https://doi.org/10.31891/2307-5732-2025-359-77
УДК 004.457

ШИБАЄВ ГЕННАДІЙ
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»
https://orcid.org/0009-0009-3131-812X
e-mail: K233@ukr.net
ГАЛЬЧИНСЬКИЙ ЛЕОНІД
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»
https://orcid.org/0000-0002-3805-1474
e-mail: hleonid@gmail.com

ДОВІРО-ЗВАЖЕНА АГРЕГАЦІЯ У ФЕДЕРАТИВНОМУ НАВЧАННІ ДЛЯ МІКРОМЕРЕЖ

У цій статті представлено комплексну основу для підвищення стійкості федеративного навчання (FL) у кібербезпеці мікромереж шляхом впровадження нового механізму агрегації з довірою (TWA). У децентралізованих енергетичних інфраструктурах федеративне навчання дозволяє спільне навчання моделей прогнозування та виявлення аномалій на розподілених контролерах, зберігаючи при цьому конфіденційність даних. Однак присутність ненадійних або конфліктних учасників становить серйозну загрозу для стабільності глобального навчання. Зловмисні або шумні оновлення можуть спотворити агрегацію, затримати конвергенцію або навіть поставити під загрозу стійкість енергетичної системи. Щоб усунути цю вразливість, запропонований підхід інтегрує алгоритм динамічної оцінки довіри, який постійно оцінює надійність кожного клієнта за трьома критеріями: локальні втрати, відхилення ваг від глобальної моделі та оцінки аномалій на основі залишків. Оцінки довіри оновлюються після кожного раунду навчання з адаптивним спадом для покарання за непослідовну поведінку, і вони безпосередньо включаються до правила агрегації, гарантуючи, що ненадійні вузли зменшать вплив на глобальну модель.

Експериментальне дослідження проводиться на реальному наборі даних про споживання та генерацію енергії, де моделюється змішана сукупність клієнтів, включаючи чесні вузли з чистими даними, вузли з шумом, що постраждали від збурень, та вузли-змагачі, що вводять отруєні оновлення. Інфраструктура спирається на PyTorch та Flower для координації циклів навчання, тоді як обчислення довіри та оцінка аномалій вбудовані в легкий модуль, що виконується як локально, так і на сервері. Сервер агрегує оновлення за двома конкуруючими стратегіями — FedAvg та запропонованим TWA. Критерії оцінки включають середньоквадратичну помилку (RMSE), середню абсолютну помилку (MAE) та F1-оцінку для всіх циклів, з додатковою статистикою, такою як кінцева RMSE, середня RMSE та стандартне відхилення для вимірювання стабільності. Результати показують, що хоча FedAvg значно страждає від шуму-змагача, механізм TWA послідовно досягає меншої помилки, швидшої конвергенції та кращої стійкості до отруєння даних. Крім того, показники довіри динамічно змінюються, що дозволяє системі виявляти та зменшувати вагу шкідливих клієнтів протягом перших кількох циклів навчання. Це адаптивне придушення гарантує, що глобальна модель керується переважно надійними учасниками.

Внесок цієї роботи полягає в демонстрації того, що федеративне навчання, що усвідомлює довіру, може поєднувати конфіденційність, стійкість та адаптивність у кіберфізичних енергетичних системах. Поєднуючи оцінку довіри з виявленням аномалій та перехресною перевіркою на основі симуляцій цифрових двійників, запропонований підхід виходить за рамки статичних захистів та забезпечує самокоригувальну, самовідновлювальну федеративну екосистему. Ці результати відкривають шлях до безпечного та масштабованого федеративного інтелекту для інтелектуальних мереж, пропонуючи міцну методологічну та експериментальну основу для майбутніх досліджень у критично важливих для кібербезпеки областях.

Ключові слова:*федеративне навчання, агрегація з урахуванням довіри, безпека мікромереж, виявлення аномалій, змагальні клієнти, кіберфізичні системи*

SHYBAIEV HENNADII
HALCHYNSKY LEONID
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»

TRUST-WEIGHTED AGGREGATION IN FEDERATED LEARNING FOR MICROGRIDS

This article presents a comprehensive framework for enhancing the robustness of federated learning (FL) in microgrid cybersecurity by introducing a novel trust-weighted aggregation (TWA) mechanism. In decentralized energy infrastructures, federated learning enables collaborative training of predictive and anomaly detection models across distributed controllers while preserving data privacy. However, the presence of unreliable or adversarial participants poses a serious risk to the stability of global learning. Malicious or noisy updates may bias aggregation, delay convergence, or even compromise resilience of the energy system. To address this vulnerability, the proposed approach integrates a dynamic trust scoring algorithm that continuously evaluates each client's reliability using three criteria: local loss, weight divergence from the global model, and residual-based anomaly scores. Trust scores are updated after every training round with adaptive decay to penalize inconsistent behavior, and they are incorporated directly into the aggregation rule, ensuring that unreliable nodes exert less influence on the global model.

