Jaal 212061IVCM

AI/ML-PÕHISTE TERVISHOIUSÜSTEEMIDE OHTUDE MODELLEERIMINE

Magistritöö

Juhendaja: Hayretdin Bahşi
PhD

Tallinn 2024

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Janno Jaal

12.05.2024

Abstract

A huge amount of data is generated by electronic health records, various biosensors and other means. To receive benefits like more effective detection of diseases from these enormous amounts of data, Artificial Intelligence (AI) and Machine Learning (ML) systems are becoming more widely used. With this technological progress comes the potential for new or previously overlooked security threats. Because of the nature of the system, the security issues within healthcare systems could bear devastating consequences. This thesis identifies the threats opposed to AI/ML-based healthcare systems by conducting comprehensive threat modelling and threat analysis. The model captures all the characteristics of a modern healthcare system that utilizes the usage of an AI/ML component with an in-house development approach. The model has different ways to gather data and interact with patients and doctors. The threat modelling is conducted based on the STRIDE methodology. In addition, STRIDE-based attack trees are used to further identify all the relevant threats that could endanger a modern healthcare system. As a result, a comprehensive list of identified threats is provided for all the components that are used in a modern healthcare AI/ML-based system. The threat list consists of conventional and AI/ML-specific threats. For AI/ML-specific threats to be successful, they need some form of a conventional attack to be carried out beforehand. The model itself and the threats identified are validated by various experts from the cybersecurity and AI/ML field. This study aims to contribute to the safe and effective implementation of AI/ML technologies in healthcare settings.

The thesis is written in English and is 71 pages long, including 6 chapters, 25 figures and 6 tables.

Annotatsioon

AI/ML-põhiste tervishoiusüsteemide ohtude modelleerimine

Elektrooniliste terviseandmete, erinevate biosensorite ja muude vahendite kaudu genereeritakse suurel hulgal andmeid. Tehisintellekt (AI) ja masinõppe (ML) süsteemid muutuvad üha laialdasemalt kasutatavaks, võimaldades selles suures andmehulgas eristada informatsiooni, mis soodustaks näiteks tõhusamat haiguste avastamist. Siiski kaasnevad selle tehnoloogilise arenguga uued või varem tähelepanuta jäetud turvariskid. Tervishoiusüsteemi olemuse tõttu võivad need kaasa tuua tõsiseid tagajärgi. Käesolev magistritöö tuvastab võimalikud ohud AI/ML-põhistele tervishoiusüsteemidele, viies läbi põhjaliku ohtude modelleerimise ja analüüsi. Koostatud mudel hõlmab kõiki tänapäevase tervishoiusüsteemi külgi, milles kasutatakse majasiseselt arendatud AI/ML komponente. Mudel demonstreerib erinevaid viise andmete kogumiseks ja suhtlemist patsientide ja arstide vahel. Ohtude modelleerimine koostati STRIDE metoodika alusel. Lisaks kasutati STRIDE-põhiseid ründepuid, et tuvastada täiendavaid probleeme, mis võiksid ohustada kaasaegset tervishoiusüsteemi. Tulemuseks on põhjalik loetelu kõigist tuvastatud ohtudest, mis on seotud kaasaegse AI/ML-põhiste tervishoiusüsteemidega. Ohtude loend jaguneb tavapärasteks ning AI/ML ohtudeks. AI/ML ohtude realiseerumiseks peab eelnevalt toimuma edukas tavapärane rünnak süsteemile. Mudeli ja tuvastatud ohud on valideerinud erinevad küberkaitse ja AI/ML valdkonna eksperdid. Käesolev töö püüab aidata kaasa AI/ML tehnoloogiate ohutule ja efektiivsele rakendamisele tervishoiusektoris.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 71 leheküljel, 6 peatükki, 25 joonist, 6 tabelit.

List of Abbreviations and Terms

AI	Artificial Intelligence
BAN	Body Area Network
DFD	Data Flow Diagram
DOS	Denial-of-service
EMR	Electronic medical record
IoT	Internet of Things
MITM	Man-In-The-Middle
ML	Machine Learning
PII	Personal Identifiable Information
STRIDE	Spoofing, Tampering, Repudiation, Information disclosure, Denial of service and Elevation of privilege
SQL	Structured query language

Table of Contents

List of Figures

List of Tables

1. Introduction

Artificial Intelligence (AI) or Machine Learning (ML) based systems are undeniably gaining popularity. AI or ML components within various systems play a key role in automated image and speech recognition, natural language processing, predictive analysis and much more. These kinds of systems can be used within various fields and they are around us all the time during our everyday lives. For example different chatbots, autonomous vehicles, healthcare diagnostics and many other applications. This has numerous benefits and creates different significant possibilities for use.

Healthcare systems play a vital role in our lives. These systems are responsible for a wide range of functions and services aimed at promoting and maintaining the health of individuals and communities. Already a few years ago different systems like IBM Watson for Oncology were introduced into the healthcare sector [1], but now they are gaining even more popularity. An enormous amount of health data is generated through electronic health records, imaging, sensor data and text [2, 3]. Because of that AI and ML-based systems are becoming more widely used in the healthcare field because of huge data handling capabilities and all the benefits they provide. The benefits could be more effective detection of diseases, management of chronic conditions, delivery of health services, and drug discovery [4, 5]. The AI/ML-based systems are considered a possible partnership for doctors to further improve clinical outcomes for the patient [6]. In addition, the systems can help develop the healthcare field itself [7, 8]. Also, the combination of AI and ML systems with different Internet of Things (IoT) devices can help solve different complex problems [9, 10]. With the advancements in technology making the micro-controllers and micro-processors smaller and faster more high-precision sensors can be made. Various health metrics could be monitored and analyzed within a very short period. These new advancements in healthcare technology do not only have benefits for the patients as well but also for doctors, nurses, pharmacists, and researchers [11].

New systems always bring new security concerns and because the systems are developing very fast the security might be overlooked. Safeguarding against potential vulnerabilities and threats should be considered during development and kept in mind throughout the deployment.

1.1 Motivation

Healthcare systems are considered a vital part of critical infrastructure in many countries. Cyber attacks targeting healthcare systems can lead to disruption or destruction of critical infrastructure and that can lead to very serious consequences with human casualties. New AI/ML-based systems are developed at a very quick pace, but with these developments, new threats emerge. These new adversarial threats to AI/ML systems aim to manipulate the model in working conditions or manipulate the training process.

Threat modelling is one of the ways to tackle this issue. Threat modelling offers a systematic approach to identifying potential cybersecurity risks and vulnerabilities in an AI/ML-based healthcare system already in the early phases of development. In addition to the conventional cybersecurity threats, the threat modelling framework has to also cover the AI/ML-specific threats. There is a need to be able to cover both of the threat types, as they can not be separated. The AI/ML-specific threats need some form of conventional attack to be carried out beforehand.

Threat modelling makes sure all the possible threats are identified and appropriate countermeasures are included in the early development phase. Through this AI/ML-based healthcare systems could be securely developed.

1.2 Research problem

The following questions this research tries to answer are:

1. How to model AI/ML-based healthcare systems?
   Sub questions:
   (a) How to model the AI/ML part of the system?
   (b) Where are the boundaries between components?
2. What kind of threats are there for healthcare AI/ML systems?

1.3 Scope, goal and novelty

This research focuses on AI/ML-based healthcare systems that provide help with detecting diseases, managing difficult conditions, discovering drugs or medical research. Out of scope are AI/ML-based systems that primarily act as chatbots or medical service delivery systems, thus providing a more targeted analysis. The AI/ML-based healthcare system used for threat modelling uses in-house development for the AI/ML model. Although there

is a possibility to acquire the pre-trained model from a third party, then in this particular case the in-house development approach is used. The in-house development approach will give insight into the development of the model and will give more AI/ML model-related content to be researched. Threat modelling is conducted in the early stages of the system development life cycle. Thus it is done on a conceptual level and no real attacks will be carried out.

The goal of the study would be to conduct threat modelling for an AI/ML-based healthcare system that uses an in-house development approach. Through threat modelling, it is possible to identify potential threats to AI/ML-based healthcare systems. Subgoal of that task would be to provide the methodology for threat modelling AI/ML-based healthcare systems. Ultimately the research can contribute to developing security measures that ensure the safe and effective implementation of AI/ML technologies in healthcare.

This study focuses on a threat modelling approach tailored specifically for AI/ML-based healthcare systems, a domain where previous research often falls short in accurately modelling the system. This study tries to provide a holistic understanding that conventional cybersecurity and AI/ML-specific threats must be assessed together. As AI/ML-specific attacks need some form of conventional attack to be carried out beforehand, they can not be separated. This is done by leveraging the STRIDE framework in combination with attack trees. This research tries to fill a crucial gap in the literature and offers valuable insights into the security of healthcare AI/ML systems, ultimately contributing to their resilience against potential threats.

The thesis consists of 6 chapters. This is the first chapter and handles the introduction. Chapter 2 gives the background information for this thesis. The model creation process is described in Chapter 3. Chapter 4 deals with the identification of threats. Chapter 5 is for discussion of the findings. The thesis ends with the summary that is provided in Chapter 6.

2. Background

The following chapter will provide additional information needed to understand this thesis. The background chapter will cover the basic definitions of threat modelling, machine learning and artificial intelligence. This is followed by a comprehensive literature review to highlight the related work for this thesis.

2.1 Threat

A "threat" signifies potential harm or danger from various sources, including conflicts, environmental issues, societal unrest, and technological weaknesses [12]. In cyber security, a threat is defined by Jensen et al. as "A possible danger to a computer system, which may result in the interception, alteration, obstruction, or destruction of computational resources, or other disruption to the system" [13]. In addition, threat can mean any impact on the organizational assets or individuals through unauthorized access, destruction, disclosure, modification of information, or denial of service [14].

2.2 Attack

An attack has multiple definitions, but the context is important. It can mean an aggressive action, criticising someone or something else. But in cyber security by definition, an attack displays "Any kind of malicious activity that attempts to collect, disrupt, deny, degrade, or destroy information system resources or the information itself" [15]. An attempt to get unauthorized access to information is also considered a cyber attack.

2.3 Threat modelling

Threat modelling is "A form of risk assessment that models aspects of the attack and defence sides of a logical entity, such as a piece of data, an application, a host, a system, or an environment" [16]. Threat modelling can help identify threats, attacks and countermeasures that could affect the system in question. Implementing threat modelling in the systems development life cycle can effectively increase the system's security. Threat modelling is also a core element of the Microsoft Security Development Lifecycle, and Microsoft defines five major threat modelling steps [17]. These five steps are (See Figure 1):

- defining security requirements,
- creating an application diagram,
- identifying threats,
- mitigating threats,
- validating that threats have been mitigated.

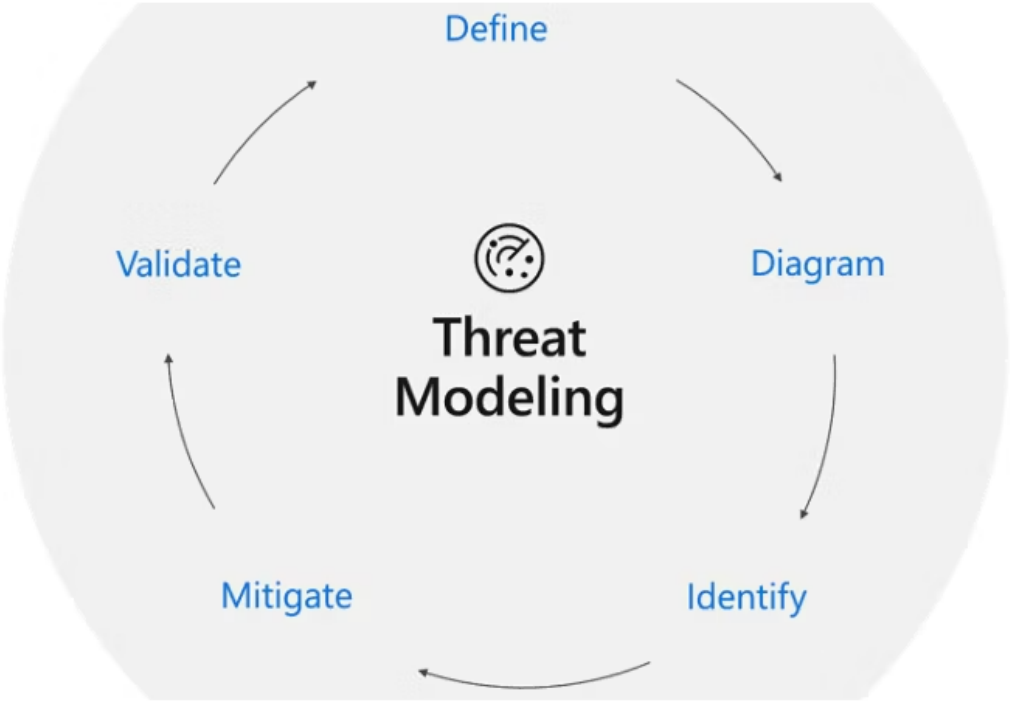


Figure 1. *Threat modelling steps [17].*

2.3.1 STRIDE

The STRIDE threat modelling method was invented by Loren Kohnfelder and Praerit Garg in 1999 [18]. Both of them were part of the security team at Microsoft where they developed the STRIDE model as a framework for analyzing and addressing potential security issues in software systems. STRIDE is an acronym that stands for Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege. Each category is described in detail in the table below (Table 1). The goal of the framework was to help people developing software identify the usual attacks that could occur. Security professionals commonly use the STRIDE model.

Table 1. *STRIDE method description [18].*

Threat	**Threat Definition**
Spoofing	Pretending to be something or someone other than yourself
Tampering	Modifying something on disk, network, memory, or elsewhere
Repudiation	Claiming that you didn't do something, or were not responsible
Information Disclosure	Providing information to someone not authorized to see it
Denial of Service	Absorbing resources needed to provide service
Elevation of Privilege	Allowing someone to do something they're not authorized to do

Microsoft Threat Modeling Tool

The Microsoft Threat Modeling Tool is a tool developed by Microsoft to make threat modelling easy for all developers [17]. The tool itself allows the users to model the system through standard notation for visualizing system components, data flows, and security boundaries. In addition to the modelling part, the tool also provides a comprehensive analysis of the model built based on the STRIDE method. The tool also makes suggestions for mitigation of security issues. The tool is free to use and is widely used for threat modelling.

2.3.2 Data Flow Diagram

Data Flow Diagram, or DFD for short, is a diagram that shows how data flows logically through the application [19]. DFD shows the processes that transform data, the data stores where data is held, and the data flows that move data between processes and data stores (See Figure 2). DFDs are used to decompose the applications. Three main strengths for using DFDs could be their simplicity of notation, their ability to manage complexity, and the fact that they are technology-agnostic [20].

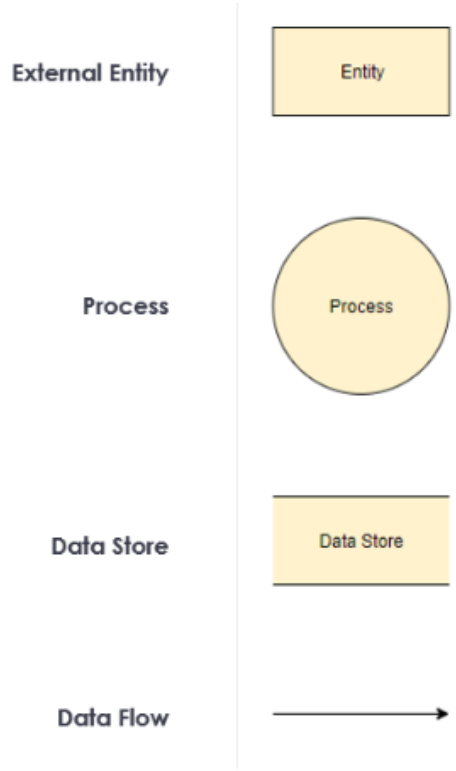


Figure 2. *DFD notation elements [21].*

2.3.3 Attack trees

In 1999 Bruce Schneier wrote about the topic "Attack trees provide a formal, methodical way of describing the security of systems, based on varying attacks. Basically, you represent attacks against a system in a tree structure, with the goal as the root node and different ways of achieving that goal as leaf nodes" [22]. Attack trees are useful for describing the security of systems or their subsystems. The possible threats or attacks are represented as a tree structure. Attack trees could be used for analyzing the system after it has been modelled with a DFD or any other diagram [18]. Attack trees go well with threat modelling and are a useful tool when identifying threats in a system. An example of an attack tree is shown below (Figure 3).

