

# IEEE Recommended Practice for Privacy and Security for Federated Machine Learning

IEEE Computer Society

Developed by the  
Artificial Intelligence Standards Committee

IEEE Std 2986™-2023

# IEEE Recommended Practice for Privacy and Security for Federated Machine Learning

Developed by the

**Artificial Intelligence Standards Committee**  
of the  
**IEEE Computer Society**

Approved 6 December 2023

**IEEE SA Standards Board**

**Abstract:** Privacy and security issues pose great challenges to the federated machine learning (FML) community. A general view on privacy and security risks while meeting applicable privacy and security requirements in FML is provided. This recommended practice is provided in four parts: malicious failure and non-malicious failure in FML, privacy and security requirements from the perspective of system and FML participants, defensive methods and fault recovery methods, and the privacy and security risks evaluation. It also provides some guidance for typical FML scenarios in different industry areas, which can facilitate practitioners to use FML in a better way.

**Keywords:** federated machine learning, FML, IEEE 2986™, machine learning, privacy, security

---

The Institute of Electrical and Electronics Engineers, Inc.  
3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2024 by The Institute of Electrical and Electronics Engineers, Inc.  
All rights reserved. Published 26 April 2024. Printed in the United States of America.

IEEE is a registered trademark in the U.S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

PDF: ISBN 979-8-8557-0705-2 STD26928  
Print: ISBN 979-8-8557-0706-9 STDPD26928

*IEEE prohibits discrimination, harassment, and bullying.*

*For more information, visit <https://www.ieee.org/about/corporate/governance/p9-26.html>.*

*No part of this publication may be reproduced in any form, in an electronic retrieval system or otherwise, without the prior written permission of the publisher.*

## Important Notices and Disclaimers Concerning IEEE Standards Documents

IEEE Standards documents are made available for use subject to important notices and legal disclaimers. These notices and disclaimers, or a reference to this page (<https://standards.ieee.org/ipr/disclaimers.html>), appear in all standards and may be found under the heading “Important Notices and Disclaimers Concerning IEEE Standards Documents.”

### Notice and Disclaimer of Liability Concerning the Use of IEEE Standards Documents

IEEE Standards documents are developed within IEEE Societies and subcommittees of IEEE Standards Association (IEEE SA) Board of Governors. IEEE develops its standards through an accredited consensus development process, which brings together volunteers representing varied viewpoints and interests to achieve the final product. IEEE standards are documents developed by volunteers with scientific, academic, and industry-based expertise in technical working groups. Volunteers are not necessarily members of IEEE or IEEE SA and participate without compensation from IEEE. While IEEE administers the process and establishes rules to promote fairness in the consensus development process, IEEE does not independently evaluate, test, or verify the accuracy of any of the information or the soundness of any judgments contained in its standards.

IEEE makes no warranties or representations concerning its standards, and expressly disclaims all warranties, express or implied, concerning this standard, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. IEEE Standards documents do not guarantee safety, security, health, or environmental protection, or guarantee against interference with or from other devices or networks. In addition, IEEE does not warrant or represent that the use of the material contained in its standards is free from patent infringement. IEEE Standards documents are supplied “AS IS” and “WITH ALL FAULTS.”

Use of an IEEE standard is wholly voluntary. The existence of an IEEE standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard.

In publishing and making its standards available, IEEE is not suggesting or rendering professional or other services for, or on behalf of, any person or entity, nor is IEEE undertaking to perform any duty owed by any other person or entity to another. Any person utilizing any IEEE Standards document, should rely upon their own independent judgment in the exercise of reasonable care in any given circumstances or, as appropriate, seek the advice of a competent professional in determining the appropriateness of a given IEEE standard.

IN NO EVENT SHALL IEEE BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

## Translations

The IEEE consensus balloting process involves the review of documents in English only. In the event that an IEEE standard is translated, only the English version published by IEEE is the approved IEEE standard.