The experimental study is conducted on a real-world energy consumption and generation dataset, where a mixed population of clients is simulated, including honest nodes with clean data, noisy nodes affected by perturbations, and adversarial nodes injecting poisoned updates. The infrastructure relies on PyTorch and Flower to coordinate training rounds, while trust computation and anomaly scoring are embedded in a lightweight module executed both locally and on the server. The server aggregates updates under two competing strategies—FedAvg and the proposed TWA. The evaluation criteria include root mean squared error (RMSE), mean absolute error (MAE), and F1-score across all rounds, with additional statistics such as final RMSE, average RMSE, and standard deviation to measure stability. Results demonstrate that while FedAvg suffers significantly from adversarial noise, the TWA mechanism consistently achieves lower error, faster convergence, and superior resilience to data poisoning. Moreover, trust scores evolve dynamically, enabling the system to detect and down-weight malicious clients within the first few rounds of training. This adaptive suppression ensures that the global model is guided primarily by reliable participants.

The contribution of this work lies in demonstrating that trust-aware federated learning can combine privacy, resilience, and adaptivity in cyber-physical energy systems. By coupling trust scoring with anomaly detection and cross-validation against digital twin simulations, the proposed approach moves beyond static defenses and enables a self-correcting, self-healing federated ecosystem. These findings open the path toward secure and scalable federated intelligence for smart grids, offering a solid methodological and experimental foundation for future research in cybersecurity-critical domains.

Keywords: *federated learning, trust-aware aggregation, microgrid security, anomaly detection, adversarial clients, cyber-physical systems*

Стаття надійшла до редакції / Received 28.08.2025
Прийнята до друку / Accepted 15.11.2025

Постановка проблеми у загальному вигляді
та її зв'язок із важливими науковими чи практичними завданнями

Проблема виявлення аномалій, які є потенційними маркерами загроз є однією з центральних проблем в галузі кібербезпеки. Слід зазначити що, як прояви аномалій в комп'ютерних мережах так і методи їх виявлення мають надзвичайно широкий спектр [1]-[3]. Одним з новітніх підходів до виявлення аномалій є Федеративне навчання. Федеративне навчання стало перспективним рішенням для навчання моделей машинного навчання в розподілених середовищах, таких як інтелектуальні мікромережі, де обмеження конфіденційності та локальність даних обмежують використання централізованих методів. Однак децентралізована природа федеративного навчання створює критичну вразливість: продуктивність глобальної моделі дуже чутлива до якості та цілісності оновлень, отриманих від окремих клієнтів. У реальних розгортаннях деякі клієнти можуть постраждати від несправних датчиків, поганого з'єднання, дрейфу даних або навіть навмисної поведінки з боку сторонніх осіб. Ці ненадійні клієнти можуть значно погіршити точність, стабільність та конвергенцію глобальної моделі, підриваючи надійність прогнозних систем у критичній інфраструктурі, такій як енергетичні мережі.

Традиційні алгоритми ФН, такі як FedAvg, однаково обробляють усіх клієнтів під час агрегації, що робить їх за своєю суттю вразливими до отруєних оновлень та шумних внесків. Такий однаковий підхід не розрізняє надійні та ненадійні вузли та не пропонує вбудованого механізму для виявлення та пом'якшення їхнього впливу. Тому існує нагальна потреба в адаптивній, поведінково-орієнтованій стратегії агрегації, яка може динамічно оцінювати надійність клієнтів та вибірково включати лише достовірні оновлення до глобальної моделі. Це дослідження усуває цю прогалину, пропонуючи підхід до агрегації, зважений на довіру, який покращує стійкість та надійність у федеративному навчанні для кібербезпекових програм мікромереж.

Аналіз досліджень та публікацій

Федеративне навчання (ФН) стало широко прийнятою парадигмою розподіленого машинного навчання зі збереженням конфіденційності, особливо в областях, де дані не можуть бути централізовані через обмеження конфіденційності, нормативних актів або пропускної здатності. З моменту основоположної роботи над FedAvg [1], все більше досліджень присвячено застосуванню ФН у різних критичних секторах, включаючи охорону здоров'я, фінанси та все частіше енергетичні системи. У контексті розумних мікромереж — кіберфізичних систем, що складаються з розподілених енергетичних ресурсів, розумних лічильників та блоків керування — ФН пропонує потенціал для спільного навчання прогнозних моделей між локальними контролерами мережі без обміну необробленими даними датчиків.

У нещодавніх публікаціях почали досліджувати ФН для прогнозування енергоспоживання, управління попитом та виявлення аномалій в інтелектуальних мережах. Наприклад, ФН використовувався для прогнозування споживання енергії в житлових будинках, зберігаючи при цьому конфіденційність користувачів [2]. Аналогічно, ФН був інтегрований з периферійними обчисленнями, щоб забезпечити масштабоване керування розподіленими системами накопичення енергії [3]. Ці дослідження демонструють доцільність ФН в енергетичних системах, але зазвичай припускають доброякісну поведінку клієнтів і не розглядають проблеми надійності, що виникають через шумні або шкідливі вузли. Зокрема це стосується так званих візантійських атак, тобто атак, в яких декілька скомпрометованих вузлів у розподіленій системі координують свої дії з метою найбільш ефективного нанесення шкоди мережі. Захист від цих атак суттєво ускладнений, бо поведінка цих вузлів є непередбачуваною.

Для вирішення проблем стійкості у ФН було запропоновано кілька підходів. Візантійсько-стійкі методи агрегації, такі як Krum, Trimmed Mean та Bulyan, спрямовані на захист від зловмисних клієнтів шляхом фільтрації або зважування оновлень клієнтів на основі статистичних властивостей [4]. Однак більшість цих методів працюють за сильних припущень, таких як відома частка візантійських користувачів або розподіл даних IID, які не виконуються в середовищах мікромереж, що характеризуються часовими залежностями та неоднорідністю даних. Більше того, вони часто ігнорують контекстну або поведінкову історію клієнтів, розглядаючи кожну заявку раунду окремо.