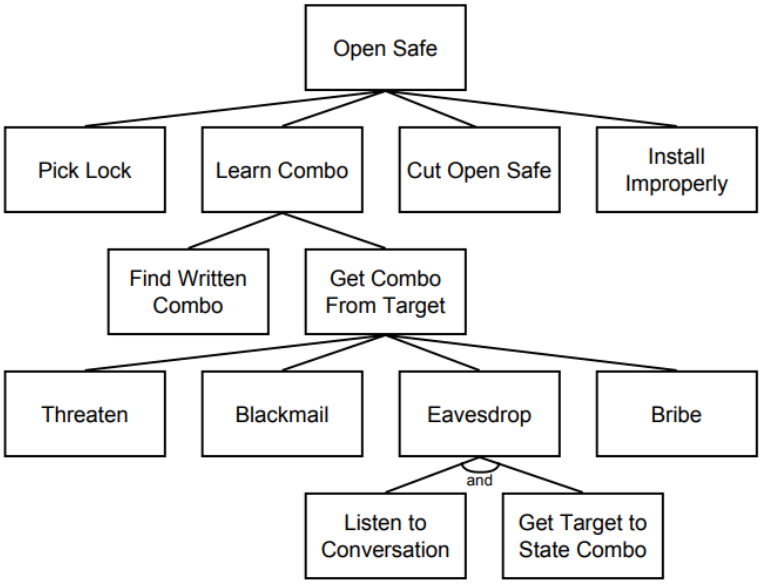


Figure 3. *Example attack tree [22].*

2.4 Artificial intelligence and machine learning

The term "Artificial Intelligence" was proposed by John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon already in 1955. The goal of McCarthy was to “find out how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" and that is quite a good explanation for the goals of modern AI as well [23]. Today we understand AI as a system that seemingly can use human intelligence to complete various tasks. AI today can be categorized into two: narrow AI and general AI. In the case of narrow AI, the system is designed to complete tasks in a limited domain. This includes voice assistants like Siri and Alexa, image recognition software, recommendation systems, and autonomous vehicles. General AI possesses human-like intelligence and can perform a wide range of tasks across different domains. General AI is still mostly theoretical. AI has applications across various industries and domains, including healthcare, finance, education, transportation and cybersecurity. Today the digital world today has enormous amounts of data, such as IoT, cybersecurity, and health data. The usage of AI, and especially machine learning (ML) could be the key to analysing the data and developing smart automated applications based on it [24].

Machine learning is a subset of artificial intelligence. Machine learning uses various algorithms like linear regression, logistic regression, decision trees, support vector machines, neural networks, deep learning, clustering algorithms and more. Using these algorithms the system learns from huge amounts of data to skillfully complete different tasks. The

main machine learning techniques are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning [25]. Each technique requires different data. Techniques and their data requirements can be seen in the figure below (Figure 4).

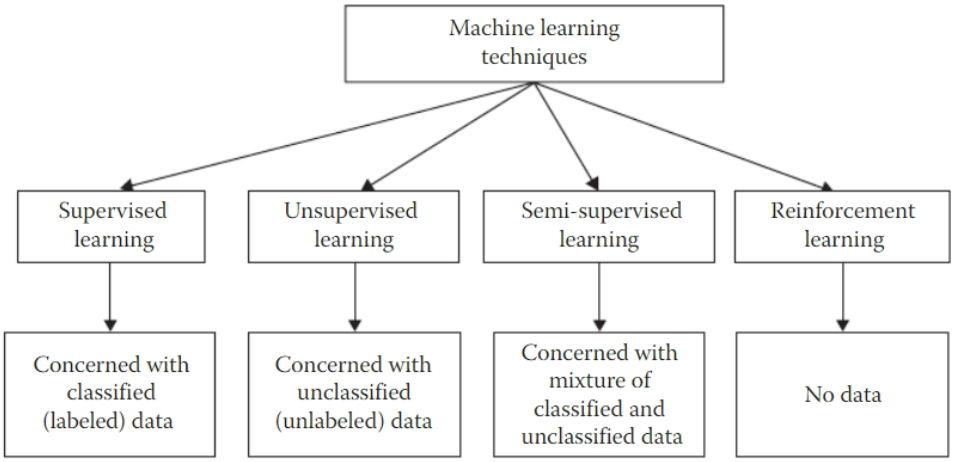


Figure 4. *ML techniques and required data [25].*

Systems with artificial intelligence and machine learning are on the rise. These systems can be used to further improve different fields like education, science, healthcare, cybersecurity and much more. AI technologies can potentially be the most powerful tool for expanding knowledge, increasing prosperity and enriching the human experience [26].

2.4.1 Adversarial machine learning

The field of adversarial machine learning studies the vulnerabilities in machine learning approaches. The goal of these studies is to develop methods to withstand adversarial manipulation for machine learning systems [27]. The attacks usually are split into two categories. The first focuses on attacks that interact with the model itself. In this case, the adversary has query access to the model and from there can craft different malicious attacks. The second approach focuses on the manipulation of the training phase. If the adversary has access to training data or similar attributes it is possible to maliciously manipulate the training process.

2.5 Related work

The following section will give a comprehensive overview of the relevant literature and research conducted in the AI/ML-based healthcare systems threat modelling domain.

In the domain of threat modelling the STRIDE framework has been widely used. STRIDE

is popular because it provides a structured approach to analyzing security risks and vulnerabilities in software applications. STRIDE framework also has been considered to be an option when threat modelling involves AI/ML-based systems. In 2020 Wilhjelm et al. explored ways to elicitate security requirements for machine learning based systems [28]. In their paper, they examined the applicability of Data Flow Diagrams and STRIDE. For the Data Flow Diagram, they provided a solution where the model is borrowed or integrated from a third party. This is a different approach to this thesis. In their diagram, the external user communicates via API with a machine learning model which has access to the data store. While the DFD gives a good general overview of a machine learning based system, it is lacking some components. There is no performance validation or explanation of how the data is stored in the data store, as the model can only query. Still, it gives some initial ideas for creating DFDs for ML-based systems.

In 2022 Alatwi et al. conducted threat modelling for machine learning based network intrusion detection systems [29]. Similarly, they used the STRIDE framework and additionally the Attack Tree approach as well. For the DFD they did not use the Microsoft Threat Modeling Tool, which when using STRIDE would make a lot more sense. Different kind of approach was used for the notation of elements. Apart from that the Data Flow Diagram uses an in-house development for the model. While the diagram is in good detail, that covers most of the machine learning lifecycle elements, it could use some more. There is no end-user or any other entities presented. There is no performance validation or operational stage elements considered. Other than that the STRIDE framework and attack tree approach seemed to give good results.

Mauri et al. created a framework to model AI/ML threats using STRIDE [30]. Their idea was to use the Failure Mode and Effects Analysis (FMEA) process to identify how assets are generated and used in various stages of an AI/ML system might fail. Then they linked the CIA3-R hexagon with it and in the end, mapped STRIDE threats to the identified assets. The presented framework could help practitioners to choose appropriate security measures for the ML assets. The authors also demonstrated this approach with a case study. The Data Flow Diagram used non-standard notation and was not very much in detail. The diagram did not include many crucial processes in the AI/ML lifecycle like data engineering, performance validation, model training, and model evaluation. While the framework itself is promising, the modelling part of the paper is lacking.

Additionally, the STRIDE method has been also used in the world of IoT sensors. Asif et al. used STRIDE to model threats for IoT-enabled precision agriculture systems. The paper gives good input on sensor modelling [31].

While there are many works on how STRIDE can be used when dealing with AI/ML-based systems there are not many works related to the healthcare domain, which is the focus of this thesis.

The article by Yaacoub et al. written in 2020 tackles the issues of securing the Internet of Things medical systems [32]. The paper does not include AI/ML-based systems in the study, but the paper still gives a really good input on how to deal with medical IoT sensors. This is good information for creating Data Flow Diagrams because AI/ML-based systems can get their data from medical IoT systems.

Recurring threats to AI/ML-based healthcare systems were presented by Sundas et al. in 2022 [33]. Even though they did not conduct threat modelling this paper gives good information on threats in the AI/ML healthcare domain. The authors discuss different attack types that are related to AI/ML systems. They also carried out different poisoning and evasion attacks and provided a comprehensive analysis. While this thesis will not conduct practical attacks, the knowledge from this paper gives good input for threat identification and evaluation. They also provided an example structure, but as the focus was not on threat modelling, that does not give much value to this thesis.

The most similar paper to this thesis was conducted by Cagnazzo et al. in 2018 [34]. The authors conducted threat modelling for mobile health systems with the STRIDE framework and also DREAD. The provided Data Flow Diagram gives a good overview of a mobile health system, but the artificial intelligence part is missing a lot of components. They decided to model the AI system as an external entity and nothing more. Even if the components are integrated from a third party still there are some trust boundary issues present. Modelling the AI part definitely has issues, as it overlooks many crucial AI/ML systems lifecycle components. STRIDE and DREAD framework usage gives good input, but the results might not be the best regarding AI threats, because of the simplicity of the Data Flow Diagram.

There have been a lot of papers that try to map STRIDE to AI/ML-based systems in various domains. There is no structured approach to the modelling part in place right now, as all the systems are modelled differently. Mainly the modelling of the AI/ML part of the system is done in various ways. Some of the models just use one component, either process or entity and call it the AI/ML model. Some of the models do not separate the different phases of the AI/ML model lifecycle. The goal of this thesis would be to conduct the modelling part more in detail so all important components of the AI/ML system lifecycle would be covered, and their threats.

3. Model creation

There are many possibilities when implementing AI/ML usage in a system. As stated in the AI and ML risks report created by the Information Systems Authority and Cybernetica, there are three different possibilities for implementing AI/ML usage in the system [35]. Implementation could be done by using an external service, using an external model or using self trained model for the system. As was defined in the scope, this study uses the third case - in-house development and training for the model. This means all steps regarding building the model and usage of the model have to be represented in the Data Flow Diagram.

The following chapter will give an overview of how the model is created and what are the points of consideration. Also, a thorough overview of the created Data Flow Diagram is given. Threat modelling was carried out using the Microsoft Threat Modeling tool. The whole model can be seen in Appendix 2 - AI/ML-based healthcare system model.

3.1 Data Flow Diagram

The Data Flow Diagram created for modelling purposes consists of two external entities, ten processes, two data stores and twenty-five data flows. In addition, the diagram has six different boundaries. All of the elements are described in detail in the subsections below. Selection for these elements and what were considered are discussed in the next sections.

3.1.1 External entities

Created DFD uses two "Human User" external entities. These entities are the Patient and Doctor. Both of them are connected to the AI/ML-based healthcare system and use it for different purposes. The goal of the patient is possibly to see personal health data that is collected centrally via IoT biosensors, get recommendations based on that data and thus make future visits more efficient. The doctor could use the system in question to more efficiently make decisions regarding treatment plans, additionally use the data gained for more efficient visits, and give other recommendations.

3.1.2 Processes

In the data flow diagram, there are 10 processes used. Four of them are for controlling sensors, one for displaying smartphone application, one for the central healthcare system and four for the model workflow. All of these processes are discussed in detail in the following matching sections.

3.1.3 Data stores

There are two data stores used in the data flow diagram. The first one of them is a regular database. This database acts as a central data repository where different processes can query data from the central healthcare system. The database responds to SQL queries containing different relevant information for the patients, doctors or the model workflow. The second data store used is a model registry. After data processing and model development, the model is ready for the operational stage and the model is deployed into the model registry. The model registry then responds to the queries coming from the central healthcare system. The responses are predictions based on the AI/ML model. Before sending the response to the central system, there is also a performance monitoring process.

3.1.4 Data flows

Different data flows are present in the modelled system. Twenty-five instances of "Generic Data Flow" were used when creating the diagram. These data flows include queries, responses, data from sensors, logins, configurations and more. Data flows that are characteristic of different processes are discussed in detail in the following matching sections.

3.2 Sensors and sensor controller

In the data flow diagram, different biosensors are represented. A biosensor is a compact analytical device that combines a biological sensing element with a transducer to detect and convert a biological response into a measurable signal for various applications (See Figure 5). These sensors are worn by the Patient entity and collect real-time relevant medical data. Sensors can capture heart rate, respiration rate, body temperature and much more [9, 10]. These sensors are typically connected to a controller unit [34, 31]. The purpose of the controller unit is to aggregate the data from the sensor, provide configuration information for the sensors and send firmware updates to add new features, fix bugs and more. From the sensor controllers, the data is extracted via different applications.

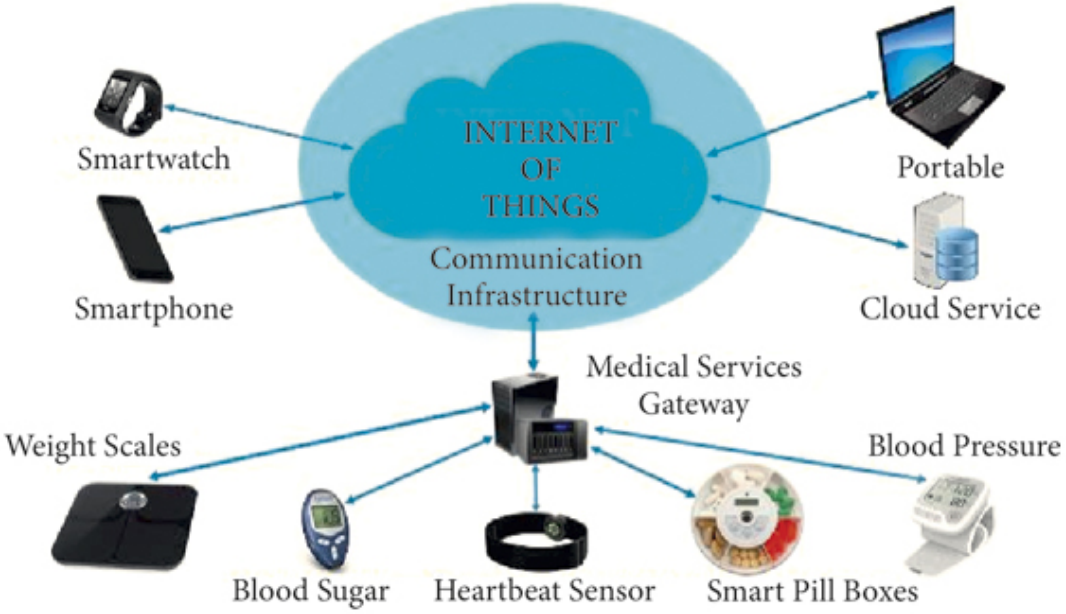


Figure 5. *IoT devices in healthcare. [36].*

The data from the sensors can be used in different ways. Firstly the data could just be stored in the central healthcare system and used whenever needed by the doctor or by the patient. This data could be used to create visualisations for different metrics or just queried from the database when needed. Secondly, the data could be used to create healthcare-related predictions by the AI/ML module. These predictions can help the patient by giving various recommendations. With these predictions, it could also be possible for doctors to make quicker and more thorough decisions involving clinical trials or suggested treatment plans. Thirdly, it could also be possible to use the data from the sensors as raw data for the training phase of the AI/ML component.

The created data flow diagram consists of 3 different sensors: blood sugar, heart rate and Sp02 (measurement of how much oxygen your blood is carrying [37]) sensors. All these sensors are communicating with the sensor controller. Sensors send data to the controller and receive configuration information or firmware updates if needed. The sensor controller can send all the data to the smartphone application.

3.3 Smartphone application

Today there are millions of smartphone applications available for download [38]. Smartphone applications are commonly used to connect to sensors and their controllers. These applications make it easy for the user to see all the relevant data in a visualised form. Also, there is a possibility to change the settings of different sensors. Very popular are different

watches or other gadgets that provide heart rate data and more for the users. The same kind of approach can be used as well to gather blood sugar levels and blood oxygen levels.

In the context of the created data flow diagram, the smartphone application is capable of receiving the data from the sensor controllers. The application can use the data to create meaningful graphs and visualisations for the user, thus helping them make better decisions regarding different activities. This could help notify the patient when to visit a doctor or give other recommendations based on the sensors used. Also, it is possible to change the settings of the sensor controller and therefore the sensors as well. The smartphone application then can communicate with a central healthcare system to provide even more metrics and relevant recommendations. The user of the application needs to be authorized by the application.

3.4 Central healthcare system

In the middle of the model, there is the Central Healthcare System which acts as a glue-like process that connects all parts of the system. All the entities and processes are connected to the central system in some way. One of the goals of the central system is to provide electronic medical record (EMR) data for patients and doctors. The EMR data can be directly queried from the central database. Data can be obtained directly using the system or via a smartphone application. Another goal of the system is to provide the data processing stage with raw data. The last goal of the system would be to query predictions from the model based on the patient data. These predictions can be used both by the doctors for better decision-making or by the patients to get recommendations or needed help. Latif et al. described a typical mobile health application architecture [39]. The approach with the central healthcare system shares different common aspects with the proposed typical architecture (See Figure 6).