## Official statements

A statement, written or oral, that is not processed in accordance with the IEEE SA Standards Board Operations Manual shall not be considered or inferred to be the official position of IEEE or any of its committees and shall not be considered to be, nor be relied upon as, a formal position of IEEE. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that the presenter's views should be considered the personal views of that individual rather than the formal position of IEEE, IEEE SA, the Standards Committee, or the Working Group. Statements made by volunteers may not represent the formal position of their employer(s) or affiliation(s).

## Comments on standards

Comments for revision of IEEE Standards documents are welcome from any interested party, regardless of membership affiliation with IEEE or IEEE SA. However, **IEEE does not provide interpretations, consulting information, or advice pertaining to IEEE Standards documents.**

Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments. Since IEEE standards represent a consensus of concerned interests, it is important that any responses to comments and questions also receive the concurrence of a balance of interests. For this reason, IEEE and the members of its Societies and subcommittees of the IEEE SA Board of Governors are not able to provide an instant response to comments, or questions except in those cases where the matter has previously been addressed. For the same reason, IEEE does not respond to interpretation requests. Any person who would like to participate in evaluating comments or in revisions to an IEEE standard is welcome to join the relevant IEEE working group. You can indicate interest in a working group using the Interests tab in the Manage Profile & Interests area of the [IEEE SA myProject system](#).<sup>1</sup> An IEEE Account is needed to access the application.

Comments on standards should be submitted using the [Contact Us](#) form.<sup>2</sup>

## Laws and regulations

Users of IEEE Standards documents should consult all applicable laws and regulations. Compliance with the provisions of any IEEE Standards document does not constitute compliance to any applicable regulatory requirements. Implementers of the standard are responsible for observing or referring to the applicable regulatory requirements. IEEE does not, by the publication of its standards, intend to urge action that is not in compliance with applicable laws, and these documents may not be construed as doing so.

---

<sup>1</sup> Available at: <https://development.standards.ieee.org/myproject-web/public/view.html#landing>.

<sup>2</sup> Available at: <https://standards.ieee.org/content/ieee-standards/en/about/contact/index.html>.

## Data privacy

Users of IEEE Standards documents should evaluate the standards for considerations of data privacy and data ownership in the context of assessing and using the standards in compliance with applicable laws and regulations.

## Copyrights

IEEE draft and approved standards are copyrighted by IEEE under U.S. and international copyright laws. They are made available by IEEE and are adopted for a wide variety of both public and private uses. These include both use, by reference, in laws and regulations, and use in private self-regulation, standardization, and the promotion of engineering practices and methods. By making these documents available for use and adoption by public authorities and private users, neither IEEE nor its licensors waive any rights in copyright to the documents.

## Photocopies

Subject to payment of the appropriate licensing fees, IEEE will grant users a limited, non-exclusive license to photocopy portions of any individual standard for company or organizational internal use or individual, non-commercial use only. To arrange for payment of licensing fees, please contact Copyright Clearance Center, Customer Service, 222 Rosewood Drive, Danvers, MA 01923 USA; +1 978 750 8400; <https://www.copyright.com/>. Permission to photocopy portions of any individual standard for educational classroom use can also be obtained through the Copyright Clearance Center.

## Updating of IEEE Standards documents

Users of IEEE Standards documents should be aware that these documents may be superseded at any time by the issuance of new editions or may be amended from time to time through the issuance of amendments, corrigenda, or errata. An official IEEE document at any point in time consists of the current edition of the document together with any amendments, corrigenda, or errata then in effect.

Every IEEE standard is subjected to review at least every 10 years. When a document is more than 10 years old and has not undergone a revision process, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE standard.

In order to determine whether a given document is the current edition and whether it has been amended through the issuance of amendments, corrigenda, or errata, visit [IEEE Xplore](#) or [contact IEEE](#).<sup>3</sup> For more information about the IEEE SA or IEEE's standards development process, visit the IEEE SA Website.

## Errata

Errata, if any, for all IEEE standards can be accessed on the [IEEE SA Website](#).<sup>4</sup> Search for standard number and year of approval to access the web page of the published standard. Errata links are located under the Additional Resources Details section. Errata are also available in [IEEE Xplore](#). Users are encouraged to periodically check for errata.

---

<sup>3</sup> Available at: <https://ieeexplore.ieee.org/browse/standards/collection/ieee>.

