Научная статья
DOI 10.66424/2071-8217-2026-2-10
УДК 004.04

ЗАЩИТА СИСТЕМ ФЕДЕРАТИВНОГО ОБУЧЕНИЯ AI/ML ОТ АТАК ОТРАВЛЕНИЯ

М. А. Полтавцева*, А. А. Васильева

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия

✉ *poltavtseva@ibks.spbstu.ru

ДЛЯ ЦИТИРОВАНИЯ

Полтавцева М. А., Васильева А. А. Защита систем федеративного обучения AI/ML от атак отравления // Проблемы информационной безопасности. Компьютерные системы. 2026. № 2. С. 121–137. DOI: 10.66424/2071-8217-2026-2-10

ПОСТУПИЛА 03.03.2026
ПРИНЯТА 27.04.2026
ОПУБЛИКОВАНА 15.06.2026

Издатель: Санкт-Петербургский политехнический университет Петра Великого

АННОТАЦИЯ

Системы федеративного обучения искусственного интеллекта подвержены атакам, позволяющим злоумышленнику изменить их поведение, как и обычные AI/ML решения. Наиболее эффективной является атака отравления. При этом защита систем федеративного обучения усложняется возможностью сговора между участниками. В таких условиях обнаруживать и предотвращать атаки становится особенно трудно. Решение этой задачи является целью работы. Исследование предлагает метод обеспечения защиты систем федеративного обучения от атак отравления с использованием сговора, основанный на комбинации известных и доказавших свою эффективность методов защиты. Выбранные методы фильтрации и надежной агрегации модифицированы для учета возможного сговора участников обучения. Корректность и эффективность предложенного метода подтверждается практическими экспериментами, позволяющими не только доказать результативность, но и выявить ограничения разработанного решения.

КЛЮЧЕВЫЕ СЛОВА

Информационная безопасность, искусственный интеллект, машинное обучение, цепочки поставок, атаки отравления

Original article
DOI 10.66424/2071-8217-2026-2-10

PROTECTION OF AI/ML FEDERATED LEARNING SYSTEMS FROM POISONING ATTACKS

M. A. Poltavtseva*, A. A. Vasilyeva

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

✉ *poltavtseva@ibks.spbstu.ru

FOR CITATION

Poltavtseva M. A., Vasilyeva A. A. Protection of AI/ML federated learning systems from poisoning attacks. *Problems of information security. Computer systems*.

ABSTRACT

Federated artificial intelligence learning systems are susceptible to attacks that allow an attacker to change their behavior, just like conventional AI/ML solutions. The most effective of such attacks today is the poisoning attack. At the same time, the protection of federated learning systems is complicated by the possibility of collusion between the

2026. No. 2, pp. 121–137.
DOI: 10.66424/2071-8217-2026-2-10
(In Russian)

RECEIVED 03.03.2026
ACCEPTED 27.04.2026
PUBLICATION 15.06.2026

participants. In such circumstances, it becomes especially difficult to detect and prevent attacks. The solution of this problem is the purpose of the presented work. The study suggests a method to ensure the protection of federated learning systems from poisoning attacks using collusion, based on a combination of known and proven protection methods. The selected methods of filtering and reliable aggregation have been modified to take into account possible collusion of the training participants. The correctness and effectiveness of the proposed method is confirmed by practical experiments, which make it possible not only to prove its effectiveness, but also to identify the limitations of the developed solution.

KEYWORDS

Information security, artificial intelligence, machine learning, supply chains, poisoning attacks

1. ВВЕДЕНИЕ

Технологии искусственного интеллекта и методы машинного обучения используются повсеместно, на данный момент уже более 45 % предприятий применяют ту или иную форму искусственного интеллекта в своей работе. Эти технологии открывают не только возможности, но и потенциальные риски. В таких системах существует риск нарушения безопасности и конфиденциальности.

Рост популярности моделей распределенного машинного обучения, в том числе с сохранением конфиденциальности, обусловлен удобством и эффективностью такого подхода в B2B решениях. Однако системы федеративного обучения искусственного интеллекта, призванные защитить приватность пользователей и конфиденциальность данных участников, сами подвержены наборам специфических атак [1]. Отдельной проблемой является проведение традиционных атак, доказавших свою эффективность, в условиях сговора двух и более участников процесса обучения общей модели.

Атаки на системы федеративного обучения AI/ML возможны всеми известными способами: через данные, модель, используемые программные пакеты, технические средства, использующиеся при создании и работе AI/ML-систем. Тем не менее одной из самых разнообразных и опасных остается атака отправления данных или модели [2]. Атаки, направленные на системы, использующие либо предварительно обученные модели, либо распределенные системы обучения, такие как федеративное обучение, являются наиболее новыми и опасными, так как в них не представляется возможным проверить сами данные, на которых происходило обучение.

Данная работа направлена на поиск и совершенствование методов обеспечения безопасности систем федеративного обучения от атаки отравления в условиях возможного сговора участников.

2. БЕЗОПАСНОСТЬ СИСТЕМ ФЕДЕРАТИВНОГО ОБУЧЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Можно выделить множество различных атак на системы федеративного обучения искусственного интеллекта [3, 4], но одной из наиболее эффективных остается атака отравления [5–7]. Систематизация характеристик атаки приведена в табл. 1.

Для федеративного обучения существует множество видов защиты, но основные способы защиты от атак отравления – это защита на основе доверия (фильтры) [8] и при помощи надежного агрегирования [9].

Методы защиты на основе доверия основаны на том, что результаты обучения, передаваемые серверу у вредоносных

клиентов, отличаются от доброкачественных. И в большинстве случаев при нахождении вредоносного клиента его влияние на модель в дальнейшем обнуляется [10]. Главный недостаток данного метода в том, что при увеличении количества вредоносных клиентов данные методы защиты становятся неэффективными.

Защита на основе надежной агрегации основана на проверке характеристик обновлений, присланных клиентами, и усреднении всех параметров на основе медианы, математическом ожидании и т.п. Недостатком данных методов является то, что зачастую влияние атакующего не удаляется полностью и в большинстве случаев точность глобальной модели понижается. Обзор основных методов защиты представлен в табл. 2.

В результате анализа принято решение разрабатывать систему защиты на основе двух методов различных категорий: фильтрации и надежной агрегации, для противодействия максимальному числу

Таблица 1 | Атака отравления в федеративном обучении AI/ML

Table 1 | Poisoning attack in AI/ML federated learning

Атака		Место внедрения	Особенности атаки	Последствия
Отравление	Данные	Через данные, поступающие на вход модели при обучении	При использовании данных от третьих лиц или при атаке на источник данных	Понижение точности, полное забывание модели, установка триггеров
	Модель	Через предварительно обученную модель	При использовании трансферного, аутсорсингового и федеративного обучения	

Таблица 2 | Методы защиты от атаки отравления в федеративном обучении AI/ML

Table 2 | Methods of protection against poisoning attacks in AI/ML federated learning

Защита	Тип защиты	Снижение точности модели, %
Защита, воздействующая на параметры моделей [11]	Модификация параметров модели	Менее 1
FL-WBC [8]	Надежное обучение	3–10
PELF [12]	Фильтр	Менее 1
FedDefender [13,14]	Фильтр	Менее 1
SignGuard [15, 16]	Надежная агрегация с фильтрацией	1–4
MultiKrum [17, 18]	Надежная агрегация	1–5
Trimmed Mean [19, 20]	Надежная агрегация	1–3
Centered clip [21]	Надежная агрегация	Менее 1
Защита на основе кластеризации (DnC) [22]	Надежная агрегация	1–9

атак отравления с использованием сговора. За основу дальнейшей работы взяты методы FedDefender и надежного агрегирования Centered Clipping, так как они являются наиболее эффективными и проработанными из рассмотренных.

3. МЕТОД ЗАЩИТЫ СИСТЕМ ФЕДЕРАТИВНОГО ОБУЧЕНИЯ ОТ АТАК ОТРАВЛЕНИЯ

В данном случае метод защиты разрабатывается для систем федеративного обучения, при следующей модели угроз: рассматривается атака отравления весами, где количество вредоносных клиентов меньше 50 %. При числе вредоносных клиентов больше этого порога обеспечение защиты уже не представляется возможным.

Идея разрабатываемой защиты заключается в объединении метода, основанного на доверии, со способом надежного агрегирования для снижения влияния атаки на точность модели. Методы FedDefender и Centered Clipping будут применяться при передаче на сервер именно весов обученных моделей клиентов, а не градиентов, что получались в каждую эпоху. Данный способ является наиболее применяемым на практике и при этом достаточно безопасным, так как не передается информация, при помощи которой можно восстановить данные клиента. Схема предложенного способа защиты приведена на рис. 1.

