FeatureCloud

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare

Project coordinator:
Professor Dr Jan Baumbach, TECHNISCHE UNIVERSITAET MUENCHEN (TUM)
info@featurecloud.eu

Deliverable D2.1
"Risk assessment methodology"

Work Package WP2
"Cyber risk assessment and mitigation"

Disclaimer

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078. Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

Document information

Grant Agreement Number 826078			**Acronym FeatureCloud**	
Full title	Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare			
Topic	Toolkit for assessing and reducing cyber risks in hospitals and care centres to protect privacy/data/infrastructures			
Funding scheme	RIA - Research and Innovation action			
Start Date	1 January 2019	**Duration**	60 months	
Project URL	https://featurecloud.eu/			
EU Project Officer	Reza RAZAVI (CNECT/H/03)			
Project Coordinator	Jan Baumbach, TECHNISCHE UNIVERSITAET MUENCHEN (TUM)			
Deliverable	D2.1 “Risk assessment methodology”			
Work Package	WP2 “Cyber risk assessment and mitigation”			
Date of Delivery	**Contractual**	M16 (30/04/2020)	**Actual**	M22 (02/10/2020)
Nature	REPORT	**Dissemination Level**	PUBLIC	
Lead Beneficiary	05 - SBA			
Responsible Author(s)	Rudolf Mayer (SBA)			
	Anastasia Pustozerova (SBA)			
	Walter Hötzendorfer (RI)			
Keywords	Risk Assessment, Machine Learning Security & Privacy Risks, Data Protection Analysis			

Table of Content

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Acronyms and definitions

AIC	availability, integrity and confidentiality
API	application program interface
ASVS	Application Security Verification Standard
BYOD	Bring Your Own Device
CIA	Confidentiality, Integrity and Availability
CIS	Center of Internet Security
concentris	concentris research management GmbH (Germany)
CSIRT	Computer Security Incident Response Team
CVE	Common Vulnerabilities and Exposures
CVSS	Common Vulnerability Scoring System
DoA	Description of the Action
DoS	Denial of service attacks
ENISA	European Union Agency for Cybersecurity
ENISA	European Union Agency for Network and Information Security
EU	European Union
FedAvg	FederatedAveraging
FIRST	Forum of Incident Response and Security Teams
GDPR	General Data Protection Regulation
GND	Gnome Design SRL (Romania)
IGs	Implementation Groups
ISO	International Organization for Standardization
IT	information technology
KPI	key performance indicator
ML	machine learning
MUG	Medizinische Universität Graz (Austria)
NIS Directive	The Directive on security of network and information systems
NIST	National Institute of Standards and Technology
non-IID	not independent and identically distributed
OWASP	Open Web Application Security Project
P2P	Peer to Peer
patients	In this deliverable, we use the term “patients” for all research subjects. In FeatureCloud, we will focus on patients, as this is already the most vulnerable case scenario and this is where most primary data is available to us. Admittedly, some research subjects participate in clinical trials but not as patients but as healthy individuals, usually on a voluntary basis and are therefore not dependent on the physicians who care for them. Thus to increase readability, we simply refer to them as “patients”.
RI	Research Institute AG & Co. KG (Austria)
RM	Risk Management
SBA	SBA Research gemeinnützige gGmbH (Austria)
SDU	Syddansk Universitet (Denmark)
TUM	Technische Universität München (Germany)
UM	Universiteit Maastricht (The Netherlands)
UMR	Philipps Universität Marburg (Germany)
WP	work package

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

1 Objectives of the deliverables based on the Description of Action

The main objective of WP2 is the definition of a security and privacy architecture **based on the cyber risk assessment and legal requirements**.

“While the approach of the FeatureCloud project by design mitigates most major concerns regarding security and privacy, the underlying foundations of the platform to be developed need to be secured thoroughly, especially considering the diversity of the local execution platforms on the hospital sites. Important attacker goals include the theft of data on a local level, as well as the theft and manipulation of results, with the inclusion of possible insider attacks.”

Objective 1 of this work package thus aims *“to develop a risk assessment methodology based on the legal and technical requirements derived from data protection law regulation on the one hand and an in-depth attacker analysis on the other.”*

The corresponding task in this work package is *Task 1: Risk Assessment Methodology :*
“The platform to be developed solves many issues which traditional cloud-based analytical platforms processing sensitive information have simply by design. Still, there are some attack vectors left that require further analysis and mitigation. In this task, SBA, MUG and RI will therefore provide a novel risk assessment methodology that targets the specific side parameters and requirements of the platform. This also includes potential insider attacks like a malicious platform, adversarial administrators (local and global), but also typical outsider threats. In addition to these technical aspects, the methodology requires the introduction of legal aspects, introducing them as side parameters for the risk and damage modelling, as well as their utilization as countermeasures.“

2 Executive Summary / Abstract

Methodology

Federated learning provides stronger privacy protection than a centralised machine learning approach, by its design. As there is no need to centralise the data at one place, as is the case in centralised machine learning, the scenarios change for an adversary. A federated learning setting, where the “raw” data (about patients, etc.) stays in its original location, e.g. at the health care provider, and the analysis algorithm instead “visits” the data, generally does not provide a larger attack surface as compared to the status quo present at the data holder. Compared to a centralised data gathering, further the impact of a successful attack (on a single node as compared to the centralised system) is generally expected to be lower, as a successful attacker gains access to a lower number of data assets.

While the federated approach chosen in FeatureCloud thus reduces many of the security and privacy risks commonly associated with “traditional” data analysis settings, where data needs to be centralised for learning the predictive models, still some risks remain, and other risks might emerge. Regarding the latter, the federated setting itself entails certain other risks that follow by its nature as a distributed system. Moreover, recently a number of threats considering the robustness and confidentiality of machine learning have been the subject of research, much of which is subsumed under the term of adversarial machine learning.

Therefore, in this deliverable, we perform an analysis on security and privacy risks from various angles, complemented by a legal perspective on these aspects.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Main results

We identify risks on several aspects, covering threats

- On single systems, and due to the distributed nature of the platform
- Against the privacy of data records and other information, namely via breaking the confidentiality of the learned machine learning model
- Against the security of the learning process, namely against the integrity and availability of the machine learning process or the resulting model
- Based on and complemented by risks and mitigations regarding data protection from law and other regulatories

Progress beyond the state-of-the-art
Especially risks associated with the machine learning process and the confidentiality of the resulting models are an emerging field, to which we can contribute with this holistic analysis. Considering also the legal context is a novel approach.

3 Challenge

Having to share and centralise data poses a major challenge in any data science endeavour that wants to analyse data from distributed sources. Especially in settings like health care and e-health, data is very often distributed between the various health care institutions, and thus collaborative analysis is an important challenge. FeatureCloud mitigates the issues associated with centralising data, starting from difficulties to obtain consent for such data sharing, to the threat these integrated and very valuable centralised databases imply, as they become an attractive target for adversaries. In the federated setting of FeatureCloud, no such central point is created, thus it becomes more difficult for adversaries to obtain the aggregated database, as they need to extract data from each contributing source.

On the other hand, some further risks might emerge due to the distributed nature of the resulting system, which not only requires one-time data transfers, but will rather continuously exchange information. Even if this information is in the form of machine learning models resp. the updates to their learned parameters, some risks for information leakage still pertain. Moreover, the learning process in the federated learning algorithm can potentially be easier manipulated than a process working on a centralised database.

These above identified types of risks will be discussed in detail in this deliverable, and will form the basis for risk analysis for all components. It further is the basis for developing mitigation strategies, for which we briefly outline potential aspects that Deliverable “D2.4 Set of (novel) attack vectors and countermeasures”, delivered in month 36, will describe in detail. In the following, we first introduce terminology related to general risks in IT. We further structure the deliverable along risks from the nature of the distributed system, from adversarial machine learning

4 Terminology and Background

This section provides an overview on common terminology and existing frameworks and models in the field of computer security, which we will build upon in the rest of this deliverable. On a generic level, **Risk Management** (RM) concerns the **assessment and control** of risks, with risk being defined as the combination of the likelihood of an event and its consequences[1]. The ultimate goal is

[1] ISO. ISO Guide 73:2009 – Risk management – Vocabulary. International Organization for Standardization, 2009.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

to manage the uncertainty associated with risks. This is achieved primarily by mitigating risks with negative consequences on objectives, often considering also the costs of the mitigation actions. A risk-based approach is a foundation for managing cybersecurity, and is recommended by the US National Institute (NIST) and also the European Union Agency for Network and Information Security (ENISA).

While different standards, methods and tools exist for targeting specific domains, ISO 31000:2009[2] describes a generic and domain-independent framework for risk management and more focused on risk identification, analysis, evaluation. The framework provides the underlying concepts and principles, along with a risk management process, which is depicted in Figure 1.

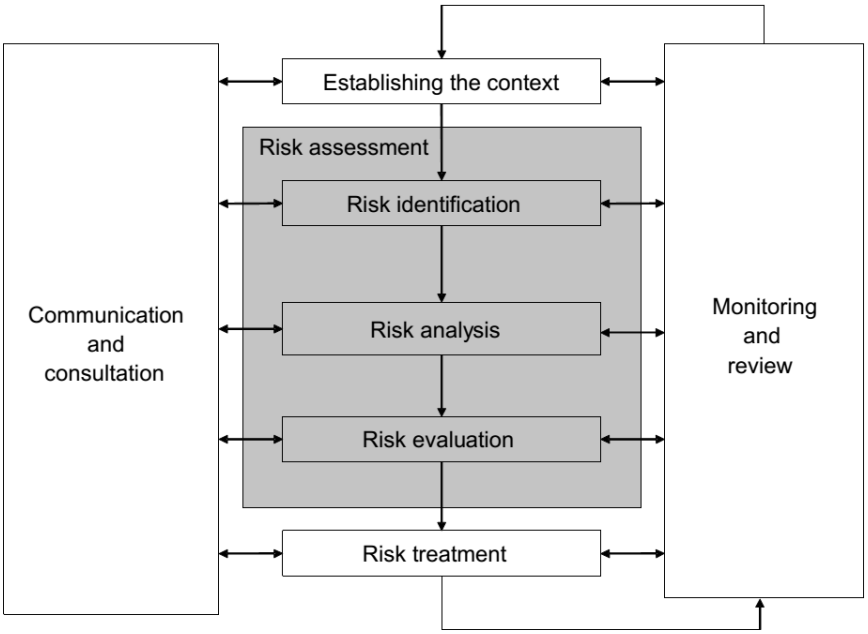

Figure 1: Risk Management Process according to ISO 31000:2009

The process is initiated by defining the internal and external context of the asset. The external context might include a description of the regulatory environment of the asset, or any other element that might affect the asset to be assessed. The internal context includes defining all the elements of the system.

In the next step, the assessment of the risks based on the collected information is performed. It is composed of three different steps: (1) risk identification, where all relevant assets, vulnerabilities, events and risks are identified; (2) risk analysis, where the value of the assets, the exposure to vulnerabilities, the likelihood of events, the risk consequence, and ultimately the risk severity are estimated; and (3) risk evaluation, where the information gathered in (1) and (2) is evaluated, culminating in a decision on whether a specific risk is acceptable or tolerable. Depending on the context of the risk assessment, different techniques can be applied to the process.
Several different variations of risk matrices have been proposed, but they generally try to depict the likelihood of a risk and the impact (severity) of it. They often differ in the granularity of the

[2] ISO/FDIS. ISO/FDIS 31000:2009 – Risk management – Principles and guidelines. International Organization for Standardization, 2009.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

assessment, i.e. the number of different levels used to distinguish between the severity and impact the risks, commonly ranging from 2 to 5 levels.

A risk is expressed by a risk severity (or risk level), that is a combination of its consequence with the likelihood of the event triggering the risk. Controls are defined as actions to be taken to mitigate risks. Controls can reduce the exposure of a vulnerability, reduce the likelihood of an event, reduce the risk consequence, or transfer the risk. A policy represents a set of controls that were applied to mitigate the risks in a specific context.

Risks which are not very likely and have a low impact, and also risks with a high cost, but which are less likely, are less likely to be defended against. Risks that have a low likelihood and high costs to mitigate against are maybe not worth defending against. Estimating the probabilities of the risks is often based on previous experience, or in special domains on published data.

The risk assessment steps result in the prioritization for risk treatment and the identification of controls. If the controls are sufficient to lower the overall risk level into acceptable values, then a risk report is defined. All the steps of the process should be communicated to the interested parties for consultation and validation. Additionally, the process should be run continuously, with constant monitoring and review of the different steps, if necessary, so that the risk management is effective.

4.1 IT Security Frameworks and Models

The NIST cybersecurity framework[3] is a risk-based framework, and consists of five basic activities starting with the *Identification* of the risks, and is more focused on the mitigation and recovery, recommending the activities *Protect, Detect Incidents, Respond*, and *Recover*, compare figure 2. We will utilise the NIST framework further for the deliverable D2.4 in the analysis of mitigation strategies.