<sup>4</sup> Available at: <https://standards.ieee.org/standard/index.html>.

## Patents

IEEE standards are developed in compliance with the [IEEE SA Patent Policy](#).<sup>5</sup>

Attention is called to the possibility that implementation of this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. If a patent holder or patent applicant has filed a statement of assurance via an Accepted Letter of Assurance, then the statement is listed on the IEEE SA Website at <https://standards.ieee.org/about/sasb/patcom/patents.html>. Letters of Assurance may indicate whether the Submitter is willing or unwilling to grant licenses under patent rights without compensation or under reasonable rates, with reasonable terms and conditions that are demonstrably free of any unfair discrimination to applicants desiring to obtain such licenses.

Essential Patent Claims may exist for which a Letter of Assurance has not been received. The IEEE is not responsible for identifying Essential Patent Claims for which a license may be required, for conducting inquiries into the legal validity or scope of Patents Claims, or determining whether any licensing terms or conditions provided in connection with submission of a Letter of Assurance, if any, or in any licensing agreements are reasonable or non-discriminatory. Users of this standard are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. Further information may be obtained from the IEEE Standards Association.

## IMPORTANT NOTICE

Technologies, application of technologies, and recommended procedures in various industries evolve over time. The IEEE standards development process allows participants to review developments in industries, technologies, and practices, and to determine what, if any, updates should be made to the IEEE standard. During this evolution, the technologies and recommendations in IEEE standards may be implemented in ways not foreseen during the standard's development. IEEE standards development activities consider research and information presented to the standards development group in developing any safety recommendations. Other information about safety practices, changes in technology or technology implementation, or impact by peripheral systems also may be pertinent to safety considerations during implementation of the standard. Implementers and users of IEEE Standards documents are responsible for determining and complying with all appropriate safety, security, environmental, health, and interference protection practices and all applicable laws and regulations.

---

<sup>5</sup> Available at: <https://standards.ieee.org/about/sasb/patcom/materials.html>.

## Participants

At the time this IEEE recommended practice was completed, the SPFML Working Group had the following membership:

**Zuping Wu, Chair**  
**Ji Feng, Vice Chair**  
**Yang Liu, Secretary**

| <i>Organization Represented</i>                                  | <i>Name of Representative</i> |
|------------------------------------------------------------------|-------------------------------|
| AsiaInfo Technologies Inc. ....                                  | Da Wang                       |
| Alipay .....                                                     | Wenting Chang                 |
| Beijing University of Posts and Telecommunications .....         | Rentao Gu                     |
| Beijing RealAI Technology Co., Ltd. ....                         | Ying Guo                      |
| Beijing Sudo Technology .....                                    | Chao Pei                      |
| China Telecom.....                                               | Zuping Wu                     |
| China Academy of Information and Communications Technology ..... | Siyuan Wang                   |
| Cloudwalk Technology .....                                       | Jun Li                        |
| Clustar.....                                                     | Liu Yang                      |
| East China Normal University .....                               | Xiangxue Li                   |
| Guangdong University of Technology .....                         | Jiewu Leng                    |
| InsightOne .....                                                 | Yukun Wang                    |
| Sinovation Ventures AI Institute.....                            | Ji Feng                       |
| SenseTime.....                                                   | Hui Jiang                     |
| Shanghai Fudata Technology Co., Ltd.....                         | Tianya Yang                   |
| Shanghai Tongtai Information Technology Co., Ltd. ....           | Xuehui Hu                     |
| Tsinghua University.....                                         | Yanhong Pu                    |
| Tencent .....                                                    | Kepeng Li                     |
| Tongdun Technology Co., Ltd.....                                 | Cuiting Huang                 |
| WeBank .....                                                     | Qiang Yang                    |
| Yixue Squirrel AI Learning .....                                 | Richard Tong                  |

The following members of the entity Standards Association balloting group voted on this recommended practice. Balloters may have voted for approval, disapproval, or abstention.