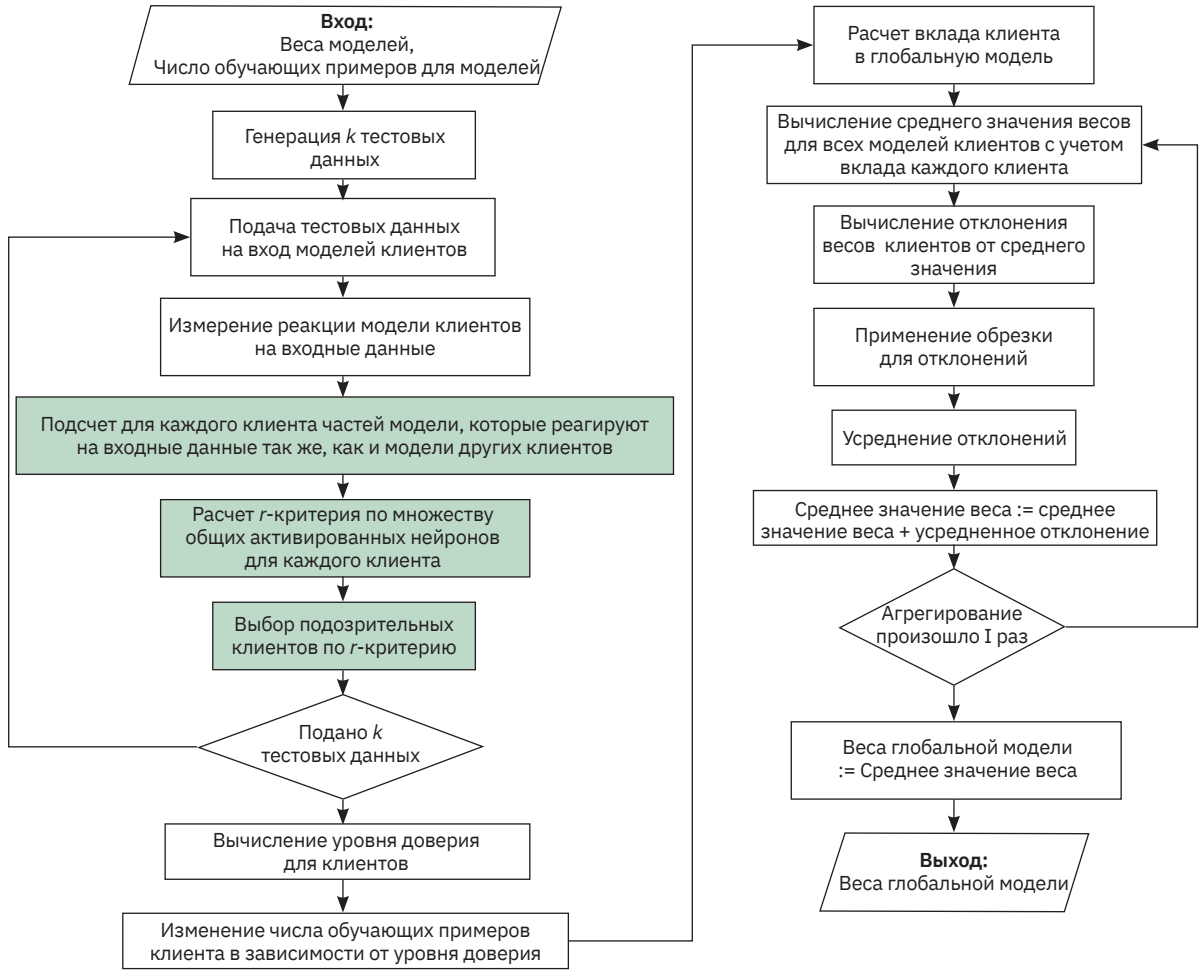

Рис. 1 | Схема предложенного способа защиты

Fig. 1 | Scheme of the proposed protection method

В предложенном способе защиты сначала происходит фильтрация клиентов на основе значения активированных нейронов. Для данного действия выполняется ряд шагов:

1. Получение тестовых данных, т.е. генерация *k* данных, подходящих для тестирования обучающейся модели.

2. Определение на каждом экземпляре тестовых данных для всех моделей клиентов активированных нейронов на каждом слое модели.

3. Определение общих активированных нейронов для каждого клиента с другими $n-1$ клиентами.

4. Определение вредоносности клиента на основе полученных данных для каждого тестового экземпляра данных.

5. Отсеивание клиента, если доверие к нему ниже 60 %, т.е. более чем по 40 % тестовых данных клиент признан вредоносным.

В данном способе защиты в четвертом пункте определяется вредоносность клиента. В работе данное действие предложено выполнить способом, основанным на исключении выбросов при неизвестном σ. Это делается при помощи *r*-критерия. Берется $x_1, \ldots, x_n$ характеризующие количество нейронов, активированных у клиента совместно с другими клиентами на одном и том же тестовом входе. Для этого проводится пересечение бинарных масок активации между всеми клиентами, затем для каждого клиента считается число нейронов, совпадающих с этим общим шаблоном. По ним вычисляется среднее значение:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad (1)$$

где n – количество клиентов; x_i – количество нейронов, активированных у клиента совместно с другими клиентами на одном и том же тестовом входе; $\overline{x}$ – среднее количество общих активированных нейронов на одного клиента.

Так же высчитывается стандартное отклонение:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}}. \quad (2)$$

Далее высчитывается *r*-критерий для значения каждого клиента. Расчет ведется по формуле

$$r_i = \frac{\left|x - \overline{x}\right|}{S\sqrt{\frac{n-1}{n}}}, \quad (3)$$

где S – стандартное отклонение.

По полученным значениям r_i определяется вредоносный клиент или нет, выяснение ведется только для клиентов, чьи значения общих активированных нейронов ниже, чем среднее значение $\overline{x}$. Клиент считается вредоносным на основе оценки следующих условий:

$$r_i > r\left(0{,}01; f = n-2\right), \quad (4)$$

$$r_i > r\left(0{,}05; f = n-2\right), \quad (5)$$

$$\left|r_i - r\left(0{,}05; f = n-2\right)\right| > \left|r_i - r\left(0{,}01; f = n-2\right)\right|, \quad (6)$$

где $r\left(0{,}05; f = n-2\right)$ и $r\left(0{,}01;\ f = n-2\right)$ являются табличными значениями и выбираются по уровню значимости и числу степеней свободы *f*, который рассчитывается с учетом числа клиентов; r_i – рассчитанный *r*-критерий для клиента *i*.

Вредоносным считается клиент, соответствующий условию (4) или условиям (5) и (6) совместно. В данном случае используется обновленный способ фильтрации, по сравнению с FedDefender, в логике определения вредоносного клиента. Вместо разбиения множества на подмножества из $n-1$ клиентов и нахождения из них доброкачественного выполняется определение клиентов с наибольшим количеством общих активированных нейронов.

После фильтрации применяется метод надежного агрегирования. Чтобы в способе агрегации были учтены результаты фильтра перед ним, делается подсчет коэффициентов для клиентов и нормализации их весов.

По результатам фильтра, изменяется число локальных примеров (num_examples), использованных клиентом при обучении, поэтому, если уровень доверия клиента меньше 0,6, то его num_examples = 0, в ином случае его num_examples = уровень_доверия * num_examples. Далее вычисляется общее число обучающих примеров по всем клиентам.

Для нормализации весов клиентов для каждого из них высчитывается параметр:

$$\alpha_j = \frac{\text{число локальных примеров клиента}}{\text{общее число обучающих примеров}}. \quad (7)$$

Вместо метода агрегации FedAvg и его модификаций [23, 24] в данной разработке использован Centered Clipping [25, 26], чтобы все атаки, которые проходят через фильтр, были либо нейтрализованы, либо смягчены. Centered Clipping выполнен с учетом входных данных в виде параметров обученных клиентских моделей, т.е. с учетом весов, а не с учетом градиентов моделей клиентов. В итоге модифицированный метод надежной агрегации включает несколько шагов.

Во-первых, получение центра полученных весов, т.е. вычисляется среднее от весов клиентов:

$$\overline{W_k} = \frac{1}{n}\sum_{j=1}^{n}\alpha_j W_j, \quad (8)$$

где W_j – параметры модели клиента; α_j – вес доверия клиента после фильтра.

Во-вторых, вычисление отклонения весов клиента от центра:

$$d_j = W_j - \overline{W_k}, \quad (9)$$

где $\overline{W_k}$ – центр усредненная модель на итерации k.

Далее применяется обрезка вида:

$$\overline{d_j} = \begin{cases} d_j & \text{если } \tau \geq \|d_j\| \\ \tau\dfrac{d_j}{\|d_j\|} & \text{иначе} \end{cases}, \quad (10)$$

где d_j – отклонение веса клиента j от центра; τ – коэффициент нормализации.

И усреднение отклонений, полученных после обрезки:

$$\bar{d} = \frac{1}{n}\sum_{j=1}^{n}\overline{d_j}, \quad (11)$$

где $\overline{d_j}$ – обрезанный вектор отклонения для клиента j.