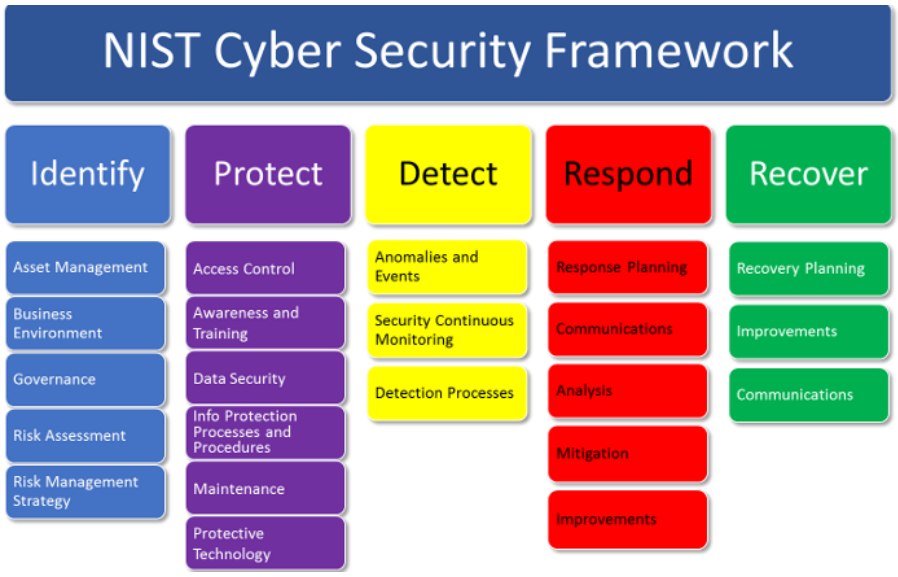

Figure 2: NIST cybersecurity framework

[3] https://www.nist.gov/cyberframework/framework

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

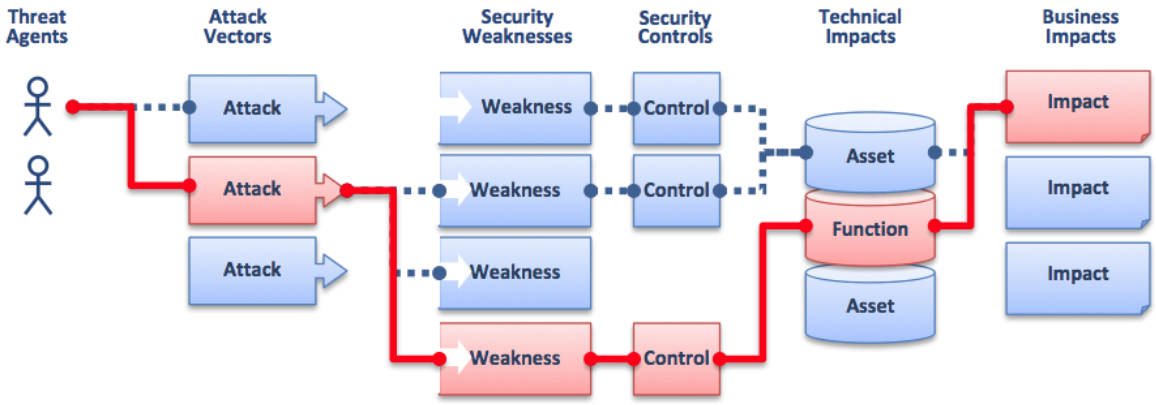

Figure 3: OWASP terminology (https://www.owasp.org/)

Specifically in IT security, the Open Web Application Security Project (OWASP) terminology can be utilised (see Figure 3). In this terminology, a **vulnerability** is a weakness that can be **exploited** by an **attacker**, to perform unauthorized actions within a system. A vulnerability is also known as the **attack surface**. To exploit a vulnerability, an attacker must have an applicable tool or technique that can connect to a system weakness. A vulnerability with one or more known instances of available attacks is thus an **exploitable vulnerability**. A **threat**[4] is the adversary’s goal, or what the adversary might try to do to the system. ISO 27005 defines an **asset** as “anything that has value to the organization, its business operations and their continuity, including information resources that support the organization's mission”[5]. An asset is thus an abstract or concrete **resource that system must protect** from misuse by an adversary. Vulnerability management is the recurring practice of identifying, classifying, remediating, and mitigating vulnerabilities.

A vulnerability is defined as “a weakness of an asset or group of assets that can be exploited by one or more threats”[6], “a flaw or weakness in a system's design, implementation, or operation and management that could be exploited to violate the system's security policy”[7], a “weakness in an information system, system security procedures, internal controls, or implementation that could be exploited by a threat source”[8]. NIST defines a vulnerability in IT context as “a flaw or weakness in system security procedures, design, implementation, or internal controls that could be exercised (accidentally triggered or intentionally exploited) and result in a security breach or a violation of the system's security policy”[9], while the European Union Agency for Cybersecurity (ENISA) defines a vulnerability as “the existence of a weakness, design, or implementation error that can lead to an unexpected, undesirable event compromising the security of the computer system, network, application, or protocol involved”[10].

[4] Frank Swiderski and Window Snyder. 2004. Threat Modeling. Microsoft Press, Redmond, WA, USA.

[5] British Standard Institute, Information technology - Security techniques - Management of information and communications technology security - Part 1: Concepts and models for information and communications technology security management BS ISO/IEC 13335-1-2004

[6] ISO/IEC, "Information technology - Security techniques-Information security risk management" ISO/IEC FIDIS 27005:2008

[7] Internet Engineering Task Force RFC 4949 Internet Security Glossary, Version 2

[8] CNSS Instruction No. 4009, http://www.cnss.gov/Assets/pdf/cnssi_4009.pdf

[9] NIST SP 800-30 Risk Management Guide for Information Technology Systems, http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf

[10] https://www.enisa.europa.eu/topics/threat-risk-management/risk-management/current-risk/risk-management-inventory/glossary/#G52

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

A security risk is often interchangeably used with vulnerability. However, a **security risk is the potential of a significant impact resulting from the exploit of a vulnerability**. It is defined as “the potential that a given threat will **exploit vulnerabilities** of an asset and thereby cause harm to the organization”[11]. If the affected asset has no value, a vulnerability might thus not have an associated risk.

A vulnerability is generally connected with a **temporal aspect**, starting from the moment when a security hole was introduced in a system, to the point when the access to it was removed, a security fix was deployed, or an attacker was disabled.

We can distinguish, on the one hand, well-known/exploit-based attacks, often characterized by a Common Vulnerabilities and Exposures number (CVE) (Tidwell et al. 2001) and related to an available exploit code. On the other hand, a **Zero-day** (also known as 0-day) vulnerability is a vulnerability that is unknown to, or unaddressed by those who should be interested in mitigating the vulnerability (including the vendor of the target component). Until the vulnerability is mitigated, hackers can exploit it to adversely affect computer programs, data, additional computers or a network. An exploit directed at a zero-day is called a zero-day exploit, or zero-day attack. "Day Zero" refers to the first day in which the vulnerability becomes known to a party. These types of vulnerabilities are especially critical, as the exploit for it might be used before the vendor of the product knows about the vulnerability, and thus many systems are unprotected.

The Common Vulnerability Scoring System (CVSS)[12] is an open framework for communicating the characteristics and severity of software vulnerabilities. It was developed by NIST and is maintained by the Forum of Incident Response and Security Teams (FIRST), a US-based non-profit organization. CVSS captures the principal technical characteristics of software, hardware and firmware vulnerabilities. Its outputs include numerical scores indicating the severity of a vulnerability relative to other vulnerabilities. CVSS consists of three metric groups: Base, Temporal, and Environmental, see figure 4. The Base group represents the intrinsic qualities of a vulnerability that are constant over time and across user environments, the Temporal group reflects the characteristics of a vulnerability that change over time, and the Environmental group represents the characteristics of a vulnerability that are unique to a user's environment. The Base metrics produce a score ranging from 0 to 10, which can then be modified by scoring the Temporal and Environmental metrics. We utilise CVSS to categorise potential threats specific to federated/machine learning, and to also give an indicative score, in Section 8.

[11] https://www.enisa.europa.eu/topics/threat-risk-management/risk-management/current-risk/risk-management-inventory/glossary/#G27, (ISO/IEC PDTR 13335-1)
[12] https://www.first.org/cvss/specification-document

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

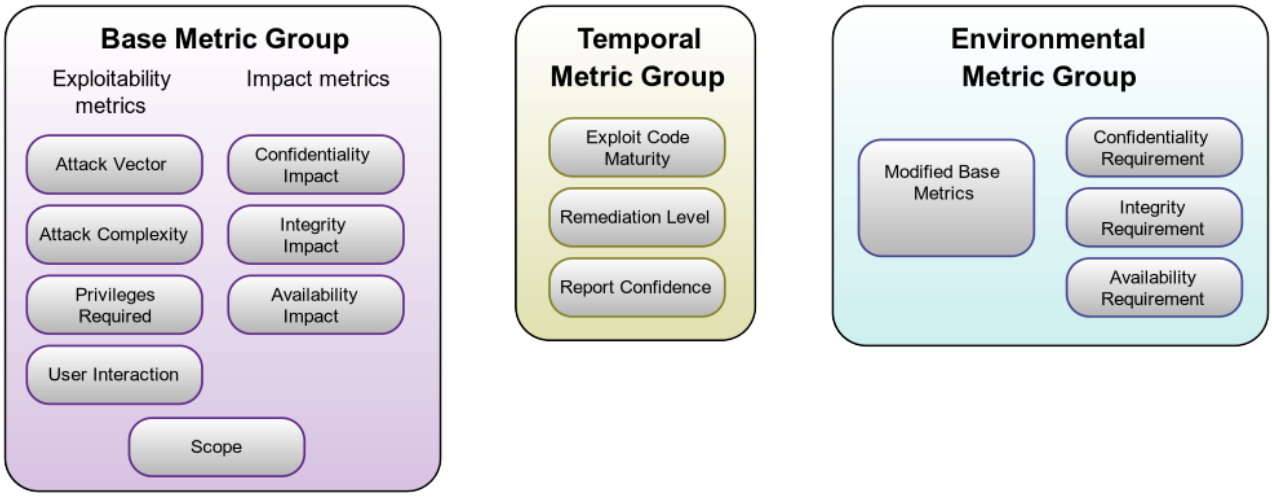

Figure 4: CVSS Metric Groups (https://www.first.org/cvss/)

As a guidance while eliciting and to achieve a large coverage of potential threats, we employ the LINDDUN[13] and STRIDE[14] frameworks. The LINDDUN privacy engineering framework defines three steps: model the system, elicit threats/risks, and manage threats, see figure 5.

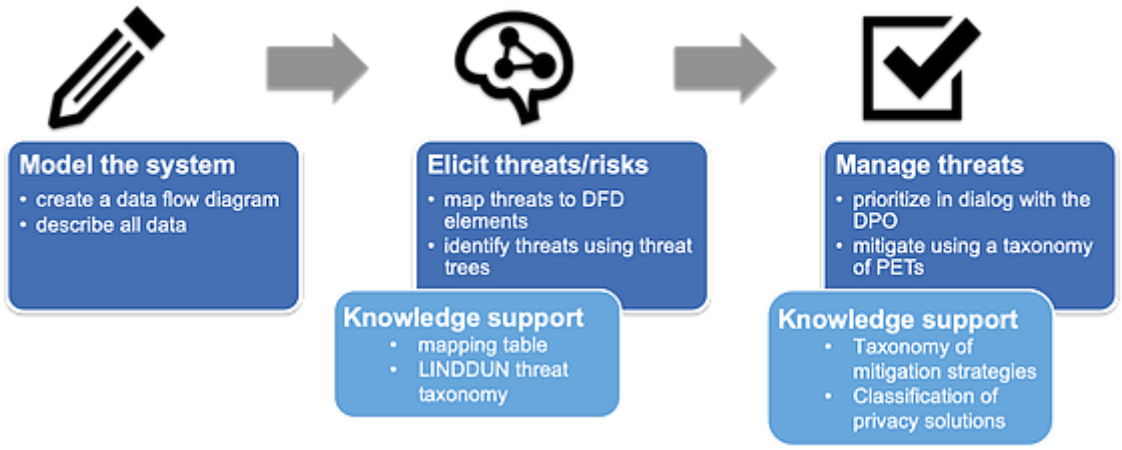

Figure 5: Steps in the LINDDUN Framework

LINDDUN allows to identify privacy threats in software systems, and categorises threats in seven groups:

1. **Linkability.** An attacker is able to distinguish whether two items of interest are linked with each other without knowing the identity of the data subject(s) involved.
2. **Identifiability.** An attacker is able to identify a data subject from a set of data subjects through an item of interest.
3. **Non-repudiation.** The data subject is unable to deny a claim
4. **Detectability.** An attacker is able to distinguish whether an item of interest about a data subject exists or not, regardless of being able to read the contents itself.
5. **Disclosure of information.** An attacker is able to learn the content of an item of interest about a data subject.