3.5 Model workflow

There are various papers published that conduct threat modelling for an AI/ML-based system. Each paper proposes a different solution for the modelling as there is no one unified way and it is case-dependent. As this study uses the in-house development option, then all different kinds of phases of development need to be modelled for proper threat evaluation. Cagnazzo et al. propose the simplest presentation - the AI/ML Model is just displayed as one external entity [34]. This way of representation does not make it possible to properly highlight all the threats to the development and usage of an AI/ML system. A similar approach, where the model is not represented with more than a few elements is

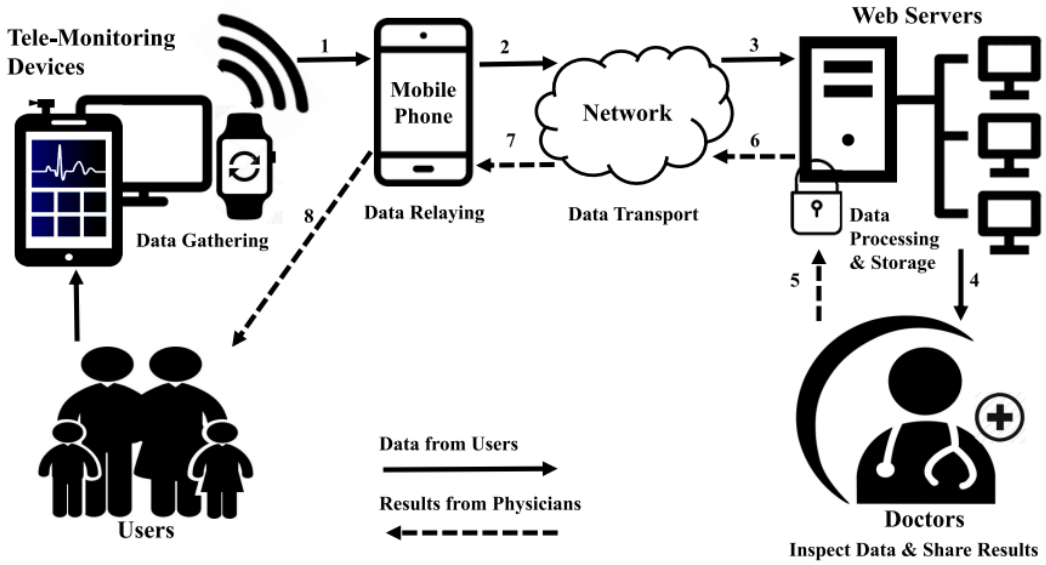


Figure 6. *Illustration of a typical mHealth architecture [39].*

used by Mauri et al. [30]. The most comprehensive way of representation is from Alatwi et al. [29]. Multiple stages are used for the development of the model, surrounded by various boundaries, but still, there could be more info about the data flows, model deployment and operational information. This study tries to represent all the parts that are needed for the development of an AI/ML model and the operational part as well.

In December 2020 the European Union Agency for Cybersecurity (ENISA) released a report called "AI CYBERSECURITY CHALLENGES" [40]. In this report, they also provide a lifecycle architecture for an AI system. According to ENISA, the lifecycle is as follows (See Figure 7):

- business goals,
- data ingestion, exploration and processing,
- feature selection,
- model selection/building, training, testing, validation and evaluation,
- model adaptation, deployment, maintenance,
- business understanding.

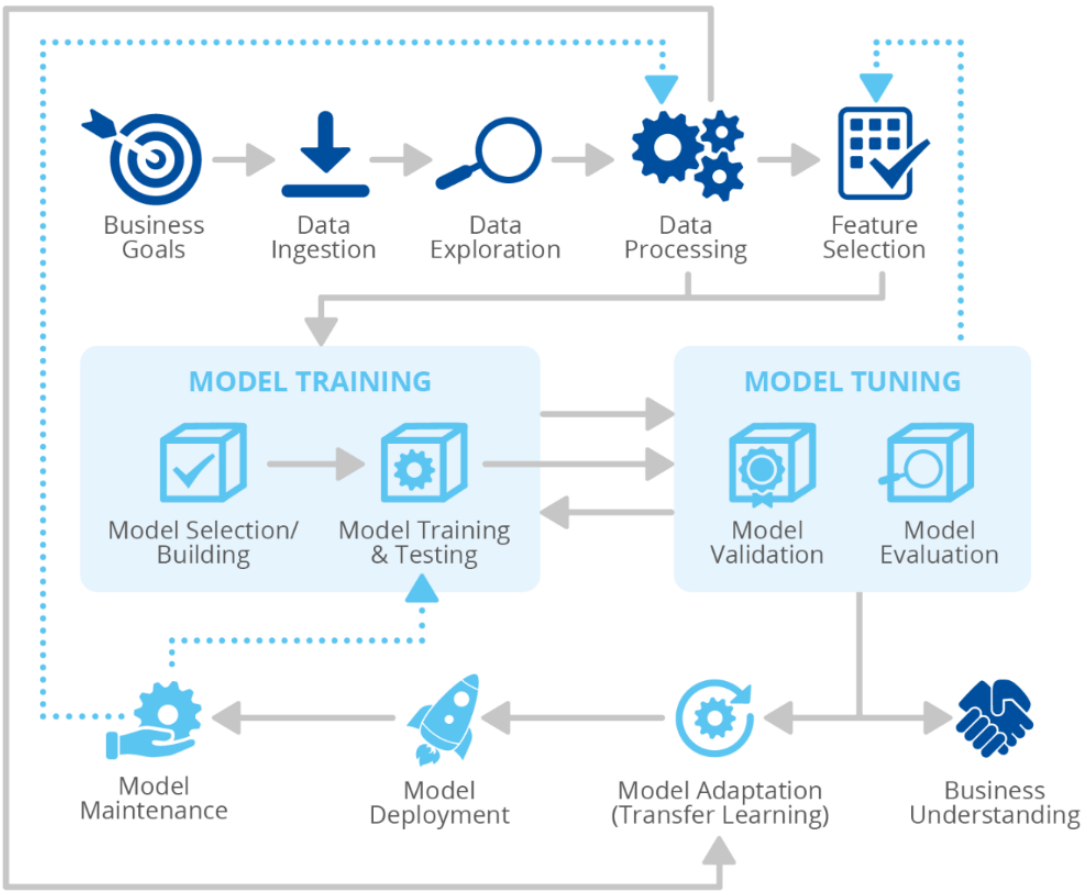


Figure 7. *ENISA generic AI lifecycle model [40].*

This model proposed by ENISA illustrates the vital parts of the AI lifecycle but also contains some information that is not relevant for conducted threat modelling. Business-related tasks can be left out as also model adaptation because the scope of the system states that the model is developed in-house rather than sourced from a third party.

The MLOps principles created by Visengeriyeva et al. state that AI/ML-based systems work is conducted in three stages [41]. These three stages are design, model development and operations (See Figure 8). The design stage focuses on gathering requirements and checking data availability thus it is out of scope for this research. The model development stage combines different data processing and model engineering tasks. The operations phase focuses on deployment, monitoring and maintenance of the already existing ML model.

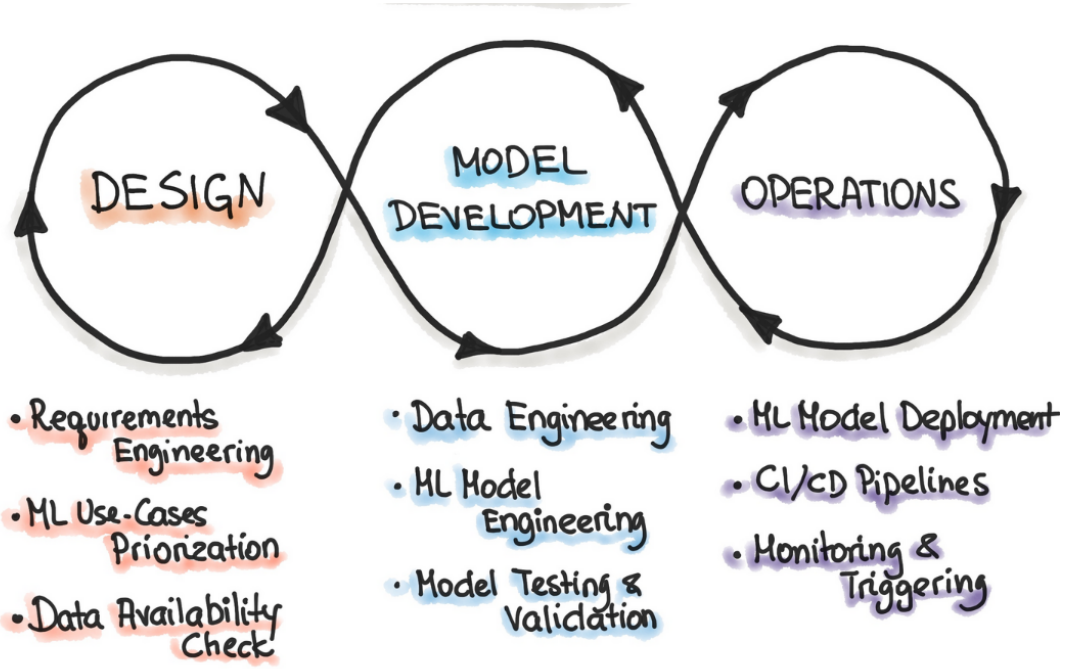


Figure 8. *MLOps processes [41].*

Combining these two principles from ENISA and MLOps it is possible to represent the lifecycle of an AI/ML model very clearly. The representation in the model is divided into three sub-categories: data processing, model development and model operation. This could be done with more granularity, but these three sub-categories are optimal for displaying the model lifecycle process in a data flow diagram. During these three stages, different tasks for data processing, building the model, training and evaluating the model, deploying the model and performance monitoring with maintenance are conducted. Saxena et al. created a framework for predicting suicidal attempts based on healthcare data [42]. Their AI architecture was presented with similar parts for the AI/ML system. This means that all the vital parts of the lifecycle are covered by the model. All of these three stages are described in detail in the sections below.

3.5.1 Data processing stage

The goal of this step is to analyse, clean and preprocess the raw data to make it suitable for the model development phase. After the initial analysis, the data can be further processed. This can include cleaning out irrelevant rows of data or handling missing values, outliers and inconsistencies. In this step also feature engineering is done. With techniques like identifying informative features and engineering new features that capture relevant data, it is possible to improve model performance. This stage is represented in the model with a data engineering process.

3.5.2 Model development

Model development can consist of multiple tasks, but for the diagram, all the tasks were categorised into two: model training and model tuning. With these two categories, it is possible to represent all the crucial parts of a model development phase without complicating the diagram too much. Both of these tasks contain vital parts for creating an AI/ML model. The goal of the model development phase would be to create, test and deliver a stable model that can be used furthermore for accurate predictions while ensuring the desired performance metrics are achieved. Model development includes different tasks related to the selection of the model, training and evaluation of the model. Model selection has to provide an appropriate learning algorithm while taking into account the model's complexity, interpretability, and scalability. After that, the model is trained and evaluated based on the data provided by data processing.

3.5.3 Model operation

After the model is developed then it moves into the operational stage. This stage is responsible for the deployment of the model, maintaining the model and also performance monitoring. After deploying the model from a test environment to a production environment it is crucial to maintain it with regular updates. Monitoring and documenting the model performance is the last step of the model lifecycle. Performance monitoring is needed to guarantee data quality and to gain visibility into the model performance [43]. In the data flow diagram data store is used for the environment where the model is deployed and the performance monitoring process takes care of the maintenance and monitoring of the model.

3.6 Trust boundaries

In data flow diagrams trust boundaries or just boundaries are there to indicate the line between different trust zones. These lines show what trust implies for the components of the system and show where the system transitions from a trusted environment to an untrusted one.

3.6.1 Sensor boundary

Different biosensors and one controller make up the sensor trust zone. Three different sensors are communicating with the sensor controller. From the sensors, the controller receives relevant health data as raw data. The sensor controller communicates with the

sensors as well. The controller sends different sensors various configuration information. All these communications from the sensors to the controller and vice versa are considered trusted interactions. Although the communication between sensors and the controller can also pose different security risks then in this system model they are trusted. The reasoning for this is that sensors usually and their controllers operate in a Body Area Network (BAN) which can be a secure environment. Poon et al. proposed a low-power bio-identification mechanism by using an inter-pulse interval (IPI) to secure the communications between different BAN sensors [44]. In [45], authors state that BAN sensors also need an intermediate device or gateway server. Using the device or the server the controller can send the data for example via Bluetooth. In the case of this model, the device is a smartphone application and it is out of the sensor trust zone. This means the interactions between the smartphone application and sensor controller are not trusted and there ends the sensor boundary environment.

3.6.2 Patient boundary

The patient boundary consists of the patient entity and the smartphone application. The only trusted communication happening in this trust zone is between the patient and the smartphone application. The patient sends login data to the application and in response, the application authenticates the user. There might be issues when the smartphone gets physically stolen or the smartphone is compromised with malware. These issues are not likely to happen and are not exactly the focus of this thesis, therefore the patient and smartphone application are represented in a trust zone together. Smartphone application also communicates with the central healthcare system. This interaction is not considered to be trusted.

3.6.3 Doctor boundary

Doctor boundary consists of the doctor entity and its communication with the central healthcare system. The doctor queries EMR data or other relevant information from the central healthcare system and the central system responds. The doctor entity can be susceptible to different spoofing and repudiation issues. Because of that none of the interactions in this zone are conducted in a trusted environment.

3.6.4 Data processing boundary

The data processing boundary only consists of one process - data engineering. Data engineering receives raw data from an untrusted zone. From data engineering training

data is passed on to the next process and this is also an untrusted interaction. In addition, performance validation data is sent to the performance monitoring process, also untrusted. This kind of approach with the boundary has been used in general AI/ML-based system threat modelling before as well. Both Alatwi et al. and Wilhjelm et al. created a separate boundary for the data processing section [29, 28].

3.6.5 Model development boundary

The model development boundary consists of two processes: model training and model tuning. The model training process receives from an untrusted data processing zone training data that can be used for training the model. After training the model the trained model moves into the model tuning process. Moving the trained model is a trusted action. Because according to ENISA the threats opposed to both of the processes are the same, they can share a trust boundary [40]. After validations and evaluations, the model is deployed to the model registry and this data flow is considered to be an untrusted data flow. Again this kind of approach has been used by Alatwi et al. [29].

3.6.6 Model operations boundary

Model operations boundary zone includes one data store and one process and their communications. The model registry data store sends predictions to the performance monitoring process. Based on these predictions the performance monitoring sends maintenance queries to the model registry data store. Both of these data flows are considered trusted interactions. The model registry is the location where the model is deployed. Deployment comes from the model tuning process and this flow is considered to be untrusted. In addition, the model registry also receives queries from the central healthcare system and responds with predictions, this interaction is also considered to be untrusted.

4. Identification of threats

The following chapter will give an overview of all relevant threats that could affect the modelled AI/ML components. Threats were identified using the STRIDE method and STRIDE-based threat tree patterns presented by Michael Howard et al. [46]. The mapping of STRIDE threat types to data flow diagram elements can be seen in the table below (Table 3).

Table 3. *Mapping STRIDE to DFD Element Types.*

Element Type	**S**	**T**	**R**	**I**	**D**	**E**
External Entity	X		X			
Data Flow		X		X	X	
Data Store		X	X	X	X	
Process	X	X	X	X	X	X

In addition, reports from The European Union Agency for Cybersecurity (ENISA), the National Institute of Standards and Technology (NIST) and Berryville Institute of Machine Learning were used for threat identification [40, 47, 48].

4.1 Threats to data flows

STRIDE method states that data flows are subject to tampering, information disclosure and denial of service.

Tampering

The data flow tampering attack tree states that tampering can take place when the integrity of the data flow channel or message is violated. Data flow channels with weak or no channel integrity can result in the attackers tampering with the channel. Additionally, channel tampering can also happen in man-in-the-middle (MITM) situations. Regarding the messages sent on the channels, weak or no message integrity could result in tampered messages. Also, if anti-replay defences such as time stamps or counters are not used then it is possible to replay valid messages. If the integrity of a message can be violated then this enables spoofing of the entity or process that receives the message.

Information disclosure

Information disclosure of data flow is possible when the channel and messages are observed or through side-channel attacks. In case of weak or no message confidentiality and no channel confidentiality, the disclosure of sensitive data is possible. The attacker could potentially view any sensitive data in this case. If the channel can be observed this opens up the possibility for an MITM attack and this could lead to listening and reading the data as well. Side channels allow the attackers to gather nonfunctional characteristics of a program, such as execution time or memory. Such attacks can end up in the disclosure of data or cryptographic keys.

Denial of service

Denial of service attacks against data flows can be the source of tampering threats. DOS is possible when incapacitating the channel through incapacitating the endpoints, consuming significantly more resources or falsifying control messages. This can also create tampering, spoofing, and DOS threats against processes. Additionally, corrupted messages with weak or no message integrity or preplay attacks can cause DOS for the data flow.

4.1.1 Raw data

The "Raw Data" data flow represents the data sent from the central healthcare system to the data processing process (See Figure 9). The data handled here is extremely sensitive, as it could contain personally identifiable information (PII).

