DOI: 10.18372/2310-5461.70.21194
UDC 004.056:004.8:004.75

Stanislava Kudrenko, Cand. Of Tech. Sc., As. Prof.
State University "Kyiv Aviation Institute", Kyiv
https://orcid.org/0000-0002-0759-3908
e-mail: stanislava.kudrenko@npp.nau.edu.ua;

Oleksii Nimych
State University "Kyiv Aviation Institute", Kyiv
https://orcid.org/0000-0003-1759-7088
e-mail: 5356349@stud.kai.edu.ua;

Ihor Makieiev
State University "Kyiv Aviation Institute", Kyiv
https://orcid.org/0009-0009-8679-5652
e-mail: 8390988@stud.kai.edu.ua

METHOD FOR DETECTING ANOMALOUS LOCAL UPDATES AND ISOLATING MALICIOUS PARTICIPANTS IN FEDERATED LEARNING SYSTEMS

Introduction

Federated learning is a distributed machine learning paradigm in which a global model is trained through the cooperation of multiple participants without direct transmission of their local datasets to a central server. This approach is especially important for systems where data privacy, limited data mobility, regulatory restrictions, or the critical nature of information make centralized data collection undesirable or impossible. Such conditions are typical for industrial Internet of Things systems, edge and fog computing environments, medical information systems, intelligent transport, autonomous platforms, and critical infrastructure monitoring systems.

In federated learning, each participant acts as an independent source of local training results. A participant trains a local model using its own data and then sends the obtained update to a central aggregation server. These updates may include model parameters, gradients, or weight changes, depending on the selected learning architecture. The aggregation server combines the received updates to form an improved global model. However, this interaction creates specific security risks. Since the server usually does not have direct access to the original local datasets or to the internal training process of each participant, it is difficult to verify whether the received updates are reliable, correct, and generated by legitimate participants.

One of the key threats in federated learning is the presence of malicious or compromised participants. Such participants may send anomalous, intentionally distorted, or statistically inconsistent local updates to the aggregation server. These updates can reduce the accuracy of the global model, slow down convergence, create hidden vulnerabilities, or lead to model poisoning. In critical infrastructure systems, this problem becomes especially dangerous because incorrect model behavior may affect monitoring, anomaly detection, decision support, and automated control processes.

Existing studies show that federated learning systems require not only privacy-preserving mechanisms, but also methods for detecting unreliable participants and filtering anomalous local updates before they influence the global model. Therefore, the development of a method for detecting anomalous local updates and isolating malicious participants is a relevant scientific and practical task.

The purpose of this study is to develop a method for detecting anomalous local updates and isolating malicious participants in federated learning systems. The proposed approach is aimed at improving the reliability of model aggregation by identifying local updates that significantly deviate from the expected behavior of the majority of participants and by preventing such updates from affecting the global model.

Analysis of recent research and publications

The general concept of federated machine learning and its practical applications are considered in [3]. The authors describe federated learning as a promising approach for building machine learning models in distributed environments where data cannot be centralized because of privacy, ownership, or regulatory limitations. This work provides the conceptual basis for understanding federated learning as a cooperation mechanism between multiple par-

ticipants that exchange model updates instead of raw data.

The main challenges, methods, and development directions of federated learning are analyzed in [4]. The study emphasizes that federated learning differs from traditional centralized learning not only by its data distribution model, but also by the presence of heterogeneous devices, non-identically distributed data, communication limitations, and security risks. These factors significantly complicate the process of training and aggregating local models.

A broader overview of federated learning problems is presented in [5], where the authors systematize open research challenges in this field. Among the key issues, they highlight privacy, robustness, fairness, communication efficiency, and resistance to adversarial behavior. This confirms that the security of federated learning cannot be reduced only to data confidentiality; it also includes the reliability of participants and the trustworthiness of model updates.

Security and privacy issues in federated learning are analyzed in detail in [6]. The authors show that federated learning systems may be vulnerable to different types of attacks, including data poisoning, model poisoning, inference attacks, and attacks performed by malicious clients. This is important for the present study because malicious participants can compromise the learning process without directly attacking the central server or accessing other participants' data.