В-третьих, проводится восстановление откорректированного веса:

$$\overline{W_{k+1}} = \overline{W_k} + \bar{d}, \quad (12)$$

где $\bar{d}$ – среднее значение обрезанных отклонений от центра.

Данный алгоритм повторяется I раз, т.е. $k = 1, \ldots, I$. Параметры K и τ назначаются разработчиком на основе предварительного анализа и подбора.

В итоге предложенный способ реализует синергию методов на основе доверия, в частности метода фильтрации, и методов агрегации, который в свою очередь основан на усреднении параметров, присылаемых клиентами серверу. Предложенный способ защиты предполагает, что количество атакующих может быть больше 1 и при этом меньше 50 % от общего количества клиентов и предполагает, что серверу от клиентов приходит только стандартная информация, а именно веса моделей и количество данных, на которых проведено обучение.

4. РАЗРАБОТКА ПРОГРАММНОГО СТЕНДА И ПОСТАНОВКА ЭКСПЕРИМЕНТА

Тестовый стенд реализован на языке Python, основными использованными библиотеками были tensorflow, Flower, torchvision и numpy. Сама система запускалась на ОС Linux с использованием CUDA. Для использования технологии распараллеливания CUDA установлен драйвер и необходимая библиотека для взаимодействия с ним. Архитектура развернутой системы федеративного обучения приведена на рис. 2.

Само обучение системы происходит по следующему сценарию:

- сервер отправляет клиенту параметры глобальной модели;
- клиент обновляет локальную модель параметрами, полученными от сервера;
- клиент обучает модель на локальных данных, что изменяет параметры модели локально;
- клиент отправляет обновленные/измененные параметры модели обратно на сервер;
- сервер получает данные с клиентов и объединяет их для получения параметров глобальной модели.

Все повторяется с шага (1) пока не будут пройдены все раунды (num_round) обучения.

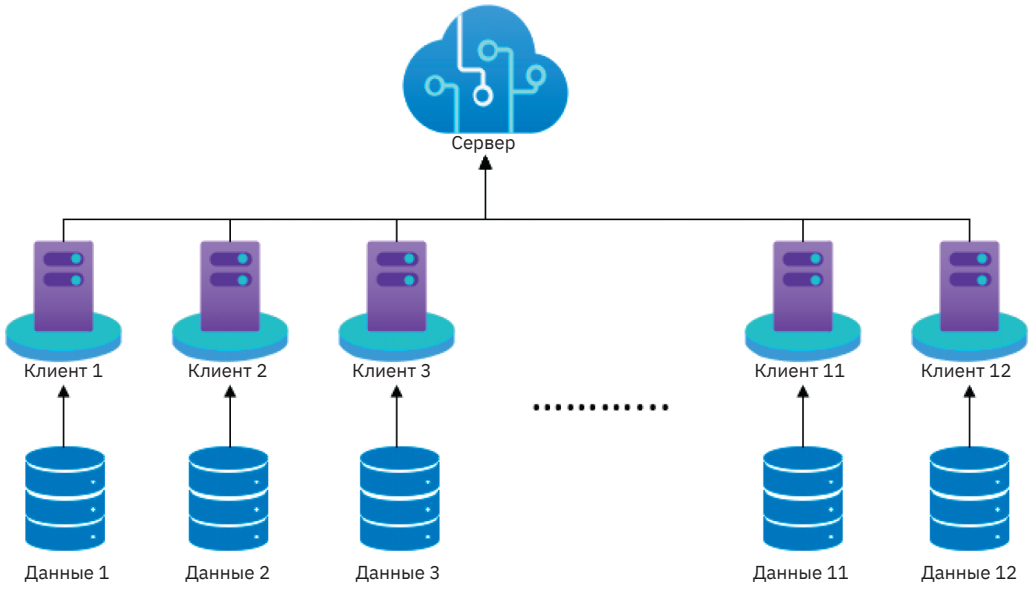

Рис. 2 | Развернутая система федеративного обучения

Fig. 2 | The expanded system of federated education

В самом коде запуска клиента создается образ локальной модели клиента, принимаются тестовые и обучающие данные для каждого клиента, проводятся атаки отравления для вредоносных клиентов и создается образ самого клиента и его запуск.

Вся симуляция федеративного обучения включает функции для инициализации клиентов, количество клиентов, участвующих в обучении, количество раундов, стратегия агрегации и задания места, где происходит обучение клиентов (GPU или CPU).

Для тестирования и обучения реализована CNN модель. Модель содержит шесть слоев. На первом слое создается шесть выходных каналов, используя ядро размером 5×5. На втором слое уменьшается размерность входных данных из первого слоя. Третий слой принимает шесть входных каналов и создает 16 выходных каналов с использованием ядра размером 5×5. На четвертом слое принимает входные данные с размерностью 16×5×5 и выводит 120 нейронов. Пятый слой принимает 120 входов и выводит 84 нейрона. Шестой слой принимает 84 входа и выводит 10 нейронов.

Реализовано несколько атак отравления для тестирования эффективности метода защиты. Первая атака – бэкдор-атака, которая реализуется изменением части существующих данных. В изображение вводится квадрат определенного цвета, например, белого, который незаметен человеческому глазу, но заметен компьютеру. Данный квадрат является триггером, при помощи которого злоумышленник может управлять действиями модели.

Второй атакой представлена одна из стандартных атак отравления – атака случайного переворота меток, не зависящая от вида модели. В ней при заданном обучающем наборе $\{x_i, y_i\} = 1$, где $x_i \in X$ и $y_i \in \{-1,1\}$, злоумышленник может случайным образом выбрать $[np]$ обучающих меток и перевернуть их.

Для предложенного способа защиты тестирование проводилось при 12 клиентах, 8 раундах и 10 эпох в раунде. Использовались датасеты MNIST и CIFAR10. Также проведен ряд тестов для определения наиболее подходящих параметров в фазе фильтрации и в фазе агрегации.

В методе агрегации есть два параметра это τ и I. Для подбора наиболее подходящих значений для них проведен ряд тестов. Сначала происходил подбор параметра τ, затем параметра I. При тестировании использовалась атака переворачивания

меток с двумя атакующими. Для сравнения параметров использовались значения точности, полученной в результате обучения модели. Подбор параметра τ велся на интервале от (0, 3], так как значения отклонения от центра в большинстве случаев не превышали значения 4 для используемых наборов данных, и после достижения значения $\tau = 2,5$ точность модели начинала падать. Все измерения проведены по три раза для получения точных результатов. На рис. 3 полученные результаты представлены в виде графиков.

По результатам тестирования сделан вывод, что данный параметр сильно зависит от данных. Его подбор необходимо