[13]https://www.linddun.org/

[14]https://docs.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20)

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

6. **Unawareness.** The data subject is unaware of the collection, processing, storage, or sharing activities (and corresponding purposes) of the data subject’s personal data.
7. **Non-compliance.** The processing, storage, or handling of personal data is not compliant with legislation, regulation, and/or policy.

For identifying security threats we use the STRIDE model, which includes the following categorization of the threats, compare figure 6:

1. **Spoofing**. An attacker is able to illegally access and use another user's authentication information, to gain an illegitimate advantage.
2. **Tampering**. Data tampering involves the malicious modification of data, e.g. unauthorized changes made to persistent data, such as that held in a database.
3. **Repudiation**. Repudiation threats are associated with users who deny performing an action without other parties having any way to prove otherwise - for example, a user performs an illegal operation in a system that lacks the ability to trace the prohibited operations.
4. **Information Disclosure**. Information disclosure threats involve the exposure of information to individuals who are not supposed to have access to it.
5. **Denial of Service**. Denial of service (DoS) attacks deny service to valid users - for example, by making a Web server temporarily unavailable or unusable.
6. **Elevation of Privilege**. In this type of threat, an unprivileged user gains privileged access and thereby has sufficient access to compromise or destroy the entire system.

	Threat	Property Violated	Threat Definition
S	Spoofing identify	Authentication	Pretending to be something or someone other than yourself
T	Tampering with data	Integrity	Modifying something on disk, network, memory, or elsewhere
R	Repudiation	Non-repudiation	Claiming that you didn’t do something or were not responsible; can be honest or false
I	Information disclosure	Confidentiality	Providing information to someone not authorized to access it
D	Denial of service	Availability	Exhausting resources needed to provide service
E	Elevation of privilege	Authorization	Allowing someone to do something they are not authorized to do

Figure 6: Threats in the STRIDE Framework

We use the LINDDUN, STRIDE and CVSS frameworks to evaluate privacy and security threats of the platform and categorize them according to these frameworks. This forms a basis to subsequently select suitable mitigation strategies to ensure platform security.

Vulnerabilities are a broader concept than just software security bugs, as they encompass also hardware, site, or personnel vulnerabilities. In this deliverable, we put a focus on the risks that stem from the *specificity* of the system that the FeatureCloud project develops, which is a federated machine learning platform. Risks that generically apply to any computing system, a standalone, local system or a distributed system alike, and how these risks need to be addressed, are briefly discussed in this deliverable. However, their exact identification, treatment and mitigation is rather delegated to state-of-the-art methodologies and frameworks. Besides these risks not being specific to FeatureCloud, the exact vulnerabilities that might trigger them are also depending on parameters that are not known during the project, such as the final setup of how the system is hosted, who operates it, etc. . Further, several of these vulnerabilities are subject to rather frequent change, and novel attack vectors to e.g. local authentication systems might appear.
For the sake of this deliverable, we therefore focus most of our analysis on vulnerabilities that are specific to federated learning.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

4.2 Threat and Attack Model

One well known model designed for information security analysis is the **Confidentiality, Integrity and Availability model,** also known as the CIA triad (the model is also sometimes referred to as the AIC triad (availability, integrity and confidentiality) to avoid confusion with the Central Intelligence Agency). The dimensions of the triad are considered among the most crucial components of IT security and privacy, and represent a foundation in IT security.

Confidentiality deals with **limiting access to information**, integrity is the assurance that the **information is trustworthy and accurate**, and availability is a guarantee of **reliable access to the information by authorized people.** The CIA triad can be used to describe generic threats to an IT system, and is also utilised for categorising threats to machine learning processes and models. Vulnerabilities can be viewed by the angle of one or more of these three concepts. A graphical representation is given in Figure 7.

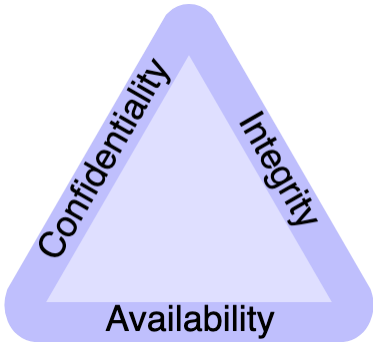

Figure 7: Confidentiality, Integrity and Availability (CIA) triangle / triad

Measures undertaken to ensure confidentiality are designed to prevent sensitive information from reaching the wrong people, while making sure that authorized people can access it. It is common for data to be categorized according to the amount and type of damage that could be done should it fall into unintended hands. More or less stringent measures can then be implemented according to those categories. In many settings, confidentiality is roughly equivalent to privacy, though that it might in general apply to any other valuable, sensitive information.

Integrity involves maintaining the consistency, accuracy, and trustworthiness of data over its entire life cycle. Data must not be changed in transit, and steps must be taken to ensure that data cannot be altered by unauthorized people (for example, in a breach of confidentiality). These measures include file permissions and user access controls.

Availability means that the system and information must be available when it is needed. This means the computing systems used to store and process the information, the security controls used to protect it, and the communication channels used to access it, must be functioning correctly. Ensuring availability also involves preventing denial-of-service attacks.

One can consider different types of attackers (adversaries) against an IT system depending on their role toward the system:

- An **insider** attacker participates in the computing system and is seen as a legitimate user or other role. This enables the attacker to launch potentially very powerful attacks, depending also on the access rights the legitimate user has.
- An **outsider** attacker has access only to externally available parts of the system, and needs to gain more access before launching more powerful attacks. Still, outsiders might be able to attack the availability, or if they are able to observe or manipulate the communication of the system to other outsiders also the confidentiality and integrity

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Another classification of adversaries is based on the attacker behaviour:

- **Semi-honest** (or honest-but-curious) adversaries comply with the protocol, but try to gather more information than the protocol allows.
- **Malicious** adversaries arbitrarily deviate from the protocol to breach security with the goal to e.g. corrupt the process or obtain confidential data.

Attack scenarios can further be distinguished based on the attacker’s knowledge of the targeted system:

- In a **white-box** access scenario, the adversary has complete knowledge about the architecture, security controls etc. of the system to be attacked. An adversary can use his knowledge to find the vulnerable parts of the system, therefore implement more powerful attacks.
- In the **grey-box** setting, the attacker has access only to some information about the system.
- In the **black-box** scenario, the attacker only has information on the externally visible parts of the system.

4.3 FeatureCloud Platform

The FeatureCloud platform consists of a number of nodes (sites) running the machine learning algorithms on the locally available data, communicating with a number of other services that orchestrate the process (see Figure 8). The platform is controlled via a web application (frontend) involving user rights management. For a more detailed description of the system, the reader is referred to deliverable D2.2 KPIs and metrics for local executions platforms.

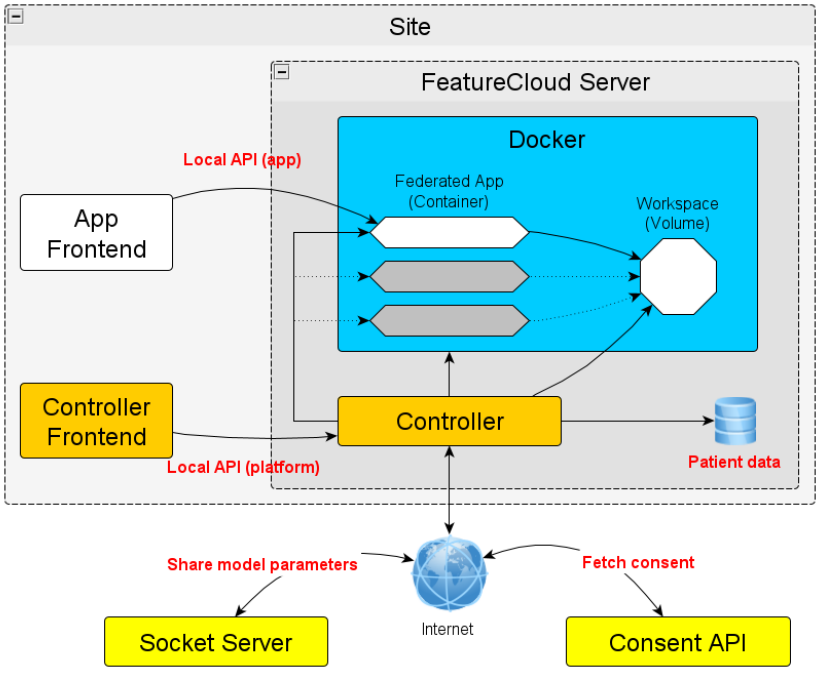

Figure 8: Local platform running in a hospital and its outside connections

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

5 Privacy Law Analysis

Due to its privacy-by-design-and-architecture approach, the platform to be developed already by design solves many issues that traditional centralized, potentially cloud-based analytical platforms processing sensitive information have. It resolves many data protection issues that arise, but not all of them. Therefore, in close interlinkage with the analysis on the distributed system and federated machine learning risks, a legal analysis is carried out. It takes a privacy perspective, and therefore solely data protection law is in the scope of the analysis.

Following the types of attackers defined in Section 4, insiders as well as outsiders are relevant for the legal analysis. It is one of the particularities of data protection law that it also regulates what the legitimate data owner is allowed to do with (personal) data.

5.1 The notion of personal data in data protection law

Art. 4 No. 1 GDPR[15] defines the subject matter of the GDPR, **personal data**, as “any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.”

The notion of personal data is the central anchor point of applicability of the GDPR. Whenever personal data is processed (within the material scope as laid down in Art. 2 GDPR and the territorial scope as laid down in Art. 3 GDPR, which can both be regarded as given for processing activities in the FeatureCloud Project), the GDPR is applicable. For a closer analysis the definition of personal data can be split into the following four elements, according to (Article 29 Data Protection Working Party 2007, pp 6 ff):

5.1.1 Information

The definition of personal data is very broad, as it covers "any information" relating to a natural person without any restrictions. (Article 29 Data Protection Working Party 2007, 6 ff). It can include both objective statements on circumstances or verifiable characteristics and subjective assessments and judgements about the person concerned. It is also irrelevant whether the information is true or not or whether it is only true with a certain statistical probability.

5.1.2 “Relating to”

Information can be considered to relate to an individual when it is about the individual, or also, when it is about an object which itself belongs to an individual or which “relates to” an individual in another way. Instances of information relating to an individual can be distinguished in a way that either a “content” element, a “purpose” element or a “result” element is present. (Article 29 Data Protection Working Party 2007, pp 10 ff)

Firstly, information relates to an individual in the sense of the term “content” element if it is about that individual. This is the most obvious and common case. Secondly, if an information is not necessarily about a particular individual but it is used or likely to be used with the purpose to evaluate or influence the status or behaviour of an individual or treat it in a certain way, then the “purpose” element is

[15] REGULATION (EU) 2016/ 679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

given. Thirdly, if the person is treated differently as a result of the evaluation or processing of the information, a “result” element is responsible for an information relating to a person.

5.1.3 Natural Person

The third element of the definition of personal data is that the person the information is relating to is a natural person, that is, a living human being. According to its recital 27 the GDPR does not apply to the personal data of deceased persons. However, data on deceased persons, especially in the medical domain, might also contain some information relating to living persons. One example in particular is regarding the relatives of the deceased person, as in the case of hereditary genetic dispositions for certain diseases. Such data falls within the scope of protection of the GDPR, not because it is relating to the deceased person, but because it is relating to another person that is still alive.

5.1.4 Identified or Identifiable

This is, in general and especially in the given context, the most crucial and most interesting of the four elements. A person may be directly identified or indirectly identifiable. Both are highly dependent on the context. (Article 29 Data Protection Working Party 2007, p 13). The most common identifier for directly identifying a person is the name. However, a very common name is in general not suitable for unambiguously identifying a person. In this case, a second piece of information such as the date of birth is required to do so. Or the name, although common, might be unique in the given context, such as within the group of participants in a particular study. On the other hand, as the Article 29 Data Protection Working Party already stated in 2007, the possibility of identifying an individual no longer necessarily means the ability to find out his or her name. (Article 29 Data Protection Working Party 2007, p 14). This is reflected in the second part of the definition of personal data cited above.

Indirect identifiability is given if the existing information about an individual is not sufficient to unambiguously identify him or her, but it is possible to identify him or her by linking the existing information with additional information, or by using additional criteria or means of identification, such as those listed in Art. 4 No. 1 GDPR. The area of indirect identifiability bears a multitude of research questions in data protection law with strong technological implications.

In this regard, recital 26 of the GDPR is of particular importance:

“The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.”