|                                        |                                                  |                                     |
|----------------------------------------|--------------------------------------------------|-------------------------------------|
| 0xSenses Corporation                   | CloudWalk Technology                             | Institute of Biomedical             |
| Alipay (China) Technology Co., Ltd.    | East China Normal University                     | Engineering                         |
| Boeing Company                         | Hunan AsiaInfo Technology Co., Ltd.              | Intel                               |
| Chongqing Changan Automobile Co., Ltd. | Inspur Electronic Information Industry Co., Ltd. | SenseTime Group Limited             |
|                                        |                                                  | Yokosuka Telecom Research Park, Inc |

When the IEEE SA Standards Board approved this recommended practice on 6 December 2023, it had the following membership:

**David J. Law**, *Chair*  
**Ted Burse**, *Vice Chair*  
**Gary Hoffman**, *Past Chair*  
**Konstantinos Karachalios**, *Secretary*

Sara R. Biyabani  
Doug Edwards  
Ramy Ahmed Fathy  
Guido R. Hiertz  
Yousef Kimiagar  
Joseph L. Koepfinger\*  
Thomas Koshy  
John D. Kulick

Joseph S. Levy  
Howard Li  
Johnny Daozhuang Lin  
Gui Lin  
Xiaohui Liu  
Kevin W. Lu  
Daleep C. Mohla  
Andrew Myles

Paul Nikolich  
Annette D. Reilly  
Robby Robson  
Lei Wang  
F. Keith Waters  
Karl Weber  
Philip B. Winston  
Don Wright

\*Member Emeritus

## Introduction

This introduction is not part of IEEE Std 2986-2023, IEEE Recommended Practice for Privacy and Security for Federated Machine Learning.

Federated machine learning (FML) allows multiple participants to collaboratively train and use machine learning models without the need to share or exchange raw data among participants. However, FML systems are still vulnerable to adversarial attacks and various design flaws. For example, FML requires multiple rounds of communication, and unsafe communication channels can be a potential security risk. There are many participants participated in FML who can carry out attacks based on model parameters and/or training data. Another malicious attack is the participant may sniff the training sample that does not belong to them. The server is responsible for data aggregation and distribution, but the server may host curious or malicious participants who can be a potential risk of harm to the FML system.

While extensive research on defense mechanisms for FML exists and various protection models have been proposed, participants should be aware of the potential risks level both in security and privacy. This standard provides a recommend practice for using FML, including common failures and defensive methods. Based on the privacy and security requirements, the privacy risks level and security risks level assessment are proposed. It also provides some guidance for typical FML scenarios in different industry areas, which can facilitate practitioners to use FML in a better way.

## Contents

|                                                                                              |    |
|----------------------------------------------------------------------------------------------|----|
| 1. Overview .....                                                                            | 11 |
| 1.1 Scope .....                                                                              | 11 |
| 1.2 Purpose .....                                                                            | 11 |
| 1.3 Word usage .....                                                                         | 12 |
| 2. Normative references .....                                                                | 12 |
| 3. Definitions, acronyms, and abbreviations .....                                            | 12 |
| 3.1 Definitions .....                                                                        | 12 |
| 3.2 Acronyms and abbreviations .....                                                         | 13 |
| 4. Common failures .....                                                                     | 14 |
| 4.1 Non-malicious failure .....                                                              | 14 |
| 4.2 Malicious failure .....                                                                  | 15 |
| 5. Privacy and security requirements .....                                                   | 18 |
| 5.1 From the perspective of system .....                                                     | 18 |
| 5.2 From the perspective of FML participants .....                                           | 21 |
| 6. Defensive methods and fault recovery methods .....                                        | 23 |
| 6.1 Fault recovery methods for non-malicious failure .....                                   | 23 |
| 6.2 Defensive methods for data attack .....                                                  | 24 |
| 6.3 Defensive methods for model attacks .....                                                | 28 |
| 7. Evaluation .....                                                                          | 32 |
| 7.1 Privacy .....                                                                            | 32 |
| 7.2 Security .....                                                                           | 36 |
| Annex A (normative) Example of a native privacy-preserving method .....                      | 40 |
| Annex B (normative) Example of a privacy protection method of member inference attacks ..... | 41 |
| Annex C (informative) Example of the quantitative evaluation method .....                    | 42 |
| Annex D (informative) FML scenarios .....                                                    | 43 |
| D.1 FML in distributed manufacturing management scenario .....                               | 43 |
| D.2 FML in financial scenario .....                                                          | 45 |
| Annex E (informative) Example of calculating the existing probability .....                  | 48 |
| Annex F (informative) Risk level for privacy and security .....                              | 49 |
| F.1 Risk level for privacy .....                                                             | 49 |
| F.2 Risk level for security .....                                                            | 50 |
| Annex G (informative) Bibliography .....                                                     | 53 |