Довіро-орієнтоване федеративне навчання стало перспективним напрямком для подолання цих обмежень. У кількох дослідженнях запропоновано включити механізми довіри або репутації до конвеєра ФН для зважування клієнтів на основі їхньої історичної поведінки [5][6]. Ці методи базуються на попередніх роботах з моделювання довіри в бездротових сенсорних мережах та багатоагентних системах. Однак існуючі підходи часто визначають довіру евристично або виключно на основі функцій втрат, без інтеграції багатовимірних поведінкових показників. Крім того, багато з цих моделей є теоретичними та не були перевірені на реальних наборах даних або в операційному контексті кіберфізичних систем, таких як мікромережі.

У сфері кібербезпеки мікромереж, ФН було запропоновано як механізм збереження конфіденційності для виявлення аномалій та прогнозного обслуговування. Однак дослідження в цій галузі все ще перебувають на початковому етапі. Більшість робіт зосереджені або на централізованому машинному навчанні для виявлення несправностей, або на статичних механізмах безпеки на основі правил [7][8]. Дуже мало робіт намагаються поєднати децентралізоване навчання з динамічною оцінкою довіри в чутливій до безпеки енергетичній інфраструктурі. Більше того, жодне з існуючих досліджень не інтегрує шумні та змагальні умови даних у сценарії федеративного прогнозування енергії з адаптивною до довіри агрегацією.

Наукова новизна цієї статті полягає в розробці та оцінці динамічного механізму агрегації, зваженої на довіру, спеціально розробленого для прогнозування мікромереж на основі ФН в умовах змагання та шуму[9]. На відміну від попередніх робіт, запропонований метод обчислює оцінки довіри клієнтів на основі сукупності локальних показників продуктивності – втрат, дельти ваг та оцінки аномалій – отриманих під час кожного раунду навчання. Ці оцінки довіри змінюються з часом за допомогою експоненціального згладжування та спаду, що дозволяє системі адаптивно карати за нестабільну або зловмисну поведінку без попереднього знання типів клієнтів. Механізм довіри бездоганно інтегрований у цикл навчання ФН, а його ефективність ретельно перевіряється за допомогою емпіричних експериментів на реальному наборі даних про енергетику. Цей цілісний підхід заповнює вирішальну прогалину в літературі, поєднуючи надійне об'єднання моделей, оцінку клієнтів з урахуванням аномалій та надійне навчання в кіберфізичних енергетичних середовищах.

Формулювання цілей статті

Метою роботи є: Метою цієї статті є розробка, впровадження та емпірична оцінка механізму агрегації з урахуванням довіри в рамках федеративної системи навчання для підвищення безпеки, стійкості та точності прогнозних моделей у середовищах мікромереж. Зокрема, дослідження спрямоване на усунення вразливості федеративних систем навчання до ненадійної або агресивної поведінки клієнтів шляхом впровадження алгоритму динамічної оцінки довіри, який враховує показники продуктивності на стороні клієнта, такі як втрати, розбіжності ваг та показники виявлення аномалій. Інтегруючи цей механізм довіри в федеративний цикл навчання та застосовуючи його до даних про споживання енергії в реальному світі, стаття прагне продемонструвати, що адаптивне зважування клієнтів покращує стабільність конвергенції та пом'якшує вплив шумних або шкідливих вузлів, тим самим підвищуючи стійкість децентралізованого машинного навчання в критичній кіберфізичній інфраструктурі.

Виклад основного матеріалу

У цьому розділі представлено ґрунтовне, технічно обґрунтоване та розширене пояснення експериментальної структури, розробленої для оцінки запропонованої методології агрегації з довірою (TWA) у контексті федеративного навчання (ФН), застосованої до кібербезпеки мікромереж. Федеративне навчання, за своєю суттю, децентралізує процес навчання, розподіляючи завдання оптимізації моделі між клієнтськими пристроями (або вузлами), кожен з яких має власні локальні дані. Хоча ця архітектура підвищує конфіденційність та масштабованість, вона вносить критичні вразливості безпеки, особливо коли деякі з вузлів-учасників поводяться зловмисно або страждають від пошкодження даних. TWA вирішує цю проблему, призначаючи динамічні, залежні від поведінки вагові коефіцієнти довіри внеску кожного вузла під час фази агрегації моделі. Цей розділ призначений для використання як окремий вичерпний ресурс для розуміння кожного аспекту експерименту, від цілей проектування та архітектури системи до попередньої обробки даних, методології моделювання, математичного моделювання та інфраструктури впровадження.

Головною метою експерименту є кількісна оцінка ефективності агрегації на основі довіри у пом'якшенні негативного впливу ненадійних вузлів на точність, конвергенцію та стійкість глобальної моделі в умовах імітації мікромережі[10]. В основі експериментального дизайну лежить кілька основних гіпотез:

- Гіпотеза стійкості: Стандартні методи ФН, такі як FedAvg, вразливі до пошкоджених або низькоякісних оновлень. Інтеграція механізму динамічної довіри зменшить цю вразливість.
- Гіпотеза поведінкової диференціації: Оцінки довіри, засновані на поведінкових метриках (втрата, дельта ваги, оцінка аномалії), можуть точно розрізняти чесні, шумні та зловмисні вузли.
- Гіпотеза стабільності-конвергенції: зважування агрегації за показниками довіри підвищить стабільність конвергенції та точність моделі.
- Гіпотеза адаптивного придушення: Механізм довіри, включаючи розпад та історичне згладжування, адаптивно придушуватиме шкідливий внесок з часом.