Firstly, it clarifies that **pseudonymised data are personal data**, as long as the additional information which makes it possible to attribute the data to a natural person is available, and that **anonymous data are not personal data**. Secondly, it makes another very important clarification: To determine whether a natural person is identifiable, not every theoretical possibility to identify the person must

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

be taken into account but only means reasonably likely to be used to do so. To ascertain whether **means are reasonably likely** to be used, all objective factors should be taken into account, such as the costs of and the amount of time required for identification, the available technology at the time of the processing, and technological developments (Esayas 2015).

From this, it can be concluded that in order to determine whether data is personal data under GDPR a practical, not a theoretical standpoint must be taken. Means reasonably likely to be used, are means that not only exist theoretically but that would be used practically.

In the context of this risk assessment methodology, this means that a practical assessment must be carried out: From the attack vectors on the anonymity of the data found in this document only those are legally relevant that are reasonably likely to be used by an actual attacker in practice. This must be assessed on the basis of objective factors such as the costs and the amount of time required, the required skills, the potential gain and the available technology but also possible technological developments in the future.

In order to assess whether an attack vector is relevant from a legal perspective, attacks that are reasonably unlikely can be ignored. An attack can be considered being reasonably unlikely if it cannot be imagined that it will happen in practice in the given context because the attacker will shy away from the effort.

5.2 Special Categories of Data

When the definition of personal data in data protection law is discussed in the context of medicine, it needs to be mentioned that the GDPR in Article 9 paragraph 1 defines special categories of personal data, also known as **sensitive data**. These data are subject to a stricter data processing regime than non-sensitive data. Health data fall within this category of data, as well as genetic data and biometric data for the purpose of uniquely identifying a natural person.

5.3 Data Protection Principles

Accountability pursuant to Art. 5 para. 2 GDPR obliges controllers to be able to demonstrate compliance with the data protection principles at any time. The data protection principles according to Art 5 para. 1 GDPR are:

- lawfulness, fairness and transparency
- purpose limitation
- data minimisation
- accuracy
- storage limitation
- integrity and confidentiality

5.3.1 Principle of lawfulness, fairness and transparency

According to Art. 5 para. 1 lit. a GDPR, personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject. The principle of the **lawfulness** of processing is a fundamental principle related to the rule of law (Heberlein in Ehmann and Selmayr, Art 5 margin no. 8).

With regard to FeatureCloud, this principle requires the existence of a legal basis for data processing, such as consent. In the context of the FeatureCloud platform, the principle of lawfulness must especially be considered when deciding which data is made available via the platform and in consent management.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

The principle of **fairness** is a criterion for the consideration of the protection purpose of the GDPR throughout the application of its provisions and prohibits an unlawful exercise of rights by the controller or the processor to the detriment of the data subject. The reasonable expectations of the data subject must be taken into account (Heberlein in Ehmann and Selmayr, Art 5 margin no. 10) .

In the context of the FeatureCloud platform, the principle of fairness must especially be considered in consent management: In order to comply with the principle of fairness, particular attention must be paid to the fact that the consent obtained from the data subject is freely given (Art. 4 No. 11 GDPR), i.e. that it meets the following requirements:

- The data subjects shall have a real choice, i.e. they must not feel pressured to give their consent or suffer negative consequences if they do not consent (Article 29 Data Protection Working Party 2017, p 6)
- Consent shall not be a non-negotiable part of the terms and conditions (Article 29 Data Protection Working Party 2011, p 12)
- There shall not be a clear imbalance between the data subjects and the controller, such as in the case of authorities or employers who act as controllers (Recital 43 of the GDPR, Article 29 Data Protection Working Party 2017, p 6 ff)
- The performance of a contract or provision of a service shall not be made conditional on consent to the processing of personal data that is not necessary for the performance of the contract (Art. 7 para. 4 GDPR, Article 29 Data Protection Working Party 2017, p 8 ff)
- Separate consent shall be obtained for each purpose (Recital 43 of the GDPR, Article 29 Data Protection Working Party 2017, p 10 ff)
- The refusal and withdrawal of consent shall be possible without adverse effects for the data subjects, i.e. without deception, intimidation, coercion or significant negative consequences (Recital 42 of the GDPR, Article 29 Data Protection Working Party 2017, p 12 ff)

It should also be noted in this context, that consent must be informed and most of the provisions of the GDPR are mandatory, which is why deviations from them - even with the consent of the persons concerned - are not permitted.

The principle of **transparency** is intended to ensure that the data subjects can understand how the personal data relating to them are processed. This principle is further elaborated in particular by the information obligations under Art. 13 and 14 GDPR and requires that all information and communications relating to the processing of this personal data are easily accessible and comprehensible and are written in clear and simple language (Recital 39 of the GDPR).

In the context of the FeatureCloud platform, the principle of transparency must especially be considered when obtaining consent from the data subject.

5.3.2 Principle of purpose limitation

According to Art. 5 para. 1 lit. b GDPR, personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes (further processing for, inter alia, scientific research purposes or statistical purposes shall, in accordance with Art. 89 para. 1 GDPR, not be considered to be incompatible with the initial purposes). Controllers are therefore required to specify the processing purpose in writing in advance. The requirement of legitimacy means that the purpose shall be assessed in relation to the entire body of law. The principle of purpose limitation prohibits the collection of data for abstract and general purposes or for purposes to be specified afterwards.

In the context of the FeatureCloud platform, the principle of purpose limitation must especially be considered when evaluating requests for using the FeatureCloud platform and the underlying personal data it makes available. It must be assured that these requests are strictly pursuing

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

scientific research purposes. Additionally, purpose limitation must especially be considered in case existing data is made available via the platform. However, this will usually be based on explicit consent and not on Art. 6 para. 4 GDPR. From the fact that the GDPR in Art. 5 as well as in other provisions explicitly acknowledges scientific research purposes it can be unambiguously concluded that these purposes can be regarded as legitimate in the sense of Art. 5 para. 1 lit. b GDPR.

5.3.3 Principle of data minimisation

The principle of data minimisation according to Art. 5 para. 1 lit. c GDPR establishes three requirements: personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

Data minimisation is also the fundamental principle for implementing the requirements of data protection by design and by default of Art. 25 GDPR.

In the context of the FeatureCloud platform, the principle of data minimisation is largely fulfilled by the architecture (privacy by design and architecture approach): It is the key feature of FeatureCloud that the data are kept in the institution where they reside and not copied to a central infrastructure in order to make them usable for research purposes. Further measures to comply with data minimisation like minimising the group of people who have access to the actual data by utilizing a full-blown authorization and access management are part of the standard security requirements for the platform. It must be noted that parts of these measures can only be fulfilled at runtime.

5.3.4 Principle of accuracy

According to Art. 5 para. 1 lit. d GDPR, personal data shall be accurate and, where necessary, kept up to date and every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay.

In the context of the FeatureCloud platform, the principle of accuracy goes hand in hand with FeatureClouds key requirement that personal data which is used as a basis for machine learning as well as the results of the learning processes must be as accurate as possible in order to achieve valuable research findings.

5.3.5 Principle of storage limitation

Personal data shall, according to Art. 5 para. 1 lit. e GDPR, be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed (personal data may be stored for longer periods insofar as the personal data will be processed solely for, inter alia, scientific research purposes or statistical purposes, in accordance with Art. 89 par. 1 subject to implementation of the appropriate technical and organisational measures required by the GDPR in order to safeguard the rights and freedoms of the data subject.

In the context of the FeatureCloud platform, the principle of storage limitation is largely fulfilled by the architecture (privacy-by-design and architecture approach): It is the key feature of FeatureCloud that the data are kept in the institution where they reside and not copied to a central infrastructure in order to make them usable for research purposes. If the data must be mirrored within the institution in order to make them available via the FeatureCloud platform (compare figure 8) the principle of storage limitation requires the deletion of that data as soon as the purpose for the mirroring is fulfilled or if consent is withdrawn. This has to be implemented as part of the consent management (WP6) respectively the platform (WP7).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

5.3.6 Principle of integrity and confidentiality

According to Art. 5 para. 1 lit. f GDPR, personal data shall be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures.

This requirement goes hand in hand with the general security requirements of FeatureCloud and is therefore fulfilled by all security measures in FeatureCloud including in particular the present risk assessment and the mitigation measures that will be developed in D2.4. See in particular chapter 8.3 and 8.4 of this report.

5.4 Technical and Organisational Data Protection Measures

In the context of risk assessment, another important aspect of data protection law needs to be discussed. The GDPR obliges data controllers to implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with the GDPR (Art. 24 and 25 GDPR), and to ensure a level of security appropriate to the risk (Art. 32 GDPR).

To assess which measures are appropriate, the controller has to take into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing (Art. 24, 25 and 32 GDPR). In particular, this means that the **cost of a particular measure** (not only in monetary terms but also in terms of reduced quality of the outcome) can be **weighed against its effectiveness** (in terms of protecting privacy and security of the data) and a balance needs to be struck.

5.5 Conclusion

FeatureCloud claims that due to the privacy-by-design-and-architecture approach no personal data are leaving data holders’ premises or, in other words, all data leaving data holders’ premises are anonymous. Consequently, from a legal point of view, it must be proven that the information leaving data holders’ premises are not personal data. Therefore, the definition of personal data in the GDPR was closely examined and applied to the context of this risk assessment methodology.

To sum up the results of the data protection law analysis, whereas from an information security perspective every theoretical possibility to attack (the anonymity) is relevant, even if it involves disproportionate effort, from a data protection law perspective only attacks (on the anonymity) that are reasonably likely to happen in practice are relevant. Although this still needs to be assessed in every individual instance, it narrows it down to a non-legal question. The analysis which attacks are theoretically possible and the assessment of their likelihood in practice require the same skill set and can therefore be made by the same people who do not need any legal skills.

In addition, the fundamental requirements of the GDPR have been mapped out in order to take them into account in every relevant further step of the FeatureCloud project.

6 Implications of the NIS Directive

The NIS Directive (Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union) was adopted in July 2016. Since it is a Directive it is not directly applicable, but

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

requires an act of transposition into the national legislation of EU Member States. This national legislation had to be enacted by each Member State by 9 May 2018.

The scope of the NIS Directive is limited to certain types of organisations: “Operators of essential services” are public or private entities providing services in the field of energy (electricity, oil, gas), transport (air, rail, water and road transport), banking and financial services, health, drinking water supply or digital infrastructure that are essential to the maintenance of critical social/economic activities and where a security incident would cause significant disruption to the provision of such a service. “Digital service providers” are providers of either an online marketplace, an online search engine or a cloud computing service.

6.1 Operators of Essential Services

Under the regime of the NIS Directive operators of essential services are obliged by the implementing acts in the Member States to take risk-based, appropriate and proportionate technical and organisational measures based on the state of the art to manage security risks of the network and information systems they use to provide their services. In addition, operators of essential services must regularly conduct security audits in order to provide evidence of the effective implementation of security policies and are obliged to report incidents.

6.2 Digital Service Providers

Digital service providers are also obliged by the regime of the NIS Directive to report incidents and to take risk-based, appropriate and proportionate technical and organisational measures based on the state of the art, to manage security risks of the network and information systems they use to provide their services. For digital services providers, further details are specified in an implementing act (Commission Implementing Regulation (EU) 2018/151 of 30 January 2018 laying down rules for application of Directive (EU) 2016/1148 of the European Parliament and of the Council as regards further specification of the elements to be taken into account by digital service providers for managing the risks posed to the security of network and information systems and of the parameters for determining whether an incident has a substantial impact).

6.3 Applicability of the NIS Regime

In the context of FeatureCloud, applicability of the NIS Regime can be relevant either with regard to providers of health services participating in the FeatureCloud network as operators of essential services or because the FeatureCloud platform could be regarded as a cloud computing service.

6.3.1 Operators of Essential Services in the context of FeatureCloud

As mentioned above, providers of health services can be classified as operators of essential services under the NIS directive. Whether a specific provider of health services is an operator of essential services and hence the NIS regime applies to this provider depends on a classification on Member State level. In Austria for example, criteria such as, in particular, the number of hospital beds a hospital provides per 1,000 inhabitants in its health care region are laid down in an ordinance by the Federal Minister for the EU, Arts, Culture and Media. In application of these criteria the Federal Chancellor decides by issuing an administrative decision whether a specific provider of health services is an operator of essential services (Sec. 16 Austrian NIS Act). Therefore, a hospital in Austria knows whether it is classified as an operator of essential services because in this case it individually has received such an administrative decision by the Federal Chancellor.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

6.3.2 Cloud Computing Services in the context of FeatureCloud

A cloud computing service is defined in Art. 4 No. 19 NIS Directive as “a digital service that enables access to a scalable and elastic pool of shareable computing resources”. This is further elaborated in recital 17 of the NIS Directive:

“For the purposes of this Directive, the term ‘cloud computing services’ covers services that allow access to a scalable and elastic pool of shareable computing resources. Those computing resources include resources such as networks, servers or other infrastructure, storage, applications and services. The term ‘scalable’ refers to computing resources that are flexibly allocated by the cloud service provider, irrespective of the geographical location of the resources, in order to handle fluctuations in demand. The term ‘elastic pool’ is used to describe those computing resources that are provisioned and released according to demand in order to rapidly increase and decrease resources available depending on workload. The term ‘shareable’ is used to describe those computing resources that are provided to multiple users who share a common access to the service, but where the processing is carried out separately for each user, although the service is provided from the same electronic equipment.”