The problem of Byzantine behavior in distributed machine learning is studied in [7]. The authors propose a Byzantine-tolerant approach to gradient aggregation and show that traditional averaging can be highly vulnerable when some participants send arbitrary or adversarial updates. This work is important because it demonstrates that even a small number of unreliable participants may significantly affect the final model if the aggregation procedure is not protected.

Further development of Byzantine-robust distributed learning is presented in [8]. The authors analyze robust aggregation methods and investigate how distributed learning can remain statistically effective in the presence of adversarial participants. Such methods form an important theoretical foundation for detecting and limiting the influence of anomalous local updates in federated systems.

The hidden vulnerabilities of distributed learning under Byzantine conditions are examined in [9]. The authors show that some distributed learning systems may remain vulnerable even when they use protection mechanisms. This means that the mere use of robust aggregation is not always sufficient. Additional detection and filtering procedures are needed to identify suspicious local updates before or during aggregation.

Local model poisoning attacks against Byzantine-robust federated learning are studied in [10]. This work is especially relevant to the present article because it demonstrates that malicious participants can adapt their attacks to bypass robust aggregation mechanisms. The study confirms that model poisoning should be considered not only as a general attack scenario, but also as a practical threat to real federated learning systems.

A specialized approach to detecting malicious clients in federated learning is proposed in [11]. The FLDetector method is based on the idea that malicious participants can be identified by analyzing the inconsistency of their model updates across training rounds. This approach is close to the logic of the present study, since the detection of anomalous local updates is considered as a basis for isolating unreliable participants from the aggregation process.

Collaborative malicious gradient filtering is considered in [12]. The authors propose an approach to Byzantine-robust federated learning based on filtering suspicious gradients. This confirms the importance of analyzing local updates not only individually, but also in relation to the behavior of other participants. The comparison of local updates with the collective distribution of updates is one of the promising directions for improving the reliability of federated learning.

The previous research of the authors [13] was devoted to predicting node compromise in edge and fog environments for critical infrastructure objects. That study considered the problem of compromised nodes in distributed computing environments. In the present article, this idea is developed in the context of federated learning, where potentially compromised elements are not only network nodes, but also learning participants that may transmit anomalous local updates to the global aggregation server.

Thus, the analysis of modern scientific literature shows that federated learning is an effective approach for privacy-preserving distributed model training, but it remains vulnerable to malicious participants and model poisoning attacks. Existing studies have proposed Byzantine-robust aggregation, gradient filtering, and malicious client detection methods. However, the problem of detecting anomalous local updates and isolating unreliable participants remains relevant, especially for federated learning systems used in edge, fog, and critical infrastructure environments.

At the same time, the analysis of existing studies shows that many approaches focus mainly on robust aggregation or filtering of individual updates during a specific training round. However, in practical fed-

erated learning systems, it is important not only to reduce the influence of a suspicious update, but also to identify the participant that repeatedly generates anomalous updates and to isolate this participant from further aggregation. Therefore, the development of a method that combines local update anomaly detection with participant-level isolation remains a relevant task for improving the resilience of federated learning systems.

Thus, the analysis of modern scientific literature shows that federated learning is an effective approach for privacy-preserving distributed model training, but it remains vulnerable to malicious participants, anomalous local updates, and model poisoning attacks. Existing studies propose Byzantine-robust aggregation, gradient filtering, and malicious client detection methods; however, many of these approaches mainly focus on reducing the influence of suspicious updates within a particular training round. In practical federated learning systems, especially in edge, fog, and critical infrastructure environments, it is important not only to filter or weaken a single anomalous update, but also to identify participants that repeatedly generate such updates and exclude them from further aggregation. Therefore, the development of a method that combines local update anomaly detection with participant-level isolation remains a relevant task for improving the security and resilience of federated learning systems.

Problem Statement

In federated learning, the global model is updated on the basis of local updates received from distributed participants. Under normal conditions, these updates reflect the learning results obtained from local datasets and should follow the general direction of model improvement. However, if some participants are malicious or compromised, they may transmit distorted local updates that differ significantly from the collective behavior of other participants and negatively affect the global model.

Let the set of local updates received by the aggregation server at training round $t$be defined as:

$$\Delta W^t = \{\Delta w_1^t, \Delta w_2^t, \dots, \Delta w_n^t\},$$

where $\Delta w_i^t$is the local update generated by the i-th participant, and $n$is the total number of participants in the federated learning system.