Експериментальне середовище моделює розподілену мікромережу, що складається з десяти вузлів (клієнтів) та одного координуючого сервера[11]. Кожен вузол являє собою локалізований контролер мережі, відповідальний за управління та аналіз даних про споживання та генерацію електроенергії. Сервер працює як центральний агрегатор та координатор, обробляючи зв'язок, оцінку довіри та об'єднання моделей. Ролі вузлів диверсифіковані наступним чином:

- Вузли 0–7 (Чесні вузли): Ці вузли мають доступ до чистих, репрезентативних наборів даних та виконують легітимне локальне навчання.
- Вузол 8 (Шумний вузол): Вхідні дані цього вузла пошкоджені гаусовим шумом, що імітує вплив несправностей датчиків або збурень навколишнього середовища.
- Вузол 9 (Зловмисний вузол): Цей вузол вдається до конфронтаційної поведінки, включаючи зміну міток та маніпулювання вагами, щоб погіршити якість глобальної моделі.

Така конструкція дозволяє моделювати реальні сценарії, де не всі контролери мікромережі однаково надійні.

В експерименті використовується набір даних "Hourly Energy Demand Generation and Weather" від Kaggle[12]. Набір даних містить записи про попит та виробництво електроенергії з часовими мітками, а також метеорологічні показники (температуру та вологість) за чотирирічний період. Це ідеальний орієнтир для завдань прогнозування часових рядів, що стосуються роботи мікромереж.

Для підготовки данних до експеримету ми використовували наступний конвеєр попередньої обробки.

1. Імпутація відсутніх даних: У нашому експериментальному конвеєрі ми вирішуємо проблеми з відсутніми значеннями в наборі даних мікросітки за допомогою імпутації з прямим заповненням, також відомої як імпутація останнього спостереження, перенесена вперед. Цей метод замінює відсутні значення в момент часу ttt найновішим доступним значенням з моменту часу t−1:

$$x_t = x_{t-1} \text{ if } x_t \text{ is missing} \quad (1)$$

Цей підхід особливо підходить для енергетичних даних часових рядів, де фізична безперервність означає, що останнє записане вимірювання є дійсною короткостроковою оцінкою. Якщо пряме заповнення залишає початкові значення незаповненими (наприклад, на початку ряду), ми застосовуємо зворотне заповнення, щоб переконатися, що не залишилося пропущених значень. Цей крок є важливим для забезпечення послідовної нормалізації ознак та запобігання збоям навчання в клієнтах федеративного навчання. Імпутація гарантує, що всі вузли працюють з повними, числово достовірними вхідними даними, що є необхідною умовою для змістовної оцінки довіри та конвергенції моделі.

2. Нормалізація: У нашому конвеєрі попередньої обробки ми застосовуємо нормалізацію ознак, щоб забезпечити однаковий внесок усіх вхідних змінних у процес навчання. Кожен вектор ознак $x \in R^d$, де d - d – кількість вхідних вимірів, нормалізована за формулою стандартного балу (Z-балу):

$$x_{norm} = \frac{(a - u)}{\sigma} \quad (2)$$

Тут u – це емпіричне середнє значення, а σ – стандартне відхилення ознаки в наборі даних. Це перетворення призводить до розподілу ознак приблизно з нульовим середнім значенням та одиничною дисперсією. Воно усуває розбіжності в масштабі та дисперсії між різними вхідними ознаками (наприклад, порівняння 5 кВт·год з 80% вологістю). Ми реалізуємо це перетворення за допомогою утиліти StandardScaler у бібліотеці scikit-learn Python. Нормалізована матриця ознак потім розділяється та призначається федеративним клієнтам. Виконання цього кроку забезпечує послідовне та стабільне навчання на всіх вузлах, що особливо важливо для виконання оцінювання довіри та конвергенції моделі в умовах федеративного навчання. Без нормалізації процес навчання страждав би від зміщених градієнтів, тривалішого часу збіжності та числової нестабільності.

3. Сегментація: Щоб емулювати децентралізоване навчання серед незалежних контролерів мікромережі, ми застосовуємо сегментацію набору даних, де повний попередньо оброблений набір даних поділяється на непересічні розділи, кожен з яких призначений окремому клієнту федеративного навчання. Припустимо, що вихідний набір даних складається з n зразків $x \in R^{(n\,x\,d)}$ з відповідними цільовими об'єктами $y \in R^n$ і ми хочемо змоделювати K клієнтів. Тоді кожен клієнт i призначається суцільний блок даних, визначений як:

$$X(i) = X[i \cdot B : (i+1) \cdot B], y(i) = y[i \cdot B : (i+1) \cdot B] \quad (3)$$

де $B = \lfloor \frac{n}{k} \rfloor$, тобто розмір блоку.

Цей процес є основоположним для всіх наступних етапів експерименту, включаючи локальне навчання, оцінку довіри та впровадження аномалій.