While the FeatureCloud platform is clearly not a pool of shareable general purpose computing resources, but a platform for accessing data in a very specific way and for a very specific purpose, it cannot be fully ruled out whether this definition applies to the FeatureCloud platform at this point. The aim of establishing the FeatureCloud platform is not to provide computing resources but to make available data for research via a common platform which are spread at different institutions. Clearly, the aim is therefore to provide something new which is practically not available at all otherwise and the aim is not to replace some existing demand for (local) computing resources with something more scalable and elastic “in the cloud”. In particular, the aim of the FeatureCloud platform is not to provide “computing resources that are flexibly allocated by the cloud service provider, irrespective of the geographical location”. The fact that the FeatureCloud platform provides computing resources is a side effect of the way the access to the data is provided, i.e. that the data stays where it resides and therefore computation needs to be carried out there. The fact that these resources are scalable is a basic feature that follows from the requirement that the FeatureCloud platform should be usable by different institutions at the same time. However, this cannot be described as an “elastic pool” of “computing resources that are provisioned and released according to demand in order to rapidly increase and decrease resources available depending on workload”.

For these reasons, we tend to the conclusion that the FeatureCloud platform is not a cloud computing service in the sense of Art. 4 No. 19 NIS Directive and hence the NIS regime would not apply to the FeatureCloud platform. However, we cannot fully rule out the opposite due to the wide and partly unclear nature of this term and the lack of a ruling by the ECJ on its interpretation. Whether the NIS regime applies to the FeatureCloud platform can therefore only be decided at a later point in time, taking into account the details of its implementation and deployment, when they are determined. Until then we will closely follow the interpretation of the term cloud computing service, possible court rulings and the requirements that would apply (see below).

6.4 Requirements by the NIS Regime

6.4.1 Requirements for Operators of Essential Services in the context of FeatureCloud

From the preceding section it can be concluded that several hospitals which will participate in the FeatureCloud network and provide data to the FeatureCloud platform will be operators of essential services and others might not be. To determine the specific implications of the applicability of the NIS regime and to meet its requirements lies within the responsibility of each individual hospital and therefore not within the responsibility of the FeatureCloud consortium.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

For the participation in the FeatureCloud network it is expected that additional information systems need to be deployed in each participating hospital, as shown in figure 4. A hospital underlying the NIS regime has to take into account and manage potential risks posed by these additional information systems (Art. 14 para. 1 NIS Directive). The same holds true for any other hospital, where the necessity for dealing with these risks is a consequence of general security considerations and other applicable laws (in particular on Member State level). The FeatureCloud architecture and platform and in particular the interfaces to the hospital’s existing information systems are designed in a way to minimise these risks, as laid down in this report and as will be further determined in Task 2.4 and 2.5 and WP7. Since the additional information systems deployed in the hospitals for the participation in the FeatureCloud network are not used for the provision of essential health services but for research purposes, the specific NIS duty to take appropriate measures to prevent and minimise the impact of incidents affecting the security of the network and information systems does not apply to these additional information systems (Art. 14 para. 2 NIS Directive). Nevertheless, irrespective of that duty, it is a core aim of the FeatureCloud project as reflected in the DoA, in Work Package 2 and beyond, to take appropriate measures to prevent and minimise the impact of incidents affecting the security of these systems.

As far as hospitals, who want to participate in the FeatureCloud network, are subject to the requirements of the NIS regime this could help manage the process of initiating their participation because of the security and security management processes, including processes to deal with data breaches, they are required to have implemented. However, as mentioned before, any other hospital is also subject to such requirements due to general security considerations and other applicable laws (in particular on Member State level).

6.4.2 Requirements for Cloud Computing Services in the context of FeatureCloud

As mentioned above, providers of cloud computing services, according to Art. 16 NIS Directive, are required to identify and take appropriate and proportionate technical and organisational measures to manage the risks posed to the security of network and information systems which they use in the context of offering their services. Having regard to the state of the art, those measures shall ensure a level of security of network and information systems appropriate to the risk posed, and shall take into account the following elements:

- the security of systems and facilities;
- incident handling;
- business continuity management;
- monitoring, auditing and testing;
- compliance with international standards

These elements are further specified in Commission Implementing Regulation (EU) 2018/151. In addition, providers of cloud computing services are required to notify the competent authority or the CSIRT without undue delay of any incident having a substantial impact on the provision of a service.

While most of these points only apply at the stage of operation of a service, the elements of security of systems and facilities, and compliance with international standards already apply at the stage of the design of a system. This is why these requirements are taken into account within this report. The latter on international standards does not refer to any specific standard and compliance with applicable international standards is a requirement in FeatureCloud anyway, which is reflected in this report and in WP3 in particular.

The element of security of systems and facilities is further specified in Art. 2 para. 1 of the Commission Implementing Regulation (EU) 2018/151 as follows:

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

(a) the systematic management of network and information systems, which means a mapping of information systems and the establishment of a set of appropriate policies on managing information security, including risk analysis, human resources, security of operations, security architecture, secure data and system life cycle management and where applicable, encryption and its management;

(b) physical and environmental security, which means the availability of a set of measures to protect the security of digital service providers' network and information systems from damage using an all-hazards risk-based approach, addressing for instance system failure, human error, malicious action or natural phenomena;

(c) the security of supplies, which means the establishment and maintenance of appropriate policies in order to ensure the accessibility and where applicable the traceability of critical supplies used in the provision of the services;

(d) the access controls to network and information systems, which means the availability of a set of measures to ensure that the physical and logical access to network and information systems, including administrative security of network and information systems, is authorised and restricted based on business and security requirements.

6.4.3 Conclusion

To conclude, the security requirements which follow from the present report are much more detailed than the requirements posed by the NIS regime, as far as the stage of the design of a system is concerned and not the stage of operation only, which is not within the scope of this analysis. If we had not already decided in the DoA to carry out a risk assessment like the one present in this report, the NIS regime – if applicable (see above) – would have required such a risk assessment.

The other fundamental requirements of the NIS have been mapped out above in order to take them into account, as far as necessary, in every relevant further step of the FeatureCloud project but mostly they go hand in hand with existing requirements following from the DoA and from the present report anyway.

7 Local System and Distributed System Risks

In this section, we consider risks that arise on a local system participating in the FeatureCloud platform, and risks originating from the nature of distributed systems - as the FeatureCloud system is a form of distributed system, these are also relevant for the project and the platform developed therein. Because on these two levels, these are rather well-known attacks and not specific to *federated learning*, this section is building on state-of-the-art methodologies and frameworks to identify and mitigate risks. The specific instances of these attacks to mitigate against depend on the final deployment of the system, the participating endpoints (e.g. hospitals with their specific setup and configuration), and the overall technologies utilised. These might change for each specific setup. This section shall guide the design of the secure platform, as also identified to be an important aspect in Section 6.4.2., and help as an aid for security aspects during implementation.
During the design phase, the Open Web Application Security Project (OWASP) Application Security Verification Standard (ASVS)[16] provides guidelines for technical security controls and also provides a list of requirements for secure development.

[16] https://owasp.org/www-project-application-security-verification-standard/

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Further practical guides on common threats are provided e.g. by the Open Web Application Security Project (OWASP), including the OWASP Top Ten security threats[17], or by the ENISA threat analysis report[18]. Also, the OWASP proactive controls[19] shall be considered during development, to address specific types of common vulnerabilities. The Center for Internet Security (CIS) offers a list of best-practice controls[20], prioritized into CIS Implementation Groups (IGs), depending on the effort an organisation can invest into security. For an overview on research on threats, (Markatos, Balzarotti, and Minchev 2013) give an overview of research problems. The above mentioned source moreover provides a good overview on risks that are inherent to and targeted at each local node, e.g. attacks from within each local node’s network, or attacks by insiders within each local node.

On the local system level, threats can be caused by weak organisational and technical policies, e.g. by the protection of physical access to computing resources, to the authentication methods and policies (e.g. password strength and rotation policies, multi-factor authentication, …), security patching practices, or enforcing of encryption of storage devices.
A specific list of potential threats and vulnerabilities, in particular operating systems and applications (such as web servers) is provided by the Center for Internet Security (CIS) benchmarks[21]. The specific CIS benchmarks provide recommendations for settings that should be considered to harden the security setup of the applications or operating systems. Further, also deliverable D2.2 “KPIs and metrics for local execution platforms” outlines some measurable indicators. Also more organisational aspects, e.g. policies regarding “Bring Your Own Device (BYOD)”, can have an impact on the emergence of security risks; social engineering attacks need to be considered as well.

It has been outlined in Section 6.4.1. that the applicability of either the NIS directive or other national data protection legislation, will require health care providers, which are eventually the data providers in the federated learning platform of FeatureCloud, to ensure a high level of security and data protection on their computing infrastructure, and thus also on the infrastructure that is specific to FeatureCloud, and running in the health care providers infrastructure.

It is to be noted that especially for the parties holding the sensitive data and participating in the learning, these policies may be enforced by and thus be specific to the organisation hosting the computing systems, e.g. at a certain hospital (influenced by national legislation, and general policies adopted). In these cases. FeatureCloud can provide additional guidelines and recommendations for these systems by providing updated frameworks and guidelines on threats and lists of the most common thread and vulnerabilities, such as the above-mentioned ones, but generally has limited influence in changing the policies already in place. In most cases it is to be expected that any additional deployment of locally required services to enable the federated learning will have to adhere to these policies. For services that FeatureCloud deploys in addition on the local level, the above mentioned frameworks shall however be applied to minimise the risks.

The **impact** of attacks on the local system (being a local or remote attack) is further worth a detailed analysis. It is generally to be expected that an attack to a single system in the FeatureCloud federation will result in a more local impact, than a breach of a centralised system that aggregates all data - in the federated setting, gaining access to a single system in first place only provides access to the system’s local data, and does cause less damage. Still, FeatureCloud as operator of a service needs to especially ensure that the additional services required to operate the federated platform are not creating an increased attack surface or novel attack vectors to gain such access.

[17] https://owasp.org/www-project-top-ten/#
[18] ENISA Threat Landscape Report 2018 15 Top Cyberthreats and Trends, https://www.enisa.europa.eu/publications/enisa-threat-landscape-report-2018

[19] https://owasp.org/www-project-proactive-controls/
[20] https://www.cisecurity.org/controls/
[21] https://www.cisecurity.org/cis-benchmarks/

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Attacks to a local system can be the starting point to attack the training process in the overall FeatureCloud platform as well, e.g. for a data poisoning data; thus, it is important to consider these attacks as a potential starting point for various types of attacks that may be targeting the machine learning process and the assets used and created therein (e.g. the models), enabling a form of insider attacks on the machine learning processes (cf. Section 8).

Regarding the distributed nature of the federated system, this setting opens up a set of attack vectors that have been extensively studied in the context of information security. Often it is distinguished between, on the one hand, decentralised (point-to-point) interactions across distributed entities without a centralised coordination service, e.g. Peer to Peer (P2P) systems. The sequential (cyclic incremental) type of federated learning can be of this distributed nature, but also parallel settings without a central coordinator will fall into this category (see Section 8.1. Types of Federated Learning for a detailed description of these modes of federated learning). On the other hand, coordinated distributed systems generally involve a central node, coordinating (a) the use of resources, or (b) services.

We outline a couple of different attacks that are frequently occurring in distributed (or generally, networked) systems as examples, and we will refer to these later on when analysing risks in the federated machine learning.

- **Eavesdropping** is an attack that tries to listen to private communication of other parties, without their consent. It can be achieved by several means. A man-in-the-middle attack is a setting where an attacker secretly relays the communication between the other parties (Conti, Dragoni, and Lesyk 2016). From the attackers point of view, eavesdropping should normally not cause disruptions on the normal operation of the systems being eavesdropped on, as that can cause that conversation (and thus the attack) to cease.
- **Masquerading** is a type of attack where the attacker pretends to be an authorized user of a system to gain access to it. It can use stolen passwords and logins to gain unauthorized access through a legitimate access identification. Masquerading can further be performed by locating gaps in programs, or by finding a way around the authentication process. If an authorization process is not fully protected, it can become vulnerable to this type of attack. The attack can be triggered either by someone within the organization or by an outsider if the organization is connected to a public network. An insufficient authentication mechanism could e.g. rely on a unique IP address being assigned to a resource that an adversary is spoofing, essentially convincing the target system to be the authorised resource. The amount of access masquerade attackers get depends on the level of authorization they have managed to attain.
- **Denial of service** attacks (DoS) aim to reduce the availability of a system or other network resource are designed to make a machine or network resource unavailable to its intended users (Hansman and Hunt 2005). Different targets can be distinguished. E.g. an individual user might be addressed, by deliberately entering a wrong password repeatedly to cause the victims account to be locked. Further, whole systems might be the target of the attack, trying to overload the capabilities of a machine or network to answer requests and thus to block all users at once. Attacks from a single source can relatively easily be identified and defended against, by e.g. blocking that source. Especially powerful are distributed denial of service attacks, where the attack comes from a larger number of attackers, and it is thus more difficult to handle all attackers. Such an attack is often performed using botnets, or by attacks that fool innocent systems into sending traffic to the target.