The main problem considered in this study is to determine whether a local update $\Delta w_i^t$is consistent with the general behavior of the other updates or should be treated as anomalous. For this purpose, each update can be compared with the reference aggregated behavior of the participants:

$$S_i^t = d(\Delta w_i^t, \bar{\Delta w}^t),$$

where $S_i^t$is the anomaly score of the i-th participant at round t, $d(\cdot)$is a distance or deviation function, and $\bar{\Delta w}^t$is the reference update calculated from the set of received local updates.

If the anomaly score exceeds a predefined threshold, the corresponding update is considered suspicious:

$$S_i^t > \theta.$$

The scientific problem is to develop a method that not only detects such anomalous local updates, but also identifies participants that repeatedly generate them and excludes these participants from further aggregation. This makes it possible to reduce the impact of malicious participants on the global model and improve the resilience of federated learning systems to model poisoning attacks.

In federated learning, the global model is updated on the basis of local updates received from distributed participants. Under normal conditions, these updates reflect the learning results obtained from local datasets and should follow the general direction of model improvement. However, if some participants are malicious or compromised, they may transmit distorted local updates that differ significantly from the collective behavior of other participants and negatively affect the global model.

Let the set of local updates received by the aggregation server at training round $t$be defined as

$$\Delta W^t = \{\Delta w_1^t, \Delta w_2^t, \dots, \Delta w_n^t\}, \quad (1)$$

where Δw_i^t is the local update generated by the i-th participant, and $n$is the total number of participants in the federated learning system.

The main problem considered in this study is to determine whether a local update Δw_i^t is consistent with the general behavior of the other updates or should be treated as anomalous. For this purpose, each update can be compared with the reference aggregated behavior of the participants:

$$S_i^t = d(\Delta w_i^t, \bar{\Delta w}^t), \quad (2)$$

where S_i^t is the anomaly score of the i-th participant at round t, $d(\cdot)$ is a distance or deviation function, and $\bar{\Delta w}^t$is the reference update calculated from the set of received local updates.

If the anomaly score exceeds a predefined threshold, the corresponding update is considered suspicious:

$$S_i^t > \theta. \quad (3)$$

The scientific problem is to develop a method that not only detects anomalous local updates according to condition (3), but also identifies participants that repeatedly generate such updates and excludes them from further aggregation. This makes it possible to reduce the impact of malicious participants on the global model and improve the resilience of federated learning systems to model poisoning attacks.

Proposed Method

The proposed method is based on the assumption that local updates generated by honest participants should have a certain degree of similarity, even if their local datasets are heterogeneous. In contrast, malicious or compromised participants may generate pdates that significantly deviate from the collective direction of model training. Therefore, the method combines three main procedures: calculation of a reference update, estimation of the anomaly score for each participant, and isolation o

f participants that repeatedly generate anomalous updates.

At each training round t, the aggregation server receives the set of local updates defined in (1). To evaluate the behavior of each participant, a reference update is first calculated. In the simplest case, it can be represented as the average update of all participants:

$$\Delta\bar{w}^t = \frac{1}{n}\sum_{i=1}^{n} \Delta w_i^t, \quad (4)$$

where $\Delta\bar{w}^t$is the reference update at round t, $\Delta w_i^t$is the local update of the i-th participant, and $n$is the total number of received updates.

Then, according to (2), an anomaly score is calculated for each participant. As a deviation function $d(\cdot)$, the Euclidean distance between the local update and the reference update can be used:

$$S_i^t = \| \Delta w_i^t - \Delta\bar{w}^t \|_2. \quad (5)$$

The value S_i^tcharacterizes how strongly the update of the i-th participant differs from the collective update pattern. A higher value of S_i^tindicates a higher probability that the corresponding update is anomalous.

To make the detection process adaptive to the current training round, the threshold θcan be determined using the mean and standard deviation of anomaly scores:

$$\theta^t = \mu_S^t + \lambda\sigma_S^t, \quad (6)$$

where $\mu_S^t$is the mean value of anomaly scores at round t, $\sigma_S^t$is their standard deviation, and λ is a sensitivity coefficient that controls the strictness of anomaly detection.