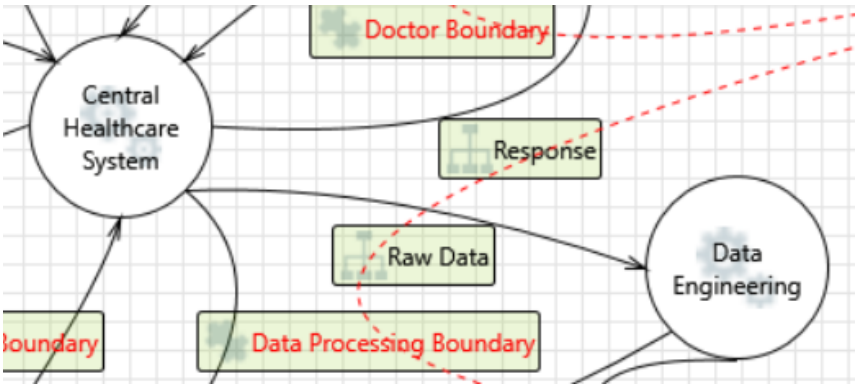


Figure 9. *Raw Data data flow.*

The raw data processed here hold very sensitive patient data. The data from sensors and medical records might not be trustworthy, reliable and suitable. For example, during MiTM attacks the attackers could tamper with the data that is later used for training or validation, making the data invalid. Data trustworthiness is in the top ten ML security risks presented in the BIML report [48].

Information disclosure at this stage can bring serious confidentiality and privacy issues. Because the data handled here is very sensitive then measures to protect it must also be very strong. Without strong counter-measures, the attackers could easily gather sensitive information about patients for malicious use. Data confidentiality is also mentioned in the top ten ML security risks by BIML [48].

Denial of service attacks for this data flow definitely can be possible, but they do not carry a big severity. Raw data is not used often as it is only needed for model training. As this data flow is not needed always, availability is not that important. If the DOS attacks invoke tampering then it has a bigger severity.

4.1.2 Training data

The "Training Data" data flow represents the data sent from the data engineering process to the model training process (See Figure 10). The data is processed by the data engineering process and is already modified to be suitable for training the model.

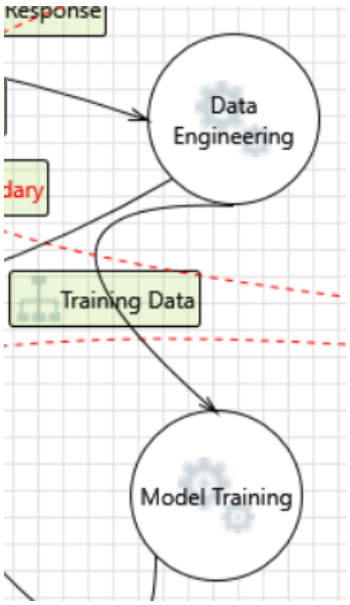


Figure 10. *Training Data data flow.*

Possible tampering with "Training Data" data flow can lead to data manipulation attacks. One possibility is when an attacker can control the data flow channel or the messages sent in this data flow. If the data flow suffers from tampering according to the Adversarial Machine Learning report it opens up the possibility for three different poisoning methods: availability poisoning, targeted poisoning and backdoor poisoning [47]. Through these methods, different label-related attacks like clean-label poisoning, where attackers have control over training examples but not their labels [47], could be carried out. Ultimately by doing the modification, the attacker can influence the behaviour of the system. Poisoning

attacks are very powerful and in the case of health data, they could also bring extreme consequences. Jagielski et al. presented how data poisoning can change the patient's dosage on average by 139% [49]. In 2015, Mozaffari-Kermani et al. poisoned five different healthcare datasets which used different algorithms [50]. Both these examples demonstrate that AI/ML models that create wrong suggestions based on poisoned data can have devastating consequences.

Information disclosure threats are also similar to the "Raw Data" data flow, because of the nature of the data. The data handled here must be secured properly or very sensitive personally identifiable information could end up in the possession of the attackers. As mentioned in the "Raw Data" data flow section, data confidentiality is a well-known issue and is represented in multiple reports.

Like the categories before the denial of service also is similar to the "Raw Data" data flow. Because this data flow is not used often, DOS does not propose a big severity. Severity increases if the DOS is bundled together with other threats like tampering.

4.1.3 Performance validation data

The "Performance Validation Data" data flow is the data used for validating the performance of the model (See Figure 11). This data is sent from the data engineering process to the performance monitoring process.

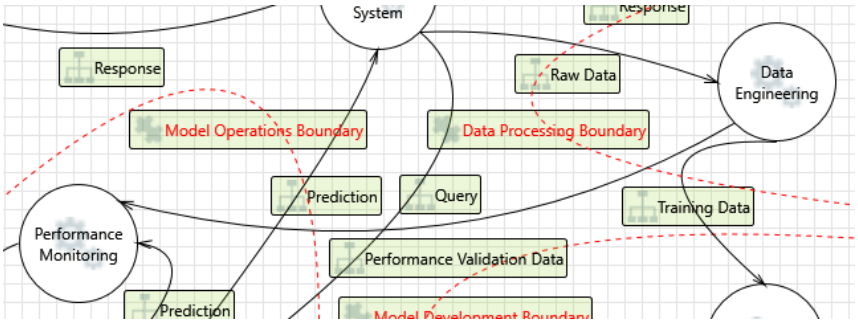


Figure 11. *Performance Validation data flow.*

The two main data-related issues present here are data manipulation and data confidentiality. Data manipulation can still happen in the case of tampering, but this does not carry an immediate effect on the functioning of the AI/ML model. Based on this data the decision to do a detailed investigation of the performance of the model or to retrain the model is conducted. This means that tampering with the validation data does not create situations where the model is malfunctioning. Still, it can create unnecessary retraining of the model to consume resources and the trust of the staff regarding performance monitoring may decrease.

In the case of information disclosure, different privacy-related threats emerge. If the data used for performance monitoring is based on real patients, therefore sensitive PII information could be disclosed. If the attacker has possession of confidential patient data, this could be used later on to better conduct an attack against the system or to learn system specifics. Additionally, the data itself could be used for attacks outside the AI/ML scope.

As this data flow is not used regularly the denial of service threat is irrelevant here, if it does not trigger other means of tampering or information disclosure.

4.1.4 Model deployment

The "Model Deployment" data flow represents how the model is sent from the model tuning process to the model registry data store (See Figure 12). The model is completed and this data flow represents how it is sent to the production setting.

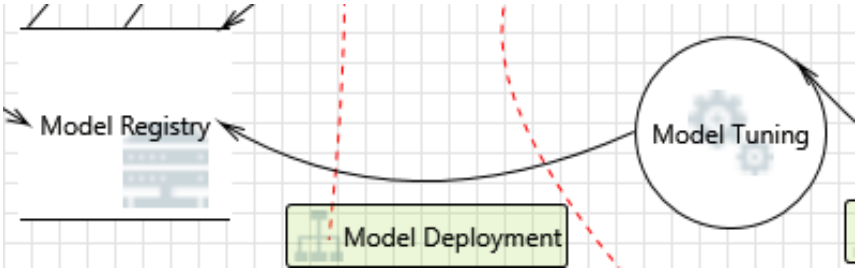


Figure 12. *Model Deployment data flow.*

The biggest threats here are related to the model itself because this data flow represents the transfer of the trained model to the operational stage. In case of tampering, the attackers could conduct model replacement. The attackers could swap the model with a malicious one or just capture the model. This could cause integrity and also availability concerns. It will take time to identify the compromised model and replace it. Additionally, this could pose risks to patient safety and treatment effectiveness.

Information disclosure opens up the possibility for model capturing attacks. Model capturing refers to the unauthorized access or theft of AI/ML models. The attackers could obtain the model if they could observe the data flow. One issue is that they can conduct tests independently to learn more about the model. Upon learning the model more attack paths are possible. The second issue would be that attackers could sell the captured model to other malicious actors who could use it for learning themselves.

Denial of service at this stage is not a severe issue, because the data flow is not used frequently. However, if the attack is carried out for an extended period this could pose some availability issues as well.

4.1.5 Query

The "Query" data flow represents the communication between the central healthcare system and the AI/ML model in the model registry (See Figure 13). The purpose of the query data flow is to send patient data to the model to get predictions based on them.

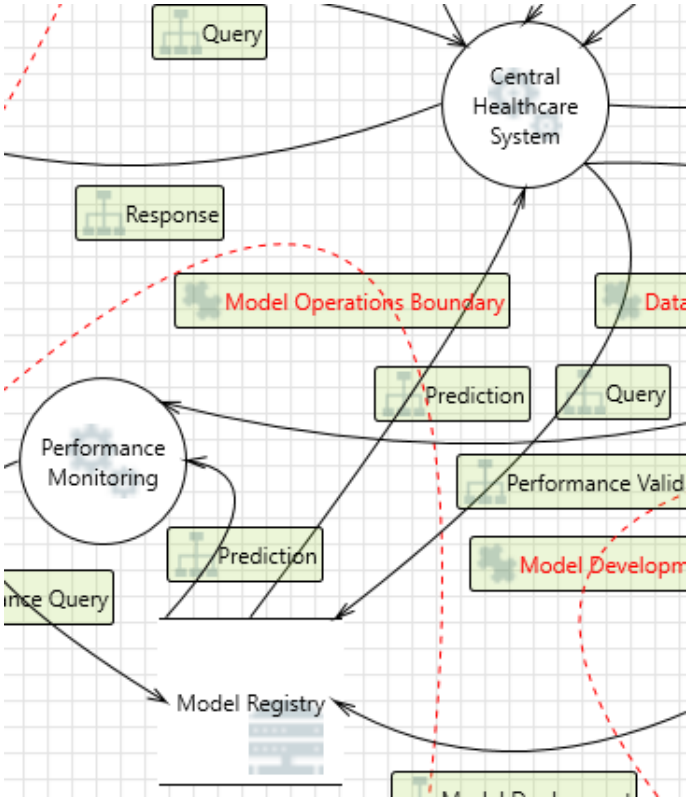


Figure 13. *Query and Prediction data flow.*

Tampering in this stage could lead to input manipulation attacks, also called adversarial examples. During input manipulation attacks the goal of the attacker is to change the inputs with small perturbations to cause the model to make wrong predictions. Sundas et al. presented various ways in which input manipulation can lead to the accuracy downfall of the model in the healthcare domain [33]. Adversarial examples are the first risk listed in the top ten ML security issues provided by BIML [48]. Rahman et al. successfully conducted evasion attacks for COVID-19 deep learning systems in medical IoT devices [51].

Data flows that carry real patient data can suffer severely from information disclosure, because of the sensitivity of the data. Upon disclosure, the attackers can use the data to craft more efficient attacks in different stages of the system. If the attackers have access to queries and predictions as well, they could extract the model or create adversarial examples. Either way, personally identifiable information is lost and that is a big privacy concern.

Means of denial of service carry a bigger importance in the case of this data flow. Denial of Service disrupts healthcare AI/ML systems by overwhelming them with requests [52]. As this data flow is used regularly then availability becomes an issue. If the services are not available this can create unwanted complications for the doctor's work and also confusion in the patients. This can potentially disrupt critical medical decision-making.

4.1.6 Prediction

The "Prediction" data flow represents how the predictions conducted by the AI/ML model in the model registry are sent to the central healthcare system or the performance validation process (See Figure 13). Based on the query from the central healthcare system the model registry responds with predictions.

Tampering with the predictions leads to wrong information ingested by the central healthcare system which can cause severe consequences. Secondly, tampering with predictions can lead to inaccurate monitoring results from performance validation. This tampering is not directly related to the model, but here usual data flow threats can occur nevertheless.

Information disclosure within this data flow may enable the attacker to reconstruct sensitive information about the training data. Additionally, the attacker can leverage knowledge about the training data to infer whether specific individuals or samples were included in the training dataset. For a membership inference attack, the adversary needs to have querying rights as well. MITM attack for the queries and predictions channel could be one way to achieve this. Shokri et al. and Breugel et al. showcase different possibilities for membership inference attacks to take place [53, 54]. Additionally, model extraction attacks are possible, where attackers try to uncover model behaviour and parameters. One way is for the attacker to passively collect data in the channel. If the attacker could have access to queries as well, then this would be more effective. These types of attacks were showcased by Oliynyk et al. in 2023 [55]. All of these issues could pose privacy risks because it is possible to disclose personally identifiable information. In 2014 Fredriksen et al. demonstrated how an attacker, given the model and some demographic information about a patient, can predict the patient's genetic markers [56]. As showcased, information disclosure at this point can cause severe confidentiality issues.

This data flow is used simultaneously with the "Query" data flow, hence the denial of service threats could impact the system here as well. If the predictions are not accessible for the doctors or patients this can cause clear availability issues.

4.2 Threats to data stores

STRIDE method indicates that data stores are susceptible to tampering, repudiation, information disclosure and denial of service.

Tampering

According to the tampering attack tree, tampering can take place when an attacker can bypass the protection scheme or bypass monitoring or overcapacity failures. Weak or no protection can lead to a protection scheme bypass. Likewise weak or no monitoring can lead to monitoring bypass. Wraparounds, discards or other failure modes can cause overcapacity failures, which is also one way that tampering can occur.

Repudiation

Repudiation is possible if the data store holds logging info or other means of auditing data. Logs could contain info about the actions conducted in the data store. This kind of data store has repudiation threats because an attacker might attempt to hide his actions by modifying or erasing the data. Repudiation can occur if the attacker can successfully repudiate messages or transactions. A weak signature system, weak logging and replay attacks cause this.

Information disclosure

One of the ways information disclosure can happen is if the attacker bypasses the protection scheme and the data is intelligible. This can occur when the data is not encrypted and the protection scheme is weak or nonexistent. Information disclosure can also happen with side-channel attacks. Storage management issues like failure to initialize storage or clear storage also can be a source of the information disclosure threat.

Denial of service

Denial of service can take place when corrupt data is sent to the data store or the container is incapacitated. The cause for corrupt data can come from a lack of monitoring means or weak monitoring processes. Corrupt data can be the source of data flow tampering as well. Problems with denied access or exceeded capacity can lead to incapacitation of the container and this creates denial of service threats as well.

4.2.1 Model registry

The "Model Registry" data store is the environment where the model is deployed after training it (See Figure 14). The model registry creates predictions based on the data ingested.

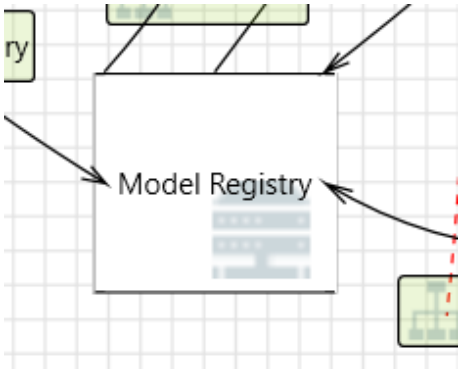


Figure 14. *Model Registry data store.*

While a model registry does not serve the purpose of a generic data store, the STRIDE method threats still apply here in some ways. The difference between a generic data store and a model registry is that the registry holds the model itself that conducts predictions based on queries. Generic data stores usually hold data and respond to queries.

Tampering in this case would mean the attacker could replace the model with a malicious one and that would cause malfunctions in the predictions. This could pose risks to patient safety and treatment effectiveness. Additionally, instead of replacing the whole mode, the attacker could change the functionalities of the existing one thus conducting model reprogramming. Model reprogramming involves altering healthcare AI/ML models to produce incorrect or biased outputs.

In the case of a model registry, all of its actions are closely monitored and logged. Because of that the possibility of repudiation is there. Repudiation threats combined with other threats like tampering or information disclosure could help the attacker cover tracks. This could mean that the issues presented in the model could be discovered a lot later and this could bring severe consequences.

Information disclosure from the model registry means that the attackers could easily obtain the model. This means that model capturing is possible. Upon capturing the malicious party could conduct investigations to learn about the model which could help with future attacks or they just could sell it to other parties for monetary gains. Either way, it can have

devastating consequences.

Denial of service for this data store means that the model can not be deployed and the model can't make predictions. Availability problems for the data store can create some issues for the doctors and patients who are heavily relying on the predictions conducted by the model.

4.3 Threats to processes

According to the STRIDE method processes are susceptible to spoofing, tampering, repudiation, information disclosure, denial of service and elevation of privileges.

Spoofing

Spoofing a process can occur when there are problems with the authentication systems or problems with credentials. If there are authentication issues like predictable credentials, null credentials or no authentication system in place then this can directly lead to spoofing the process. The other category is problems with credentials themselves. Attackers could falsify credentials by guessing. Also, there is an option for the attackers to obtain the legitimate credentials. Obtaining legitimate credentials could take place when weak storage or weak change management systems are in place.

Tampering

Tampering with a process can be provoked by providing false credentials or corrupting the state of the process. Corruption can be sourced from input validation failure or unauthorized memory access. Tampering by using false credentials can occur if there is a failure to check the call chain. Tampering with a process can also lead to tampering with a subprocess and spoofing an external entity.

Repudiation

Repudiation can occur if the attacker can successfully repudiate messages or transactions. A weak signature system, weak logging and replay attacks can be the cause of this. Logs not containing sufficient data or if the logs are unauthenticated or possess weak authentication means this can be a source of transaction repudiation. Repudiation threats can also lead to spoofing entities and tampering with logs.