# IEEE Recommended Practice for Privacy and Security for Federated Machine Learning

## 1. Overview

Privacy and security issues pose great challenges to the machine learning community. Federated machine learning (FML), a distributed machine learning system, has gained more and more attention in both academia and industry due to its ability of obtaining a shared model between multiple parties without sharing raw training data directly. To benefit from the FML system, practitioners should be aware of the potential risks. On the one hand, these risks can include privacy issues that could reveal important training data because of the misuse of the FML system; on the other hand, the distributed training system could be fragile with respect to facing either non-malicious or malicious failure, which could lead to performance drop and even undetectable backdoors in the federated model. This document provides a recommended practice for using FML, including common failures and defensive methods. It also provides some guidance for typical FML scenarios in different industry areas, which can facilitate practitioners to use FML in a better way.

### 1.1 Scope

This document provides recommended practices related to privacy and security for FML, including security and privacy principles, defense mechanisms against non-malicious failures and examples of adversarial attacks on a FML system. This document also defines an assessment framework to determine the effectiveness of a given defense mechanism under various settings.

Data privacy and security are highly complex and increasingly regulated areas of law, and no recommended practice can provide unconditional consistency with all applicable laws and regulations, which may also vary at the local, state and regional level. Users of this document should evaluate any implementation for considerations of data privacy, security and data ownership in the context of FML, and are responsible for conformance with all such laws and regulations.

### 1.2 Purpose

The purpose of this recommended practice is to provide a resource on the topics of security and privacy for designers and users of FML systems and to accelerate the deployment of FML technology across industries.

### 1.3 Word usage

The word *shall* indicates mandatory requirements strictly to be followed in order to conform to the standard and from which no deviation is permitted (*shall* equals *is required to*).<sup>1, 2</sup>

The word *should* indicates that among several possibilities one is recommended as particularly suitable, without mentioning or excluding others; or that a certain course of action is preferred but not necessarily required (*should* equals *is recommended that*).

The word *may* is used to indicate a course of action permissible within the limits of the standard (*may* equals *is permitted to*).

The word *can* is used for statements of possibility and capability, whether material, physical, or causal (*can* equals *is able to*).

## 2. Normative references

The following referenced documents are indispensable for the application of this document (i.e., they must be understood and used, so each referenced document is cited in text and its relationship to this document is explained). For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments or corrigenda) applies.

GSMA TS.47, AI Mobile Device Requirements Specification.<sup>3</sup>

## 3. Definitions, acronyms, and abbreviations

### 3.1 Definitions

For the purposes of this document, the following terms and definitions apply. The *IEEE Standards Dictionary Online* should be consulted for terms not defined in this clause.<sup>4</sup>

**federated machine learning (FML):** A framework or system that enables multiple participants to collaboratively build and use machine learning models without disclosing the raw and private data owned by the participants while achieving good performance.

**privacy preservation:** Methods, products or services that protect the personally identifiable information (PII) of individuals and the native (or private) data of participants in federated machine learning.

**privacy preferences:** Specific choices made by an individual participant in federated machine learning about how personally identifiable information should be processed for a particular purpose.

**privacy risk assessment:** Overall process of risk identification, risk analysis, and risk evaluation with regard to the processing of personally identifiable information on native (or private) datasets of participants.

<sup>1</sup> The use of the word *must* is deprecated and cannot be used when stating mandatory requirements; *must* is used only to describe unavoidable situations.

<sup>2</sup> The use of *will* is deprecated and cannot be used when stating mandatory requirements; *will* is only used in statements of fact.