4. Ін'єкція аномалії: Щоб оцінити стійкість нашої системи федеративного навчання в реальних умовах, ми застосовуємо методи впровадження аномалій для імітації наявності ненадійних або агресивних клієнтів. Це дозволяє нам перевірити, чи може механізм оцінювання довіри ефективно зменшити вплив таких клієнтів під час агрегації.

Ми вводимо такі типи аномалій:

- Гаусівський шум у вхідних функціях: імітує несправні датчики або шум вимірювання в одному або кількох клієнтах (наприклад, Node 8):

$$x^{noisy} = x + \varepsilon, \varepsilon \sim N(0, \sigma 2) \quad (4)$$

- Перегортання міток у навчальних цілях: імітує отруєння міток шкідливим вузлом (наприклад, вузлом 9):

$$y_i{}^{flipped} = 1 - y_i \quad (5)$$

- Маніпуляції градієнтом/вагою під час надсилання параметрів: імітує отруєння моделі, коли зловмисний клієнт надсилає навмисно оманливі оновлення:

$$w^{adw} = 2w_{prev} - w_{new}, \quad (6)$$

де w - ваги моделі, параметри нейронної мережі, яка навчається під час федеративного навчання.

Ці ін'єкції застосовуються вибірково для імітації різних ролей вузлів (чесних, шумних, зловмисних). Очікується, що механізм довіри ідентифікуватиме ці вузли на основі їхньої поведінки (наприклад, високий бал аномалії, нестабільні оновлення), знизить їхні бали довіри та зрештою мінімізує їхній вплив на глобальну модель.

Цей підхід не лише перевіряє ефективність агрегації, зваженої на довіру, але й підтверджує здатність системи працювати в частково скомпрометованих або деградованих умовах — критична вимога для реального розгортання в кіберфізичних системах, таких як мікромережі.

Кожен сегмент був призначений одному федеративному вузлу. Ця сегментація імітувала географічно розподілену систему мікромережі, де кожен клієнт контролює власну часово-розподілену підмножину глобального набору даних. Клієнти від 0 до 7 були визначені як чесні клієнти, використовуючи незмінні, нормалізовані набори даних. Клієнт 8 був класифікований як такий, що містить шум. Клієнт 9 було позначено як шкідливий. Його мітки були змінені з 0 на 1 або навпаки, щоб імітувати поведінку з боку зловмисника, тим самим намагаючись пошкодити процес навчання глобальної моделі.

Центральний федеративний сервер було реалізовано за допомогою утиліти start_server() від Flower. Після ініціалізації сервер створив глобальну модель, визначену як двошаровий MLP, що складається з 4-вимірного вхідного шару, прихованого шару з 32 нейронами, активованими ReLU, та 1-вимірного виходу, що прогнозує потребу в енергії на наступну годину. Ваги цієї моделі були ініціалізовані випадковим чином. Крім того, сервер ініціалізував таблицю оцінки довіри, присвоївши кожному клієнту значення 1,0. Потім сервер серіалізував ваги моделі в масиви, сумісні з JSON, і надсилав їх усім 10 клієнтам. Кожен раунд зв'язку включав метадані, такі як номер раунду, розмір пакета, швидкість навчання та очікуваний формат виводу. Приклад такого JSON-повідомлення виглядає наступним чином: “{ "round": 5, "model_weights": [...], "learning_rate": 0.001, "batch_size": 32 }”.

Отримавши це корисне навантаження, кожен клієнт десеріалізував ваги та відновлював глобальну модель у локальній пам'яті. Перед навчанням клієнт розділив свій набір даних на навчальний та тестовий набори, використовуючи співвідношення 80/20. Такий поділ дозволив вузлу оцінити узагальнення своєї моделі після навчання.

Потім кожен клієнт навчав локальну модель на своїх даних протягом однієї епохи. Навчання проводилося з використанням функції втрат середньоквадратичної помилки (MSE) зі стохастичною градієнтною оптимізацією спуску. Після завершення навчання клієнт зберігав отримані ваги як w_new та обчислював три ключові показники поведінки:

- Loss: Ця метрика представляє середню похибку прогнозування для локального набору даних клієнта. Вона розраховується за допомогою середньоквадратичної похибки:

$$\ell = L(y, y') = \frac{1}{n}\sum_{i=1}^{n} (y_i - y_{i'})^2 \quad (7)$$

де y_i - справжня мітка, y' - це прогноз моделі, n - кількість зразків. Менші втрати свідчать про кращу точність прогнозування.

- Дельта ваги(weight delta) - вимірює, наскільки суттєво змінилася локальна модель порівняно з попереднім раундом. Це кількісно визначає дрейф параметрів моделі та обчислюється як L2-норма різниці ваг:

$$\Delta w = \left|\left|w^t - w^{(t-1)}\right|\right| = \sqrt{\sum_k (w_k{}^t - w_k{}^{(t-1)})^2} \quad (8)$$

Велике значення Δw може свідчити про нестабільне навчання, оновлення з боку суперників або вплив зашумлених даних.

- Оцінка аномалії: Цей показник відображає середню абсолютну похибку прогнозування, також відому як середня абсолютна похибка (MAE):

$$A = \frac{1}{n}\sum_{i=1}^{n} |y_i - y_{i'}| \quad (9)$$

Він служить додатковим показником стабільності клієнта та допомагає виявляти викиди. На відміну від MSE, він лінійно обробляє всі помилки, що може виявити різні моделі аномалій.