Information disclosure

Similarly to spoofing threats, information disclosure can take process when the process is corrupt. This again can happen by input validation failure or by accessing the memory. In addition, threats from side-channel attacks also could cause information disclosure. Information disclosure against a process can lead to tampering threats against persistent processes or the spoofing of an external entity.

Denial of service

Denial of service against a process can occur in various ways. One of the sources can be that the process consumes too much of the application-specific resources. Another issue could be that the process consumes too much of fundamental resources. Additionally, input validation can also create denial of service threats against a process.

Elevation of privilege

Elevation of privilege is possible through corruption or authorization system issues. Dynamic corruption through input validation failure or memory access can lead to the elevation of privileges. Additionally, static corruption can achieve the same result. Attackers could also leverage insufficient authorization means. Cross-domain issues or call-chain issues can be the source of insufficient authorization.

4.3.1 Data engineering

The "Data Engineering" process is the step where all the data processing happens (See Figure 15). The process gains ingests raw data from the central system and outputs training data to the model training process.

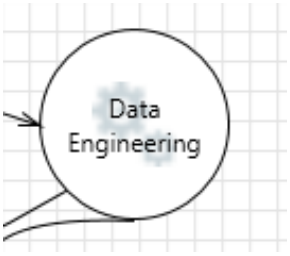


Figure 15. *Data Engineering process.*

Different spoofing threats can lead to serious consequences in this process. Possible types of spoofing could be identity, data spoofing or session hijacking. This can result in falsified data, malicious actions conducted by the attacker and access to sensitive data or other

critical resources. Spoofing can be the source of other threats.

As this process mostly handles data, the biggest tampering threats are against the data. Different poisoning attacks can take place here, as the attacker could modify the data. If the raw data or training data has integrity issues, this can cause problems for the entire AI/ML application. BIML report brings out data poisoning and data confidentiality threats in their top ten security issues report that are relevant for this process [48].

If repudiation is possible all the other threats like tampering or disclosure carry more severity. If the attacker could mask their actions when tampering with the data the end product suffers malfunctions and the source can be harder to find. Also not knowing that there have been other attacks that result in information disclosure, elevation of privileges or other threats could lead to serious consequences as the attackers can freely gather valuable information.

Information disclosure threats here are also all data-related. If the attackers gain access to the data being processed this results in loss of confidentiality for the data. As this data is very sensitive personally identifiable information, then measures to protect it must be set in place.

Denial of service against the data engineering process does not carry that much severity, because the process might not be used often. As the process is only needed when data engineers are working with data the availability constraints are a bit more flexible.

With elevated privileges the attackers could conduct previously mentioned attacks more efficiently or gather even more valuable information about the system or the data.

4.3.2 Model training and model tuning

The "Model Training" and "Model Tuning" processes conduct the training and evaluation phases of the AI/ML model (See Figure 16). Because both of these processes share the same trust boundary and are similar processes they share the same threats against the system.

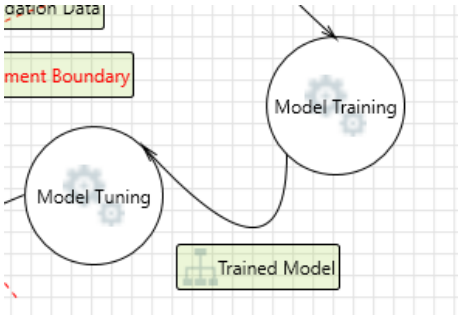


Figure 16. *Model Training and Tuning processes.*

Identity spoofing or session hijacking are some of the possible spoofing threats for these processes. Spoofing the processes could result in a malfunctioning AI/ML model that gives wrong predictions. If the model is giving wrong input that can cause problems for all the users.

A similar situation is with tampering. If the attackers have access to the training phase of the model, they can carry out model reprogramming attacks that result in a malfunctioning AI/ML model. The neural net reprogramming threat was also brought out by Microsoft in their AI/ML-specific threats report [57].

Repudiation here can again hide all the actions conducted by attackers. After the model is poisoned then finding the reason behind it can be hard if the integrity of logs can not be trusted.

While attackers can access data used for training and all the training process then information disclosure threats are dangerous. Attackers could access the training data and reveal sensitive patient information. Attackers also could just capture the model for further studying or selling. With access to the model, it is easier to make adversarial examples that are used in model evasion attacks.

Denial of service against the model training and tuning processes does not carry that much severity, because the processes might not be used often. As the process is only needed when data engineers are working on creating the model the availability constraints are a bit more flexible.

With elevated privileges the attackers could conduct previously mentioned attacks more efficiently or gather even more valuable information about the system or the data.

4.3.3 Performance monitoring

The "Performance Monitoring" process handles all the tasks that are performed to measure that the model is working at an expected level (See Figure 17). The performance monitoring process is in the model operation boundary.

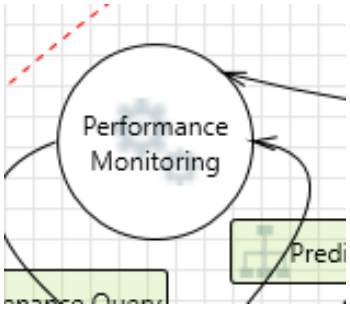


Figure 17. *Performance Monitoring process.*

Spoofing threats that could occur here might be session hijacking or identity spoofing. The performance monitoring process has access to different assets, so various threats emerge.

Tampering with the process can lead to inaccurate validation data. If the validation data is inaccurate this means wrong results from the monitoring process. Wrong results from performance monitoring can result in the model not being re-trained if needed, so the model that produces sub-par results still is active in a production environment.

Repudiation issues can help hide the attacker's tracks and lead to the malfunctioning model staying longer in a production environment.

From this process, sensitive personally identifiable data can be obtained by the attackers. Because the performance monitoring process uses validation data, which can be based on real patient data, then data confidentiality issues pose a big risk.

Denial of service threats for the performance monitoring process carries similar consequences as repudiation threats. Attacker's denial of service attacks lead to the malfunctioning model staying longer in a production environment.

With elevated privileges the attackers could conduct previously mentioned attacks more efficiently or gather even more valuable information about the system or the data.

4.3.4 Central healthcare system

The "Central Healthcare System" process is the process that sends raw data for training and is the main user of the AI/ML model predictions (See Figure 18). Through this process, different entities get their info from the model.

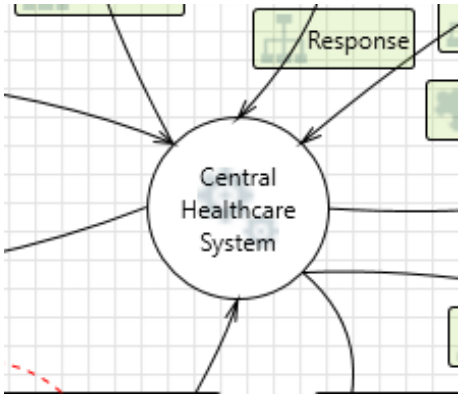


Figure 18. *Central Healthcare System process.*

While the process itself is not part of the development of operation phases of the model, it still interacts with it in various ways. Because of the interactions some AI/ML-related threats still occur for this process in addition to the STRIDE ones.

Different data-related threats are present for this process, meaning this can result in loss of confidentiality. This could include data confidentiality and data trustworthiness issues. While having access to queries and predictions attackers could conduct membership inference or data reconstruction attacks.

Although the process itself is not directly involved in the AI/ML lifecycle, the process still has access to different data and querying rights. While controlling queries and having access to predictions model extraction or evasion attacks could be a possibility. Because of that this process is susceptible to various AI/ML-related threats that need to be addressed.

4.4 Attack tree implementation

Attack trees help identify common attack patterns and help developers think about security conditions. This chapter will give an overview of how to implement the STRIDE-based attack trees to the model while keeping machine learning specifics in mind. Previously identified three types of modelling elements that are needed for machine learning threat modelling will be used. Threat trees used are created by Michael Howard et al. [46].

4.4.1 Data flow

For this example, the "Training Data" data flow is used. This data flow handles the training that is being sent to the model training process. The mapping of the STRIDE threats to Data Flow elements indicates that data flows are susceptible to tampering, information disclosure and denial of service.

As described previously for the "Training Data" flow, in the case of tampering threats of data manipulation or poisoning emerge [49, 50, 48]. The two threats presented in the tampering tree are violations of integrity for the message and for the channel itself (See Figure 19). Poisoning attacks could happen through the means of Man-In-The-Middle (MITM) or when weak security measures are used. This can result in inaccurate training outcomes and potentially biased or unreliable machine learning models. Similar outcomes can result from non-existent or weak message integrity measures as well. Replay attacks are relevant as well because if the attacker replays fraudulent or biased data, it could skew the model's behaviour or predictions. If certain types of data are replayed more frequently than others, the model may inadvertently learn to prioritize or overemphasize these patterns, leading to biased predictions. Also spoofing of the processes can be the cause of data flow tampering. An attacker may forge the source address of data packets containing training data to make it appear as if the packets originate from a trusted source within the central healthcare system. By spoofing the source, the attacker can bypass authentication means, tricking the model training process into accepting and processing unauthorized data. Ultimately the tampering tree for data flow has all the needed aspects to map relevant AI/ML-related threats that may occur.

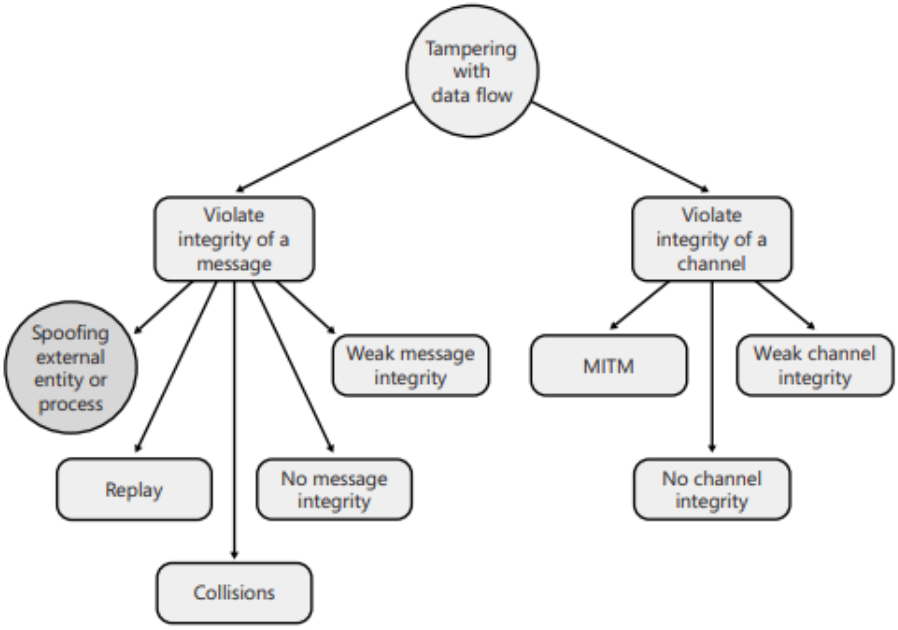


Figure 19. *Data flow tampering attack tree [46].*

As the "Training Data" flow handles sensitive information then the data confidentiality-related issues are the most impactful. Data confidentiality is also listed as one of the top ten security risks for ML models [48]. The attack tree for information disclosure can be seen below (Figure 20). In the context of a training data flow used for machine learning model training, side-channel attacks are typically not directly applicable. Observation of the data flow channel or messages is quite relevant here. Observation can happen through no security measures, weak security measures or MITM. In the case of observation, the attackers could use the data to construct more targeted attacks or sell the data for monetary gains. Sensitive medical data must be encrypted using strong means. Wood et al. provided different approaches to this topic in 2020 [58]. The attack tree used covers all the relevant threats to AI/ML-based data flow well and gives good input for the design process.

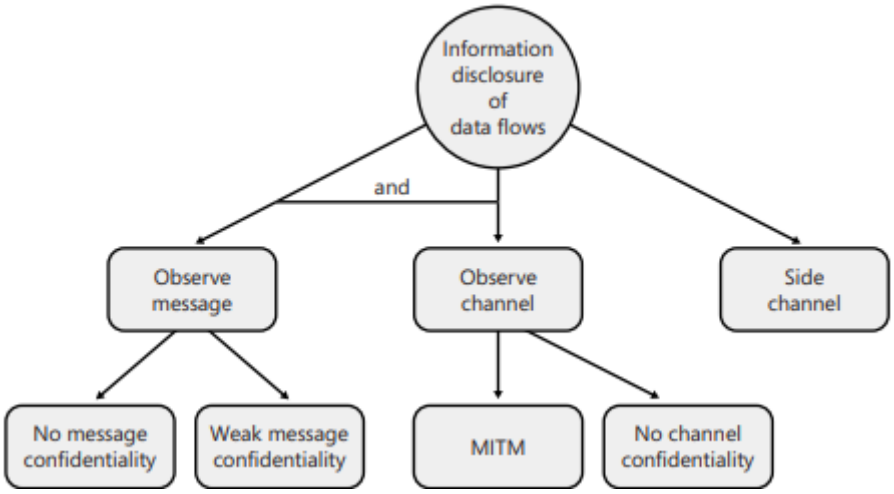


Figure 20. *Data flow information disclosure attack tree [46].*

Denial of service attacks here are a possibility but they are not that severe, as this data flow is not used often. There are no AI/ML-specific denial of service attacks, as the default threats apply. From the threat tree for denial of service, the most crucial point is if it invokes tampering against a data flow or process. Other than that the tree itself covers the basic means of denial of service attacks well.

4.4.2 Data store

The threat tree is implemented using the only data store element associated with ML/AI from the model. The "Model Registry" is a data store and its main goal is to hold the trained model to respond to the queries. While the model registry is not a classical relational database, the data store element still represents the functionalities of an AI/ML model the best. Modelling it as a data store gives the option to showcase the model storing capabilities with also the option to respond to incoming queries based on the information stored. The STRIDE methodology indicates that data stores are susceptible to tampering, repudiation, information disclosure and denial of service. The biggest AI/ML-related threats to the model registry are model reprogramming, model replacement, and model capturing [47, 48, 57].

Tampering with a data store can happen if the protection scheme or monitoring is bypassed or if the data store suffers from overcapacity failures (See Figure 21). Three reasons why protection schema could be bypassed are no or weak protection or canonicalization failures. Canonicalization refers to an issue when data is not properly standardized or normalized before being stored. As this data store is not a regular data store that handles data, but it handles only models, then it is not that useful here. In the case of "Monitor

Bypass", the attack scenario involves an adversary attempting to circumvent or evade monitoring mechanisms put in place to detect and prevent unauthorized activities within a data store. Overcapacity failures might happen if numerous models are sent to the registry or the registry itself has low capacity from the start. Even though it is a possibility it is a rather small chance of this happening. Except for a few options, the attack tree covers the reprogramming cases that could occur at this stage.

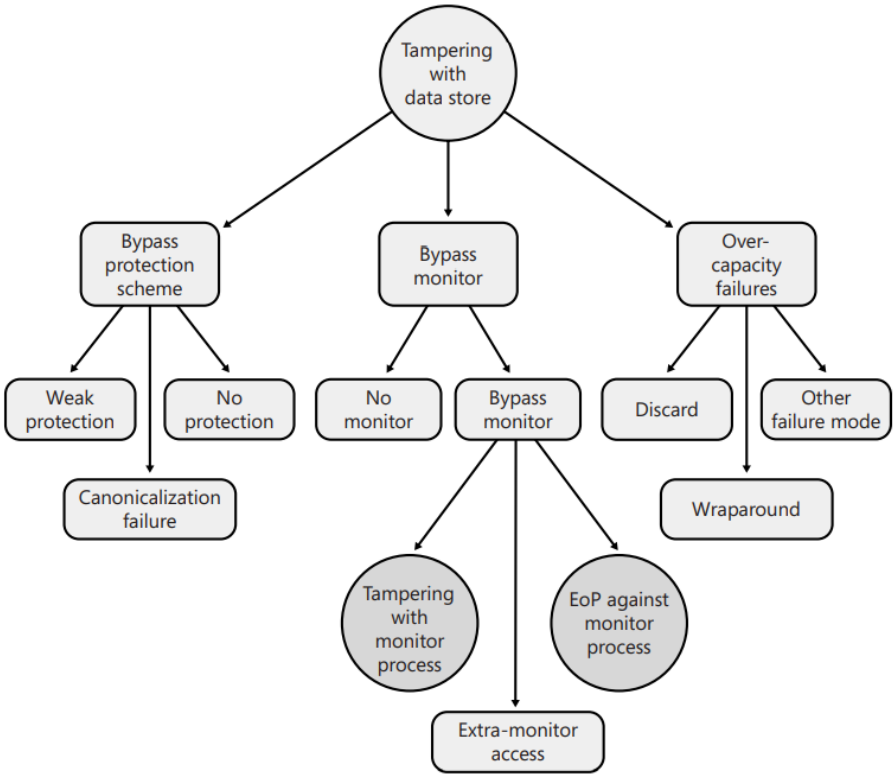


Figure 21. *Data store tampering attack tree [46].*

An information disclosure-related threat here is model capturing. This is quite similar to just data capturing so the attack tree can cover these model-related threats as well. In the attack tree, some paths are not relevant for AI/ML systems like side channels or canonicalization failures (See Figure 22). While storage management can be an issue, it is quite unlikely as well if the storage is with low capacity and configured improperly. Encrypting the features used by the model is an approach provided by Wood et al. in 2020 [58]. If the model or its parameters are not encrypted, then upon information disclosure important information about the model itself may leak. This can lead to model capturing, where the adversaries could also construct the model themselves. Altogether the attack tree covers the main threat opposed but could do with less.