One specific threat in the FeatureCloud approach is that of an attacker that manages to have a **malicious application** executed at the remote sites. Such an application could in the most powerful form simply transmit the data to an outside destination, in the form of a covert channel (Zander, Armitage, and Branch 2007). In the most convenient form for the attacker this could be a direct

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

communication to the external destination. A more disguised form could be hiding the information to be transmitted along with the lawful communication about the model updates, and being able to extract that information afterwards. This can be seen as a form of **steganography**, which is a form of **information hiding** that conceals the existence of the secret data hidden in a cover medium (the model updates). The machine learning model itself can be the vehicle to hide information, as will be detailed in the data exfiltration attacks described in Section 8 Machine Learning Risks.

Another aspect to consider will be in the user rights management and consent provisioning of the users, which will be developed in Work Package 6 “Blockchains and user right management”. On the one hand, this system itself might be a target of attack to circumvent the user rights and consent giving, i.e. by granting more access to user data than specified by the users. On the other hand, the confidentiality of the users managing their rights with this system must be protected, as an attacker (even in the semi-honest (honest-but-curious) form) might be able to infer specific data and information about the individual users. This can include information about participation in studies which might reveal information, or this can be information for a doctor that a patient being treated by her has withdrawn consent from her study.

8 Machine Learning Risks

Beside the advantages machine learning offers, several attacks that try to manipulate or otherwise disturb the training, or exploit a trained model, have recently been described. While federated learning offers advantages in reducing the amount of information that needs to be exchanged, distributing the training process to potentially insecure or malicious clients creates new entry points for attackers. The earlier introduced CIA triad (see Figure 7) can also be used to categorize different attacks on a machine learning process.

Attacks threatening the integrity and availability of machine learning models can mostly be described as security risks. Attacks on the confidentiality of a model are, eventually, mostly (but not exclusively) attacks on the privacy of the underlying training data or other (meta-)data associated with the model training (such as hyper-parameters), or on the users of the model. We describe these two types of attacks in detail below. For each type of attack, we relate it to the fitting concepts of the above described frameworks STRIDE, LINDDUN and CVSS; for the later, we suggest fitting base metric values, and where reasonable also provide example values of vulnerability scores, obtained with the CVSS calculator[22]. First, we start with a description of different types of federated learning, as these can impact the available attack vectors.

8.1 Types of Federated Learning

The main goal of machine learning applications is to train a model to make predictions based on sample data. In many real-world problems, this data is **distributed** among different sources, e.g. in health care where several health care providers might hold complementary information about a patient. Traditionally, aggregation of data in one place is needed for subsequent machine learning processing. Federated learning allows **training machine learning models on data distributed** over different nodes avoiding communication of this data which may contain private information. We denote these data holders as nodes. In federated learning, each node independently and locally learns its own machine learning model on its data, and then **shares only the model** to the global aggregator, or to other participants of the federated learning process. While this greatly reduces the associated privacy risks, there are still potential attack vectors left. In principle, federated learning can be implemented in a sequential or parallel manner, each of which offers slightly different attack vectors:

[22] https://www.first.org/cvss/calculator/3.0

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

- **Sequential** federated learning (see Figure 9), also referred as *cyclic incremental learning* (Sheller et al. 2018), allows training a global model sequentially at each node in the setting. The benefit of this approach is that no aggregator is needed in this setting. Chang et al. 2018 showed that this approach allows machine learning models to have performance close to the models trained on centralized data. However, the approach is not efficient with a rather large amount of nodes. A randomly initialized model is sent to one of the learning participants for local training. After the training, this node transfers the updated model to the next party in the sequence. This process may take several cycles, similar to batch training in neural networks. Regarding risks, if there are malicious nodes in a federated learning setting, they can corrupt the training process or threaten the privacy of other nodes.

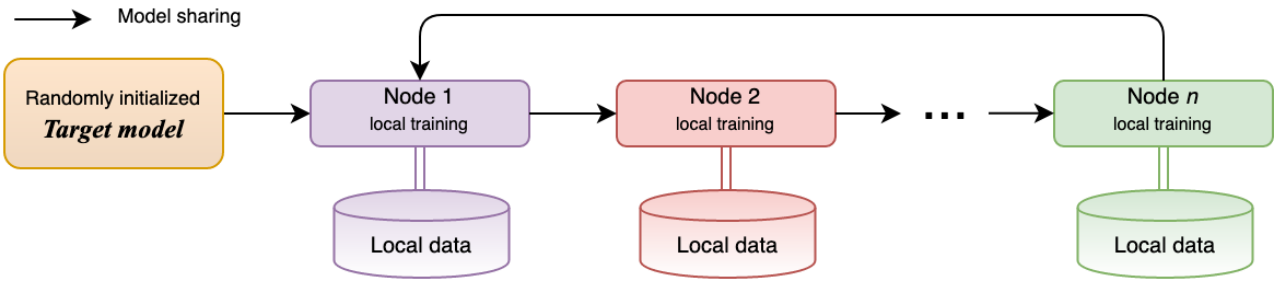

Figure 9: Sequential federated learning. A randomly initialized model is locally trained at the first node and then passed to the next node in the sequence. After completing a full cycle of n nodes, the model is passed again to the first node for repeating the training process.

- **Parallel** federated learning (see Figure 10) has an aggregator that collects local models from different nodes and aggregates them to a global model. The majority of the research works consider this approach while referring to federated learning. Aggregation of the models can be implemented in different ways, e.g. the averaging of models' weights in the case of neural networks, or averaging of the gradients (FederatedAveraging (FedAvg), by McMahan et al. 2017). This approach allows training several models in parallel, therefore it is more efficient than sequential federated learning with a large number of nodes. However, if an aggregator is compromised all the nodes of the setting are under a threat.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

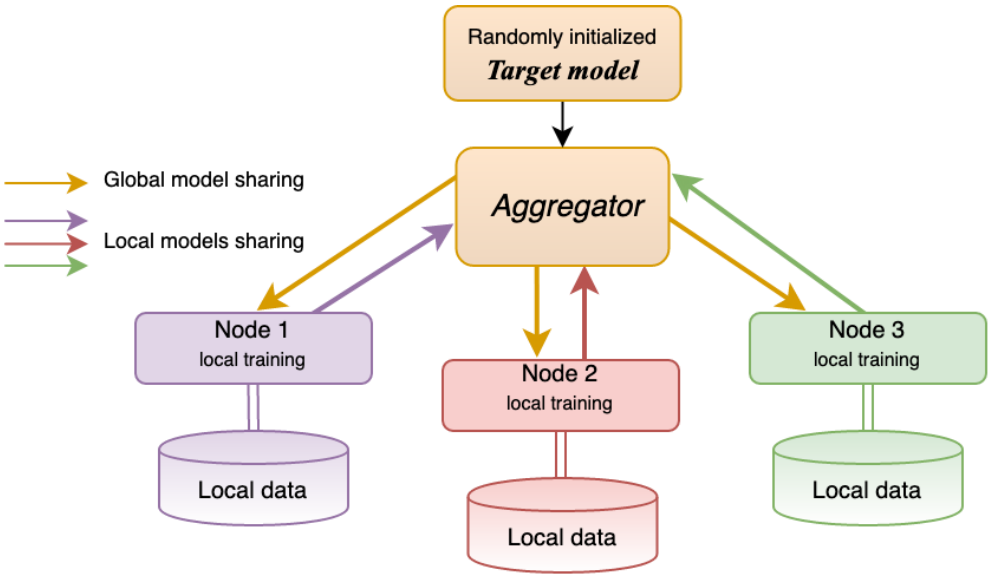

Figure 10: Parallel federated learning. The aggregator initializes a global model with random weights and shares it to every node in the setting. Each node trains the model in parallel on its local data and then returns it to the aggregator. From the locally trained models, a new global model is aggregated and shared to the clients for the following training cycle.

Federated learning implementation includes several challenges, such as communication costs, systems and data heterogeneity, unbalanced data, privacy and security risks. As machine learning models are the only data that is communicated during the federated learning process, they become the main surface for attacks. There are several approaches on how an adversary can perform an attack on a machine learning model, e.g. by manipulating the training data, corrupting the model or distorting the output (see examples of particular attacks in Section 7.3 Security Risks). Non independent and identically distributed (non-IID) data, which is another issue in federated learning, can cause additional security and privacy vulnerabilities, as attacks might be more difficult to detect.

8.2 Threat Model in Federated Machine Learning

Similar to the generic attacker model, one can consider different types of attackers in a federated learning setting depending on their role in the training process:

- An **insider** attacker participates in the federated learning process and has **access to the models** during training. We distinguish between the following two types:
  - A participant insider attacker represented by one or several nodes, the owners of the data, who train models locally (e.g. described by Bagdasaryan et al. 2018).
  - A coordinator playing an aggregator role in parallel federated learning and that collects locally trained models from data owners (e.g. considered by Wang et al. 2019).
- An **outsider** attacker has access only to the final model after the federated learning process is finished. Outsider attacks can be seen roughly equivalent to privacy and security attacks on machine learning models trained in a centralized manner, as they have no additional information on the different model updates stemming from individual nodes.

Based on the attacker goal in federated learning, we distinguish between two typically considered adversary models, based on the general adversary models defined in Section 4:

- **Semi-honest** (or honest-but-curious) adversaries perform a “passive” attack, following the protocol, but trying to gather more information than the protocol allows, e.g. information about

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

the training data or de-anonymized federated learning participants. **Malicious** adversaries perform “active” attacks and arbitrarily deviate from the protocol, e.g. with the goal to corrupt the learning process or also to infer confidential information.

Regarding the attacker’s knowledge of the targeted system, we can observe the following specialisation of the generic IT security setting (e.g. considered by Shokri et al. 2017, Truex et al. 2019):

- In the **white-box setting**, the adversary has full access to the model, and **knowledge about its architecture**, model weights and hyperparameters.
- In a **grey-box** access, the attacker has access only to some information about the model, e.g. information on specific layers of the model, or some intermediate results.
- In the **black-box** scenario, the attacker can **use the model for making predictions** (via some API / service), but there is no access to any other information about the model. In this case, the attacker can use the machine learning model only as a service to query output for some specific input, to infer some valuable information for implementing attacks. Black-box attacks are limited by the amount of information that can be extracted from the output, and the type and amount of queries the attacker can perform before the service will stop answering.

Figures 11 and 12 represent two types of the possible attack scenarios in sequential and parallel federated learning. In Figure 11, an attacker is aiming to intrude a federated learning process in order to gain access to machine learning models and use them to break the integrity of machine learning, infer some private or confidential information about training data, model architecture or parameters. In this case, the attacker is an outsider (i.e. not a node directly participating in the learning), and performs e.g. an eavesdropping attack to observe the exchanged information.