<sup>3</sup> GSMA standards are available from the GSM Association at <https://www.gsma.com/>.

<sup>4</sup> *IEEE Standards Dictionary Online* is available at: <http://dictionary.ieee.org>. An IEEE Account is required for access to the dictionary, and one can be created at no charge on the dictionary sign-in page.

**privacy risk identification:** Identify the list of privacy risk list based on participants' initial environment information in federated machine learning.

**privacy risk analysis:** Analyze the extent of privacy leakage based on identified risks.

**privacy risk evaluation:** Assess privacy risk level based on the analytical risk.

**security:** Protection of the data, model and intermediate model parameters of federated machine learning system during all processes, including data storage, data access, data transferring, model training, model evaluation, and model prediction.

**security risk assessment:** Overall process of risk identification, risk analysis and risk evaluation with regard to all the processes (including data storage, data access, data transferring, model training, model evaluation, and model prediction) of participants in federated machine learning.

**security risk identification:** Identify the list of security risk based on participants' behavior and initial environment information in federated machine learning.

**security risk analysis:** Analyze the potential consequences and threats based on identified risks.

**security risk evaluation:** Assess security risk level based on the analytical risk.

### 3.2 Acronyms and abbreviations

|         |                                         |
|---------|-----------------------------------------|
| DDP     | distributed differential privacy        |
| DP      | differential privacy                    |
| FML     | federated machine learning              |
| GAN     | generated adversarial network           |
| GC      | garbled circuit                         |
| HE      | homomorphic encryption                  |
| HFL     | horizontal federated machine learning   |
| IPR     | intellectual property rights            |
| LDP     | local differential privacy              |
| MC      | malicious-collude                       |
| MPC     | secure multiparty computation           |
| MS      | malicious-single                        |
| non-IID | non-independent identically distributed |
| OT      | oblivious transfer                      |
| PII     | personally identifiable information     |

|        |                             |
|--------|-----------------------------|
| PSI    | private set intersection    |
| SecAgg | secure aggregation          |
| SS     | secret sharing              |
| SVM    | support vector machine      |
| VFL    | vertical federated learning |

## 4. Common failures

### 4.1 Non-malicious failure

Compared to malicious attacks, non-malicious failures are generally less damaging but are more common potentially (Kairouz, et al. [B23]).<sup>5</sup> In the context of FML, non-malicious failures can be caused by unreliable participants outside the control of the service provider, systems factor as well as data constraints/noise. Generally, there are three possible non-malicious failures that are commonly seen—participant reporting failure, noisy model updates, and non-independent identically distributed data (non-IID).

#### 4.1.1 Participant reporting failure

Participant reporting failures occur when participants take too much time to respond or when participants drop out unexpectedly. In the context of cross-device FML, participant reporting failures may happen at any step during the training process (Chen, et al. [B14]). Failures can be introduced due to system factors such as constraint network connection, participant device's limited computational power, and participant withdrawal. The training process will be slowed down since the transmission time or participant computation time is prolonged, causing a longer responsive time from participant to server. In addition, there are potential failures when implementing secure aggregation (SecAgg), especially if the participants drop out SecAgg during the protocol and become unresponsive.

#### 4.1.2 Noisy model updates

Model updates sent to the server may become distorted because of network and/or the FML architectural factors where separate entities control the sever, participants, and network. A participant's data can be distorted due to the same reasons. Also, noise can be unintentionally introduced into the participant's data (e.g., lower camera resolution) or into the data labels. These network factors and noisy data may harm the convergence of the FML process. These corruptions can be regarded as a mild form of model update and data poisoning attack.

#### 4.1.3 Non-IID

In addition, the training data used in the cross-device FML scenario could be uneven in quantity and/or follow a non-IID. This is because the training data in this scenario mainly comes from users' daily data, which may include the users' age, region, daily habits and other information, resulting in uneven numbers

---

<sup>5</sup> The numbers in brackets correspond to those of the bibliography in Annex G.

and(or) non-IID. Therefore, large difference in model updates could lead to a decrease on a model's convergence rate or even cause a degradation on shared models' performance and stability.