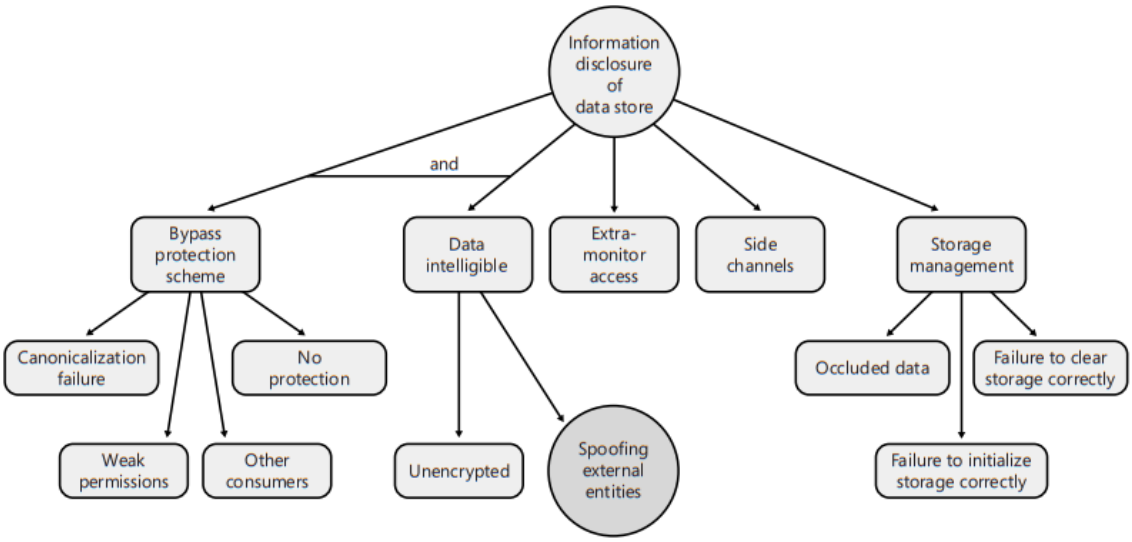


Figure 22. *Data store information disclosure attack tree [46].*

Repudiation and denial of service threats for data stores are not that AI/ML specific and are covered well enough by the attack trees provided.

4.4.3 Process

For the process example, the "Data Engineering" process is used. The data engineering process processes raw data to make it into training data that is used for the AI/ML model learning phase. According to the STRIDE framework, processes are susceptible to spoofing, tampering, repudiation, information disclosure, denial of service and elevation of privileges.

Spoofing this process can be the source of different threats for other elements connected to it. For example, the attacker could spoof this process and all the "Raw Data" data flow will be sent to the adversary. As there are no AI/ML-related spoofing threats then the spoofing attack tree provided covers all the needed aspects. The process must be protected with a proper authentication system and the credentials must be kept safe.

The biggest threats in the case of tampering are different data-related issues like data poisoning or data corruption. If the attacker can gain access to the process and tamper with the data, then it would be possible to send tampered training data onwards to change the intended functionalities of the model [49, 50, 48, 47]. From the attack tree, the most relevant part is the failure to check the call chain (See Figure 23). An attacker could control or have influence on some libraries that might be used and through that inject malicious code into the process. If the process calls that malicious code this can lead to process tampering. Other than that this tree is not that relevant for the "Data Engineering" process,

as the data-related issues are not handled properly.

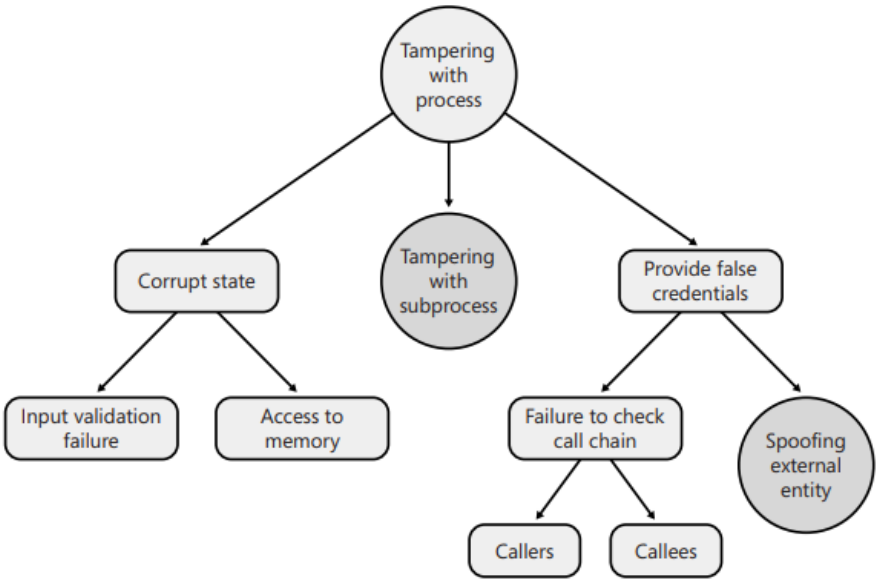


Figure 23. *Process tampering attack tree [46].*

A similar case is with the information disclosure attack tree as well, while it provides some means on how this could happen it is lacking data-specific options (See Figure 24). The tree suggests that information disclosure could come from spoofing external entities, corrupt processes or side channels. This process does not have external entities tied to it and side channel attacks are not relevant here. Information disclosure could happen through data leakage, data interception, malware and cyber-attacks. This process's most valuable asset is the data and it must be protected with adequate security measures.

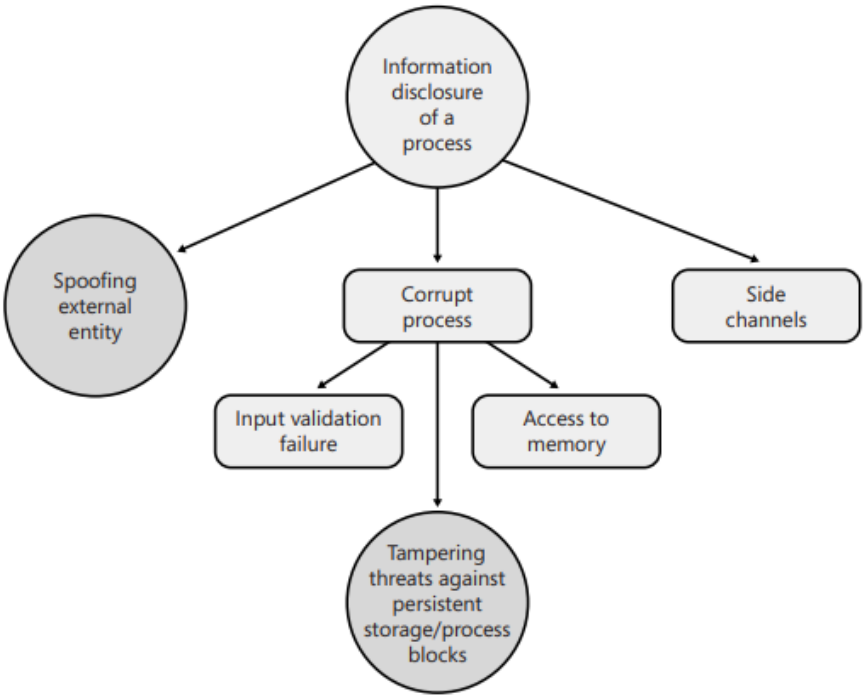


Figure 24. *Process information disclosure attack tree [46].*

Repudiation, denial of service and elevation of privileges can be the source input for other threats that may influence the AI/ML system. For example, tampering with a data store could bring elevated privileges and as a result, may enable model capturing or replacement attacks. Even though repudiation, denial of service and elevation of privileges threats can be connected to AI/ML attacks they are not AI/ML specific. Because they are more general threats then they are covered well enough by the attack trees provided.

4.5 List of threats identified

This section provides an overview of all the AI/ML specific threats identified for AI/ML systems through threat modelling, reviewing research papers and implementing STRIDE-based attack trees. The following three tables will give a list of identified threats for different AI/ML system-related components. These tables are divided as data flows (See Table 4), data stores (See Table 5) and processes (See Table 6).

Additionally, the threats provided are divided into two categories, conventional cybersecurity threats and AI/ML-specific cybersecurity threats.

Conventional cybersecurity threats are well-known threats that have been addressed in traditional cybersecurity practices. They may include common threats such as malware, phishing, denial of service attacks, MiTM attacks, and unauthorized access. For example,

capturing the model file by cyber means is considered a conventional threat. These threats are marked with a “Conv” tag.

AI/ML-specific cybersecurity threats are threats that are specifically made to exploit vulnerabilities in AI/ML models, algorithms, or data used by AI/ML systems. They may include threats such as adversarial attacks, data poisoning, and inference attacks. For these threats to be enabled they may also need some form of conventional cybersecurity attacks to be successfully carried out beforehand. These threats are marked with an “AI/ML” tag.

Each threat was judged if it was conventional or AI/ML-specific. If the attack only consists of conventional elements, then it is classified as a conventional cybersecurity threat. If the attack consists of AI/ML attack elements, then it is classified as an AI/ML-specific threat.

Table 4. *Identified threats for data flows.*

Type	**Data flow**	**Identified threats**	**Threat description**
AI/ML	Training Data	Data poisoning	Data poisoning in healthcare means intentionally manipulating medical data to influence AI/ML model training. This could be achieved through MiTM or spoofing attacks and modification of data features or labels. This could result in inaccurate diagnoses and compromised patient care [49, 50, 47, 48]
Conv	Raw Data, Training Data, Performance Validation Data	Data confidentiality threats	Data confidentiality threats in healthcare involve unauthorized access or disclosure of sensitive medical data. This could be achieved by means like packet sniffing. This could lead to potential privacy breaches and misuse of personal information [48]
Continued on next page			

Table 4 – continued from previous page

Type	Data flow	Identified threats	Threat description
Conv	Raw Data	Data trustworthiness	The data from sensors and medical records might not be trustworthy, reliable and suitable. For example, during MiTM attacks the attackers could tamper with the data that is later used for training or validation, making the data invalid [48]
Conv	Model Deployment	Model replacement	Model replacement involves replacing whole healthcare AI/ML models with malicious ones. Attackers could replace the model with a malicious one by conducting a MiTM attack. This could pose risks to patient safety and treatment effectiveness
Conv	Model Deployment	Model capturing	Model capturing refers to the unauthorized access or theft of AI/ML models. This could be achieved by packet sniffing. This could lead to model misuse (using the model in other attacks) and loss of intellectual property
AI/ML	Query	Evasion attacks	Evasion attacks aim to deceive healthcare AI/ML systems. Attackers provide malicious input with small perturbations that cause the system to make false predictions. After a successful MiTM attack, the adversaries can introduce small perturbations and send them to the model. This could compromise patient safety and treatment accuracy as the diagnosis could be inaccurate [33, 48, 51, 47]

Continued on next page

Table 4 – continued from previous page

Type	Data flow	Identified threats	Threat description
Conv	Query	Denial of service	Denial of Service disrupts healthcare AI/ML systems by overwhelming them with requests. Spoofing the source of the request could be done for this. This can potentially disrupt critical medical decision-making [52]
AI/ML	Prediction	Model extraction	Model extraction exploits healthcare AI/ML models by trying to extract information about the model architecture and parameters. By spoofing healthcare system or performance monitoring processes the attacker could send various queries to uncover model architecture and parameters [55, 48, 47]
AI/ML	Prediction	Membership inference	During membership inference attacker tries to find out if a particular record or sample was part of the training, by querying the model. Spoofing the processes and sending several queries can be one way to achieve this. This could uncover sensitive patient data, posing risks to privacy and confidentiality [53, 54]
AI/ML	Prediction	Data reconstruction	During data reconstruction, the attacker tries to reconstruct a subset of training data. Spoofing the central healthcare system and gaining access to querying rights can be one way to achieve this. This can lead to the loss of sensitive patient data, posing risks to privacy and confidentiality [47]

Table 5. *Identified threats for data stores.*

Type	**Data store**	**Identified threats**	**Threat description**
AI/ML	Model Registry	Model reprogramming	Model reprogramming involves altering healthcare AI/ML models to produce incorrect or biased outputs. Attackers could gain access to the model from weak access control mechanisms in the model registry. This could pose risks to patient safety and treatment effectiveness [57]
Conv	Model Deployment	Model replacement	Model replacement involves replacing whole healthcare AI/ML models with malicious ones. Attackers could gain access to the model from weak access control mechanisms in the model registry. This could pose risks to patient safety and treatment effectiveness
Conv	Model Registry	Model capturing	Model capturing refers to the unauthorized access or theft of AI/ML models. For example, faulty access control issues could be the cause. This could lead to model misuse (using the model in other attacks) and loss of intellectual property

Table 6. *Identified threats for processes.*

Type	**Process**	**Identified threats**	**Threat description**
AI/ML	Data Engineering	Data poisoning	Data poisoning in healthcare means intentionally manipulating medical data to influence AI/ML model training. Attackers could use backdoors, privilege escalation, malware or software vulnerabilities to access the process to carry out the poisoning. This could result in inaccurate diagnoses and compromised patient care [49, 50, 47, 48]
Conv	Data Engineering, Performance Monitoring, Central Healthcare System	Data confidentiality threats	Data confidentiality threats in healthcare involve unauthorized access or disclosure of sensitive medical data. Attackers could use backdoors, malware or software vulnerabilities to compromise the process. This could lead to potential privacy breaches and misuse of personal information [48]
Conv	Central Healthcare System	Data trustworthiness	The data from sensors and medical records might not be trustworthy, reliable and suitable. For example, if the attackers could compromise the integrity of medical records, the raw data provided for data engineering by the central system could result in disrupted patient care or mistreatment [48]

Continued on next page

Table 6 – continued from previous page

Type	Process	Identified threats	Threat description
AI/ML	Model Training, Model Tuning	Model reprogramming	Model reprogramming involves altering healthcare AI/ML models to produce incorrect or biased outputs. Attackers could achieve access to the model from weak access control policies and maliciously fine-tune the model. This could pose risks to patient safety and treatment effectiveness [57]
Conv	Model Training, Model Tuning	Model capturing	Model capturing refers to the unauthorized access or theft of AI/ML models. For example, faulty access control issues could be the cause. This could lead to model misuse (using the model in other attacks) and loss of intellectual property

5. Discussion

The following chapter will give insights into the investigation of threat modelling for AI/ML-based healthcare systems. The thesis can be divided into two phases - creating the model and identifying relevant threats. Discussion about complications and solutions for these two phases will be conducted below.

The first decision for modelling was to select the way the AI/ML model is used by the healthcare system. As mentioned in the Model creation chapter there are three main ways to implement the AI/ML model into the healthcare system. Two options of the three share some similarities. These two choices are to use a plugin that is connected to a model or use the whole model received from a third party. While it's easier to implement these options into a system, both of these do not have insight into the actual development of the model itself. Some of the risks are transferred to a third party. That is the reason why in this thesis it was decided to use an in-house development approach. The in-house development approach will give insight into the development of the model and will give more AI/ML model-related content to be researched, as the risks are not transferred. Most of the threats present in an in-house development still are present in the two previously mentioned methods, so theoretically, the in-house approach would cover a lot more.

After deciding the approach the next thing in line was to see how a healthcare system that utilizes AI/ML models could be modelled. As there are numerous papers on all the possible architectures of a modern healthcare system then the question of how it works was solved quickly. The healthcare system utilizes data input from different smart biosensors that are already commonly used around the world. Additionally, the system receives input from the patients and doctors themselves. All of these inputs are combined into a central healthcare system. The more interesting and harder question was how the AI/ML part itself can be modelled and incorporated into the healthcare system. There are no straight guidelines on how to conduct threat modelling for an AI/ML in general. The research involving threat modelling presents various ways how to display the AI/ML components. The easiest way found in papers is to just use one component, either process or entity and call it the model. As this thesis intended to take a deeper dive into the AI/ML-based components this approach was not sufficient. For this thesis, it was decided to model the AI/ML model as a data store element. While the model registry is not a classical data store, the element still represents the functionalities of an AI/ML model the best. Modelling it as a data store gives the option to utilize the model storing capabilities while

also giving the option to respond to incoming queries based on the information stored. The other big obstacle was actually how to define the boundaries in the system. Which elements trust which and when is the trust broken? The boundaries between the sensors, entities and applications are quite clear, but the complication is to define the boundaries between AI/ML components. In the end, three main boundaries for AI/ML components were used. The first one is the data processing boundary. Data engineering can not trust the data communication blindly before processing and that is why it deserves its boundary. The second one is for the development phase and the third one is for model operations. For each of these stages, different kinds of threats and threat actors can oppose different threats and for that reason, the flows between development and operational phases can not always be trusted. After some further investigation, the Data Flow Diagaram used for threat modelling was conducted. The key aspect to highlight from the diagram is that all components of a complete AI/ML model lifecycle are present. This is not always the case when looking at threat modelling papers. In addition, the components are divided into development and operational phases, quite similar to how it is done in practice. This diagram closely mimics the real processes of an in-house development approach.