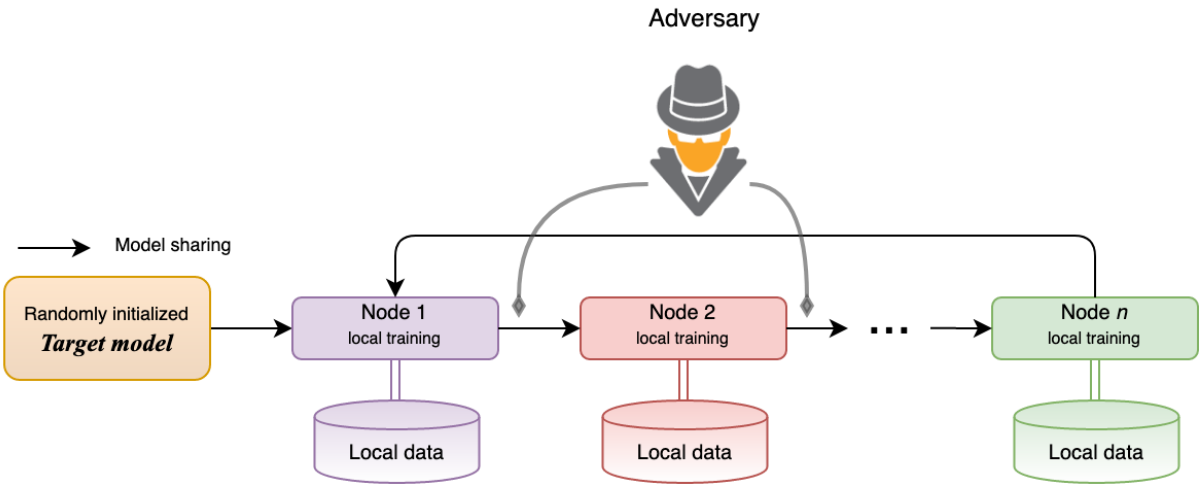

Figure 11: Sample attack scenario in sequential federated learning, in the form of an eavesdropping attack, e.g. with the goal to breach the confidentiality

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

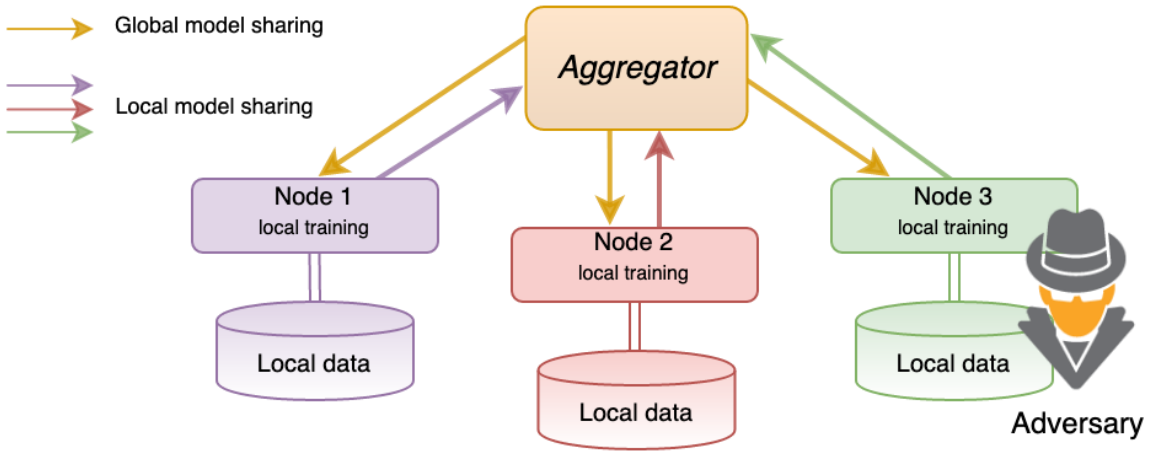

Figure 12: Sample attack scenario in parallel federated learning with one of the nodes being malicious (or being impersonated), with the goal of e.g. performing an attack on the integrity

Training models with federated learning may require several cycles of training for higher effectiveness, when data holders have to send and receive models from the aggregator several times. Therefore, the attacks can be performed in the training phase in the form of inference of sensitive information from the intermediate results (intermediate global models), or intrusion of the training process with malicious goals e.g. by poisoning training data. The attack during the training is e.g. the attack performed by an insider, such as a malicious node or aggregator.

Nodes (and especially aggregators) may also have access to the final global model for local usage, depending on the federated learning setting and the contract underlying the federated learning process. In this case, malicious nodes can use inference attacks on the final model. Some attacks require only access to the model for inference of the sensitive information (e.g. model inversion), other attacks can require additionally some knowledge about the training data (e.g. membership inference requires the knowledge of potential dataset member).

Attacks on the machine learning process or models, subsumed as adversarial machine learning, is a relatively new field, which started to evolve due to the recent increase of usage of machine learning approaches in different industries. To evaluate privacy and security risks in the machine learning setting, we researched existing attacks on machine learning models and mitigation strategies which can be applied to reduce privacy and security risks in federated learning.

8.3 Threats to availability and integrity

Security risks of federated learning include the attacks with goals such as a reduction in the model’s confidence, or causing a misclassification, which can be further distinguished into targeted or universal misclassification. Using the STRIDE framework, these attacks related to denial of a service, data tampering or repudiation threats. An example is an attack on a facial recognition system when an attacker tries to cause false-positive outputs to undermine authentication integrity. One can distinguish the following types of security attacks on machine learning models:

1. **Poisoning** attacks include scenarios when adversaries are trying to **manipulate training data** (e.g. editing, inserting or removing data instances) to change the model’s behaviour, e.g. dropping the model’s accuracy or achieving a particular misclassification (Gu et al. 2019; Bagdasaryan et al. 2018). One type of poisoning attacks are backdoor attacks, where a certain pattern is embedded in the final learned model to trigger specific behaviour. This type

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

of attack is executed during the model training, and can be performed by a malicious participant of the federated learning process. Neural networks are especially vulnerable to this type of attacks, as it is harder to interpret these types of models, and due to their behaviour to overfit, they are further more likely to learn the backdoor pattern. In terms of STRIDE categorisation, poisoning attacks can be related to the data tempering category and also can cause a denial of a service, when the machine learning model, for instance, gets corrupted and gives (primarily) false predictions. Using CVSS base metrics, poisoning attacks can be rather impactful with a score around 8, as the attack vector can be only adjacent and the scope of security can be changed.

2. **Evasion** attacks include scenarios when an attacker feeds the network with adversarial input to reach the goal of e.g. (targeted) misclassification or confidence reduction. Applying certain perturbation to the input can cause the network to misclassify (Szegedy et al. 2014; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Eykholt et al. 2018; Papernot et al. 2016). For instance, in the case of an image classification task, these perturbations may include adding some specific pixels to sample images, which are **not visible to the human eye**. This attack is generally carried out during the prediction phase of the machine learning process, and can be executed even with only black-box access to the model. With grey-box and white-box access to the model, the attack can become more powerful. Evasion attacks can be related to tampering and repudiation threats in STRIDE, and within CVSS base metrics is characterized by base metric values like low required privilege and low attack complexity.

8.3.1 Threats to availability and integrity in Federated Learning

Several recent works show that federated learning has specific vulnerabilities compared to centralized machine learning. Such aspects as distributed learning and heterogeneous data provide a new surface for attacks, and make defense even more difficult compared to the centralised case. (Xie et al. 2019) developed an attack named **distributed backdoor attack**, which is specifically aimed to use the distributed nature of federated learning to perform more powerful attacks. (Bhagoji et al. 2019) perform **model poisoning** attacks **in federated learning**. They show that even when an adversary can manipulate only one participant of the training, the attack still can be successful. This implies that for the FeatureCloud platform, a strong security of all local participants is a key aspect to ensuring overall security of the federated learning process.

8.4 Threats to confidentiality

The data of all clients participating in the federated learning process remains with the data owners. Only locally trained models are shared with an aggregator (in parallel learning, without a decentralised aggregation in place) or other clients (in sequential learning). Therefore, privacy risks in federated learning are mostly connected to models leaking information about their training data. Machine learning models capture structures in data and they can be used by adversaries to infer **statistical properties** of this data, perform **membership or attribute inference attacks**, and in the most powerful attack, a **model inversion** attack that aims at creating an approximation of the original training data. Using the LINDDUN categorization model, these attacks can be related to threats like disclosure of information, detectability, or identifiability. This information, which can be inferred from the models, can potentially have a high market value, or be particular information about specific individuals, thus potentially relevant in terms of data protection regulations, as described above.

There are several attacks that have recently been described, threatening the confidentiality of the models or the clients’ data, or the clients participating in a federated learning process themselves:

1. The **membership inference** attack (Shokri et al. 2017) refers to the scenario when an adversary has a sample record of a form of training set data, and a "black-box" access to the

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

model. The attacker can then infer if this record was in the training set of the model, or not, which can reveal certain meta-data about the individual, e.g. if the training set was on a study about a certain disease the individual would not like to reveal being infected with. The attack is challenging to mitigate (Song, Ristenpart, and Shmatikov 2017; Truex et al. 2019). Membership inference can be categorised (using LINDDUN notation) as detectability threat or disclosure of information threat. According to CVSS base metric, the attack vector can be Network or Adjacent, without any privileges required and result in a sample score of 7.5.

2. In the **model inversion** attack (Fredrikson, Jha, and Ristenpart 2015), an adversary tries to recreate data samples that represent the underlying original objects. This has been shown to work in very specific settings, such as in the case of recreating pictures of the people to be identified by a facial recognition system. It is more difficult to achieve in other settings, where an individual does not correspond to one of the classes distinguished by the machine learning system. Model inversion is related to the identifiability threat in LINDDUN. Model inversion would have similar CVSS basic metric values as membership and property inference attacks.

3. **Property inference** attacks can be performed on machine learning models to infer information about training data. Specifically, global properties of the training data can be inferred from the model (Ganju et al. 2018). In (Ateniese et al. 2015), the authors showed how to infer statistical properties of the training data, by comparing the difference of the model before and after training on this data. Machine learning models also can leak users' private information when the adversary has access to their public data. (Weinsberg et al. 2012) showed how to infer the gender of a user from a recommendation system, based on ratings which the user has given. Using LINDDUN categorisation the attack can cause threats like detectability, disclosure of information or identifiability.

4. **Model stealing** refers to the attack scenarios when a machine learning model has a high commercial value. Training of such machine learning models can take much resources, such as computing resources, money and human effort e.g. for generating the training data and tuning hyper-parameters, as well as evaluation. Further, these tasks might be time-consuming. Thus, an attacker might be interested in creating a copy of a model that is available as a service. It was shown in several works (Orekondy, Schiele, and Fritz 2019; Tramèr et al. 2016), that having a black-box access to the model (e.g. using a model as a service), an attacker may steal functionality of this model. This directly affects the **confidentiality of the model**. CVSS metric values would be also similar to the previously described attacks as the black box access to the model is also the minimal requirement for the attack.

5. **Data exfiltration** via machine learning (ML) models refers to embedding information in the models or model updates, as described e.g. by (Song et al, 2017) It is vital that ML models trained on sensitive inputs (e.g., personal images or documents) not leak (too much) information about the training data. An operator of a machine learning model who supplies model-training code to the data holder, does not observe the training. An adversarial operator might then obtain white- or black-box access to the resulting model. If the algorithm is designed in such a way that it “memorizes” information about the training dataset, the operator can extract that information from the model. This attack is in some way similar to the model inversion attack, just with the differentiation that in this setting, the attacker is able to influence the amount and type of information embedded in the model. The attackers goal is to let the model appear unsuspicious, i.e. train it to be as accurate and predictive as a conventionally trained model. Data exfiltration attacks therefore can cause a number of privacy threats e.g. disclosure of information or identifiability within LINDDUN categories. According to CVSS frameworks, the metric values are more restricted than previously

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

mentioned attacks, as the attacker in this case is an operator of the model - the attacker requires some privileges and potentially also user interaction, therefore the score would be lower, with values of 5 being reasonable.

8.4.1 Threats to confidentiality in federated learning

In this section, we specifically highlight some of the confidentiality attacks that have already been investigated within a federated setting. This does not exclude other of the above mentioned attacks being also feasible in a federated setting.

- Deanonymization of participants of the federated learning process. An aggregator in parallel federated learning with the power to deanonymize the nodes can easier address the malicious behaviour of adversarial nodes (e.g. nodes performing poisoning attacks). However, deanonymization can cause privacy leaks in case of a malicious coordinator, who will be able to perform more targeted attacks against particular nodes (including attacks like membership inference, model inversion, attribute inference and others) (Orekondy et al. 2019).
- Membership inference attack was considered in federated learning settings in (Nasr, Shokri, and Houmansadr 2019). The authors showed that insider attackers can perform a powerful membership inference attack. The repeated parameter updates in federated learning are the main factors for increasing the accuracy of the attacks.

These specific threats to federated learning imply that the design of the FeatureCloud platform shall consider secure aggregation and in general protection of the exchanged parameters, to prevent any potential inference and attacks on confidentiality that might stem from the federated learning setting. Also limiting the access that individual participants have to other nodes’ models should be minimised.

9 Mitigations

In this section, we give a brief outline on possible mitigation actions towards the risks identified in the previous assessment. This can serve as a basis for the upcoming deliverable ***D2.4 Set of (novel) attack vectors and countermeasures (Month 36)***, but is not a complete and taxative list, as some mitigation actions will still be developed in the course of this project.

9.1 Local and Distributed System

For mitigation countermeasures against risks associated with the local and distributed system aspects of the FeatureCloud platform, we mostly rely on well-known concepts and techniques, amyn of which have been outlined in Section 4. As such, strong authentication and authorisation mechanisms, encryption of data storage and communication, and minimisation of data exchange will be considered with high priority.

Specifically, availability will not be a primary goal to achieve in the prototypically development of the system, as this depends a lot on the final deployment and resources available in the eventual (commercial) setting the platform is utilised after the conclusion of the project. However, these aspects shall be **considered** strongly for possible impact on the **architecture and design** of the **platform**.