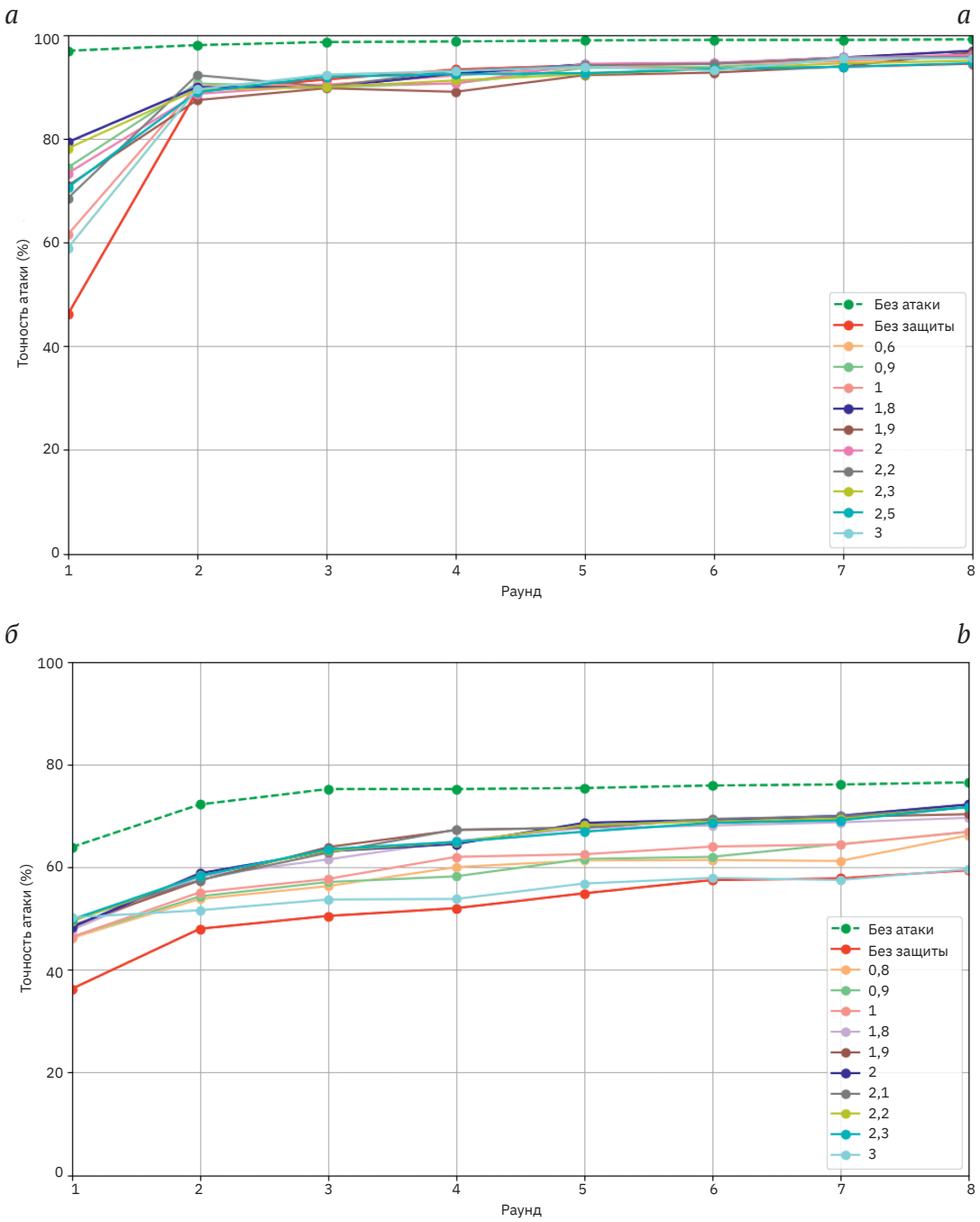

Рис. 3 | Подбор параметра τ для набора данных MNIST (*а*) и CIFAR10 (*б*)

Fig. 3 | Selection of the parameter τ for the MNIST (*a*) and CIFAR10 (*b*) data sets

вести в зависимости от значений отклонения от центра, т.е. от высчитываемого среднего. Замечено, что брать слишком маленькие значения, т.е. из интервала (0, 0,5], не имеет смысла, так как слишком сильно уменьшается влияние на модель не только вредоносных клиентов, но и нормальных. Наилучшими параметрами для датасетов MNIST и CIFAR10 оказались параметры из интервала (1, 2], а именно $\tau = 1{,}8$ и $\tau = 2{,}0$.

Подбор параметра I велся на интервале от [1, 10], так как далее изменения в работе алгоритма были не существенны. При этом параметр τ был равен для каждого датасета значениям, которые выбраны при помощи предыдущих тестов. На рис. 4

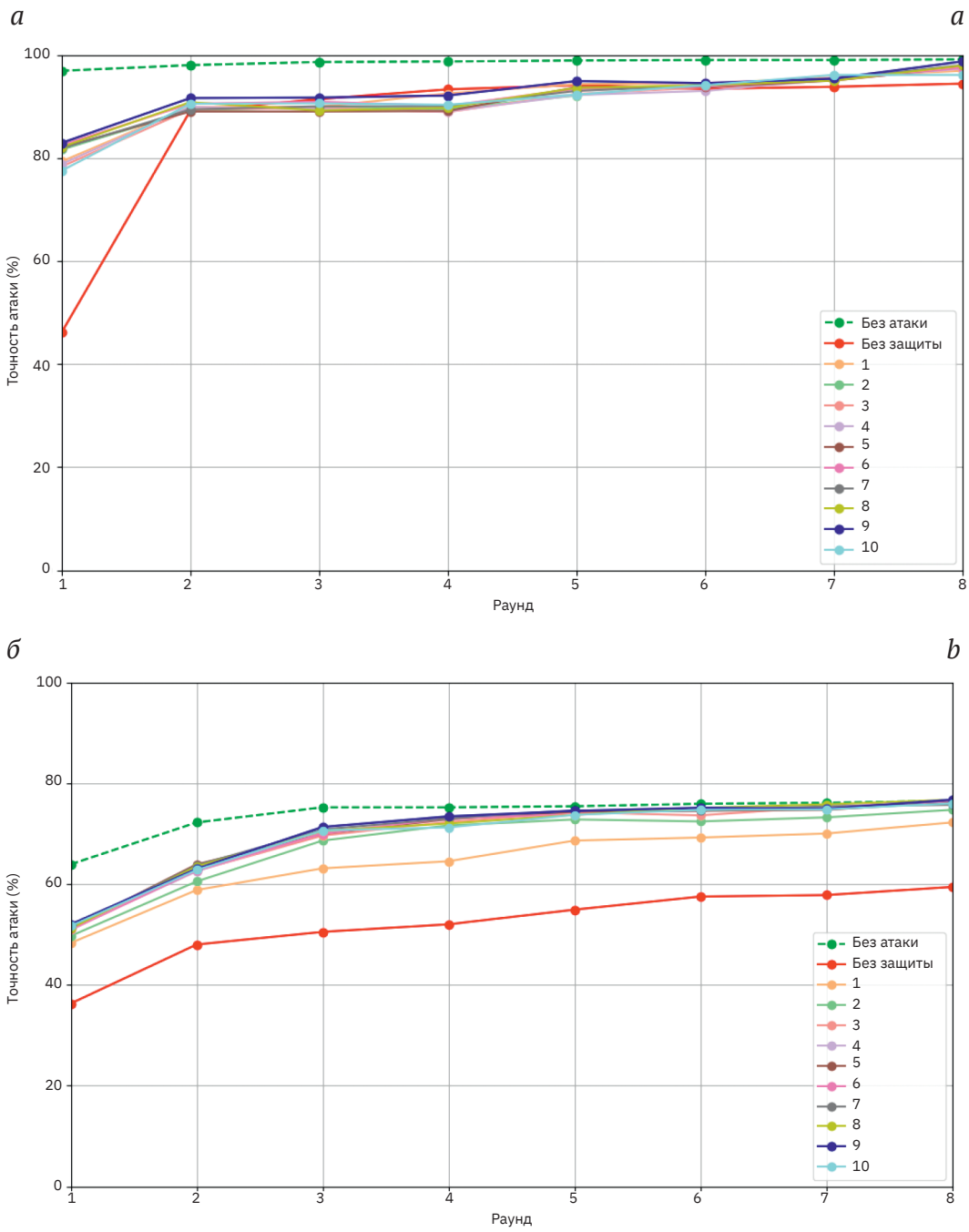

Рис. 4 | Подбор параметра I для набора данных MNIST (*а*) и CIFAR10 (*б*)

Fig. 4 | Selection of parameter I for the MNIST (*a*) and CIFAR10 (*b*) data sets

полученные результаты представлены в виде графиков.

По результатам тестирования сделан вывод, что данный параметр не зависит от данных. При его увеличении влияние атаки уменьшается, особенно это видно на результатах первого раунда. Но сильное увеличение данного параметра ведет к увеличению времени работы, затрачиваемого на агрегирование моделей клиентов. Для датасетов MNIST и CIFAR10 наилучшим параметром оказался $i = 9$.

5. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Проведено тестирование способа защиты. Сравнение велось со способами защиты, на которых основан разрабатываемый способ защиты и со случаем, когда защита не применялась вообще, также не отмечено эталонное значение точности, которое получается без проведения атаки.

Для датасета MNIST проведены тесты с 1–4 атакующими с двумя различными атаками – перевертывание меток и бэкдор-атака. Результаты представлены в виде графиках на рис. 5, на них приведено сравнение точности модели в каждом раунде при атаке перевертывания меток для MNIST для ситуаций, когда защиты нет, нет атаки, применяется предложенный способ защиты и Centered Clipping.

По полученным результатам видно, что предложенный способ защиты эффективен и снижает влияние атаки почти во всех случаях как минимум на 40 %.

Далее на графиках приведено сравнение точности модели в каждом раунде при бэкдор-атаке для MNIST, для ситуаций, когда защиты нет, нет атаки, применяется предложенный способ защиты и Centered Clipping. Результаты приведены на рис. 6.

Проведено сравнение точности бэкдор-атаки для MNIST для ситуаций, когда защиты нет, нет атаки, применяется предложенный способ защиты и Centered Clipping. На рис. 7 показаны графики точности атаки.