Upon completing the diagram the focus moved into the second phase of the thesis, which is the identification of relevant threats based on relevant literature and different methodologies. For threat modelling, there are various methodologies, but for this thesis, the STRIDE framework was chosen. The selection of STRIDE for threat modelling in AI/ML-based healthcare systems was justified due to its comprehensive coverage of six essential threat categories. These six categories of Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege cover a wide range of threats. In addition, STRIDE is a widely used and well-established framework. The Data Flow Diagrams used in STRIDE are intuitive and good for system modelling. STRIDE provides an easy-to-understand approach to identifying and addressing potential security risks. These are the reasons it is favoured by experts. With this, it is possible to ensure thorough protection against a wide range of threats in the complex healthcare environment. In addition to STRIDE, the attack trees based on the same methodology were used. Attack trees stand out by their ability to offer a detailed and hierarchical representation of threats, aligning closely with the complicated nature of AI/ML-based healthcare systems. By using attack trees based on the STRIDE framework, the analysis could systematically explore various scenarios, enabling a comprehensive understanding of potential attack paths.

This approach with STRIDE methodology, attack trees and relevant literature paid off. All the relevant threats mentioned in various reports from different organizations and researchers dealing with AI/ML security were covered by the combination of STRIDE and the attack trees. Here it is worth highlighting that the combination of them worked the best

to cover those AI/ML specific threats. While the STRIDE methodology itself gave good input about the potential threats and the potential outcomes, the attack trees accompanied it well in giving a lot more detail and a few extra paths to be considered. It could be possible to achieve good results while only looking at STRIDE elements, but this could not be done by only looking at the attack trees. Because the STRIDE methodology itself is more general and does not go into details that much, then it is still quite relevant and can be applied to newer technologies. The attack trees help, but some of the paths can be a little bit outdated and do not consider the complications of the newer technologies used. In conclusion, the best results were achieved by combining the base methodology and the attack trees. This way it effectively offers a systematic and comprehensive approach to identifying and mitigating security threats in AI/ML-based healthcare systems.

Moving forward, continued research in this field is essential to stay ahead of evolving threats and maintain the security and trustworthiness of AI/ML applications in healthcare.

5.1 Qualitative validation results

For this kind of research, the validation of modelling and threat identification is one key limitation. There are various approaches to AI/ML threat modelling, and each of them is doing something different. How to decide if the system modelled captures all the relevant data flows and if all the parts are represented correctly? Additionally, how to make sure that the threats identified via STRIDE and attack trees are even relevant?

For validation, it was decided to use the qualitative approach. According to Uwe Flick, the aim of qualitative analysis is to make general statements by comparing various materials or various texts [59]. For this research, the materials or text come from a survey. Various experts with backgrounds in cybersecurity and/or AI/ML were asked to fill out a questionnaire that was based on the research. The whole questionnaire can be seen in Appendix 3 - Questionnaire for experts. The questionnaire consisted of twenty-seven questions. Some closed questions, some open-ended questions. The questionnaire was divided into two sections - the modelling part and the threat identification part. The experts were given materials that included the whole Data Flow Diagram, only the AI/ML specific Data Flow Diagram and the table of threats identified with a quick introduction to the table.

Ultimately four different experts took part in the questionnaire with a fitting background for this kind of thesis.
Expert one (EX1) has professional experience in both cybersecurity and AI/ML systems. The experience has mostly been related to the defence and healthcare industry.
Exprt two (EX2) has professional experience in AI/ML system development. Mainly has

researched securing AI systems and more practical experience from the ML side.
Expert three (EX3) has professional experience in both cybersecurity and AI/ML systems. Expert three is a postdoctoral associate in privacy-preserving machine learning for healthcare.
Expert four (EX4) has professional experience in both cybersecurity and AI/ML systems. The experience has mostly been related to threat modelling and AI security. Expert four has the most experience in both domains.
On average the experts have more than five years of professional experience in their respectable domain or even both of the domains mentioned before.

The first part of the questionnaire aimed to get validation for the modelling itself. All the questions were about the representation of the AI/ML model components in the Data Flow Diagram. The first of the questions was if the flow of data within the healthcare system makes sense. All experts agreed that in general, the flow makes sense and looks good. Some minor structure issues were brought out by EX1 and EX3 to further make the model more understandable at first glance. All of the experts agreed that representing all the data processing tasks as one process "Data Engineering" and using a data store element for the "Model Registry" is a valid approach. Additionally, EX3 mentioned some steps that could be done with the data store to improve clarity. The suggestion by EX3 was to split the registry for the deployed model and the registry for older models and metadata into two separate elements. One piece of feedback that came from three different experts was that the "Performance Monitoring" process should also be included already earlier in the model tuning phase. This would give an option to re-tune the model for better performance. Other than this the experts agreed with the representation of the model development phase and boundaries. In their opinion, all the important steps of an AI/ML model lifecycle were represented.

The second part of the questionnaire aimed to get validation for the identified threats. The main questions were whether the identified threats were relevant and if some threats were overlooked. There were eleven unique threats identified and for each threat, the experts were asked if they are relevant in this context. All of the experts agreed with the relevancy of all the threats but also provided some additional info. EX1 brought out that ideally, the threats would have an impact tied to them as well. Also, EX1 mentioned that more privilege escalation, human errors, rouge employees and physical attacks could be mentioned. EX2 also mentioned insider threats as a possibility. EX3 also mentioned backdoor attacks. While these mentioned threats are present in the system, they are not that AI/ML-specific and that's why they were excluded from the list. These threats are more conventional cybersecurity threats and could be used to enable AI/ML-related attacks. EX4 feedback also mentioned that if the system should use generative AI model functionalities

then that should be added and with that many other issues come into play. While it is possible, this system in question does not intend to utilize the capabilities of generative AI.

In conclusion, the feedback from experts was mostly positive. The threat identification part was very positive as validation that the threats are relevant was acquired. The modelling part still got mostly positive feedback, but also some areas definitely could be improved to improve clarity and the data flow. The improvements could involve dividing the model registry into two separate components or integrating the "Performance Validation" earlier in the lifecycle. After the results from the questionnaire, it is clear that the model is up to standards and all the relevant AI/ML-specific threats have been identified.

6. Summary

AI/ML-based systems are undeniably gaining popularity in various domains. These systems are trying to extract benefits from the enormous amounts of data generated every day. With these technological advancements, new or previously overlooked security threats emerge. Security issues within healthcare systems could bear devastating consequences.

This thesis has conducted an in-depth threat modelling that has captured all the characteristics of a modern healthcare system that utilizes the usage of AI/ML systems. The Data Flow Diagram conducted uses an in-house development approach for the AI/ML components and uses all the core components of an AI/ML model lifecycle.

Additionally, this thesis has provided a comprehensive analysis of the AI/ML-specific security threats affecting these AI/ML-based systems in healthcare settings. By using the STRIDE methodology and STRIDE-based attack trees, a thorough examination of potential threats has been conducted. Through this analysis, numerous AI/ML-specific and conventional cybersecurity threats have been identified showcasing the vulnerability of the AI/ML model lifecycle. These identified threats showcase the need for proper cybersecurity defences to protect from malicious attacks, privacy breaches and other threats that could compromise patient safety or treatment accuracy.

Both the Data Flow Diagram and the threats identified were validated by four experts from the cybersecurity and AI/ML field. The experts were given a questionnaire that consisted of twenty-seven questions. The questionnaire was split into two sections, first covering the modelling part, and second covering the threats identified part. After the validation from experts, it is clear that the model is up to standards and all the relevant AI/ML-specific threats have been identified.

As AI/ML-based systems are becoming more commonly used in healthcare, security considerations must be integrated into the early stages of the development process to tackle all the opposing threats. When taking benefits from these new technological advancements, it is important to make sure that the safety and privacy of the users are maintained. The model creation could be researched even further to improve clarity and coverage. Additionally, new AI/ML-specific threats may surface or the threats presented in this thesis could be studied further. Continued research in this field is essential to stay ahead of the evolving threats opposing AI/ML-based healthcare systems.

References

[1] IBM. *IBM Watson for Oncology*. Last accessed 12.05.2024. 2015. URL: https://www.ibm.com/docs/en/announcements/watson-oncology.

[2] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. "Deep learning in bioinformatics". In: *Briefings in bioinformatics* 18.5 (2017), pp. 851–869.

[3] Riccardo Miotto et al. "Deep learning for healthcare: review, opportunities and challenges". In: *Briefings in bioinformatics* 19.6 (2018), pp. 1236–1246.

[4] Sudipto Datta, Ranjit Barua, and Jonali Das. "Application of Artificial Intelligence in Modern Healthcare System". In: *Alginates*. Ed. by Leonel Pereira. Rijeka: IntechOpen, 2019. Chap. 8. DOI: 10.5772/intechopen.90454. URL: https://doi.org/10.5772/intechopen.90454.

[5] Aditi Govindu and Sushila Palwe. "Early detection of Parkinson's disease using machine learning". In: *Procedia Computer Science* 218 (2023), pp. 249–261.

[6] Vivek Kaul, Sarah Enslin, and Seth A. Gross. "History of artificial intelligence in medicine". In: *Gastrointestinal Endoscopy* 92.4 (2020), pp. 807–812. ISSN: 0016-5107. DOI: https://doi.org/10.1016/j.gie.2020.06.040. URL: https://www.sciencedirect.com/science/article/pii/S0016510720344667.

[7] Tingting Han, Fan Yang, and Kesui Deng. "Application and Development Prospect of Artificial Intelligence in Healthy Pension Industry". In: *Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare*. CAIH2020. Taiyuan, China: Association for Computing Machinery, 2020, pp. 79–83. ISBN: 9781450388641. DOI: 10.1145/3433996.3434364. URL: https://doi-org.ezproxy.utlib.ut.ee/10.1145/3433996.3434364.

[8] Srinivasa Rao Burri et al. "Predictive Intelligence for Healthcare Outcomes: An AI Architecture Overview". In: *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*. 2023, pp. 1–6. DOI: 10.1109/ICSTSN57873.2023.10151477.

[9] Tariq Ahamed Ahanger Anil Audumbar Pise Khalid K. Almuzaini. "Enabling Artificial Intelligence of Things (AIoT) Healthcare Architectures and Listing Security Issues". In: *Computational Intelligence and Neuroscience*. 2022.

[10] Lakis Christodoulou. "AI Remote Vital Signs Monitoring and Diagnostics based on Wireless Wearable Bio-sensors-Systems-Devices". In: Jan. 2021.

[11] M. M. Kamruzzaman. “Architecture of Smart Health Care System Using Artificial Intelligence”. In: *2020 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 2020, pp. 1–6. DOI: 10.1109/ICMEW46912.2020.9106026.

[12] Hans Günter Brauch. “Concepts of Security Threats, Challenges, Vulnerabilities and Risks”. In: *Coping with Global Environmental Change, Disasters and Security: Threats, Challenges, Vulnerabilities and Risks*. Ed. by Hans Günter Brauch et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 61–106. ISBN: 978-3-642-17776-7. DOI: 10.1007/978-3-642-17776-7_2. URL: https://doi.org/10.1007/978-3-642-17776-7_2.

[13] Wayne A Jansen, Theodore Winograd, and Karen Scarfone. “Guidelines on active content and mobile code”. In: *NIST Special Publication* 800 (2001), p. 28.

[14] FIPS Pub. “Minimum security requirements for federal information and information systems”. In: (2006).

[15] Michael Nieles, Kelley Dempsey, Victoria Yan Pillitteri, et al. “An introduction to information security”. In: *NIST special publication* 800.12 (2017), p. 101.

[16] Joint Task Force Transformation Initiative Interagency Working Group et al. “Security and Privacy Controls for Information Systems and Organizations”. In: *NIST Special Publication* (2020), pp. 800–53.

[17] Microsoft. *Threat Modeling*. Last accessed 12.05.2024. URL: https://www.microsoft.com/en-us/securityengineering/sdl/threatmodeling.

[18] Adam Shostack. *Threat modeling: Designing for security*. John Wiley & Sons, 2014.

[19] Larry Conklin. *Threat Modeling Process*. Last accessed 12.05.2024. URL: https://owasp.org/www-community/Threat_Modeling_Process.

[20] Laurens Sion et al. “Security threat modeling: are data flow diagrams enough?” In: *Proceedings of the IEEE/ACM 42nd international conference on software engineering workshops*. 2020, pp. 254–257.

[21] Visual Paradigm. *DFD Using Yourdon and DeMarco Notation*. Last accessed 12.05.2024. URL: https://online.visual-paradigm.com/knowledge/software-design/dfd-using-yourdon-and-demarco.

[22] Bruce Schneier. “Attack trees”. In: *Dr. Dobb’s journal* 24.12 (1999), pp. 21–29.

[23] Gary Anthes. “Artificial intelligence poised to ride a new wave”. In: *Communications of the ACM* 60.7 (2017), pp. 19–21.

[24] IH Sarker. *Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2: 160*. 2021.

[25] Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine learning: algorithms and applications*. Crc Press, 2016.

[26] Eric Schmidt et al. "National Security Commission on artificial intelligence". In: (2021).

[27] Yevgeniy Vorobeychik et al. "Adversarial machine learning". In: (2018).

[28] Carl Wilhjelm and Awad A. Younis. "A Threat Analysis Methodology for Security Requirements Elicitation in Machine Learning Based Systems". In: *2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*. 2020, pp. 426–433. DOI: 10.1109/QRS-C51114.2020.00078.

[29] Huda Ali Alatwi and Charles Morisset. "Threat Modeling for Machine Learning-Based Network Intrusion Detection Systems". In: *2022 IEEE International Conference on Big Data (Big Data)*. 2022, pp. 4226–4235. DOI: 10.1109/BigData55660.2022.10020368.

[30] Lara Mauri and Ernesto Damiani. "Modeling Threats to AI-ML Systems Using STRIDE". In: *Sensors* 22.17 (2022). ISSN: 1424-8220. DOI: 10.3390/s22176662. URL: https://www.mdpi.com/1424-8220/22/17/6662.

[31] Md. Rashid Al Asif et al. "STRIDE-based Cyber Security Threat Modeling for IoT-enabled Precision Agriculture Systems". In: *2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI)*. 2021, pp. 1–6. DOI: 10.1109/STI53101.2021.9732597.

[32] Jean-Paul A Yaacoub et al. "Securing internet of medical things systems: Limitations, issues and recommendations". In: *Future Generation Computer Systems* 105 (2020), pp. 581–606.

[33] Amit Sundas and Badotra. "Recurring Threats to Smart Healthcare Systems Based on Machine Learning". In: *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. 2022, pp. 1–8. DOI: 10.1109/ICRITO56286.2022.9964783.

[34] Matteo Cagnazzo et al. "Threat modeling for mobile health systems". In: *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. 2018, pp. 314–319. DOI: 10.1109/WCNCW.2018.8369033.

[35] Information System Authority. *Tehisintellekti ja masinõppe tehnoloogia riskide ja nende leevendamise võimaluste uuring*. 2024.

[36] Fahad Taha Al-Dhief et al. "A Survey of Voice Pathology Surveillance Systems Based on Internet of Things and Machine Learning Algorithms". In: *IEEE Access* 8 (2020), pp. 64514–64533. DOI: 10.1109/ACCESS.2020.2984925.

[37] Cleveland Clinic. *Blood Oxygen Level*. Last accessed 12.05.2024. 2022. URL: https://my.clevelandclinic.org/health/diagnostics/22447-blood-oxygen-level.

[38] Ishan Gupta. *Mobile App Industry Statistics 2023*. Last accessed 12.05.2024. 2023. URL: https://ripenapps.com/blog/mobile-app-industry-statistics/.

[39] Siddique Latif et al. "Mobile Health in the Developing World: Review of Literature and Lessons from A Case Study". In: *IEEE Access* PP (June 2017), pp. 1–1. DOI: 10.1109/ACCESS.2017.2710800.

[40] European Union Agency for Cybersecurity (ENISA). *Artificial Intelligence Cybersecurity Challenges*. 2020.

[41] Isabel Bär Dr. Larysa Visengeriyeva Anja Kammer. *MLOps Principles*. Last accessed 12.05.2024. URL: https://ml-ops.org/content/mlops-principles.

[42] Pradumn Saxena and Sandeep Prabhu. "Framework For Predicting Suicidal Attempts Using Healthcare Data and Artificial Intelligence". In: *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*. 2023, pp. 1085–1089. DOI: 10.1109/ICIDCA56705.2023.10099967.

[43] Evidently AI. *The open-source ML observability platform*. Last accessed 12.05.2024. URL: https://www.evidentlyai.com/.