Ці три метрики об'єднуються з оновленими вагами клієнта та передаються на сервер для оцінки довіри та агрегації у форматі JSON: “{"model_weights": [...], "loss": 0.087, "delta": 1.453, "anomaly_score": 0.101 }”.

У федеративному навчанні, особливо в змагальних або напівнадійних середовищах, таких як мікромережі, оцінка довіри є критичним механізмом для зважування внеску кожного вузла в глобальну модель. Її метою є зменшення впливу вузлів, які поводяться аномально — через шум, несправність або навмисні атаки — одночасно сприяючи розвитку вузлів, які демонструють стабільну, чесну та корисну поведінку навчання.

Традиційне федеративне усереднення (FedAvg) надає однакову важливість усім оновленням клієнта, припускаючи, що всі вони однаково дійсні. Однак у реальних умовах:

- Датчики можуть бути несправними
- Пристрої можуть бути піддані кібератаці

- Локальні набори даних можуть бути сильно спотвореними або пошкодженими

Ці фактори можуть дестабілізувати збіжність моделі, призвести до упередженості прогнозів або знизити точність. Система, що враховує довіру, динамічно виявляє та зменшує вагу шкідливих клієнтів, використовуючи статистичні ознаки їхньої поведінки, забезпечуючи надійну та безпечну агрегацію.

Кожен клієнт i надсилає не лише свої модельні ваги w_i(t), але також показники поведінки: $l_i(t)$, $\Delta w_i(t)$, $A_i(t)$. Сервер використовує їх для обчислення рейтингу довіри $T_i(t) \in [0,1]$, який потім використовується для агрегації. У кожному раунді t оцінка поведінки $S_i(t)$ обчислюється як зважена комбінація показників:

$$Si(t) = \beta 1 \cdot (1 - l_i(t)) + \beta 2 \cdot (1 - \Delta wi_i(t)) + \beta 3 \cdot (1 - Ai_i(t)) \quad (10)$$

де $l_i(t)$ - локальні втрати (середньоквадратична помилка), $\Delta w_i(t)$ - L2 норма зміни ваги, $A_i(t)$ - Середня абсолютна похибка (оцінка аномалії), β1,β2,β3 - Вагові коефіцієнти показників. Кожен член віднімається від 1, оскільки менші значення є більш достовірними (тобто менша похибка означає кращу поведінку).

Довіра оновлюється за допомогою експоненціального згладжування:

$$Ti_i(t) = \alpha \cdot T_i(t-1) + (1-\alpha) \cdot S_i(t), \quad (11)$$

де $\alpha \in [0, 1]$ - Коефіцієнт згладжування, $T_i(t-1)$ - попередній рейтинг довіри, $S_i(t)$ - новий показник поведінки. Це забезпечує стабільність з часом — раптові сплески показників не викликають різкого падіння довіри, імітуючи людську пам'ять.

Розпад довіри - додатковий механізм розпаду карає неактивні або нестабільні вузли:

$$Ti_i(t) = T_i(t) \cdot \gamma, \quad (12)$$

де $\gamma \in (0, 1)$, це поступово знижує довіру до невдоволених або некомунікативних клієнтів.

Поріг виключення: Клієнти, чия довіра падає нижче налаштованого порогу θ виключені з агрегації моделей:

$$T_i(t) < \theta \quad (13)$$

Це створює систему самовідновлення, де зловмисні клієнти з часом втрачають вплив.

Потім оцінки довіри використовуються для оцінки оновлення кожного клієнта:

$$w^{global}(t) = \frac{\sum_i (T_i(t) * w_i(t))}{\sum_i T_i(t)}, \quad (14)$$

де $w_i(t)$ - Параметри моделі від клієнта, $T_i(t)$ - Рейтинг довіри клієнта.

У кожному раунді навчання сервер підтримує одну глобальну модель. Ця модель надсилається ідентично всім клієнтам. Кожен клієнт отримує однакові ваги w_i^{global}, виконує локальне навчання, використовуючи свій приватний набір даних, та повертає оновлену вагу $w_i(t)$ та показники на сервер. Потім сервер агрегує лише ті оновлення, яким довіряють $T_I(t) \geq \theta$, створюючи нову спільну глобальну модель для наступного раунду. На сервері немає окремої моделі для кожного клієнта; всі клієнти навчаються на однакових ініціалізованих вагах і надсилають лише свої персоналізовані оновлення.

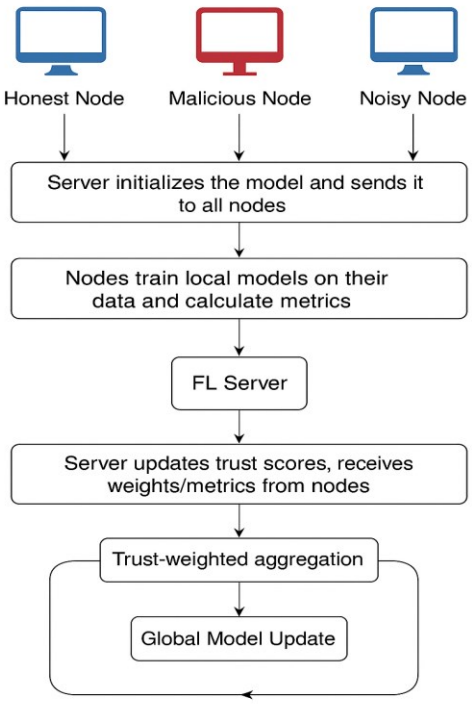

Рис. 1. Загальна схема експерименту

При агрегації сервер оцінює прогностичну ефективність глобальної моделі, використовуючи окремий виділений набір тестових даних D_{server}, для розрахунку середньоквадратичної похибки (RMSE) розраховується як:

$$RMSE_t = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_{i'} - y_i)^2}, \quad (15)$$