According to condition (3), the local update of the i-th participant is considered anomalous if

$$S_i^t > \theta^t. \quad (7)$$

However, a single anomalous update is not always sufficient to classify a participant as malicious. In practical federated learning systems, deviations may also appear due to non-identically distributed local data, temporary noise, or unstable network conditions. Therefore, the proposed method uses an accumulated anomaly counter for each participant:

$$C_i^t = C_i^{t-1} + I(S_i^t > \theta^t), \quad (8)$$

where $C_i^t$is the accumulated number of anomalous updates generated by the i-th participant up to round t, and $I(\cdot)$is an indicator function that equals 1 if condition (7) is satisfied and 0 otherwise.

A participant is isolated from further aggregation if the number of detected anomalous updates exceeds the allowed limit:

$$C_i^t \geq C_{\max}, \quad (9)$$

where $C_{\max}$is the maximum acceptable number of anomalous updates for one participant.

After detecting suspicious updates, the aggregation server forms a set of trusted participants:

$$P_T^t = \{p_i \mid S_i^t \leq \theta^t \text{ and } C_i^t < C_{\max}\}. \quad (10)$$

Only updates from this trusted set are used to form the global model:

$$w^{t+1} = w^t + \frac{1}{|P_T^t|}\sum_{p_i \in P_T^t} \Delta w_i^t. \quad (11)$$

Thus, unlike conventional federated averaging, the proposed method does not include all received local updates in the global aggregation. Each update is first evaluated according to (5)–(7), and each participant is additionally assessed according to its accumulated behavior using (8) and (9). This allows the system to reduce the influence of single anomalous updates and to isolate participants that repeatedly demonstrate suspicious behavior.

The general sequence of the proposed method is as follows. First, the aggregation server receives local updates from all participants according to (1). Second, the reference update is calculated using (4). Third, anomaly scores are determined according to (5). Fourth, the adaptive threshold is calculated according to (6), and anomalous updates are detected using (7). Fifth, the anomaly counters are updated according to (8). Finally, participants that exceed the allowed anomaly limit are isolated according to (9), and the global model is updated only on the basis of trusted participants according to (10) and (11).

The main advantage of the proposed method is that it combines update-level anomaly detection with participant-level isolation. This is important because model poisoning attacks may be performed gradually, when a malicious participant sends moderately distorted updates over several training rounds. In such cases, the analysis of only one training round may be insufficient. The accumulated anomaly counter makes it possible to detect repeated suspicious behavior and prevent unreliable participants from continuing to influence the global model.

Illustrative Example

To demonstrate the logic of the proposed method, a simplified illustrative example is considered. The purpose of this example is not to reproduce a full-scale federated learning experiment, but to show how anomalous local updates can be detected using the proposed scoring procedure. Therefore, synthetic numerical data are used. Such data make it possible to clearly trace each calculation step, including the formation of the reference update, the calculation of anomaly scores, and the identification of a suspicious participant.

The example assumes a federated learning system with five participants. Each participant sends a local update to the aggregation server during one training round. For simplicity, each update is represented as a two-dimensional vector. This representation is chosen only for clarity of explanation, although in real federated learning systems local updates are usually high-dimensional vectors of model parameters, gradients, or weight changes.

The first four updates are selected to be close to each other because they represent normal participants whose local training results follow a common direction of model improvement. The fifth update is intentionally chosen to be significantly different from the others in order to imitate the behavior of a malicious or compromised participant. This makes it possible to verify whether the proposed method can detect an update that deviates from the collective update pattern.

Let the local updates received by the aggregation server be as follows:

$$\Delta w_1^t = (0.12,0.08), \Delta w_2^t = \\ = (0.10,0.09), \Delta w_3^t = (0.11,0.07), \\ \Delta w_4^t = (0.13,0.08), \Delta w_5^t = (0.80, -0.45).$$

These values model a situation in which four participants generate consistent local updates, while one participant generates an anomalous update. According to (4), the reference update is calculated as the average value of all received local updates:

$$\Delta\bar{w}^t = \left(\frac{0.12+0.10+0.11+0.13+0.80}{5}, \frac{0.08+0.09+0.07+0.08-0.45}{5}\right) = \\ (0{,}252, -0{,}026).$$

It should be noted that the reference update is already shifted because the anomalous vector is included in the averaging procedure. Nevertheless, the deviation of the fifth update remains much larger than the deviations of the other updates.