[44] C.C.Y. Poon, Yuan-Ting Zhang, and Shu-Di Bao. "A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health". In: *IEEE Communications Magazine* 44.4 (2006), pp. 73–81. DOI: 10.1109/MCOM.2006.1632652.

[45] Ovunc Kocabas, Tolga Soyata, and Mehmet K. Aktas. "Emerging Security Mechanisms for Medical Cyber Physical Systems". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13.3 (2016), pp. 401–416. DOI: 10.1109/TCBB.2016.2520933.

[46] Michael Howard and Steve Lipner. *The Security Development Lifecycle*. USA: Microsoft Press, 2006. ISBN: 0735622140.

[47] Apostol Vassilev et al. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. en. 2024. DOI: https://doi.org/10.6028/NIST.AI.100-2e2023. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=957080.

[48] Harold Figueroa Gary McGraw. *AN ARCHITECTURAL RISK ANALYSIS OF MACHINE LEARNING SYSTEMS: Toward More Secure Machine Learning*. Berryville Institute of Machine Learning, 2020.

[49] Matthew Jagielski, Alina Oprea, and Biggio. "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning". In: *2018 IEEE Symposium on Security and Privacy (SP)*. 2018, pp. 19–35. DOI: 10.1109/SP.2018.00057.

[50] Mehran Mozaffari-Kermani and Sur-Kolay. "Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare". In: *IEEE Journal of Biomedical and Health Informatics* 19.6 (2015), pp. 1893–1905. DOI: 10.1109/JBHI.2014.2344095.

[51] Abdur Rahman et al. "Adversarial Examples—Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices". In: *IEEE Internet of Things Journal* 8.12 (2021), pp. 9603–9610. DOI: 10.1109/JIOT.2020.3013710.

[52] Jiaqi Ruan et al. "Applying Large Language Models to Power Systems: Potential Security Threats". In: *IEEE Transactions on Smart Grid* (2024).

[53] Reza Shokri et al. *Membership Inference Attacks against Machine Learning Models*. 2017. arXiv: 1610.05820 [cs.CR].

[54] Boris van Breugel et al. *Membership Inference Attacks against Synthetic Data through Overfitting Detection*. 2023. arXiv: 2302.12580 [cs.LG].

[55] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. "I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences". In: 55.14s (2023). ISSN: 0360-0300. DOI: 10.1145/3595292. URL: https://doi.org/10.1145/3595292.

[56] Matthew Fredrikson et al. "Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing". In: *Proceedings of the 23rd USENIX Conference on Security Symposium*. SEC'14. San Diego, CA: USENIX Association, 2014, pp. 17–32. ISBN: 9781931971157.

[57] Jugal Parikh Andrew Marshall. *Threat Modeling AI/ML Systems and Dependencies*. 2022.

[58] Alexander Wood, Kayvan Najarian, and Delaram Kahrobaei. "Homomorphic encryption for machine learning in medicine and bioinformatics". In: *ACM Computing Surveys (CSUR)* 53.4 (2020), pp. 1–35.

[59] Uwe Flick. *The SAGE handbook of qualitative data analysis*. Sage, 2013.

Appendix 1 – Non-exclusive license for reproduction and publication of a graduation thesis[1]

I Janno Jaal

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Threat Modelling for AI/ML-based Healthcare Systems", supervised by Hayretdin Bahşi
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

12.05.2024

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 - AI/ML-based healthcare system model

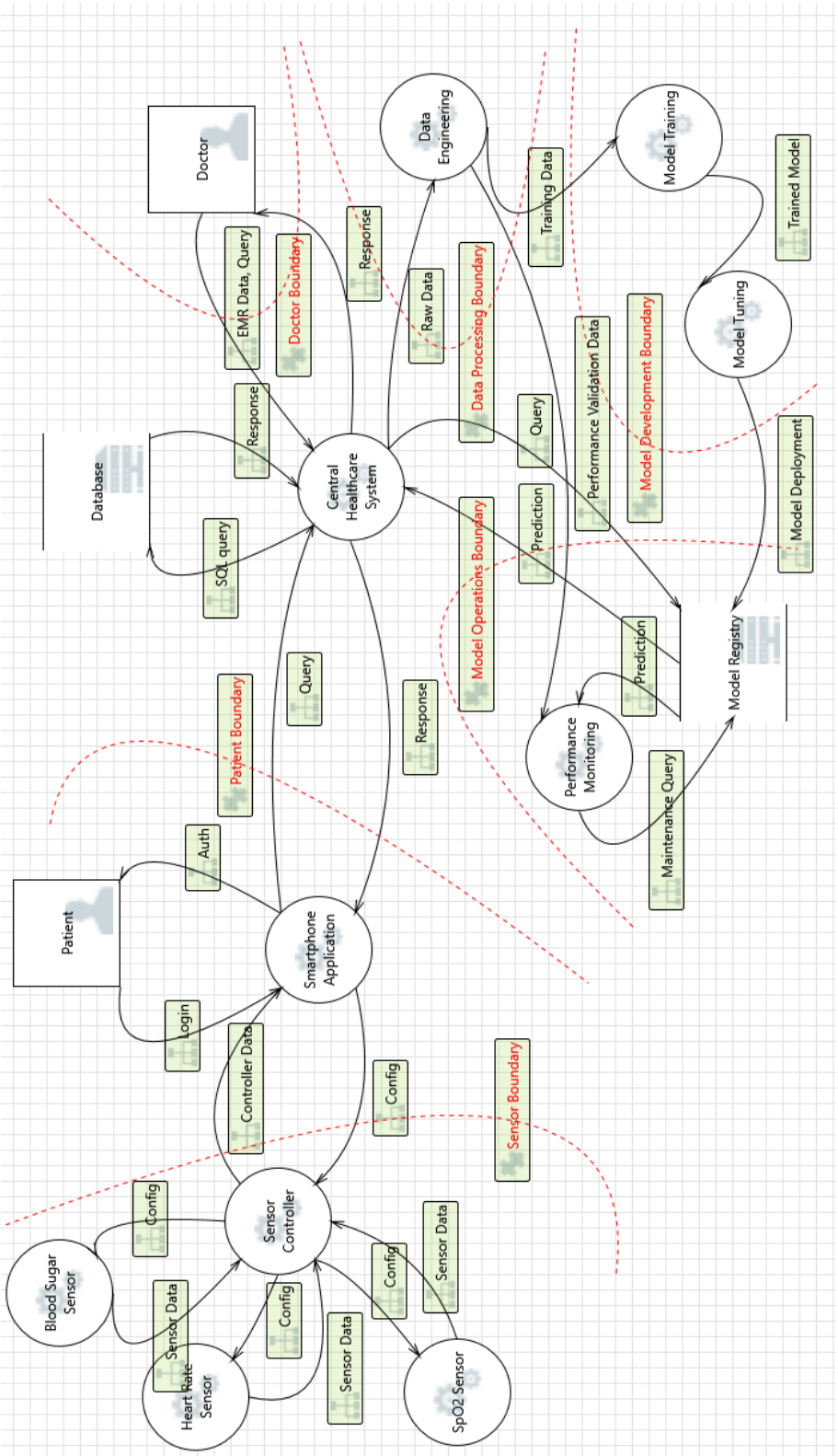


Figure 25. *AI/ML-based healthcare system.*

Appendix 3 - Questionnaire for experts

Threat modelling for AI/ML-based healthcare systems

My name is Janno Jaal and for my Masters Thesis I conducted threat modelling for AI/ML-based healthcare systems. This questionnaire aims to get experts' input to validate the threat model and threats identified. The Data Flow Diagram and the table containing the identified threats are provided separately.

Artificial Intelligence (AI) and Machine Learning (ML) continue to revolutionize healthcare, providing more effective detection of diseases, management of chronic conditions, delivery of health services, and drug discovery. As an enormous amount of health data is generated through electronic health records, imaging, sensor data and text, AI/ML-based systems are more widely used.

The model captures all the characteristics of a modern healthcare system that utilizes the usage of an AI/ML component. The threat modelling is conducted based on the STRIDE methodology. In addition, STRIDE-based attack trees are used to further identify all the relevant threats that could endanger a modern healthcare system.

* Required

Intro

This questionnaire consists of 27 questions. Some are choice, some are text based. It is divided into 2 sections - the modeling part and the threat identification part. Altogether this should not take more than 20-30 minutes.

All the required materials for completing the questionnaire can be found here:
https://drive.google.com/drive/folders/1_PFeTBnj7d4_KrFPvfj1zOPwpPnumrQm?usp=drive_link

wholeModel.png - The Data Flow Diagram of a healthcare system that utilizes AI/ML components.

AI_ML_Components.png - The narrowed down Data Flow Diagram, only consists of AI/ML related components. All the components are given unique ID's. Processes = P, Data Flows = DF and Data Stores = DS.

ThreatsIdentified.pdf - List of the AI/ML related threats identified.
The first column displays the threat ID. Process threats are marked as T-P, data flow threats are marked as T-DF and data store threats marked as T-DS. Additionally, the threats provided are divided into two categories, conventional cybersecurity threats (Conv) and AI/ML-specific cybersecurity threats (AI/ML). This is explained further in the file.

1

Please enter you name *

The results represented in the thesis will be anonymous, but for feedback processing purposes the name is needed.

2

Years of professional experience in Cyber Security *

- ◯ No experience
- ◯ Less than 3 years
- ◯ More than 3 years
- ◯ More than 6 years
- ◯ More than 9 years

3

Years of professional experience with AI/ML systems *

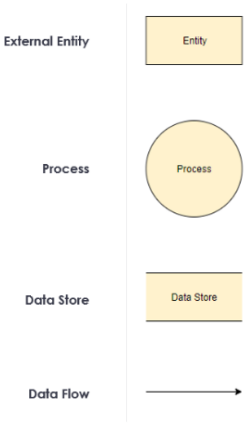
- ◯ No experience
- ◯ Less than 3 years
- ◯ More than 3 years
- ◯ More than 6 years
- ◯ More than 9 years

4

Please provide additional details about prior experience. *

For example experience with adversarial attacks, threat modeling or what type of projects and so on.

Data Flow Diagram



In this section, there are questions about the representation of the AI/ML model components. In addition to the whole model, these components are also brought out in a separate file (**AI_ML_Components.png**).

The AI/ML components are considered to be raw data, training data, performance validation data, query, model registry, data engineering, model training and model tuning, performance monitoring, and central healthcare system.

Data Flow Diagram uses five different elements for notations (explanations from OWASP Threat Modeling Process article by Larry Conklin).
The external entity shape is used to represent any entity outside the application that interacts with the application via an entry point.
The process shape represents a task that handles data within the application. The task may process the data or perform an action based on the data.
The data store shape is used to represent locations where data is stored. Data stores do not modify the data, they only store data.
The data flow shape represents data movement within the application. The direction of the data movement is represented by the arrow.
The privilege boundary (or trust boundary) shape is used to represent the change of trust levels as the data flows through the application. Boundaries show any location where the level of trust changes.

This research focuses on AI/ML-based healthcare systems that provide help with detecting diseases, managing difficult conditions, discovering drugs or medical research. Out of scope are AI/ML-based systems that primarily act as chatbots or medical service delivery systems. The AI/ML-based healthcare system used for threat modelling uses in-house development for the AI/ML model.

5

Does the AI/ML related flow of data within the healthcare system make sense? *

6

Are there any additional data flows that should be included or if any unnecessary flows can be removed? *

7

Do you agree with the approach that all the data processing tasks are represented as one process "Data Engineering"? *

The "Data Engineering" process aims to clean and preprocess the raw data to make it suitable for the model development phase. In this step also feature engineering is done. All the data engineering tasks are gathered into this one process and are represented at a more abstract level (ID = **P2**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] Yes
- [ ] No
- [ ] Other

8

Do you agree that the "Model Registry" is best represented as a data store? *

After data processing and model development, the model is ready for the operational stage and the model is deployed into the model registry. The model registry then responds to the queries coming from the central healthcare system (ID = **DS1**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] Yes
- [ ] No
- [ ] Other

9

Do you agree with the representation of "Model Training" and "Model Tuning" processes? *

The “Model Training” and “Model Tuning” processes make up the model development phase. The goal of the model development phase would be to create, test and deliver a stable model that can be used furthermore for accurate predictions (ID = **P3,P4**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] Yes
- [ ] No
- [ ] Other

10

Do you agree with how "Performance Monitoring" is represented? *

After the model is developed then it moves into the operational stage. Monitoring and documenting the model performance is the last step of the model lifecycle. Performance monitoring is needed to guarantee data quality and to gain visibility into the model performance (ID = **P5**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] Yes
- [ ] No
- [ ] Other

11

Are the important steps of a AI/ML model lifecycle represented? *

The lifecycle of an AI/ML model is represented by combining two principles from ENISA and MLOps. The representation in the model is divided into three sub-categories: data processing, model development and model operation. ENISA provides all the steps for the lifecycle and MLOps adds the operational stage tasks.

For additional comments about the answer please select and write in the "Other" box.

- [ ] Yes
- [ ] No
- [ ] Other

12

Are the data flows and system components appropriately isolated with boundaries? *

The three phases in the model lifecycle are divided into three sub-categories: data processing, model development and model operation. All of these phases are separated by trust boundaries.

13

Additional comments

Identified threats

In this section, questions are based on the table of the threats identified for AI/ML systems (**ThreatsIdentified.pdf**).

14

Is data poisoning a relevant threat? *

Data poisoning in healthcare means the intentional manipulation of medical data to corrupt AI/ML model training, resulting in erroneous diagnoses and compromised patient care.

Threat is present in the **Training Data, Performance Validation Data** data flows (Threat ID = **T-DF1**).
Threat is present in the **Data Engineering, Performance Monitoring** processes (Threat ID = **T-P1**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

15

Is data confidentiality threats a relevant threat? *

Data confidentiality threats in healthcare involve unauthorized access or disclosure of sensitive medical data, posing risks of privacy breaches and potential misuse of patients' personal information.

Threat is present in the **Raw Data, Training Data, Performance Validation Data** data flows (Threat ID = **T-DF2**).
Threat is present in the **Data Engineering, Performance Monitoring, Central Healthcare System** processes (Threat ID = **T-P2**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

16

Is data trustworthiness a relevant threat? *

The data from sensors and medical records might not be trustworthy, reliable and suitable.

Threat is present in the **Raw Data** data flow (Threat ID = **T-DF3**).
Threat is present in the **Central Healthcare System** process (Threat ID = **T-P3**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

17

Is data reconstruction a relevant threat? *

During data reconstruction, the attacker tries to reconstruct a subset of training data. This can lead to the loss of sensitive patient data, posing risks to privacy and confidentiality

Threat is present in the **Prediction** data flow (Threat ID = **T-DF10**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

18

Is denial of service a relevant threat? *

Denial of Service disrupts healthcare AI/ML systems by overwhelming them with requests. This can potentially disrupt critical medical decision-making.

Threat is present in the **Query** data flow (Threat ID = **T-DF7**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

19

Is evasion attack a relevant threat? *

Evasion attacks aim to deceive healthcare AI/ML systems, jeopardizing patient safety and treatment accuracy.

Threat is present in the **Query** data flow (Threat ID = **T-DF6**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

20

Is membership inference a relevant threat? *

During membership inference attacker tries to find out if a particular record or sample was part of the training, by querying the model. This could uncover sensitive patient data, posing risks to privacy and confidentiality

Threat is present in the **Prediction** data flow (Threat ID = **T-DF9**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

21

Is model extraction a relevant threat? *

Model extraction exploits healthcare AI/ML models by trying to extract information about the model architecture and parameters.

Threat is present in the **Prediction** data flow (Threat ID = **T-DF8**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

22

Is model capturing a relevant threat? *

Model capturing refers to the unauthorized access or theft of AI/ML models. This could lead to model misuse and compromise patient care and confidentiality.

Threat is present in the **Model Deployment** data flow (Threat ID = **T-DF5**).
Threat is present in the **Model Registry** data store (Threat ID = **T-DS3**).
Threat is present in the **Model Training, Model Tuning** processes (Threat ID = **T-P5**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

23

Is model replacement a relevant threat? *

Model replacement involves replacing healthcare AI/ML models with malicious ones. This could pose risks to patient safety and treatment effectiveness.

Threat is present in the **Model Deployment** data flow (Threat ID = **T-DF4**).
Threat is present in the **Model Registry** data store (Threat ID = **T-DS2**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

24

Is model reprogramming a relevant threat? *

Model reprogramming involves altering healthcare AI/ML models to produce incorrect or biased outputs, posing risks to patient safety and treatment effectiveness.

Threat is present in the **Model Registry** data store (Threat ID = **T-DS1**).
Threat is present in the **Model Training, Model Tuning** processes (Threat ID = **T-P4**).

For additional comments about the answer please select and write in the "Other" box.

- [ ] No
- [ ] Yes
- [ ] Other

25

Are the threats clearly defined and understood, including its potential impact on the healthcare system? *

26

Are there any additional threats that were not identified and should be added? *

27

Additional comments

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

Microsoft Forms