Regarding threats from attacks that utilise **malicious application code**, one approach can be based on auditing and certification of these applications, i.e. they need to undergo a review process before they are added to and distributed from the FeatureCloud repository of analysis applications (“app

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

store”). Further mitigations against information hiding attacks will need to be developed, in close relation to other mitigation countermeasures that deal with specific machine learning related attacks.

9.2 (Federated) Machine Learning

There are privacy-preserving and secure computation techniques that can be applied to mitigate privacy and security risks in federated learning. However, different types of attacks can behave differently under these mitigations. The goal is to find a mitigation strategy which would ensure security and privacy of federated learning under the most probable privacy and security attacks. The challenge in finding a proper mitigation strategy is not only finding the safest and secure technique, but also the one that does not decrease the quality of the model. The efficiency of the training is also a critical factor, especially while considering different encryption techniques which can be computationally costly.

For these reasons, the aim should not be to mitigate all theoretically thinkable attacks and to reduce the risk to the absolute minimum, but to strike a balance between model quality and efficiency on the one hand and minimising the likelihood of practically thinkable attacks alongside the legal considerations laid down above.

9.2.1 Integrity and availability

To mitigate a poisoning attack one can use different mechanisms based on detection of outliers (see (Steinhardt, Koh, and Liang 2017)). (Y. Liu, Xie, and Srivastava 2017) considered filtering methods for the input as a defence against poisoning. Also, methods like pruning the network were considered in the literature (K. Liu, Dolan-Gavitt, and Garg 2018).

Mitigation against adversarial attacks can be based on the strategy of modifying training samples, model structure or combining the model with other models. (Papernot et al. 2016) proposed adversarial distance as a defense mechanism against adversarial samples, so the defender can detect vulnerable input. However, all proposed mitigations trying to prevent the creation of adversarial inputs do not guarantee full security. Defining mechanisms against security attacks in machine learning remains an open problem (Papernot et al. 2018).

9.2.2 Confidentiality

To decrease risks to confidentiality in federated learning one can use cryptographic mechanisms like secure multiparty computation, homomorphic encryption, or anonymization techniques like differential privacy. Implementation of these techniques makes data protection regulations stricter and reduces risks of data breaches (Argaw et al. 2020).

Secure multiparty computation allows several parties to jointly and securely compute a function over their inputs. It can be used in parallel federated learning to compute models' average (Bonawitz et al. 2017), instead of relying on a central, trusted coordinator.

Homomorphic encryption schemes, either fully or partially homomorphic, can be a potential solution to mitigate privacy risks in federated learning. The idea is to encrypt models’ parameters before sending to the aggregator, who performs operations on them. Additively homomorphic encryption was shown to ensure the security of federated learning for an honest-but-curious coordinator while preserving identical accuracy of a federated learning system without homomorphic encryption (Phong et al. 2017). Several works proposed privacy preserving federated learning with homomorphic encryption on different regression models, e.g. ridge regression (Chen et al. 2018), logistic regression (Hardy et al. 2017).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Differential privacy is another technique that can ensure a chosen level of privacy by adding noise. The technique can be applied to different machine learning algorithms (Shokri and Shmatikov 2015). This can be a defence mechanism for specific data analysis applications. Further, even single participants in the learning process may choose to employ this measure if they want to additionally protect their input.

10 Conclusion

In this deliverable, we detailed a security and privacy risk assessment specifically for the case of Federated Learning, which is utilised in the FeatureCloud project. While the design of the project mitigates risks that are associated with centralising (potentially) large amounts of data, still a number of risks pertain, as information can also be extracted from intermediate or final results of a machine learning process. Further, tampering with the process is potentially easier, as an attacker can successfully infer confidential information with the result of the training process (e.g. via poisoning attacks). In several settings, it is sufficient to manipulate one (or at least very few) of the participants to perform a successful attack. Thus, the system protected by the weakest security is to be considered the weakest link.

We also outlined a number of potential mitigation strategies that will be further elaborated in the deliverable D2.4 “Set of (novel) attack vectors and countermeasures” (Month 36). Further, the results of this analysis will influence the deliverables D2.3 “Working PAML-Layer with low distortion” (Month 32) and D2.5 “Secure Architecture and safety evaluation” (Month 60). Some types of attacks can also be controlled by achieving desirable values for the KPIs defined in D2.2 “KPIs and metrics for local execution platforms” (Month 16).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

11 References

Argaw, Salem T, Juan R. Troncoso-Pastoriza, Darren Lacey, Marie-Valentine Florin, Franck Calcavecchia, Denise Anderson, Wayne Burleson, Jan-Michael Vogel, Chana O’Leary, Bruce Eshaya-Chauvin, and Antoine Flahault. 2020. “Cybersecurity of Hospitals: discussing the challenges and working towards mitigating the risks.” *BMC Med Inform Decis Mak* **20,** 146. https://doi.org/10.1186/s12911-020-01161-7

Article 29 Data Protection Working Party. 2007. “Opinion 4/2007 on the Concept of Personal Data, WP 136.” https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf.

Article 29 Data Protection Working Party. 2011. “Opinion 15/2011 on the definition of consent, WP 187.” https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2011/wp187_en.pdf.

Article 29 Data Protection Working Party. 2017. “Article 29 Working Party Guidelines on consent under Regulation 2016/679, WP 259 rev.01.” https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=51030.

Ateniese, Giuseppe, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 2015. “Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers.” *Int. J. Secur. Netw.* 10 (3): 137–150. https://doi.org/10.1504/IJSN.2015.071829.

Bagdasaryan, Eugene, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2018. “How To Backdoor Federated Learning.” *CoRR* abs/1807.00459. http://arxiv.org/abs/1807.00459.

Bhagoji, Arjun Nitin, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. 2019. “Analyzing Federated Learning through an Adversarial Lens.” In *International Conference on Machine Learning*, 634–643. http://proceedings.mlr.press/v97/bhagoji19a.html.

Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. “Practical Secure Aggregation for Privacy-Preserving Machine Learning.” In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191. (CCS ’17). New York, NY, USA: ACM. https://doi.org/10.1145/3133956.3133982.

Chang, Ken & Balachandar, Niranjan & Lam, Carson & Yi, Darvin & Brown, James & Beers, Andrew & Rosen, Bruce & Rubin, Daniel & Kalpathy-Cramer, Jayashree. 2018. “Distributed deep learning networks among institutions for medical imaging.” Journal of the American Medical Informatics Association : JAMIA. 25. 10.1093/jamia/ocy017.

Chen, Yi-Ruei, Amir Rezapour, and Wen-Guey Tzeng. 2018. “Privacy-preserving ridge regression on distributed data”. Information Sciences, Volumes 451–452,

Conti, Mauro, Nicola Dragoni, and Viktor Lesyk. 2016. “A Survey of Man in the Middle Attacks.” *IEEE Communications Surveys & Tutorials* 18 (3): 2027–51.

Ehmann, Eugen and Martin Selmayr. 2018, *Datenschutz-Grundverordnung*, 2nd Edition, C.H. Beck, Munich, Germany.

Esayas, Samson Yoseph. 2015., "The role of anonymisation and pseudonymisation under the EU data privacy rules: beyond the 'all or nothing' approach." *European Journal of Law and Technology*, Vol 6, No 2.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. “Robust Physical-World Attacks on Deep Learning Visual Classification.” In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1625–34.

Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. 2015. “Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures.” In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Denver, Colorado, USA: ACM. https://doi.org/10.1145/2810103.2813677.

Ganju, Karan, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. 2018. “Property Inference Attacks on Fully Connected Neural Networks Using Permutation Invariant Representations.” In

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 619–633. CCS ’18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3243734.3243834.

Gu, Tianyu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. “BadNets: Evaluating Backdooring Attacks on Deep Neural Networks.” *IEEE Access* 7: 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068.

Hansman, Simon, and Ray Hunt. 2005. “A Taxonomy of Network and Computer Attacks.” *Computers & Security* 24 (1): 31–43.

Hardy, Stephen, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. 2017. “Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption.” *ArXiv, abs/1711.10677*.

Liu, Kang, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. “Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks.” In *Research in Attacks, Intrusions, and Defenses*, edited by Michael Bailey, Thorsten Holz, Manolis Stamatogiannakis, and Sotiris Ioannidis, 273–294. Cham: Springer International Publishing.

Liu, Yuntao, Yang Xie, and Ankur Srivastava. 2017. “Neural Trojans.” *IEEE 35th International Conference on Computer Design (ICCD)*, Boston, MA, USA, 2017 pp. 45-48. doi: 10.1109/ICCD.2017.16

Markatos, Evangelos, Davide Balzarotti, and Zlatogor Minchev. 2013. *The Red Book–A Roadmap in the Area of Systems Security.* The SysSec Consortium.

McMahan, H Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. “Communication-Efficient Learning of Deep Networks from Decentralized Data.” Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. 2016. “DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks.” In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2574–82.

Nasr, Milad, Reza Shokri, and Amir Houmansadr. 2019. “Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Inference Attacks against Centralized and Federated Learning.” *2019 IEEE Symposium on Security and Privacy (SP)*, 739–53.

Orekondy, Tribhuvanesh, Bernt Schiele, and Mario Fritz. 2019. “Knockoff Nets: Stealing Functionality of Black-Box Models.” In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Papernot, Nicolas, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016. “The Limitations of Deep Learning in Adversarial Settings.” In , 372–87. https://doi.org/10.1109/EuroSP.2016.36.

Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman. 2018. “SoK: Security and Privacy in Machine Learning.” In *2018 IEEE European Symposium on Security and Privacy (EuroS P)*, 399–414.

Phong, Le Trieu, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai. 2017. “Privacy-Preserving Deep Learning: Revisited and Enhanced.” International *Conference* on Applications and Techniques for Information Security.

Sheller, Micah J., G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. 2016. “Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation.” *Brainlesion : glioma, multiple sclerosis, stroke and traumatic brain injuries : second International Workshop, BrainLes 2016, with the challenges on BRATS,*

Shokri, Reza, and Vitaly Shmatikov. 2015. “Privacy-Preserving Deep Learning.” In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1310–1321. (CCS ’15). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2810103.2813687.

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. “Membership Inference Attacks Against Machine Learning Models.” In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. https://doi.org/10.1109/SP.2017.41.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Song, Congzheng, Thomas Ristenpart, and Vitaly Shmatikov. 2017. “Machine Learning Models That Remember Too Much.” In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 587–601. CCS ’17. New York, NY, USA: ACM. https://doi.org/10.1145/3133956.3134077.

Steinhardt, Jacob, Pang Wei Koh, and Percy Liang. 2017. “Certified Defenses for Data Poisoning Attacks.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3520–3532. NIPS’17. Red Hook, NY, USA: Curran Associates Inc.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna Estrach, Dumitru Erhan, Ian Goodfellow, and Robert Fergus. 2014. “Intriguing Properties of Neural Networks.” In *2nd International Conference on Learning Representations, ICLR 2014 ; Conference Date: 14-04-2014 Through 16-04-2014*.

Tidwell, Terry, Robert Larson, Kenneth Fitch, and John Hale. 2001. “Modeling Internet Attacks.” In *Proceedings of the 2001 IEEE Workshop on Information Assurance and Security*. Vol. 59. United States Military Academy West Point, NY.

Tramèr, Florian, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. “Stealing Machine Learning Models via Prediction APIs.” In *Proceedings of the 25th USENIX Conference on Security Symposium*, 601–618. SEC’16. USA: USENIX Association.

Truex, Stacey, Ling Liu, Mehmet Gursoy, Lei Yu, and Wenqi Wei. 2019. “Demystifying Membership Inference Attacks in Machine Learning as a Service.” *IEEE Transactions on Services Computing* PP: 1–1. https://doi.org/10.1109/TSC.2019.2897554.

Wang, Zhibo, Song Mengkai, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. 2019. “Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning.” IEEE INFOCOM 2019 - IEEE Conference on Computer Communications

Weinsberg, Udi, Smriti Bhagat, Stratis Ioannidis, and Nina Taft. 2012. “BlurMe: Inferring and Obfuscating User Gender Based on Ratings.” In *Proceedings of the Sixth ACM Conference on Recommender Systems*, 195–202. RecSys ’12. New York, NY, USA: ACM. https://doi.org/10.1145/2365952.2365989.

Xie, Chulin, Keli Huang, Pin-Yu Chen, and Bo Li. 2019. “DBA: Distributed Backdoor Attacks against Federated Learning.” In *DBA*. https://openreview.net/forum?id=rkgyS0VFvr.

Zander, Sebastian, Grenville Armitage, and Philip Branch. 2007. “A Survey of Covert Channels and Countermeasures in Computer Network Protocols.” *IEEE Communications Surveys & Tutorials* 9 (3): 44–57. https://doi.org/10.1109/COMST.2007.4317620.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.