## 4.2 Malicious failure

Malicious failure contains possible risks when the FML is under the malicious setting, that is, there are some malicious attackers who will deliberately attack the learning system. The attackers could be internal dishonest participants or external attackers.

It is important to remember that the main goal of FML is to keep data privacy and improve model performance. With that in mind, the malicious attack can be divided into two parts—data attacks and model attacks.

### 4.2.1 Data attacks

In data attacks, valuable private data are often the attacker's main goal. Usually, it is hard for attackers to get private datasets without touching the data provider. The more realistic threats are getting access to one or more data points used in the FML system. There have been some attack methods where attackers are dishonest training participants who use the model gradients. Some attacks could be even more threatening, e.g., those which do not depend on the gradient information.

#### 4.2.1.1 Training sample sniffing

Training sample sniffing is one typical threat during FML. Under this scenario, the attackers could be one or more malicious participants of FML whose goal is to get some specific data record that does not belong to them. Generally, there are two possible attacks that are commonly seen—membership inference attack and gradient leakage attack.

- *Membership inference attack* (Shokri, et al. [B36]) proposes a query-based black-box method for a more general setting: given a data record and black-box access to a model, determine if the record was used in the model's training phase. By discovering the fact that the machine learning model is often overfitting in the training data, the attacker can train a classifier that only uses output logits to determine where a data sample was used to train this model. Under the FML setting, the membership inference attack could even be stronger (Nasr, et al. [B30]). Since the attacker can get full model parameters during training, the attack success rate could be improved by a large margin, which shows the possibility of private data leakage in FML.
- *Gradient leakage attack* is another typical method of training sample sniffing. In some FML scenarios, to improve the performance of a FML model, not only can the participants get the models parameters, but also the gradients of these parameters. That may introduce more risks. Deep leakage (Zhu, et al. [B49]) is a recently proposed attack target to this setting. The malicious attacker can iteratively recover the training sample from scratch by repeatedly querying the gradients of the same inputs. The method can be successfully adapted to any federated setting that did not do any protection on the model's gradients. What's worse, by utilizing the leaked gradients, the attacker can not only recover the private data that does not belong to them, they can even design a vulnerable test example of the federated model.

#### 4.2.1.2 Recovery of training data properties

Besides training sample sniffing, another goal of data attack is to recover some specific properties of training data. For example, in medical image data, if the attacker could know that the training data that is labeled with “disease” usually corresponds to “having blue eyes,” the valuable commercial data could be leaked. For FML application, such an attack often happens in a vertical setting when some participants try to infer unknown labels of their data.

- *Server-side GAN-based attack:* Server-side generated adversarial network (GAN)-based attack (see He, et al. [B20], Melis, et al. [B27], Wang, et al. [B40], Hitaj, et al. [B21]) aims to infer participant-level privacy data (i.e., the category of samples, real or fake samples, participant identity) and reconstruct victim training data in the FML scenario. Instead of directly inverting the federated model, it usually uses the generated example training by a GAN to interact with the federated system. Compared with the attack launched by the participant side, it can only infer a certain category of representation. To achieve the previous threat goals, the attacker designs a multi-task discriminator based on a variation of GAN (Wang, et al. [B40]). The structure of the discriminator is the same as the shared model except for the last output layer owing to different multi-task output formats. Additionally, participant-level retrieval of private data possibly happened in the FML server-side’s ability owing to the threat of distinguishing the participant’s identity. When the federated model finished training, the GAN model would collect enough information of training data, so it may generate training data of a certain class. Multiple existing defense methods are applied to reduce or mitigate such attacks [e.g., SecAgg (Bonawitz, et al. [B9]), secure multiparty computation (MPC) (Bonawitz, et al. [B10]), and so on].
- *Model inversion attack:* Model inversion attack (Fredrikson, et al. [B17]) is carried out to reconstruct training data by inverting the target model based on the known model information (e.g., intermediate gradient, predictions confidence score, and so on) under the white box or black box access. The defined model structure for each participant is the same in the FML scenario, meaning naturally white box setting. The honest and curious server launched the attack, completely following the FML protocol while curiously inferring participant-side private data. Different from training sample sniffing, since the model inversion attack only input a label, it can only recover some properties of the class but not a specific individual. Participants’ private data is reconstructed by inverting the participant-side model based on minimization optimization of the distance (Geiping, et al. [B18], Zhu, et al. [B49]), based on the server-side aggregated result (Jayaraman and Evans [B22]) and participant-side intermediate gradient updated (He, et al. [B20]). For example, utilizing the model inversion attack, the attacker may get an input recovered from the label 0, but they might not have the similar image of the recovered one. In FML, since the attacker may get the parameters of the model during the training phase, the model inversion attack would be riskier.
- *Label inference attack for vertical federated learning (VFL):* In a vertical FML setting, not all participants obtain the information of label. Label inference attack is carried out by the latter kind of participants and aims to infer/reconstruct the local private label information of the former type of participants. The process of label inference is often separated from the original VFL training or predicting process. Multiple methods are applied to infer labels, including extracting information from communicated gradients and fine-tuning local model with auxiliary labeled data.