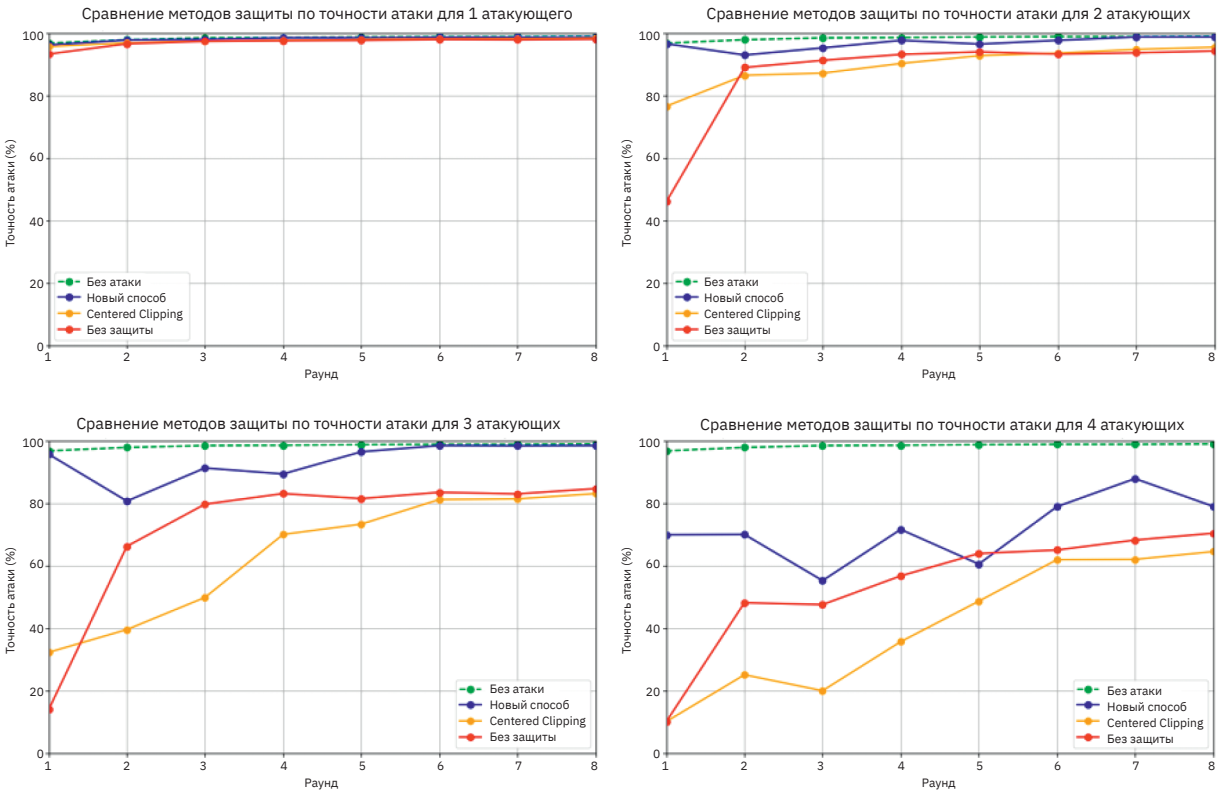

Рис. 5 | Графики точности модели для MNIST при атаке переворота метки

Fig. 5 | Graphs of model accuracy for MNIST under a label flip attack

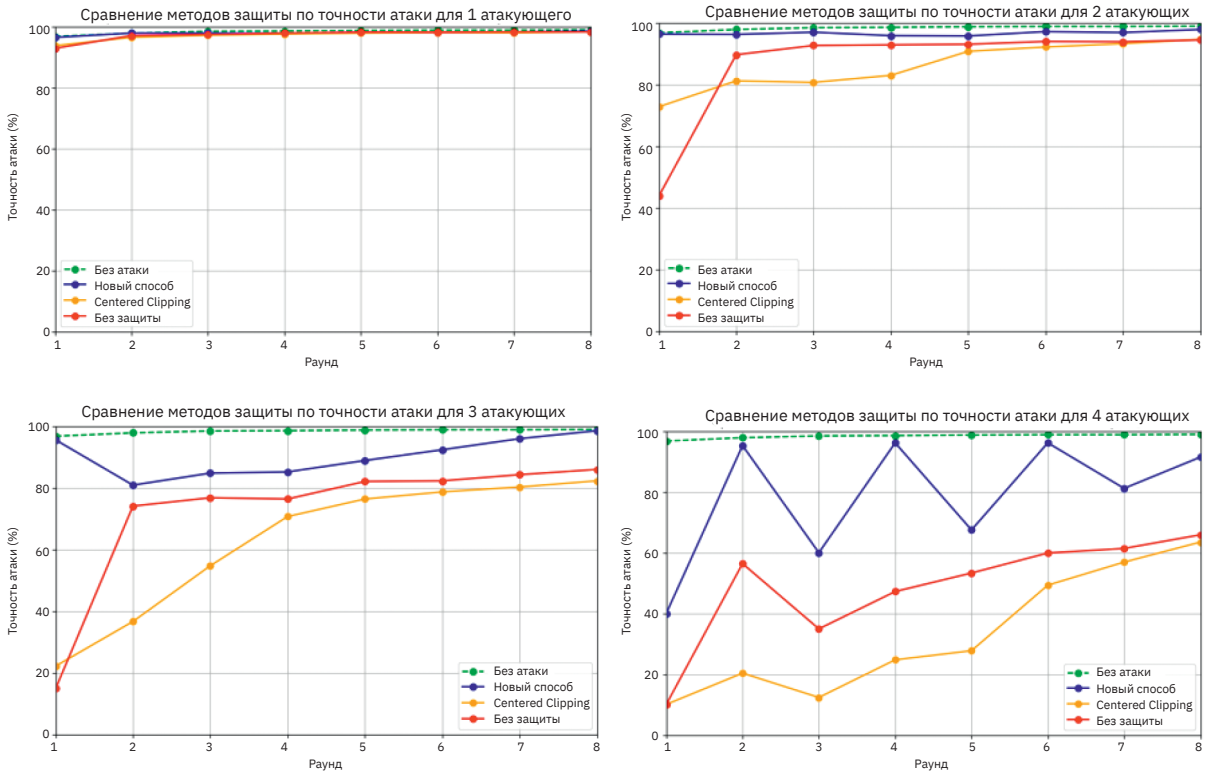

Рис. 6 | Графики точности модели для MNIST при бэкдор-атаке

Fig. 6 | MNIST model accuracy graphs during backdoor attack

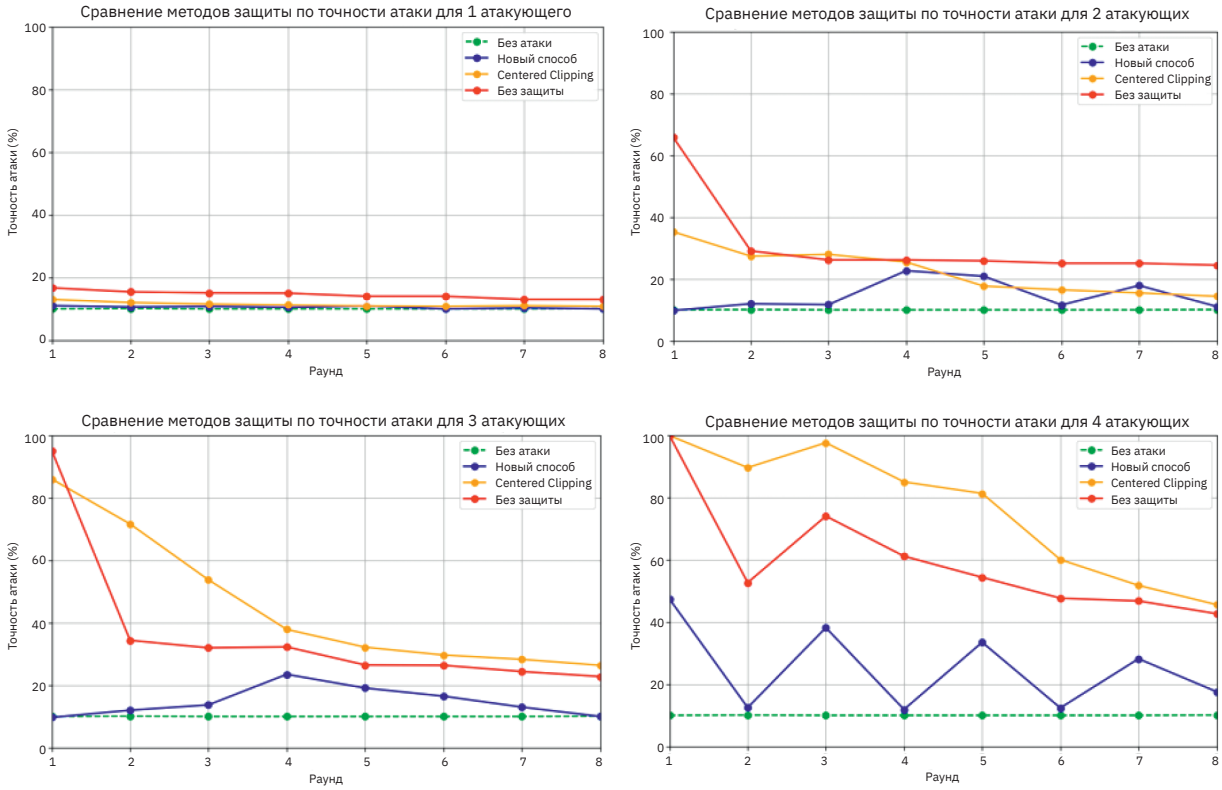

Рис. 7 | Графики точности атаки для MNIST при бэкдор-атаке

Fig. 7 | MNIST attack accuracy graphs for backdoor attacks

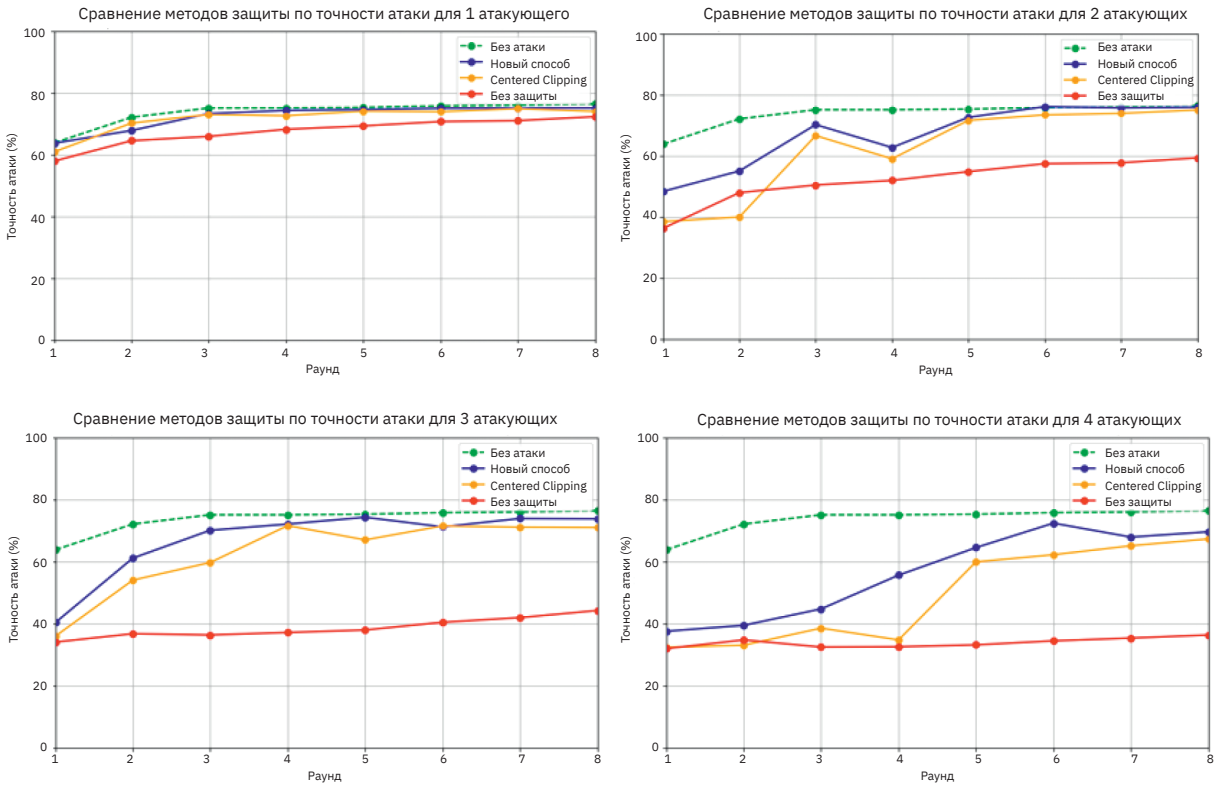

Рис. 8 | Графики точности модели для CIFAR10 при атаке переворота метки

Fig. 8 | Graphs of model accuracy for CIFAR10 during a label flip attack

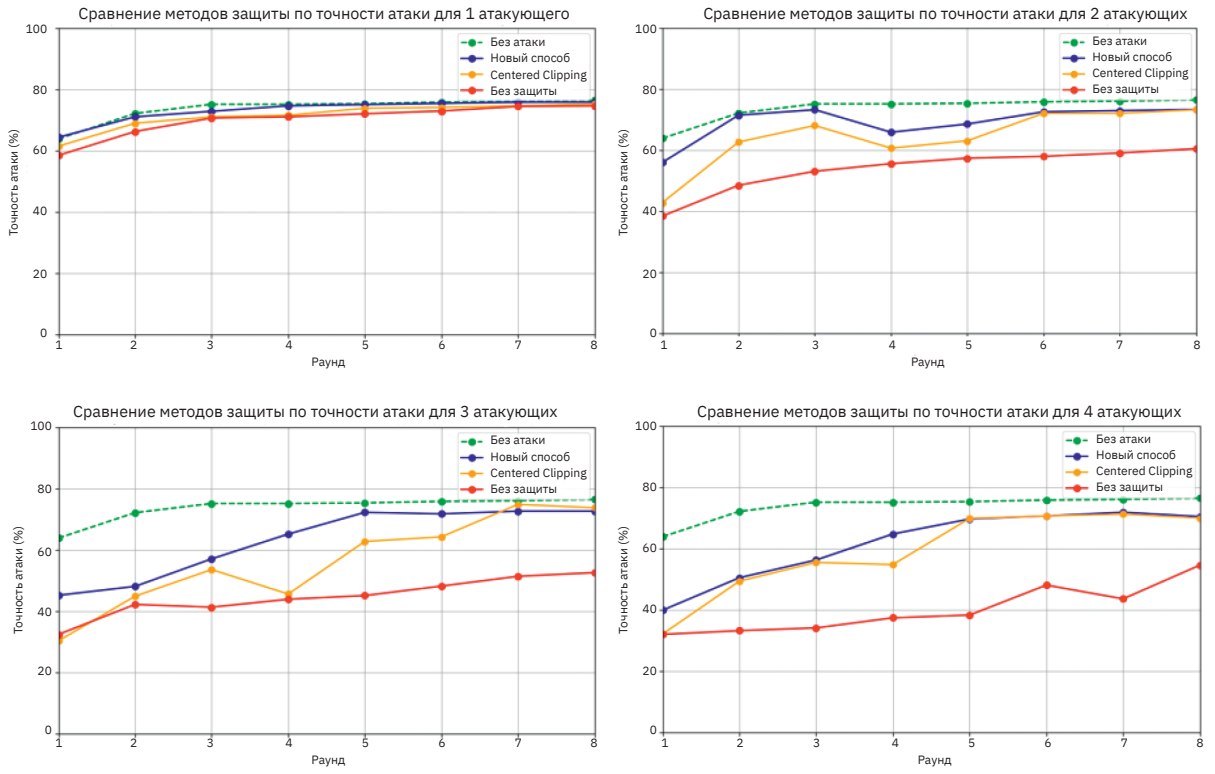

Рис. 9 | Графики точности модели для CIFAR10 при бэкдор-атаке

Fig. 9 | Graphs of model accuracy for CIFAR10 during backdoor attack

По полученным результатам можно сделать вывод, что предложенный метод почти во всех раундах снижает влияние атаки на точность модели как минимум на 30 %, при этом точность атаки не превышает значения 24 % для ситуаций, когда атакующих меньше четырех, и точность атаки не превышает значения 40 % при четырех атакующих, но при этом есть раунды, когда влияние атаки на точность модели не снижается.

Для набора данных CIFAR10 проведены тесты с 1–4 атакующими с двумя различными атаками перевертывание меток и бэкдор-атака. На графиках (рис. 8) приведено сравнение точности модели в каждом раунде при атаке перевертывания меток для CIFAR10 для ситуаций, когда защиты нет, нет атаки, применяется предложенный способ защиты и Centered Clipping.

По полученным результатам можно сделать вывод, что предложенный метод хорошо снижает влияние атаки на точность модели при атаке перевертывания модели. При количестве атакующих менее трех влияние атаки на точность модели снижается как минимум на 30 %, при четырех атакующих же влияние атаки на точность модели снижается более чем на 10 % во всех раундах. При этом при увеличении количества атакующих предложенный способ работает лучше, чем метод надежной агрегации Centered Clipping.

Далее на графиках (рис. 9) показано сравнение точности модели в каждом раунде при бэкдор-атаке для CIFAR10 для ситуаций, когда защиты нет, нет атаки, применяется предложенный способ защиты и Centered Clipping.

Для бэкдор-атаки получены графики для точности атаки. В них сравнивается точность бэкдор-атаки для CIFAR10 для ситуаций, когда защиты нет, нет атаки, применяется предложенный способ защиты и Centered Clipping (рис. 10).