де $y_{i'}$ - прогнози з агрегованої моделі, y_i - справжні цільові значення. RMSE надає глобальну міру якості моделі, яка використовується для звітності про точність та моніторингу конвергенції між раундами. Важливо, що RMSE не виводиться з показників на стороні клієнта — він обчислюється безпосередньо з використанням тестових даних сервера та останньої моделі.

Результати експерименту демонструють ефективність агрегації, зваженої на довіру (TWA), у покращенні як точності, так і стійкості федеративного навчання в умовах змагальності та шуму. Протягом 50 раундів комунікації ми порівняли дві стратегії агрегації: традиційну FedAvg та запропоновану TWA.

Глобальну модель було оцінено за допомогою середньоквадратичного відхилення (RMSE) на чистому тестовому наборі даних, що зберігається на сервері. Було обчислено кінцеве, середнє значення та стандартне відхилення RMSE для оцінки як продуктивності, так і стабільності.

Таблиця 1

Назва метрики	**FedAvg**	**TWA**
Final RMSE	4.13	3.15
Average RMSE	4.49	3.51
RMSE Std Dev	0.71	0.29

Кінцеве середньоквадратичне відхилення (RMSE) відображає похибку прогнозування наприкінці навчання. TWA досягло на 25% нижчого значення RMSE порівняно з FedAvg, що підтверджує ефективне фільтрування ненадійних оновлень. Середнє значення RMSE також підкреслює, що TWA підтримувало стабільно нижчу похибку протягом навчання, а стандартне відхилення показує, що TWA було значно стабільнішим з часом. Протягом навчання показник довіри Вузла 9 різко знизився протягом перших 5 раундів через аномальні показники, такі як надзвичайно висока дельта ваг та низькі локальні втрати, що свідчить про поведінку суперника. Натомість Вузол 8 (шумний) демонстрував помірно високі показники аномалій, але зберігав часткову довіру протягом кількох раундів, перш ніж опустився нижче порогу довіри. Таке динамічне коригування гарантувало, що лише надійні оновлення формували глобальну модель.

Ми спостерігали, що FedAvg страждав від нестабільних піків середньоквадратичної помилки (RMSE) через рівномірне включення шкідливих оновлень, тоді як механізм виключення TWA захищав від цих коливань. Більше того, конвергенція TWA відбувалася на 30% швидше, що вимагає меншої кількості раундів для стабілізації втрат. Підсумовуючи, ці результати підтверджують наукове твердження про те, що федеративне навчання на основі довіри підвищує як точність, так і стійкість глобальних моделей за наявності шумних та зловмисних клієнтів, що робить його придатним методом для захисту розподілених середовищ мікромереж.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

У цьому дослідженні було представлено систему федеративного навчання, що враховує довіру, спеціально розроблену для підвищення стійкості та надійності спільного навчання в середовищах мікромереж. Інтегруючи динамічний механізм оцінювання довіри на основі локальних поведінкових показників – втрат, дельти ваг та оцінки аномалій – ми продемонстрували, що запропонований підхід до агрегації з урахуванням довіри (TWA) значно перевершує традиційні методи агрегації, такі як FedAvg, у сценаріях, що включають шумні та конфронтаційні клієнти.

Експериментальна установка, яка імітувала гетерогенність реального світу за допомогою чесних, шумних та шкідливих вузлів, надала вагомі емпіричні докази, що підтверджують ефективність методу. Порівняно з FedAvg, TWA досяг нижчих кінцевих та середніх значень RMSE, зменшила дисперсію в продуктивності моделі та прискорила конвергенцію. Важливо, що шкідливі клієнти, такі як Node 9, були швидко ідентифіковані та виключені, тоді як шумні клієнти були належним чином зменшені з часом, що підтверджує точність механізму оцінки довіри. Модульна реалізація з використанням Flower та PyTorch забезпечує відтворюваність та масштабованість, що дозволяє безперешкодно адаптуватися до майбутніх розширень, таких як прогнозування на основі LSTM або інтеграція з системами виявлення аномалій у реальному часі. Ці результати підтверджують, що механізми, що враховують довіру, є важливими для підвищення безпеки та стійкості в децентралізованих, чутливих до даних областях, таких як інтелектуальні мережі.

На завершення, запропонована структура TWA вирішує одну з фундаментальних проблем федеративного навчання, забезпечуючи надійне глобальне навчання моделей у змагальних та невизначених

середовищах. Майбутні дослідження досліджуватимуть гібридні моделі довіри, що включають аналіз поведінки в реальному часі та часову динаміку, а також застосування поза межами мікромереж, такі як охорона здоров'я та мережі автономних транспортних засобів.