#### 4.2.2 Model attacks

Different from data attacks, model attacks aim at the federated model. The goals of the model attacks could be various. Sometimes, model attacks are executed to reduce the performance of the federated model. There are also some scenarios that the model could be backdoor during training, so that the attacker could manipulate the federated model after training. Different attacks could happen during each stage of the federate learning algorithm, from the data collecting to model deployment.

#### 4.2.2.1 Before training

There are two types of model attacks before training, as follows:

- *Data poisoning attack for FML*: While the FML environment motivates participants to actively contribute training data and send model parameters to the server, FML systems trained on user-supplied data are vulnerable to data poisoning (Biggio, et al. [B6]). The poisoning attack of the FML system would affect integrity and availability of the FML model by contaminate samples in the training set (Steinhardt, et al. [B37], Lyu, et al. [B26]), and it most often occurs at the stage of data acquisition. Meanwhile, the attacker modifies the dataset to implement a “white-box attack,” which could be implemented by the malicious subset of participants by sending malicious model updates derived from their locally-created malicious data (Xiao, et al. [B42]). This kind of attack is often seen in horizontal federated machine learning (HFL) systems and can be launched by any single participant. However, this does not mean it is hard to apply in VFL systems. If the participant that controls labels in VFL system flip some of the labels, the federated model will be totally powerless facing samples from that class.

Except for the stage of data acquisition, the (re-)training phase involves the data poisoning attack as well. Since the quality of the data at this point is higher compared to the initial (untrained), it might provide an opportunity for an attacker to efficiently manipulate this data.

- *Backdoor attack for FML*: Backdoor attack is to hide the backdoor (with a predefined trigger) during the local model training before updating the server. As the backdoor is activated, the target label would be replaced with the pre-specified label from the attacker. The backdoor attack retains the normal performance of FML in the untriggered case. Therefore, the accuracy of FML tasks might not be decreased.

#### 4.2.2.2 During training

Byzantine attack is the most common malicious attack during training.

- *Byzantine attack*: Byzantine attack means that malicious participants can disobey FML rules during training and upload randomly generated or even specifically modified gradient information to the central node, thus misleading the optimization process and leading to serious degradation of learning performance or other potential risks in FML system (Bagdasaryan, et al. [B4]).

Byzantine attack can be divided into two categories according to their effect, namely untargeted attack and targeted attack.

- a) *Untargeted attack*: Malicious participants attempt to damage the global model accuracy or make the model unusable on main task (Bhagoji, et. al. [B5]). During the attack, the parameters of the model may not converge to the optimum and finally the model’s predictions may not be accurate enough to meet the needs of the actual scenario.
- b) *Targeted attack*: Malicious participants’ goal is not necessarily to compromise the global model accuracy, but to keep the model maintaining good prediction performance on the main task while implanting backdoors into the model (Kairouz, et al. [B23]).

From the distribution of malicious gradient messages, there are two types of Byzantine attacks, i.e., explicit boosting attack and stealthy boosting attack, which are described as follows:

- c) *Explicit boosting attack*: Malicious participants upload forged gradient values significantly greater than those by honest participants. This kind of attack is very obvious and can usually