По полученным результатам можно сделать вывод, что влияние атаки на точность модели снижается более чем на 27 %

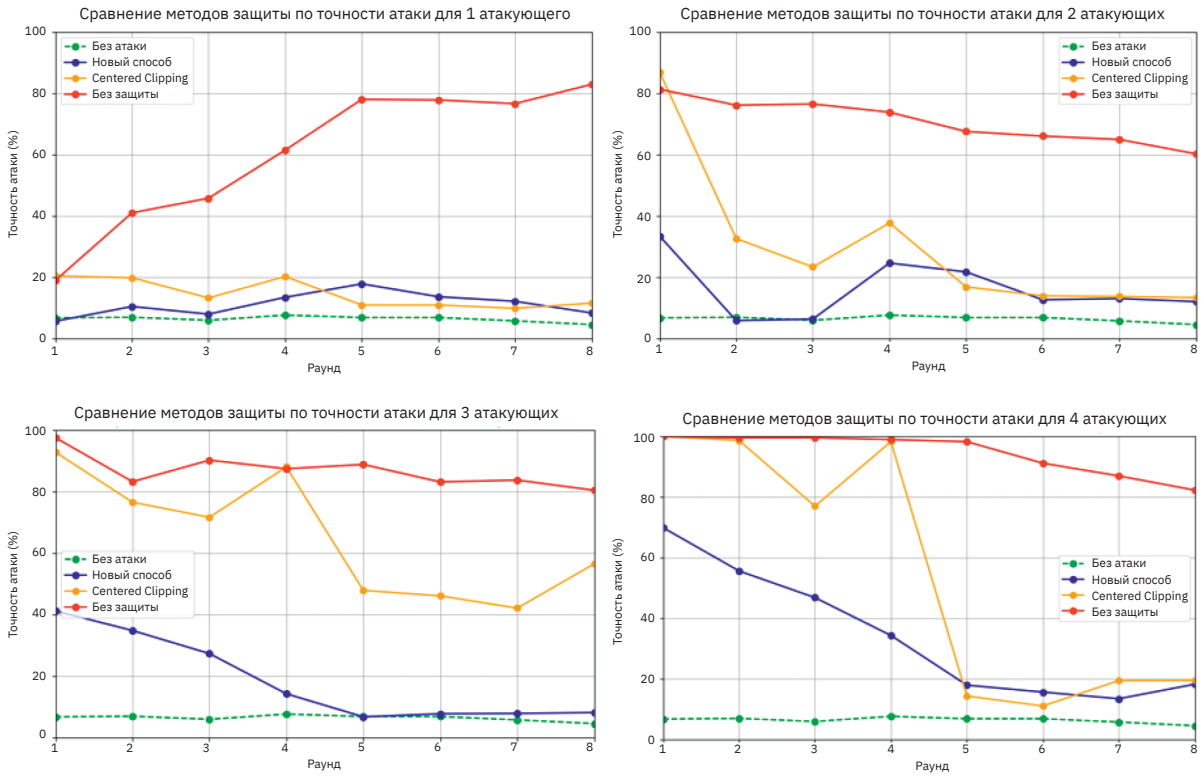

Рис. 10 | Графики точности атаки для CIFAR10 при бэкдор-атаке

Fig. 10 | Attack accuracy graphs for CIFAR10 in a backdoor attack

для ситуации, когда атакующих меньше трех, при четырех атакующих же влияние атаки снижается более чем на 22 % на всех раундах. Точность модели для предложенного метода и для Centered Clipping почти одинаковы, но при этом точность атаки для предложенного метода падает для любого количества атакующих достаточно сильно, как минимум на 20 %, когда при Centered Clipping точность атаки для ситуаций, когда количество атакующих больше одного, доходит до максимальных значений.

6. ЗАКЛЮЧЕНИЕ

При рассмотрении методов защиты от атак отравления изучены методы защиты, применяющиеся при возможности доступа к подозрительным данным и без доступа к ним. Также рассмотрены способы защиты от атаки отравления модели, которая возможна только при использовании предварительно обученных моделях или распределенных типах обучения. По результатам анализа атак и методов защит предложен метод защиты от атаки отравления для систем федеративного обучения, основанный на объединении модификаций известных методов. В предложенном решении использован метод фильтрации FedDefender, измененный для определения более чем одного вредоносного клиента, и метод надежной агрегации Centered Clipping, примененный не к градиентам, а к самим весам моделей.

Для тестирования данного способа разработан тестовый стенд, включающий в себя: систему федеративного обучения с сервером, которому от клиентов приходят только параметры обученных моделей и количество данных, на которых велось обучение; 12 клиентов; две реализованные атаки отравления – бэкдор-атака и атака переворота метки. Предложенное решение протестировано при различном числе атакующих на двух разных атаках. Выполнено сравнение с ситуациями, когда атака на систему не осуществляется, атака проводится, но защита от нее отсутствует или в защите применяется только Centered Clipping. На основе тестов сделан вывод, что предложенный метод эффективен против отравляющих атак. На представленных тестах он выполняет свою задачу не хуже отдельных методов, на которых он основан не смотря на модификации, а в ситуациях, когда атакующих много его эффективность взрастает по сравнению с исходными методами. Недостатком данного решения является рост числа ложных срабатываний при увеличении числа вредоносных клиентов.

КОНФЛИКТ ИНТЕРЕСОВ / CONFLICT OF INTERESTS

Авторы заявляют об отсутствии конфликта интересов / The authors declare no conflict of interests.

СПИСОК ИСТОЧНИКОВ

1. **Александрова Е. Б., Гадисова В. А.** Проблемы безопасности федеративных систем, использующих криптографическую защиту // Проблемы информационной безопасности. Компьютерные системы. 2025. № 4. С. 76–88. DOI: 10.48612/jisp/rp53-1tp9-n87g
2. **Безбородов П. Д., Лаврова Д. С.** Защита нейросетевых моделей от угроз нарушения конфиденциальности в федеративном обучении с использованием методов оптимизации // Проблемы информационной безопасности. Компьютерные системы. 2025. № 1. С. 21–29. DOI: 10.48612/jisp/fpvk-xpna-9hx5
3. **Kai Hu, Sheng Gong, Qi Zhang et al.** An overview of implementing security and privacy in federated learning // Artificial intelligence review. 2024. Vol. 57. № 8. P. 204.

4. **Hallaji E., Razavi-Far R., Saif M. et al.** Decentralized federated learning: A survey on security and privacy // IEEE Transactions on Big Data. 2024. Vol. 10. № 2. P. 194–213.
5. **Geming Xia, Jian Chen, Chaodong Yu, Jun Ma.** Poisoning attacks in federated learning: A survey // IEEE Access. 2023. Vol. 11. P. 10708–10722.
6. **Yazdinejad A., Dehghantanha A., Karimipour H. et al.** A robust privacy-preserving federated learning model against model poisoning attacks // IEEE Transactions on Information Forensics and Security. 2024. Vol. 19. P. 6693–6708.
7. **Kasyap H., Tripathy S.** Beyond data poisoning in federated learning // Expert Systems with Applications. 2024. Vol. 235. P. 121192.
8. **Jingwei Sun, Ang Li, DiValentin L. et al.** Fl-wbc: Enhancing robustness against model poisoning attacks in federated learning from a client perspective // Advances in Neural Information Processing Systems. 2021. Vol. 34. P. 12613–12624.
9. **Wei W., Liu L.** Trustworthy distributed AI systems: Robustness, privacy, and governance // ACM Computing Surveys. 2025. Vol. 57. № 6. P. 1–42.
10. **Крундышев В. М., Ческидов В. К., Калинин М. О.** Метод защиты глобальных моделей в системах федеративного обучения на основе модели доверия // Проблемы информационной безопасности. Компьютерные системы. 2024. № 4. С. 94–108. DOI: 10.48612/jisp/mf2n-fb13-p7p6
11. **Shammar E., Cui X., Al-qaness M. A. A.** Swarm learning: a survey of concepts, applications, and trends // arXiv preprint arXiv: 2405.00556. 2024.
12. **Ramirez M. A., Sangyoung Yoon, Damian E. et al.** New data poison attacks on machine learning classifiers for mobile exfiltration // arXiv preprint arXiv:2210.11592. 2022.
13. **Sungwon Park, Sungwon Han, Fangzhao Wu et al.** Feddefender: Client-side attack-tolerant federated learning // Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining. 2023. P. 1850–1861.
14. **Gill W., Anwar A., Gulzar M. A.** Feddefender: Backdoor attack defense in federated learning // Proceedings of the 1st International Workshop on Dependability and Trustworthiness of Safety-Critical Systems with Machine Learned Components. 2023. P. 6–9.
15. **Jian Xu, Shao-Lun Huang, Linqi Song et al.** Byzantine-robust federated learning through collaborative malicious gradient filtering // 2022 IEEE 42nd international conference on distributed computing systems (ICDCS). IEEE, 2022. P. 1223–1235.
16. **Chenghao Yang, Zhengchun Zhou, Yihuai Liang, Chunming Tang.** Sign-Based Privacy-Preserving and Communication-Efficient Federated Learning in Large-Scale Edge Computing // IEEE Transactions on Vehicular Technology. 2026.
17. **Zhang J., Li Q.** Federated Learning Against Dynamic Mixed Poisoning Attack and Defense // International Symposium on Cyberspace Safety and Security. Singapore: Springer Nature Singapore, 2025. P. 316–331.
18. **Xie Y., Fang M., Gong N. Z.** Model poisoning attacks to federated learning via multi-round consistency // Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. P. 15454–15463.
19. **Wang T., Zheng Z., Lin F.** Federated learning framework based on trimmed mean aggregation rules // Expert Systems with Applications. 2025. Vol. 270. P. 126354.
20. **Omer A. R., Khan M. S., Yousafzai A.** Robust federated learning: Defence against model poisoning using mean filtering // 2024 Horizons of Information Technology and Engineering (HITE). IEEE, 2024. P. 1–5.
21. **Wang X., Xia H., Zhang Y.** Defending against model poisoning attacks in federated learning via client-guided trust // 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2024. P. 1749–1755.
22. **Shejwalkar V., Houmansadr A.** Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning // Network and Distributed System Security Symposium. 2021. DOI: 10.14722/ndss.2021.24498.
23. **Xing S., Ning Z., Zhou J. et al.** N-FedAvg: Novel federated average algorithm based on FedAvg // 2022 14th International Conference on Communication Software and Networks (ICCSN). IEEE, 2022. P. 187–196.
24. **Mehta S., Aneja A.** Securing data privacy in machine learning: The fedavg of federated learning approach // 2024 4th Asian Conference on Innovation in Technology (ASIANCON). IEEE, 2024. P. 1–5.

25. **Yang C., Ghaderi J.** Byzantine-robust decentralized learning via remove-then-clip aggregation // Proceedings of the AAAI Conference on Artificial Intelligence. 2024. Vol. 38. № 19. P. 21735–21743.

26. **Partohaghighi M., Marcia R., West B. J., Chen Y. Q.** When Gradient Clipping Becomes a Control Mechanism for Differential Privacy in Deep Learning // arXiv preprint arXiv: 2602.10584. 2026.

REFERENCES

1. **Aleksandrova E. B., Gadisova V. A.** Security issues in federated learning systems. *Problems of information security. Computer systems*. 2025. No. 4, pp. 76–88. DOI: 10.48612/jisp/rp53-1tp9-n87g.
2. **Bezborodov P. D., Lavrova D. S.** Protecting neural network models from privacy violation threats in federated learning using optimization methods. *Problems of information security. Computer systems*. 2025. No. 1, pp. 21–29. DOI: 10.48612/jisp/fpvk-xpna-9hx5
3. **Kai Hu, Sheng Gong, Qi Zhang et al.** An overview of implementing security and privacy in federated learning. *Artificial intelligence review*. 2024. Vol. 57. No. 8, pp. 204.
4. **Hallaji E., Razavi-Far R., Saif M. et al.** Decentralized federated learning: A survey on security and privacy. *IEEE Transactions on Big Data*. 2024. Vol. 10. No. 2, pp. 194–213.
5. **Geming Xia, Jian Chen, Chaodong Yu, Jun Ma.** Poisoning attacks in federated learning: A survey. *IEEE Access*. 2023. Vol. 11, pp. 10708–10722.
6. **Yazdinejad A., Dehghantanha A., Karimipour H. et al.** A robust privacy-preserving federated learning model against model poisoning attacks. *IEEE Transactions on Information Forensics and Security*. 2024. Vol. 19, pp. 6693–6708.
7. **Kasyap H., Tripathy S.** Beyond data poisoning in federated learning. *Expert Systems with Applications*. 2024. Vol. 235, pp. 121192.
8. **Jingwei Sun, Ang Li, DiValentin L. et al.** Fl-wbc: Enhancing robustness against model poisoning attacks in federated learning from a client perspective. *Advances in Neural Information Processing Systems*. 2021. Vol. 34, pp. 12613–12624.
9. **Wei W., Liu L.** Trustworthy distributed AI systems: Robustness, privacy, and governance. *ACM Computing Surveys*. 2025. Vol. 57. No. 6, pp. 1–42.
10. **Krundyshev V. M., Cheskidov V. K., Kalinin M. O.** A protection method for the global model of the federated learning systems based on a trust model. *Problems of information security. Computer systems*. 2024. No. 4, pp. 94–108. DOI: 10.48612/jisp/mf2n-fb13-p7p6.
11. **Shammar E., Cui X., Al-qaness M. A. A.** Swarm learning: a survey of concepts, applications, and trends. *arXiv preprint arXiv:2405.00556*. 2024.
12. **Ramirez M. A., Sangyoung Yoon, Damian E. et al.** New data poison attacks on machine learning classifiers for mobile exfiltration. *arXiv preprint arXiv:2210.11592*. 2022.
13. **Sungwon Park, Sungwon Han, Fangzhao Wu et al.** Feddefender: Client-side attack-tolerant federated learning. Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining. 2023, pp. 1850–1861.
14. **Gill W., Anwar A., Gulzar M. A.** Feddefender: Backdoor attack defense in federated learning. Proceedings of the 1st International Workshop on Dependability and Trustworthiness of Safety-Critical Systems with Machine Learned Components. 2023, pp. 6–9.
15. **Jian Xu, Shao-Lun Huang, Linqi Song et al.** Byzantine-robust federated learning through collaborative malicious gradient filtering. 2022 IEEE 42nd international conference on distributed computing systems (ICDCS). IEEE, 2022, pp. 1223–1235.
16. **Jian Xu, Shao-Lun Huang, Linqi Song et al.** Sign-Based Privacy-Preserving and Communication-Efficient Federated Learning in Large-Scale Edge Computing. *IEEE Transactions on Vehicular Technology*. 2026.
17. **Zhang J., Li Q.** Federated Learning Against Dynamic Mixed Poisoning Attack and Defense. International Symposium on Cyberspace Safety and Security. Singapore: Springer Nature Singapore, 2025, pp. 316–331.
18. **Xie Y., Fang M., Gong N. Z.** Model poisoning attacks to federated learning via multi-round consistency. Proceedings of the Computer Vision and Pattern Recognition Conference. 2025, pp. 15454–15463.

19. **Wang T., Zheng Z., Lin F.** Federated learning framework based on trimmed mean aggregation rules. *Expert Systems with Applications*. 2025. Vol. 270, pp. 126354.
20. **Omer A. R., Khan M. S., Yousafzai A.** Robust federated learning: Defence against model poisoning using mean filtering. 2024 Horizons of Information Technology and Engineering (HITE). IEEE, 2024, pp. 1–5.
21. **Wang X., Xia H., Zhang Y.** Defending against model poisoning attacks in federated learning via client-guided trust. 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2024, pp. 1749–1755.
22. **Shejwalkar V., Houmansadr A.** Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. Network and Distributed System Security Symposium. 2021. DOI: 10.14722/ndss.2021.24498.
23. **Xing S., Ning Z., Zhou J. et al.** N-FedAvg: Novel federated average algorithm based on FedAvg. 2022 14th International Conference on Communication Software and Networks (ICCSN). IEEE, 2022, pp. 187–196.
24. **Mehta S., Aneja A.** Securing data privacy in machine learning: The fedavg of federated learning approach. 2024 4th Asian Conference on Innovation in Technology (ASIANCON). IEEE, 2024, pp. 1–5.
25. **Yang C., Ghaderi J.** Byzantine-robust decentralized learning via remove-then-clip aggregation. Proceedings of the AAAI Conference on Artificial Intelligence. 2024. Vol. 38. No. 19, pp. 21735–21743.
26. **Partohaghighi M., Marcia R., West B. J., Chen Y. Q.** When Gradient Clipping Becomes a Control Mechanism for Differential Privacy in Deep Learning. *arXiv preprint arXiv:2602.10584*. 2026.

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

ПОЛТАВЦЕВА Мария Анатольевна – д-р техн. наук, доцент, профессор, Санкт-Петербургский политехнический университет Петра Великого, Россия, 195251, Санкт-Петербург, ул. Политехническая, д. 29
E-mail: poltavtseva@ibks.spbstu.ru
ORCID: 0000-0001-9659-1244

POLTAVTSEVA Maria A. – Doctor of Engineering Sciences, Associate Professor, Professor, Peter the Great St. Petersburg Polytechnic University, Russia, 195251, St. Petersburg, Polytechnicheskaya str., 29

ВАСИЛЬЕВА Анастасия Александровна – студент, Санкт-Петербургский политехнический университет Петра Великого, Россия, 195251, Санкт-Петербург, ул. Политехническая, д. 29
E-mail: vamp.be.live@gmail.com
ORCID: 0009-0007-3203-6007

VASILYEVA Anastasia A. – Student, Peter the Great St. Petersburg Polytechnic University, Russia, 195251, St. Petersburg, Polytechnicheskaya str., 29