Using (5), the anomaly score is calculated as the Euclidean distance between each local update and the reference update:

$$S_1^t = \sqrt{(0.12-0.252)^2 + (0.08+0.026)^2} \\ = 0.169,$$
$$S_2^t = \sqrt{(0.10-0.252)^2 + (0.09+0.026)^2} \\ = 0.191,$$
$$S_3^t = \sqrt{(0.11-0.252)^2 + (0.07+0.026)^2} \\ = 0.171,$$
$$S_4^t = \sqrt{(0.13-0.252)^2 + (0.08+0.026)^2} \\ = 0.162,$$
$$S_5^t = \sqrt{(0.80-0.252)^2 + (-0.45+0.026)^2} \\ = 0.693.$$

The obtained values show that the fifth participant has the largest deviation from the collective update pattern. Even when the reference update is significantly shifted due to the influence of an anomalous vector, the anomaly score S_5^t remains high enough to provide unambiguous identification of the deviation.

For this example, the mean value and standard deviation of anomaly scores are:

$$\mu_S^t = 0{,}277, \\ \sigma_S^t = 0{,}208.$$

If the sensitivity coefficient in (6) is set to $\lambda = 1.1$, the adaptive threshold is:

$$\theta^t = \mu_S^t + \lambda\sigma_S^t = 0{,}277 + 1{,}1 \cdot 0{,}208 = 0{,}506.$$

In this illustrative example, the value $\lambda = 1{,}1$ is selected only to demonstrate the operation of the adaptive threshold. In practical federated learning systems, this parameter should be selected experimentally depending on the expected level of data heterogeneity, the number of participants, and the acceptable balance between false positives and false negatives. For example, λ may be tuned using validation experiments, cross-validation, or statistical rules similar to the three-sigma principle.

According to the detection condition (7), a local update is considered anomalous if its anomaly score exceeds the threshold. In this case, only the fifth participant satisfies this condition:

$$S_5^t = 0{,}693 > 0{,}506.$$

Therefore, the update generated by participant p_5 is classified as anomalous. The updates of the other participants are not classified as anomalous because their scores are below the threshold:

$$S_1^t, S_2^t, S_3^t, S_4^t < \theta^t.$$

As a result, the trusted set of participants for the current training round is formed as:

$$P_T^t = \{p_1, p_2, p_3, p_4\}.$$

The suspicious update generated by participant $p_5$is excluded from the aggregation process. There-

fore, the global model update is calculated only on the basis of trusted participants:

$$\Delta w_T^t = \frac{1}{4}(\Delta w_1^t + \Delta w_2^t + \Delta w_3^t + \Delta w_4^t).$$

Substituting the values of the trusted updates gives:

$$\Delta w_T^t = \left(\frac{0.12+0.10+0.11+0.13}{4}, \frac{0.08+0.09+0.07+0.08}{4}\right) = (0.115, 0.080).$$

As can be seen, the final aggregated vector is almost identical to the updates generated by the honest participants. This confirms that excluding the anomalous update of participant $p_5$prevents its negative influence on the global model.

Thus, after excluding the anomalous update, the aggregated update corresponds to the common direction of the normal participants. If the fifth participant continues to generate anomalous updates in subsequent training rounds, its anomaly counter will increase according to (8). When the counter reaches the maximum allowed value $C_{\max}$, this participant will be isolated from further aggregation according to (9).

This example demonstrates that the proposed method can detect a local update that significantly deviates from the collective behavior of other participants and prevent it from directly influencing the global model. It also shows the practical role of participant-level isolation: a single suspicious update may be filtered during one round, while repeated anomalous behavior becomes the basis for excluding the corresponding participant from the federated learning process.

Comparative Analysis of Existing Approaches

The illustrative example presented above demonstrates how the proposed method detects anomalous local updates and excludes suspicious participants from the aggregation process. However, this example describes only one simplified training round. To better understand the role of the proposed method in the broader context of federated learning security, it is necessary to compare it with existing protection mechanisms that are commonly used to reduce the influence of unreliable or malicious participants.

Figure 1 presents a comparative taxonomy of several representative protection approaches in federated learning systems. The considered mechanisms include standard aggregation, Byzantine-robust aggregation, behavioral detection methods, and the proposed dual-layer defense approach.

As shown in Fig. 1, one of the basic and widely used aggregation approaches is Federated Averaging (FedAvg) [1]. Its main advantage is simplicity and relatively low computational complexity. In this approach, the aggregation server directly averages the local updates received from participating clients. However, FedAvg does not include a mechanism for checking the reliability of these updates. Therefore, if one or several participants transmit distorted or intentionally poisoned updates, such updates may be included in the global model and negatively affect its accuracy, convergence, or stability.

Byzantine-robust aggregation methods were proposed to reduce the influence of unreliable or adversarial updates in distributed learning. One of the well-known approaches is Krum [7], which selects the update that is closest to the majority of other updates. This makes it possible to limit the influence of strongly deviating updates. Another group of approaches includes coordinate-wise robust aggregation methods, such as Median and Trimmed Mean [8]. These methods reduce the influence of extreme parameter values before forming the aggregated update. In comparison with FedAvg, Byzantine-robust aggregation provides a higher level of protection against abnormal local updates.

At the same time, Byzantine-robust aggregation methods mainly operate within a single training round. They can reduce the impact of suspicious updates during aggregation, but they do not usually accumulate information about the behavior of each participant across several rounds. As a result, these methods may not explicitly identify a participant that repeatedly sends suspicious updates. In addition, robust aggregation may sometimes discard useful updates from honest participants, especially when local data are highly heterogeneous.

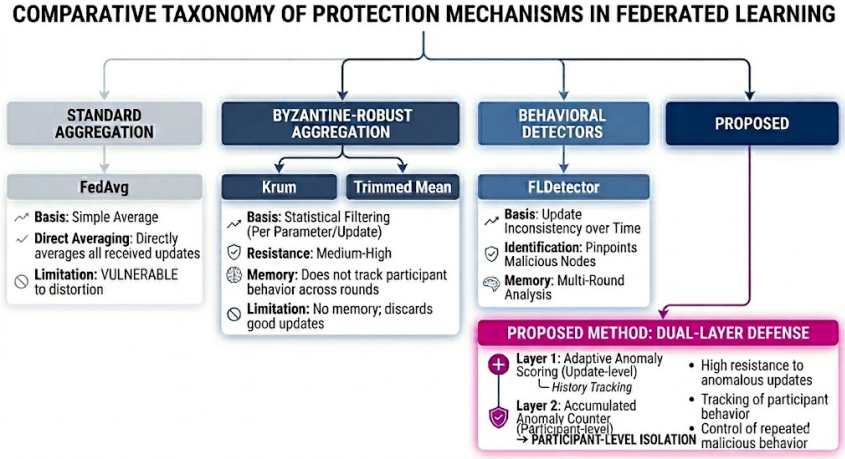

Fig. 1. Comparative taxonomy of protection mechanisms in federated learning systems

Behavioral detection approaches extend the protection logic by analyzing the consistency of participant behavior over time. For example, FLDetector [10] detects malicious clients by analyzing inconsistencies in their model updates across training rounds. This direction is especially important because malicious behavior in federated learning is not always expressed as a single strong deviation. A malicious participant may act gradua-lly and send moderately distorted updates over several rounds, making such behavior more difficult to detect using only one-round filtering.

The proposed method combines update-level anomaly detection with participant-level isolation. At the first level, each local update is evaluated using an anomaly score that reflects its deviation from the collective update pattern. At the second level, repeated abnormal behavior is accumulated using an anomaly counter. If a participant repeatedly generates anomalous updates, this participant can be isolated from further aggregation. Therefore, the method does not only reduce the influence of a suspicious update in one training round, but also limits the long-term influence of unreliable participants.

Thus, the proposed approach occupies an intermediate position between robust aggregation and malicious client detection. Unlike conventional aggregation methods, it does not assume that all participants are equally reliable. Unlike one-round robust aggregation methods, it tracks repeated anomalous behavior. Unlike purely update-filtering approaches, it provides an explicit participant-level isolation mechanism. This makes the method suitable for federated learning systems used in edge and fog environments, where participants may be heterogeneous, dynamically connected, and potentially vulnerable to compromise.

Conclusions

This paper proposes a method for detecting anomalous local updates and isolating malicious participants in federated learning systems. The method is based on the assumption that local updates generated by honest participants should remain relatively consistent with the collective direction of model training, while malicious or compromised participants may generate updates that significantly deviate from this pattern.

The proposed approach combines two protection levels. The first level is update-level anomaly detection, where each received local update is evaluated using an anomaly score. The second level is participant-level isolation, where repeated anomalous behavior is accumulated by an anomaly counter and used as a basis for excluding unreliable participants from further aggregation.

An illustrative numerical example was provided to demonstrate the logic of the proposed method. The example showed how an anomalous local update can be detected even when it shifts the reference update used for comparison. It also demonstrated that excluding the suspicious update allows the aggregated vector to remain close to the updates generated by honest participants.

The comparative analysis showed that the proposed method differs from conventional aggregation and Byzantine-robust approaches because it not only filters suspicious updates in a particular training round, but also tracks participant behavior over time. This makes the method potentially useful for federated learning systems operating in edge and fog environments, where participants may be heterogeneous, dynamically connected, and vulnerable to compromise.

At the same time, this study has a conceptual and analytical character. Full-scale simulation experiments with real or benchmark datasets were not conducted in the present work. Therefore, further research should include experimental evaluation of the proposed method under different levels of data heterogeneity, different numbers of malicious participants, and different values of the sensitivity coefficient $\lambda$and the isolation threshold $C_{\max}$. Such experiments will make it possible to estimate detection accuracy, false positive rate, false negative rate, and the influence of the proposed method on the convergence and final accuracy of the global model.

REFERENCES

[1] McMahan B., Moore E., Ramage D., Hampson S., Arcas B. Y. A. Communication-Efficient Learning of Deep Networks from Decentralized Data // Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). 2017. P. 1273–1282. URL: https://proceedings.mlr.press/v54/mcmahan17a.html

[2] Kairouz P. et al. Advances and Open Problems in Federated Learning // Foundations and Trends in Machine Learning. 2021. Vol. 14, No. 1–2. P. 1–210. DOI: 10.1561/2200000083.

[3] Li T., Sahu A. K., Talwalkar A., Smith V. Federated Learning: Challenges, Methods, and Future Directions // IEEE Signal Processing Magazine. 2020. Vol. 37, No. 3. P. 50–60. DOI: 10.1109/MSP.2020.2975749.

[4] Mothukuri V., Parizi R. M., Pouriyeh S., Huang Y., Dehghantanha A., Srivastava G. A Survey on Security and Privacy of Federated Learning // Future Generation Computer Systems. 2021. Vol. 115. P. 619–640. DOI: 10.1016/j.future.2020.10.007.

[5] Cui L., Suh S. C., Tan Y. et al. A Survey on Federated Learning for Cyber Security // IEEE Communications Surveys & Tutorials. 2024. Vol. 26,

No. 1. P. 565–596. DOI: 10.1109/COMST.2023.3326741.

[6] Blanchard P., El Mhamdi E. M., Guerraoui R., Stainer J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30. URL: https://proceedings.neurips.cc/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html

[7] Yin D., Chen Y., Kannan R., Bartlett P. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates // Proceedings of the 35th International Conference on Machine Learning (ICML). 2018. P. 5650–5659. URL: https://proceedings.mlr.press/v80/yin18a.html

[8] El Mhamdi E. M., Guerraoui R., Rouault S. The Hidden Vulnerability of Distributed Learning in Byzantium // Proceedings of the 35th International Conference on Machine Learning (ICML). 2018. P. 3521–3530. URL: https://proceedings.mlr.press/v80/mhamdi18a.html

[9] Fang M., Cao X., Jia J., Gong N. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning // Proceedings of the 29th USENIX Security Symposium. 2020. P. 1605–1622. URL: https://www.usenix.org/conference/usenixsecurity20/presentation/fang

[10] Zhang X., Hong M., Dhople S., Yin W., Liu Y. FedPD: A Federated Learning Framework with Adaptivity to Non-IID Data // IEEE Transactions on Signal Processing. 2021. Vol. 69. P. 6055–6070. DOI: 10.1109/TSP.2021.3115952.

[11] Kudrenko S., Nimych O., Makieiev I. Method for Predicting Node Compromise in Edge and Fog Environments for Critical Infrastructure // Information Protection. 2025. Vol. 27, No. 2. P. 87–95. DOI: 10.18372/2410-7840.27.21183.

[12] Kozlovsky V., Pavlov V., Kozlovska D., Kudrenko S. Development of Chain Models of Irregular Antiradiolocation Coatings // Information Protection. 2025. Vol. 27, No. 2. DOI: 10.18372/2410-7840.27.21176.

[13] Yang Q., Liu Y., Chen T., Tong Y. Federated Machine Learning: Concept and Applications // ACM Transactions on Intelligent Systems and Technology. 2019. Vol. 10, No. 2. P. 1–19. DOI: 10.1145/3298981.

Кудренко С. О., Німич О. В., Макєєв І. Г.

МЕТОД ВИЯВЛЕННЯ АНОМАЛЬНИХ ЛОКАЛЬНИХ ОНОВЛЕНЬ ТА ІЗОЛЯЦІЇ ЗЛОВМИСНИХ УЧАСНИКІВ У СИСТЕМАХ ФЕДЕРАТИВНОГО НАВЧАННЯ

У роботі запропоновано метод виявлення аномальних локальних оновлень та ізоляції зловмисних учасників у системах федеративного навчання. Метод орієнтований на підвищення стійкості розподілених моделей машинного навчання до атак типу model poisoning та передачі викривлених локальних оновлень. Запропонований підхід поєднує два рівні захисту: оцінювання аномальності локальних оновлень на основі їх відхилення від колективного шаблону оновлень та ізоляцію учасників, які повторно демонструють підозрілу поведінку. Для оцінювання локальних оновлень використовується показник аномальності, а для контролю довготривалої поведінки учасників — накопичувальний лічильник аномалій. У роботі наведено формалізацію запропонованого методу, описано механізм адаптивного визначення порогу аномальності та процедуру формування множини довірених учасників для подальшої агрегації глобальної моделі. Для демонстрації логіки роботи підходу наведено ілюстративний числовий приклад, який показує можливість виявлення аномального локального оновлення та зменшення його впливу на глобальну модель. Проведено порівняльний аналіз запропонованого підходу з існуючими методами агрегації та виявлення зловмисних учасників у системах федеративного навчання. Показано, що запропонований метод, на відміну від традиційних схем агрегації, забезпечує не лише фільтрацію підозрілих оновлень, а й контроль повторюваної аномальної поведінки учасників. Запропонований підхід може бути використаний у системах федеративного навчання для edge та fog середовищ, а також в інформаційних системах об'єктів критичної інфраструктури.

Ключові слова: виявлення аномалій, машинне навчання, глибоке навчання, штучний інтелект, кібербезпека, інформаційна безпека, кібератаки, кіберзагрози, критична інфраструктура, захист даних, нейронна мережа, розподілене навчання, edge computing, fog computing.

Kudrenko S., Nimych O., Makieiev I.

METHOD FOR DETECTING ANOMALOUS LOCAL UPDATES AND ISOLATING MALICIOUS PARTICIPANTS IN FEDERATED LEARNING SYSTEMS

This paper proposes a method for detecting anomalous local updates and isolating malicious participants in federated learning systems. The method is aimed at improving the resilience of distributed machine learning models to model poisoning attacks and distorted local updates. The proposed approach combines two protection levels: anomaly assessment of local updates based on their deviation from the collective update pattern and isolation of participants that

repeatedly demonstrate suspicious behavior. An anomaly score is used to evaluate local updates, while an accumulated anomaly counter is applied to control long-term participant behavior. The paper presents the formalization of the proposed method, describes the adaptive anomaly threshold mechanism, and defines the procedure for forming a trusted participant set for further global model aggregation. An illustrative numerical example is provided to demonstrate the operation of the proposed approach and to show the possibility of detecting anomalous local updates while reducing their influence on the global model. A comparative analysis of the proposed approach with existing aggregation and malicious participant detection methods in federated learning systems is also presented. It is shown that, unlike conventional aggregation schemes, the proposed method provides not only suspicious update filtering but also control of repeated anomalous participant behavior. The proposed approach can be applied in federated learning systems for edge and fog environments, as well as in information systems of critical infrastructure facilities.

Keywords: anomaly detection, machine learning, deep learning, artificial intelligence, cybersecurity, information security, cyberattacks, cyber threats, critical infrastructure, data protection, neural network, distributed learning, edge computing, fog computing.

Received: 18.04.2026 p.
Accepted: 18.05.2026 p.
Published: 28.05.2026 p.