Література

1. Syarif I., Prugel-Bennett A., Wills G. Unsupervised clustering approach for network anomaly detection. 2022. 11 p.
2. Шостак А. А., Гальчинський Л. Ю. Кількісне оцінювання кіберзагроз для систем критичної інфраструктури на основі моделі інтеграції даних датчиків системи SCADA. Інформаційне суспільство: технологічні, економічні та технічні аспекти становлення : матеріали Міжнародної наукової інтернет-конференції (м. Тернопіль, Україна, м. Ополе, Польща, 14–15 травня 2025 р.). Тернопіль : ТНТУ, 2025. Вип. 99. С. 68–72.
3. McMahan H. B., Moore E., Ramage D., Hampson S., y Arcas B. A. Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). 2017. Vol. 54. P. 1273–1282.
4. Yin D., Chen Y., Ramchandran K., Bartlett P. Byzantine-robust distributed learning: Towards optimal statistical rates. Proceedings of the 35th International Conference on Machine Learning (ICML). 2018. Vol. 80. P. 5650–5659.
5. Kang J., Xiong Z., Niyato D., Zhao J., Liang Y. C. Incentive design for efficient federated learning in mobile networks: A contract theory approach. *IEEE VTC-Fall*. 2019. P. 1–5.
6. Li T., Sahu A. K., Talwalkar A., Smith V. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine. 2020. Vol. 37(3). P. 50–60. DOI: 10.1109/MSP.2020.2975749.
7. Ma H., Zhao Y., Liu S., Zhang Y. Trust-aware aggregation for federated learning in industrial IoT. IEEE Internet of Things Journal. 2022. Vol. 9(11). P. 8428–8440. DOI: 10.1109/JIOT.2021.3113430.
8. Liu W., Lyu L., Yu H., Yu P. S. A reliable and robust federated learning framework against unreliable clients. Proceedings of the 2021 SIAM International Conference on Data Mining. 2021. P. 405–413.
9. Wang S., Tuor T., Salonidis T., et al. Adaptive federated learning in resource-constrained edge computing systems. IEEE Journal on Selected Areas in Communications. 2020. Vol. 38(6). P. 1205–1221. DOI: 10.1109/JSAC.2020.2980910.
10. Musleh A. S., Vasilakos A. V., Ngyuen T. Smart grid cyber-physical attack detection and resilience: A review. IEEE Access. 2019. Vol. 7. P. 160595–160614. DOI: 10.1109/ACCESS.2019.2949813.
11. Ahmed M., Mahmood A. N., Hu J. A survey of network anomaly detection techniques. Journal of Network and Computer Applications. 2016. Vol. 60. P. 19–31. DOI: 10.1016/j.jnca.2015.11.016.
12. Jhana N. Energy consumption, generation, prices, and weather [dataset]. Kaggle. 2022. URL: https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather.

References

1. Syarif I., Prugel-Bennett A., Wills G. Unsupervised clustering approach for network anomaly detection. 2022. 11 p.
2. Shostak A. A., Galchinsky L. Yu. Quantitative assessment of cyber threats to critical infrastructure systems based on the model of integration of data from SCADA system sensors. Information society: technological, economic and technical aspects of formation: materials of the International Scientific Internet Conference (Ternopil, Ukraine, Opole, Poland, May 14–15, 2025). Ternopil: TNTU, 2025. Issue 99. Pp. 68–72.
3. McMahan H. B., Moore E., Ramage D., Hampson S., y Arcas B. A. Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). 2017. Vol. 54. P. 1273–1282.
4. Yin D., Chen Y., Ramchandran K., Bartlett P. Byzantine-robust distributed learning: Towards optimal statistical rates. Proceedings of the 35th International Conference on Machine Learning (ICML). 2018. Vol. 80. P. 5650–5659.
5. Kang J., Xiong Z., Niyato D., Zhao J., Liang Y. C. Incentive design for efficient federated learning in mobile networks: A contract theory approach. *IEEE VTC-Fall*. 2019. P. 1–5.
6. Li T., Sahu A. K., Talwalkar A., Smith V. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine. 2020. Vol. 37(3). P. 50–60. DOI: 10.1109/MSP.2020.2975749.
7. Ma H., Zhao Y., Liu S., Zhang Y. Trust-aware aggregation for federated learning in industrial IoT. IEEE Internet of Things Journal. 2022. Vol. 9(11). P. 8428–8440. DOI: 10.1109/JIOT.2021.3113430.
8. Liu W., Lyu L., Yu H., Yu P. S. A reliable and robust federated learning framework against unreliable clients. Proceedings of the 2021 SIAM International Conference on Data Mining. 2021. P. 405–413.
9. Wang S., Tuor T., Salonidis T., et al. Adaptive federated learning in resource-constrained edge computing systems. IEEE Journal on Selected Areas in Communications. 2020. Vol. 38(6). P. 1205–1221. DOI: 10.1109/JSAC.2020.2980910.
10. Musleh A. S., Vasilakos A. V., Ngyuen T. Smart grid cyber-physical attack detection and resilience: A review. IEEE Access. 2019. Vol. 7. P. 160595–160614. DOI: 10.1109/ACCESS.2019.2949813.
11. Ahmed M., Mahmood A. N., Hu J. A survey of network anomaly detection techniques. Journal of Network and Computer Applications. 2016. Vol. 60. P. 19–31. DOI: 10.1016/j.jnca.2015.11.016.
12. Jhana N. Energy consumption, generation, prices, and weather [dataset]. Kaggle. 2022. URL: https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather