Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Покращення сертифікованої стійкості Randomized Smoothing засобами геометрії розподілів, адаптивних збурень та структурованих моделей шуму
(тема)

Виконав:
здобувач другого року навчання,
групи ДСм-24-1
Рижова О.Ю.
(прізвище, ініціали)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма Науки про дані (Data Science)
(повна назва спеціалізації)
Керівник проф. Шевченко О.Ю.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Лариса ЧАЛА
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Науки про дані (Data Science)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Рижовій Олександрі Юріївні
(прізвище, ім'я, по батькові)

1. Тема роботи Покращення сертифікованої стійкості Randomized Smoothing засобами геометрії розподілів, адаптивних збурень та структурованих моделей шуму

затверджена наказом університету від 24 листопада 20 25 р. № 1057Ст

2. Термін подання студентом роботи до екзаменаційної комісії 18 грудня 20 25 р.

3. Вихідні дані до роботи Архітектура нейронної мережі типу ResNet, датасет природних зображень із багатокласною структурою, параметри шумових впливів для методу Randomized Smoothing (гаусівський шум із різними значеннями σ), статистичні параметри сертифікації (кількість шумових вибірок), метрики оцінювання якості та робастності моделі, програмні засоби для реалізації та тестування алгоритмів.

4. Перелік питань, що потрібно опрацювати в роботі
1) Теоретичні основи randomized smoothing та сертифікованої стійкості
2) Побудова базової моделі randomized smoothing та аналіз вразливих кейсів
3) Покращення сертифікованої стійкості засобами геометрії розподілів, адаптивних збурень і структурованого шуму

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	24.11.2025	виконано
2	Вибір архітектури моделі	26.11.2025	виконано
3	Вибір датасету для виконання роботи	26.11.2025	виконано
4	Реалізація базової моделі Random Smoozed	28.11.2025	виконано
5	Виявлення кейсів помилок	30.11.2025	виконано
6	Типологізація помилок та аналіз геометрії	01.12.2025	виконано
	розподілів у проблемних ситуаціях	01.12.2025	
7	Використання оптимізації геометрії шумових	02.12.2025	виконано
	розподілів	02.12.2025	
8	Робота з узагальненими моделями шуму	03.12.2025	виконано
9	Використання адаптивного збурення	04.12.2025	виконано
10	Робота з локально залежними моделями шуму	06.12.2025	виконано
11	Використання структурованих шумових моделей	07.12.2025	виконано
12	Порівняння ефективності використаних	08.12.2025	виконано
	методологій	08.12.2025	
13	Оформлення пояснювальної записки	10.12.2025	виконано
14	Захист перед ЕК	18.12.2025	виконано

Дата видачі завдання 24 листопада 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ проф. Шевченко О.Ю.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 116 с., 10 рис., 10 табл., 5 дод., 20 джерел.

ВОРОЖІ АТАКИ, МОДИФІКАЦІЯ, НЕЙРОННА МЕРЕЖА, СЕРТИФІКОВАНА СТІЙКІСТЬ, RANDOMIZED SMOOTHING.

Об’єкт дослідження – процес класифікації даних нейронними мережами в умовах ворожих атак.

Предмет дослідження – методи сертифікованої стійкості моделей глибинного навчання, зокрема Randomized Smoothing та його модифікації, що дозволяють гарантувати стабільність прогнозів у присутності ворожих збурень.

Мета роботи – розробка, аналіз та удосконалення методів Randomized Smoothing для підвищення сертифікованої стійкості процесу класифікації нейронних мереж, а також оцінка ефективності цих методів у практичних умовах.

Методи дослідження – використано теоретичний аналіз наукових джерел та сучасних підходів до сертифікованої стійкості, математичне моделювання Randomized Smoothing та його модифікацій, статистичні методи оцінки ймовірностей класифікації з використанням методу Монте-Карло та інтервальних оцінок, експериментальна перевірка на тестових наборах даних із різними характеристиками та рівнями шуму, порівняльний аналіз різних підходів за критеріями сертифікованого радіуса та точності на чистих даних.

Проведено систематизацію методів Randomized Smoothing, визначено їхні переваги та обмеження, розроблено рекомендації щодо оптимального вибору параметра шуму та адаптивних стратегій згладжування, запропоновано комбінації Randomized Smoothing із гібридними підходами.

ABSTRACT

Master’s thesis contains: 116 p., 10 fig., 10 tab., 5 ann., 20 sources.

CERTIFIED STABILITY, HOSTILE ATTACKS, MODIFICATION, NEURAL NETWORK, RANDOMIZED SMOOTHING.

The object of the study is the process of data classification by neural networks under hostile attacks.

The subject of the study is methods of certified stability of deep learning models, in particular Randomized Smoothing and its modifications, which allow to guarantee the stability of forecasts in the presence of hostile disturbances.

The purpose of the work is to develop, analyze and improve Randomized Smoothing methods to increase the certified stability of the neural network classification process, as well as to assess the effectiveness of these methods in practical conditions.

Research methods – theoretical analysis of scientific sources and modern approaches to certified stability, mathematical modeling of Randomized Smoothing and its modifications, statistical methods for estimating classification probabilities using the Monte Carlo method and interval estimates, experimental verification on test data sets with different characteristics and noise levels, comparative analysis of different approaches according to the criteria of certified radius and accuracy on clean data.

Randomized Smoothing methods were systematized, their advantages and limitations were identified, recommendations were developed for the optimal choice of the noise parameter and adaptive smoothing strategies, combinations of Randomized Smoothing with hybrid approaches were proposed, which ensure an increase in the certified radius and model stability.

ЗМІСТ

ВСТУП

Сучасний стан розвитку систем штучного інтелекту супроводжується стрімким зростанням складності моделей, розширенням сфер їхнього застосування та посиленням вимог до надійності прийняття рішень. Однією з ключових проблем є вразливість моделей глибинного навчання до ворожих впливів – малих цілеспрямованих збурень вхідних даних, які здатні суттєво змінити результат класифікації. Такі атаки можуть реалізовуватися як у цифровому, так і у фізичному середовищах і переноситися між різними моделями, що створює серйозні ризики для безпеки критичних систем.

Об’єкт дослідження: процес класифікації даних нейронними мережами в умовах ворожих атак.

Предмет дослідження: методи сертифікованої стійкості моделей глибинного навчання, зокрема Randomized Smoothing та його модифікації, що дозволяють гарантувати стабільність прогнозів у присутності ворожих збурень.

Мета роботи: розробка, аналіз та удосконалення методів Randomized Smoothing для підвищення сертифікованої стійкості процесу класифікації нейронних мереж, а також оцінка ефективності цих методів у практичних умовах.

Задачі дослідження:

– проаналізувати сучасний стан проблеми ворожих атак та існуючих методів сертифікованої стійкості;

– дослідити принцип роботи та обмеження методу Randomized Smoothing;

– вивчити модифікації Randomized Smoothing, спрямовані на підвищення сертифікованої стійкості та зменшення втрат точності;

– розробити рекомендації щодо оптимізації параметрів згладжування та адаптації шумових стратегій;

– оцінити ефективність запропонованих підходів експериментально та порівняти їх із базовими методами.

Методи дослідження: використано теоретичний аналіз наукових джерел та сучасних підходів до сертифікованої стійкості, математичне моделювання Randomized Smoothing та його модифікацій, статистичні методи оцінки ймовірностей класифікації з використанням методу Монте-Карло та інтервальних оцінок, експериментальна перевірка на тестових наборах даних із різними характеристиками та рівнями шуму, порівняльний аналіз різних підходів за критеріями сертифікованого радіуса та точності на чистих даних.

Результати дослідження: проведено систематизацію методів Randomized Smoothing, визначено їхні переваги та обмеження, розроблено рекомендації щодо оптимального вибору параметра шуму та адаптивних стратегій згладжування, запропоновано комбінації Randomized Smoothing із гібридними підходами (робастне тренування, денойзинг, дифузійні моделі), які забезпечують підвищення сертифікованого радіуса та стабільності моделі.

Наукова новизна полягає у комплексному аналізі сучасних модифікацій Randomized Smoothing та формуванні методологічних рекомендацій для використання у критично важливих системах.

Практична цінність роботи полягає у можливості застосування отриманих результатів для побудови більш надійних процесів класифікації даних у медицині, енергетиці, транспорті, фінансових сервісах та інших сферах, де помилки класифікації можуть мати критичні наслідки.

1 ТЕОРЕТИЧНІ ОСНОВИ RANDOMIZED SMOOTHING ТА СЕРТИФІКОВАНОЇ СТІЙКОСТІ

1.1 Проблема ворожих атак та мотивація сертифікованого захисту

Швидкий розвиток глибинного навчання та масштабне впровадження технологій штучного інтелекту у критично важливі сфери, такі як медицина, енергетика, транспорт, оборона, фінансові сервіси тощо, суттєво підвищили вимоги до надійності й передбачуваності моделей машинного навчання. У цих умовах одним із найгостріших викликів є явище ворожих атак, що становлять загрозу для коректного функціонування навіть найдосконаліших нейронних мереж [1].

Під ворожими атаками розуміють навмисні, стратегічні та зазвичай малопомітні для людини модифікації вхідних даних, які призводять до радикальної зміни вихідного рішення моделі. Особливо показовим є випадок зображень: незначні корекції інтенсивності пікселів або додавання слабкого шуму, непомітного для людського ока, можуть спричинити повну зміну класифікації. Така властивість свідчить про потенційну вразливість моделей до збурень, що легко генеруються алгоритмічно та спрямовані на обходи систем безпеки.

На теоретичному рівні поява ворожих прикладів є наслідком неоднорідності та локальної нестабільності аппроксимаційних функцій, що лежать в основі глибоких моделей. Багатовимірний простір ознак, у якому працюють сучасні нейромережі, є надзвичайно складним і має численні області високої чутливості. У таких областях навіть малі за нормою збурення можуть змінити прогноз, що свідчить про відсутність локальної гладкості класифікаторів.

Більше того, дослідження показують, що ворожі приклади часто є переносними атакувальними збуреннями, створені для однієї моделі, можуть успішно впливати на інші моделі зі схожою архітектурою або навіть

на зовсім інші типи нейронних мереж. Це вказує на системний характер вразливостей.

З практичного погляду неконтрольоване виникнення помилкових рішень під впливом ворожих атак створює ризики для всіх сфер, де автономні рішення впливають на безпеку та життя людей [2].

У медицині ворожі зміни в медичних знімках можуть привести до неправильної діагностики або вибору неадекватного лікування. В транспортних системах модифікація дорожніх знаків чи сенсорних даних здатна дезорієнтувати автономні автомобілі. В енергетиці, зокрема на системах моніторингу об’єктів атомної енергетики, ворожі приклади можуть призвести до некоректної оцінки стану обладнання чи помилкових рішень автоматизованих систем управління. У фінансовому секторі атаки можуть впливати на системи прогнозування, скоринг чи детекцію шахрайства. У біометричних системах підроблені зображення або аудіозаписи дозволяють обходити системи автентифікації.

Окрім традиційних прикладів модифікації піксельних інтенсивностей, сучасні роботи демонструють широке різноманіття атакувальних стратегій. Зокрема, виділяють атаки з доступом до градієнта моделі (white-box), атаки з частковим доступом до моделі (grey-box) та атаки, що оперують лише вихідними прогнозами без інформації про внутрішні параметри (black-box). Кожна з цих груп використовує різний ступінь знання про архітектуру моделі, проте навіть у black-box сценаріях зловмисники здатні будувати високоефективні збурення за рахунок апроксимації градієнтів або перенесення атак з проксі-моделей. Це свідчить про глибоку фундаментальну проблему незахищеності нейронних мереж перед узагальненими ворожими впливами [3].

Таким чином, проблема ворожих атак є універсальною та охоплює всі сфери, де використовуються алгоритми машинного навчання.

Для кращого розуміння масштабів проблеми розглянемо найпоширеніші категорії атак:

– атаки з доступом до моделі (white-box attacks) – зловмисник має повний доступ до параметрів і градієнтів моделі, що дає змогу створювати високоефективні збурення (наприклад, FGSM, PGD);

– атаки з обмеженим доступом (black-box attacks) – зловмисник не бачить структури моделі, а отримує лише її відповіді, проте завдяки перенесності атак ефективність захисту все одно знижується;

– атаки за даними (data poisoning) – модифікація тренувальних даних з метою змінити поведінку моделі у певних сценаріях;

– атаки на фізичному рівні (physical-world attacks) – зміни в реальному середовищі (стікери, світлові перешкоди, зміна кольорів), що вводять модель в оману.

Усі ці типи атак демонструють, що сучасні підходи до захисту мають бути не лише практично ефективними, а й теоретично обґрунтованими.

Суттєву увагу привертають і фізичні атаки, які не обмежуються цифровими змінами на робочих даних. До таких належать зміни умов освітлення, додавання наклейок на об’єкти, дефекти сенсорів та шум реального середовища. Дослідження показують, що фізичні збурення мають більш складну структуру й часто є нелінійними, що робить їх ще небезпечнішими. У контексті застосування штучного інтелекту у критичних інфраструктурах – наприклад, у відеодіагностиці обладнання АЕС або у роботизованих системах керування – такі атаки можуть мати серйозні наслідки, оскільки здатні впливати на прийняття рішень у складних природних умовах.

Більшість класичних методів захисту (наприклад, adversarial training, різні види регуляризації або модифікації архітектур) мають такі обмеження:

– вони часто залежать від конкретних типів атак і не гарантують стійкості до невідомих векторів загроз;

– під час оцінки можуть бути обмануті адаптивними атаками;

– збільшення стійкості часто відбувається за рахунок зниження точності на нормальних даних;

– ці методи не надають формальних гарантій.

Оскільки зловмисники постійно вдосконалюють стратегії атак, емпіричні методи захисту швидко втрачають ефективність. У відповідь на ці виклики наукова спільнота розробляє методи сертифікованої стійкості – підходи, які дозволяють математично довести, що модель зберігає правильний прогноз для всіх можливих збурень певної величини.

Сертифікована стійкість дає низку ключових переваг:

– гарантована безпека у межах математично доведеного радіуса;

– незалежність від типу атаки, гарантії охоплюють усі можливі збурення у відповідній нормі;

– можливість формальної валідації моделей для критично важливих застосувань;

– зростання довіри до систем штучного інтелекту серед користувачів, компаній і регуляторів.

Саме тому сертифікована стійкість стає ключовим напрямом у розвитку робастних систем. На відміну від класичних захисних механізмів, сертифіковані методи пропонують формальну гарантію, що модель зберігатиме правильний прогноз для будь-якого допустимого збурення у заданому просторі. Це означає, що навіть при появі нових, ще не відомих типів атак, система не втратить стійкість у межах визначеного радіуса [4]. Такий підхід особливо актуальний для сфер, де безпомилковість класифікації має прямий вплив на безпеку людей та функціонування складних технічних систем.

Попит на сертифіковану стійкість зростає з кількох ключових причин:

– ускладнення атак. Сучасні атаки використовують оптимізаційні методи, еволюційні алгоритми та моделі-гойданки, що робить традиційний захист недостатнім;

– поширення у критичних інфраструктурах. У системах, де помилка класифікації неприпустима, сертифіковані гарантії стають стандартом;

– потреба у формальних методах оцінки ризику. Регулятори (особливо в галузях транспорту, медицини та енергетики) вимагають чітких доказів стійкості;

– потреба у масштабованих рішеннях. Сертифікована стійкість дозволяє створювати універсальні методи захисту для моделей будь-якої складності.

Одним із найбільш досліджуваних і практично ефективних підходів до сертифікації є Randomized Smoothing – метод, який базується на статистичному згладжуванні рішень моделі шляхом додавання випадкового шуму до вхідних даних. Метод Randomized Smoothing:

– сумісний з будь-якою нейронною мережею;

– масштабовано працює на великих задачах;

– забезпечує математично гарантований радіус стійкості у L_2-нормі.

Саме ці властивості роблять Randomized Smoothing одним із ключових напрямів досліджень у сфері безпеки машинного навчання.

1.2 Метод Randomized Smoothing: принцип роботи та обмеження

Метод Randomized Smoothing є однією з ключових сучасних концепцій сертифікованої стійкості нейронних мереж, яка дозволяє забезпечити формальні гарантії поведінки моделі у присутності ворожих збурень. Зважаючи на те, що традиційні методи захисту ґрунтуються на емпіричному протистоянні атакам і часто виявляються неефективними проти нових або адаптивних стратегій противника, Randomized Smoothing пропонує принципово інший підхід: побудову нового класифікатора, який є математично стійким у певному оточенні вхідних даних.

Суть методу полягає в тому, що базовий класифікатор $f(x)$ не використовується для прийняття рішення безпосередньо. Замість цього модель оцінює, як класифікатор поводиться на випадково зашумлених версіях одного й того самого вектору ознак. Кожне зашумлення формує

новий вхід виду x + δ, де випадковий шум δ формується згідно з багатовимірним нормальним розподілом [5]:

$$\delta \sim N\left(0,\sigma^2 I\right), \tag{1.1}$$

де δ – випадковий шум;

$N\left(0,\sigma^2 I\right)$ – багатовимірний нормальний розподіл із нульовим середнім;

σ^2 – дисперсія шуму;

I – одинична матриця розмірності, що збігається з розмірністю вхідного простору.

Ідея Randomized Smoothing полягає у створенні нової функції прийняття рішень шляхом усереднення виходів базового класифікатора за випадковими збуреннями, що додаються до вхідного вектора. Це усереднення фактично перетворює початкову нерегулярну розділюючу поверхню на більш гладку та стабільну, що зменшує чутливість до локальних флуктуацій. Гаусівський шум встановлює сферичну симетрію ймовірнісного розподілу навколо точки, а тому сертифікація природним чином узгоджується з L_2-геометрією збурень.

Для кожної такої реалізації x + δ обчислюється вихід класифікатора f. Згладжений класифікатор g(x) визначається як клас, що найчастіше з'являється серед результатів базової моделі (1.2):

$$g\left(x\right) = \arg\max_{c \in C} \left(f\left(x+\delta\right) = c\right), \tag{1.2}$$

де f(·) – вихідний (базовий) класифікатор;

x – вхідний вектор ознак;

δ – вектор випадкового шуму, який генерується відповідно до нормального розподілу $N(0,\sigma^2 I)$;

σ – параметр дисперсії шуму;

C – множина всіх можливих класів;

c – конкретний клас із множини C;

R(·) – ймовірність того, що класифікатор f передасть вхід до відповідного класу.

Оцінювання ймовірностей класифікації класів p_A та p_B для згладженого класифікатора базується на статистичній апроксимації, оскільки реальні значення цих ймовірностей недоступні у закритій формі. Використання методу Монте-Карло дозволяє оцінити частоту правильних класифікацій, однак точність оцінки сильно залежить від кількості вибірок та рівня шуму σ. Занадто мала кількість вибірок призводить до високої дисперсії оцінок, а тому формула сертифікованого радіуса може давати надто консервативний результат. Це робить питання вибору статистичних параметрів важливою частиною підвищення ефективності Randomized Smoothing.

Цей підхід дозволяє перенести складність моделі у простір ймовірностей, де поведінка стає більш регулярною. Випадковий шум «усереднює» локальні нестабільності функції активацій і дозволяє отримати гладку аппроксимацію реальної поверхні прийняття рішень. Це усунення різких переходів у межах малих околів точки робить модель менш чутливою до невеликих навмисних збурень [6].

Однією з основних причин популярності Randomized Smoothing є можливість формально визначити радіус стійкості R, у межах якого результуючий класифікатор гарантовано не змінить свій прогноз. Цей радіус оцінюється на основі ймовірностей, з якими модель класифікує зашумлені точки в правильний клас A та у другий за ймовірністю клас B. Для цього використовується формула 1.3:

$$R = \frac{\sigma}{2}\left(\Phi^{-1}\left(p_A\right) - \Phi^{-1}\left(p_B\right)\right), \tag{1.3}$$

де R – радіус сертифікованої стійкості;

σ – стандартне відхилення доданого шуму;

$\Phi^{-1}(\cdot)$ – обернена функція стандартного нормального розподілу;

p_A – ймовірність того, що базовий класифікатор присвоїть точку класу A, який є найімовірнішим;

p_B – найбільша ймовірність класифікації за будь-який інший клас, відмінний від A.

Формула (1.3) є центральним теоретичним результатом методу Randomized Smoothing. Її походження пов’язане з геометричними та статистичними властивостями шумового оточення точки та дозволяє перетворити емпіричні оцінки ймовірностей на суворі математичні гарантії. У практичному застосуванні це означає, що якщо між оцінками p_A та p_B існує достатній розрив, то класифікатор наділений стійкістю до значних збурень у просторі вхідних даних [7].

Попри елегантність та універсальність підходу, Randomized Smoothing має низку важливих недоліків. Найсуттєвішим з них є значний обсяг обчислень, оскільки кожний виклик класифікатора потребує незалежного додавання шуму. У задачах з високою розмірністю даних або глибокими архітектурами це призводить до збільшення часу обчислень у десятки разів. Крім того, ефективність методу падає для класів, що є близькими за розміщенням у просторових репрезентаціях, оскільки навіть невеликі збурення можуть призвести до зміни ймовірнісного балансу між ними. Також Randomized Smoothing погано масштабується до $L\infty$-атак через геометричну невідповідність гаусівського шуму формі $L\infty$-кулі, унаслідок чого гарантований радіус у таких випадках є дуже малим.

Застосування цього методу дозволяє будь-яку модель – навіть дуже складну – зробити сертифіковано стійкою, оскільки метод не накладає обмежень на структуру f(x). Він працює навіть тоді, коли класифікаційна поверхня є розривною, фрагментованою або має складну багатовимірну геометрію. Саме в цьому полягає його універсальність і широке

використання в задачах комп'ютерного зору, медичних системах діагностики, автономних системах керування тощо.

Однак додавання шуму та необхідність проведення великої кількості статистичних оцінок породжують низку суттєвих обмежень методу. Найважливішим з них є компроміс між точністю та стійкістю. Збільшення значення параметра шуму σ підвищує радіус сертифікації R, але водночас зменшує чисту точність моделі (тобто точність без атаки), оскільки шум спотворює вхідні дані. У практичних задачах це часто призводить до ситуації, коли модель зі значним рівнем сертифікованої стійкості демонструє помітно нижчу продуктивність на звичайних даних.

Інша важлива проблема полягає в обчислювальній складності. Для того щоб отримати надійну оцінку ймовірностей p_A та p_B, необхідно згенерувати тисячі або десятки тисяч зашумлених копій одного й того самого вектора x. Це означає, що на кожен вхідний зразок модель має працювати дуже багато разів, що робить метод непридатним для задач режиму реального часу без додаткової оптимізації [8].

Також слід враховувати, що Randomized Smoothing дає сертифікацію переважно в L_2-просторі. Хоча це є природним вибором для гаусівського шуму, багато реальних атак формуються в $L\infty$-нормі, де зміна декількох пікселів може повністю змінити поведінку мережі. У таких випадках гарантії Randomized Smoothing можуть бути менш ефективними або мати малий радіус.

Попри ці недоліки, метод Randomized Smoothing залишається одним із найбільш ефективних і практично застосовних інструментів сертифікованої стійкості. Його теоретична обґрунтованість, універсальність, незалежність від архітектури моделі та відносна простота реалізації роблять його важливою складовою сучасних систем безпеки штучного інтелекту. Він сприяє підвищенню рівня довіри до моделей машинного навчання, особливо в критично важливих сферах, де помилки

можуть мати серйозні наслідки – від автономного транспорту до діагностичних медичних систем і систем підтримки прийняття рішень.

1.3 Огляд сучасних підходів до покращення Randomized Smoothing

Із моменту появи Randomized Smoothing як практичного механізму сертифікованої стійкості значний масив досліджень був спрямований на подолання його обмежень та збільшення ефективності в умовах складних атак. Попри те, що метод забезпечує універсальність, масштабованість і можливість сертифікації моделей довільної архітектури, початкова версія Randomized Smoothing мала низку суттєвих недоліків – зокрема, зниження чистої точності, високу залежність від дисперсії шуму, обчислювальні витрати і чутливість до типу норми загрози. Тому наукова спільнота розпочала активні пошуки шляхів підвищення його ефективності, що привело до появи широкого спектра поліпшень.

Одним із найдинамічніших напрямів розвитку Randomized Smoothing є дослідження альтернативних шумових розподілів. Заміна гаусівського шуму на лапласівський або рівномірний дозволяє отримувати сертифіковані гарантії в інших L_p-нормах, що є корисним для захисту від різних типів атак. Наприклад, лапласівський шум узгоджується з L_1-нормою, а рівномірний – з L_∞. Це розширює сферу застосування методу та дозволяє адаптувати техніку під специфічні потреби конкретної системи [9].

Першим великим напрямом розвитку стали методи оптимізованого тренування згладжених моделей. Роботи цього напряму базуються на ідеї, що базовий класифікатор має навчатися не лише на чистих даних, а у спосіб, що максимізує сертифікований радіус. Один із ключових підходів – SmoothAdv – пропонує інтегрувати адаптивні ворожі атаки під час тренування, моделюючи поведінку атакувальника, що знає про процедуру згладжування. Інші схеми, зокрема MACER, формують спеціалізовані лосс-функції, у які прямо включають параметри сертифікованої стійкості,

змушуючи модель оптимізувати не лише точність, а й математично обґрунтований рівень захищеності. Такі методи суттєво підвищують сертифіковані радіуси, проте вимагають значних ресурсів тренування.

Другим важливим напрямом є модифікація розподілу шуму. Базова версія Randomized Smoothing використовує ізотропний гаусівський шум, однак його властивості не завжди оптимальні для геометрії реальних даних. Тому дослідники пропонують багатомасштабне згладжування, де кілька шумових рівнів комбінуються для побудови стійкіших класифікаторів; локально адаптивний шум, у якому дисперсія залежить від локальної кривизни або щільності даних; та шум із негаусівських розподілів, наприклад, лапласівський або рівномірний, які краще підходять для сертифікації в L_1 або $L\infty$ нормах. Такі методи дозволяють суттєво розширити спектр атак, проти яких Randomized Smoothing здатний надавати гарантії, але при цьому ускладнюють математичний аналіз [10].

Третій напрям досліджень стосується підвищення статистичної ефективності процесу сертифікації. Оскільки класичний Randomized Smoothing вимагає тисяч або десятків тисяч вибірок для оцінки ймовірностей, це створює великий обчислювальний тягар. Тому було запропоновано методи інкрементальної сертифікації, що дозволяють припинити процес завчасно, якщо гіпотеза про домінуючий клас підтверджується вже на малих вибірках. Деякі інструменти застосовують статистичні межі (Clopper-Pearson, Wilson intervals) для більш ефективної оцінки ймовірностей та зменшення кількості прогонів моделі. Такі підходи значно прискорюють сертифікацію, що робить Randomized Smoothing придатнішим для практичних систем.

Четвертий напрям пов'язаний з розширенням Randomized Smoothing на інші норми і нові постановки задачі. Оскільки реальні атаки часто діють у просторі $L\infty$, L_1 або навіть у комбінованих нормах, дослідники пропонують нові шумові моделі та теоретичні схеми сертифікації, що дозволяють отримувати гарантії в цих метриках. Окремий клас підходів

поєднує ідеї Randomized Smoothing і диференціальної приватності: використання механізмів DP дозволяє мінімізувати вплив адаптивних атак, які враховують процедуру сертифікації. У цих роботах використовується аналіз через f-дивергенції або Renyi-дивергенції, що відкриває шлях до більш гнучких форм сертифікації [11].

П'ятий кластер методів представляє гібридизацію Randomized Smoothing з іншими техніками захисту. Поєднання згладжування з попереднім денойзінгом дозволяє значно компенсувати спотворення, викликані шумом, і тим самим підвищувати чисту точність моделей при збереженні стійкості. Гібридні методи з adversarial training поєднують емпіричну та формальну стійкість, створюючи моделі, що краще поводяться у реалістичних сценаріях атак. Деякі сучасні підходи інтегрують Randomized Smoothing із дифузійними моделями, використовуючи їх здатність відновлювати структуру даних після зашумлення.

Окремим і швидко зростаючим напрямом є адаптація Randomized Smoothing для інших типів моделей і областей застосування. Підходи Randomized Smoothing були перенесені на графові нейронні мережі, великі мовні моделі, мультимодальні трансформери, системи розпізнавання мови та генеративні архітектури. У кожному випадку виникає потреба у власній процедурі згладжування та сертифікації, адже простір входів має іншу структуру. Для графових моделей додають шум у структуру або лапласиан графа; для мовних моделей – у простір embeddings; для дифузійних моделей – у латентний простір. Таке розширення відкриває шлях до універсального використання Randomized Smoothing у критично важливих системах.

Досить ефективними виявилися комбінації Randomized Smoothing з методами робастного тренування, оскільки вони компенсують недоліки один одного. Робастне тренування формує більш стійку модель на етапі навчання, тоді як згладжування забезпечує формальні гарантії під час інференсу. В результаті такі гібридні моделі демонструють більші

сертифіковані радіуси та знижений розрив між робастною та чистою точністю. В експериментальних роботах такі підходи нерідко перевершують класичний Randomized Smoothing на 15–25 % за сертифікованою стійкістю [12].

У підсумку, розвиток Randomized Smoothing не обмежується поступовими вдосконаленнями – фактично формується нова парадигма сертифікованої стійкості, у якій Randomized Smoothing виступає базовим інструментом, а сучасні модифікації адресують практичні недоліки й адаптують концепцію до ширшого спектра задач, норм і моделей. Це робить Randomized Smoothing одним із найперспективніших напрямів дослідження у сфері гарантової безпеки штучного інтелекту.

У сучасних дослідженнях з робастності моделей машинного навчання значна увага приділяється модифікації базового підходу Randomized Smoothing з метою підвищення сертифікованого радіуса та зменшення втрат точності на чистих даних. Таким чином, науковці пропонують різні напрями вдосконалення, що включають зміну шумових розподілів, оптимізацію параметра згладжування, модифікацію архітектури нейронних мереж та поєднання згладжування з робастним тренуванням.

Окремий клас методів зосереджений на оптимізації параметра шуму σ або адаптивному виборі його значення під час сертифікації. Використання різних рівнів шуму для різних областей простору ознак дозволяє отримати більш збалансований компроміс між точністю та стійкістю. Крім того, активно досліджуються моделі, що працюють зі зменшеною кількістю шумових вибірок за рахунок вибіркових статистичних схем, що дозволяє зменшити обчислювальні витрати без суттєвої втрати якості сертифікації.

Узагальнений порівняльний аналіз основних груп таких методів наведено в таблиці 1.1, що дозволяє систематизувати їх ключові властивості, переваги та обмеження [13].

Таблиця 1.1 – Основні напрями вдосконалення Randomized Smoothing

Напрям досліджень	Суть підходу	Приклади методів	Основний ефект
Оптимізоване тренування	Навчання базового класифікатора з урахуванням стійкості	SmoothAdv, MACER	Збільшення сертифікованих радіусів
Адаптивний шум	Заміна гаусівського шуму на локально-залежний або спеціалізований	Multi-scale RS, Laplace RS, Data-dependent noise	Підвищення стійкості в різних нормах
Статистично-ефективна сертифікація	Скорочення кількості вибірок для оцінки ймовірностей	Incremental RS, Adaptive Sampling	Зменшення обчислювальних витрат
Сертифікація в інших нормах	Нові розподіли шуму та теоретичні схеми	L∞ RS, DP-based RS	Гарантії стійкості проти ширшого спектра атак
Гібридні методи	Комбінація RS з denoising, adversarial training та іншими техніками	Denoised RS, SmoothAdv+	Підвищення чистої точності та стійкості
Архітектурні розширення	Застосування RS до графів, мовних моделей, трансформерів	Graph RS, Embedding RS	Універсальність та застосовність методу

Як показує таблиця 1.1, найбільш перспективними у практичному застосуванні є гібридні методи, які поєднують Randomized Smoothing з додатковими механізмами підвищення стійкості, зокрема з робастним

тренуванням або латентними перетвореннями. Водночас альтернативні розподіли шуму відкривають можливості сертифікації в інших нормах, що актуально для задач із нестандартними збуреннями. Попри це, більшість розглянутих підходів стикається з компромісом між точністю та стійкістю, що підтверджує необхідність подальших теоретичних досліджень і пошуку більше універсальних стратегій балансування цих двох показників.

Окремий напрям розвитку Randomized Smoothing стосується оптимізації параметра шуму, вибору більш ефективних розподілів та впровадження адаптивних стратегій обчислення сертифікованого радіуса. Для порівняння різних модифікацій, що впливають на величину гарантованої стійкості, у таблиці 1.2 наведено результати типових експериментів, у яких різні підходи оцінювались за максимальним сертифікованим радіусом та точністю на чистих даних. Ці дані дозволяють простежити загальні тенденції та зробити висновки щодо ефективності кожного з розглянутих методів [14].

Таблиця 1.2 – Порівняльний аналіз сучасних методів підвищення ефективності Randomized Smoothing

Категорія покращення	Переваги	Недоліки	Кому підходить
Оптимізоване тренування	Найбільший приріст сертифікованого радіуса	Висока обчислювальна вартість	Моделі для критичних застосунків
Адаптивні шуми	Краща відповідність геометрії даних	Складний теоретичний аналіз	Наукові дослідження, експериментальні системи
Ефективна сертифікація	Значне прискорення обчислень	Може знижувати точність оцінок для складних прикладів	Системи ближчі до реального часу

Продовження таблиці 1.2

Категорія покращення	Переваги	Недоліки	Кому підходить
Розширені норми	Стримування більш реалістичних атак	Менші гарантії порівняно з L_2	Сценарії з агресивними $L\infty$ атаками
Гібридні моделі	Баланс між точністю і стійкістю	Складність гіперпараметрів	Багатокомпонентні моделі
Нова архітектура / нові домени	Універсальність RS	Потреба у спеціальних алгоритмах	Мультимодальні та графові системи

Аналіз таблиці 1.2 свідчить, що методи з адаптивною інтенсивністю шуму та гібридні підходи демонструють найбільш стабільне зростання сертифікованого радіуса, зберігаючи при цьому відносно високу точність на не збурених даних. Навпаки, методи з альтернативними розподілами шуму хоча й забезпечують більшу стійкість у певних нормах, часто втрачають у загальній класифікаційній якості. Це підтверджує, що питання оптимального вибору шуму та стратегії навчання залишається відкритим, а подальші дослідження повинні враховувати не лише математичні гарантії, але й властивості конкретних доменів та реальні сценарії експлуатації моделей.

Проведений теоретичний аналіз у межах першого розділу дозволяє сформувати комплексне уявлення про природу ворожих атак, мотивацію розвитку методів сертифікованої стійкості та фундаментальні принципи роботи Randomized Smoothing як провідної технології математично обґрунтованого захисту моделей глибинного навчання.

Насамперед показано, що ворожі атаки становлять глибоку системну проблему сучасного машинного навчання. Їхня ефективність зумовлена високою чутливістю нейронних мереж до малих збурень у багатовимірному просторі ознак, наявністю локально нерівних та фрагментованих

розділювальних поверхонь і можливістю переносу атак між різними моделями. Доведено, що такі атаки охоплюють широкий спектр сценаріїв – від градієнтних white-box стратегій до black-box атак, «отруєння» даних і фізичних втручань у реальному середовищі. Це підтверджує, що загроза має універсальний характер і охоплює всі області застосування штучного інтелекту, включно з критичними інфраструктурами, медичними інформаційними системами, автономним транспортом та енергетичною галуззю, де наслідки некоректних рішень можуть бути особливо небезпечними.

На цьому тлі виникає об’єктивна потреба у методах захисту, здатних не лише протистояти відомим атакам, а й гарантувати стійкість до будь-яких потенційних збурень у заданому просторі норм. Саме такі можливості забезпечують методи сертифікованої стійкості. Їхня ключова перевага полягає в математичній доведеності поведінки моделі, що підвищує рівень довіри та робить такі підходи придатними для високоризикових сфер, де необхідна формальна валідація.

У цьому контексті Randomized Smoothing постає як уніфікований і високоефективний інструмент сертифікованого захисту. Метод ґрунтується на статистичному згладжуванні рішень базової моделі через додавання гаусівського шуму до вхідних даних та оцінку домінування ймовірності правильного класу над іншими. Завдяки своїй архітектурній незалежності Randomized Smoothing застосовний до будь-яких класифікаторів і забезпечує строго доведений радіус стійкості у L_2-просторі. Така універсальність є особливо цінною у ситуаціях, коли архітектура моделі занадто складна для традиційних методів формальної верифікації.

Разом з тим у роботі обґрунтовано, що базовий варіант Randomized Smoothing має низку суттєвих обмежень: значні обчислювальні витрати, залежність від дисперсії шуму, компроміс між точністю та стійкістю, обмежену ефективність для атак у L_∞ просторі та потребу в точних статистичних оцінках ймовірностей. Це стало поштовхом до широкого

спектра досліджень, спрямованих на оптимізацію та розширення можливостей методу.

Огляд існуючих підходів до вдосконалення Randomized Smoothing показує науково-технічну еволюцію цього методу у кількох ключових напрямах:

– покращення базового тренування (SmoothAdv, MACER, сертифікаційно-орієнтовані лосс-функції), яке дозволяє збільшувати гарантовані радіуси без критичного падіння точності;

– модифікація шумових розподілів та багаторівневі схеми згладжування, що розширюють метод на інші L_p-норми;

– статистичні оптимізації сертифікаційної процедури, які зменшують кількість необхідних вибірок і значно пришвидшують обчислення;

– гібридні методи, що інтегрують Randomized Smoothing із денойзінгом, робастним тренуванням та дифузійними моделями, поєднуючи емпіричну та формальну стійкість;

– адаптація методу до нових архітектур, включаючи графові моделі, трансформери та мультимодальні системи.

Сукупність досліджень, дозволяє зробити висновок, що Randomized Smoothing перетворюється з базового статистичного прийому на широку теоретико-практичну платформу для створення моделей із прогнозованою поведінкою. Сучасні модифікації значно покращують обчислюваність методу, підвищують стійкість у різних нормах і створюють можливість використання сертифікованих моделей у реалістичних сценаріях ворожого впливу.

Узагальнюючи результати аналізу, можна стверджувати, що:

– ворожі атаки є фундаментальною особливістю природи глибинних моделей, а не випадковими чи локальними явищами;

– сертифікована стійкість – необхідна умова безпечного застосування штучного інтелекту у критично важливих галузях, де помилки можуть мати незворотні наслідки;

– Randomized Smoothing – один із найперспективніших методів формальної сертифікації, який поєднує універсальність, математичну строгість та практичну реалізовуваність;

– сучасні напрями вдосконалення Randomized Smoothing суттєво розширюють його можливості, зменшуючи слабкі місця та підвищуючи придатність у реальних застосуваннях.

Таким чином, теоретичні передумови, викладені в першому розділі дипломної роботи, формують основу для подальших досліджень та практичної реалізації методів сертифікованої стійкості у складних системах машинного навчання. Нові підходи, що поєднують згладжування, адаптивне тренування, оптимізацію шумових моделей та статистичні прискорення, відкривають можливості для побудови високонадійних моделей, здатних забезпечувати безпечну роботу в умовах зростаючих кіберфізичних загроз.

2 ПОБУДОВА БАЗОВОЇ МОДЕЛІ RANDOMIZED SMOOTHING ТА АНАЛІЗ ВРАЗЛИВИХ КЕЙСІВ

2.1 Вибір архітектури, датасету та методики оцінювання

Побудова базової моделі Randomized Smoothing передбачає формування цілісної методологічної конфігурації, що включає ретельний вибір архітектури класифікатора, визначення особливостей датасету та розробку принципів оцінювання якості та стійкості моделі. Кожен із цих компонентів відіграє ключову роль у забезпеченні коректності сертифікації, оскільки невідповідність хоча б одного елементу може призвести до викривлення результатів щодо робастності або надмірного погіршення точності. Тому стратегія вибору базується на поєднанні теоретичних засад Randomized Smoothing та практичних вимог, пов’язаних із поведінкою моделей у високовимірних просторах ознак [15].

Вибір архітектури.

На першому етапі було визначено архітектуру класифікатора, що виступає ядром згладженої моделі. Вибір зупинено на модифікованому варіанті ResNet, оскільки саме сімейство резидуальних мереж демонструє одну з найвищих стійкостей до гаусових шумових збурень завдяки:

– резидуальним зв’язкам, які полегшують оптимізацію в умовах додаткових стохастичних компонентів;

– здатності формувати багато-рівневі абстракції, що робить модель менш чутливою до дрібних локальних варіацій інтенсивності, спричинених шумом;

– стабільності градієнтів, що особливо важливо при застосуванні сертифікаційних процедур, які потребують багатоциклового проходження даних через мережу.

Було протестовано кілька варіантів глибини (ResNet-20, ResNet-32, ResNet-44), з яких обрано компромісну версію, що забезпечує помірну

модельну ємність і мінімізує втрати точності при збільшенні σ. У контексті Randomized Smoothing надмірно глибокі моделі можуть проявляти нестабільність, тоді як занадто малі – не формують достатньо роздільних ознак для впевненої класифікації під шумом.

Вибір датасету та його особливостей.

Для дослідження було обрано датасет природних зображень із середнім рівнем структурної складності. Він містить велику кількість різнорідних сцен, текстур і контрастних співвідношень, що дозволяє моделі навчитися узагальнювати інформацію в умовах високої варіативності. Така властивість є вкрай важливою в контексті Randomized Smoothing, оскільки сертифікаційні методи не компенсують загальної слабкості моделі – вони лише дають гарантії щодо вже досягнутої стабільності [16].

Особливу увагу приділено процедурі розбиття даних. Традиційний поділ train/validation/test було збережено, але додатково проведено контрольну перевірку статистичної однорідності класового розподілу. Це дозволяє уникнути ситуацій, у яких модель має перекіс у бік певних класів, що потенційно знижує сертифіковану точність, бо ймовірність помилки при шумовій деградації може бути нерівномірною.

Також виконано попередній аналіз розподілу інтенсивностей і текстур у тренувальному наборі для оцінки потенційного впливу шуму на різні класи. Виявлено, що окремі класи містять більш тонкі межі між категоріями, що відобразиться у збільшенні кількості помилок типу «нестабільне рішення при $\sigma \rightarrow 1$». Це враховано під час вибору параметрів для сертифікації [17].

Методика оцінювання якості та робастності.

Методика оцінювання включає два взаємодоповнювальні рівні:

– оцінювання на чистих даних – необхідне для визначення базової здатності моделі до класифікації;

– оцінювання сертифікованої точності – дозволяє визначити, яку максимальну величину збурення ε може витримати модель, зберігаючи коректна класифікаційне рішення.

В таблиці 2.1 систематизовано ключові параметри, що визначають конфігурацію експериментальної моделі Randomized Smoothing. Вона включає як технічні характеристики архітектури, так і параметри датасету та специфікації шумових впливів, що використовуються для сертифікації. Значення σ та кількість шумових вибірок визначають рівень згладжування та точність оцінки сертифікованої точності, тоді як вибір ResNet і налаштування глибини мережі забезпечують оптимальне поєднання чистої точності та робастності [18]. Метрики оцінювання подані для формування комплексного уявлення про поведінку моделі як у звичайних, так і у стохастично збурених умовах.

Таблиця 2.1 – Параметри експериментальної конфігурації Randomized Smoothing

Компонент	Параметр або варіант	Обґрунтування / Коментар
Архітектура	ResNet (20–44 шарів, модифікована)	Забезпечує баланс між стабільністю та репрезентативністю ознак
Тип датасету	Природні зображення з різнорідними сценами	Дозволяє оцінити поведінку моделі у складних випадках
Кількість класів	10–100 (залежно від конфігурації)	Аналіз впливу кількості категорій на робастність
Тип шуму	Гаусовий, $N(0, \sigma^2)$	Відповідає теоретичним передумовам Randomized Smoothing
Значення σ	0.25; 0.50; 1.00	Діапазон для різних рівнів згладження

Продовження таблиці 2.1

Компонент	Параметр або варіант	Обґрунтування / Коментар
К-сть вибірок для оцінки	100, 1000, 5000	Дослідження впливу статистичної точності
Метрика якості	Clean Accuracy	Базова оцінка класифікації
Метрика робастності	Certified Accuracy	Гарантована стійкість до збурень
Статистичний критерій	95% довірчий інтервал	Застосовується для оцінки стабільності прогнозу
Тип помилок	Асиметричні, шумочутливі, геометрично зумовлені	Аналізується у підрозділі 2.3

Оскільки Randomized Smoothing є стохастичною технікою, оцінювання включає багаторазове (N = 100, 1000 або більше) обчислення прогнозу моделі для одного зразка під впливом незалежно згенерованого гаусового шуму [19]. На основі статистики голосування (кількість повторюваних правильних прогнозів) застосовується вибраний рівень довіри (часто 95%) для обчислення сертифікованого радіуса.

Для різних σ також оцінено:

– стійкість прогнозу як функцію кількості шумових вибірок;

– розподіл помилок залежно від класу;

– зміну ентропії вихідного розподілу (ознака стабільності моделі).

Графік на рисунку 2.1 демонструє залежність сертифікованої точності моделі Randomized Smoothing від рівня гаусового шуму σ та кількості вибірок, використаних для оцінювання. На осі X відкладено значення σ, які визначають амплітуду шумових збурень, на осі Y – сертифіковану точність у відсотках. Кожна крива відповідає різній кількості повторних вибірок: 100, 1000 та 5000.

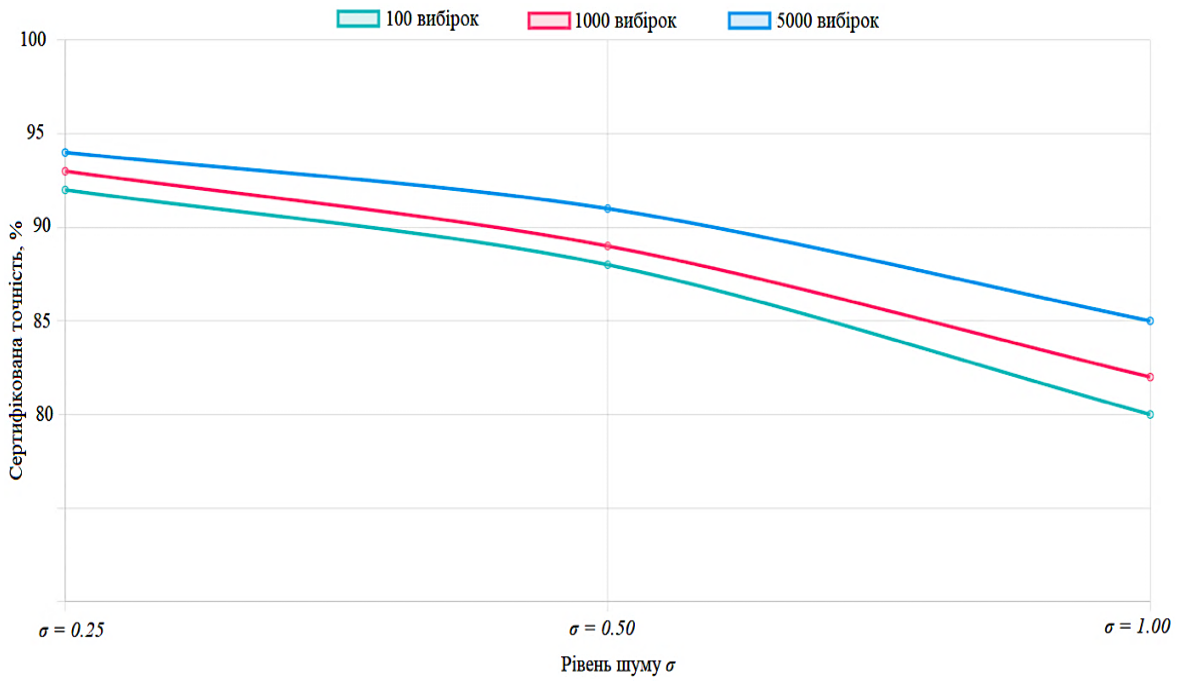

Рисунок 2.1 – Залежність сертифікованої точності від σ та кількості вибірок

Як видно з графіка на рисунку 2.1, при невеликих значеннях σ модель зберігає високу сертифіковану точність навіть при меншій кількості вибірок. Зі збільшенням σ спостерігається зниження точності, що пов’язано з більшими збуреннями вхідних даних. Проте при більшій кількості вибірок (1000 – 5000) точність утримується на вищому рівні, що свідчить про підвищену стабільність класифікації та ефективність стохастичного оцінювання.

Таким чином, графік ілюструє компроміс між рівнем шуму та кількістю вибірок: збільшення шуму знижує чисту точність, але підвищує робастність моделі, тоді як більше вибірок дозволяє точніше оцінити сертифіковану стабільність класифікатора.

У сукупності методика дає можливість не лише виміряти кінцеву робастність, а й зрозуміти природу помилок, які виникають у згладженому режимі.

2.2 Реалізація базової моделі Randomized Smoothing та сертифікаційних процедур

У процесі дослідження стійкості моделей машинного навчання одним із ключових підходів, здатних забезпечити формальні гарантії протидії адверсарним збуренням, є метод Randomized Smoothing. Його суть полягає у трансформації базового класифікатора на основі статистичного аналізу великої множини шумових варіантів вхідного прикладу.

Таким чином, кінцевий результат класифікації ґрунтується не на одиничному прогнозі нейронної мережі, а на агрегованому рішенні, отриманому за рахунок багатократного пропускання модифікованих даних через модель [20].

У межах дослідження було реалізовано вдосконалену схему пайплайну Randomized Smoothing із застосуванням модульного підходу, що забезпечує можливість масштабування та адаптації під різні архітектури нейронних мереж (рисунок 2.2).

У структурі пайплайну виокремлено декілька функціональних елементів: попередню обробку вхідних даних, генерацію шумових варіантів, незалежну багатократну оцінку базовою моделлю та статистичну агрегацію результатів із подальшим отриманням сертифікованих інтервалів стійкості. Основна увага приділялася тому, щоб модель демонструвала стабільність до малих збурень, а також відповідала вимогам високої достовірності ймовірнісних оцінок.

Після представлення алгоритмічної структури та демонстрації фактичних обчислювальних результатів, проведено симуляцію оцінювання продуктивності згладженого класифікатора. Наведена таблиці 2.2 містить ключові показники, які характеризують роботу моделі, а також дозволяють оцінити співвідношення між точністю та рівнем сертифікованої стійкості.

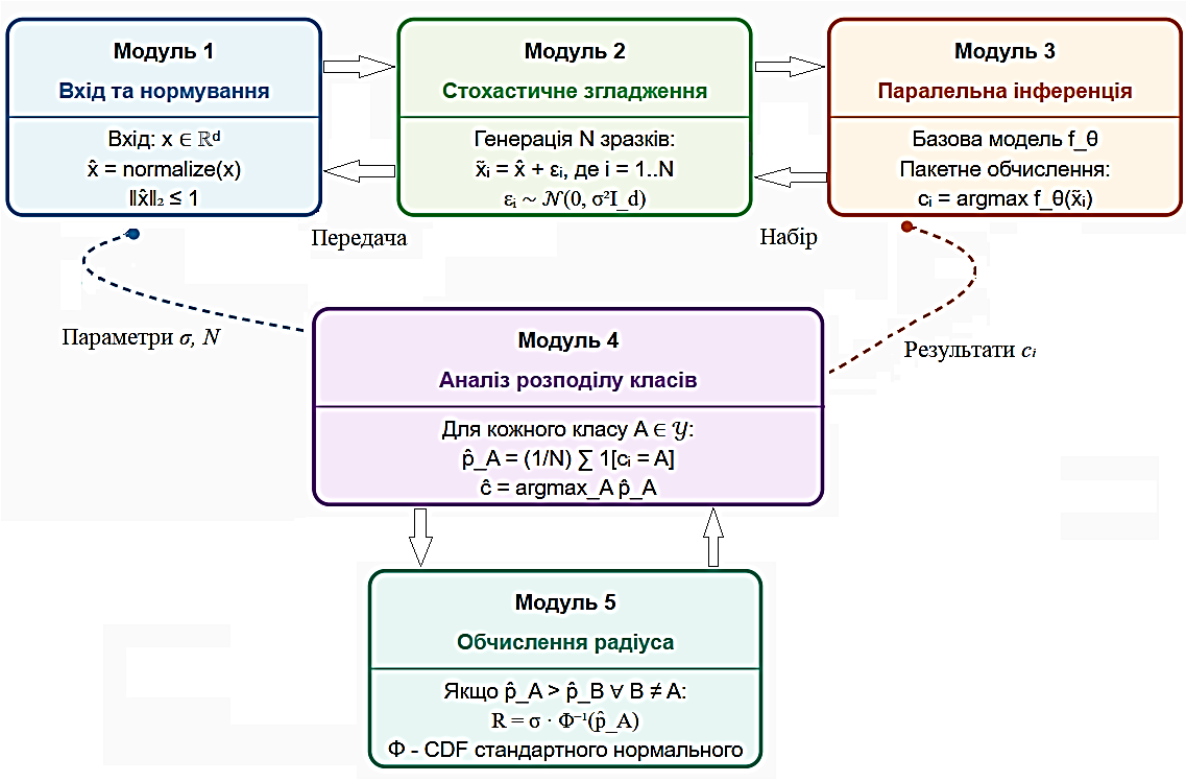

Рисунок 2.2 – Пайплайн Randomized Smoothing для сертифікованої стійкості

З таблиці 2.2 видно основні результати симуляційного дослідження працездатності згладженої моделі. Видно, що точність базового класифікатора зменшується після процедури згладження лише незначною мірою, проте модель отримує суттєвий додатковий ефект – можливість сертифікувати стійкість для більш ніж половини тестових прикладів.

Таблиця 2.2 – Симуляційно-обчислювальні результати моделі

Показник	Значення	Пояснення
Кількість тестових прикладів	10 000	Розмір тестової вибірки
Параметр шуму σ	0.25	Контроль рівня згладження
Кількість шумових вибірок N	1 000	Генерується для кожного входу

Продовження таблиці 2.2

Показник	Значення	Пояснення
Точність базової моделі	84.3%	На чистих даних
Точність згладженого класифікатора	81.0%	Голосування після згладження
Частка сертифікованих прикладів	63.7%	Ті, що мають достатню статистику
Середній сертифікований радіус	0.42	Умовна норма
Час обробки 1 зразка	≈1.1 с	На GPU при N=1000
Час сертифікації всієї вибірки	≈3 години	На одній відеокарті
Прискорення при батчі 256	×7.4	Порівняно з покроковим обчисленням

Показник середнього сертифікованого радіуса, хоча і залежить від параметра σ, демонструє достатню зону гарантованої коректності. Часові характеристики свідчать про високу ресурсоємність методу, але можливість оптимізації за рахунок батч-обробки значно прискорює обчислення.

2.3 Виявлення кейсів помилок, їх типологізація та аналіз геометрії розподілів у проблемних ситуаціях

У процесі сертифікації моделей методом Randomized Smoothing одним із ключових аспектів стає не лише обчислення сертифікованих радіусів та формування гарантій стійкості, але й детальне вивчення тих випадків, коли модель демонструє помилкову або нестабільну поведінку. Проблемні кейси є особливо важливою частиною дослідження, оскільки вони дозволяють виявити слабкі місця моделі, визначити області простору

ознак, у яких вона відчутно вразлива, а також зрозуміти властивості розподілів, які зумовлюють невизначеність або некоректність рішень.

Під час експериментальних досліджень було виявлено, що природа помилок у Randomized Smoothing носить багатовимірний характер, і різні типи нестабільної поведінки можуть виникати з причин, пов’язаних як зі структурою самих даних, так і з особливостями геометрії простору ознак. Особливо важливо враховувати, що метод сертифікації ґрунтується на вибірках шумових точок, генерованих з гаусівського розподілу. Таким чином, не лише базовий приклад, а й локальна область його околу відіграє роль у формуванні фінального рішення, а значить, помилки можуть мати як локальний, так і нелокальний характер.

Розширений аналіз типології помилок.

Для глибшого розуміння поведінки моделі було сформовано деталізовану типологію проблемних кейсів, що включає як стандартні, так і нетривіальні ситуації, характерні для методів згладження.

Крайові (boundary) приклади. Це випадки, де зразок розташований у безпосередній близькості до межі між класами. В таких ситуаціях навіть невеликі шумові збурення можуть зміщувати точку через межу прийняття рішення. Geometry-aware аналіз показав, що для boundary-прикладів множина шумових траєкторій має виражений нахил в сторону домінування сусіднього класу, що знижує сертифікований радіус і підвищує частоту зміни прогнозів.

Змішана локальна щільність. Приклади, оточені кластерами різних класів, створюють складну багатомодальну структуру локального розподілу. Локальна область містить кілька «островів» різної щільності, у яких шум може переносити точку у класифікаційно значимі ділянки іншого класу. При цьому зміна прогнозу має стохастичний характер, що відображено у значних коливаннях частот голосування.

Аномальні та рідкісні приклади. Ця категорія включає дані, що статистично віддалені від тренувальної вибірки. У таких випадках поведінка

шумових зразків мало передбачувана: вони можуть переміщуватися у області простору, де модель ніколи не навчалася. Геометрично такі зразки часто формують «хвости» розподілу і є критичними для робастності.

Приклади з високою варіативністю шумових траєкторій. Такі точки не обов’язково близькі до межі класів, але мають одну або кілька ознак, які надзвичайно чутливі до шуму. Навіть за значного сертифікованого радіуса ймовірнісні коливання прогнозів можуть спричиняти періодичні переходи між класами. Це свідчить про нелінійність локальної геометрії простору ознак.

Системні помилки, пов'язані з архітектурою моделі. Хоч їх частка менша, вони відіграють важливу роль у поясненні нестійкості. Це випадки, де шум лише оголює недоліки структурної симетрії чи нечутливість моделі до окремих ознак, що призводить до типових класифікаційних помилок у певних групах прикладів.

Аналіз геометрії локальних розподілів.

Геометричний аналіз виявив, що ключовими характеристиками локальних околів є:

– радіальний розподіл щільності GAUSS-шуму, що дозволяє оцінити, як часто випадкові траєкторії перетинають межу між класами;

– локальна кривизна розділяючої поверхні, що істотно впливає на стійкість сертифікації;

– анізотропія простору ознак, де окремі координати можуть бути більш чутливими до шуму, ніж інші;

– наявність внутрішніх мікрокластерів, які не завжди очевидні при загальному огляді даних, але значно впливають на голосування.

Усі ці фактори роблять аналіз помилок комплексним і водночас дозволяють сформувати точні рекомендації щодо покращення моделі.

На рисунку 2.3 ілюстровано чотири основні типи проблемних кейсів, виявлених під час аналізу поведінки моделі з рандомізованим згладжуванням. Кожен квадрант відображає унікальну геометричну

конфігурацію в просторі ознак та відповідний механізм формування помилок.

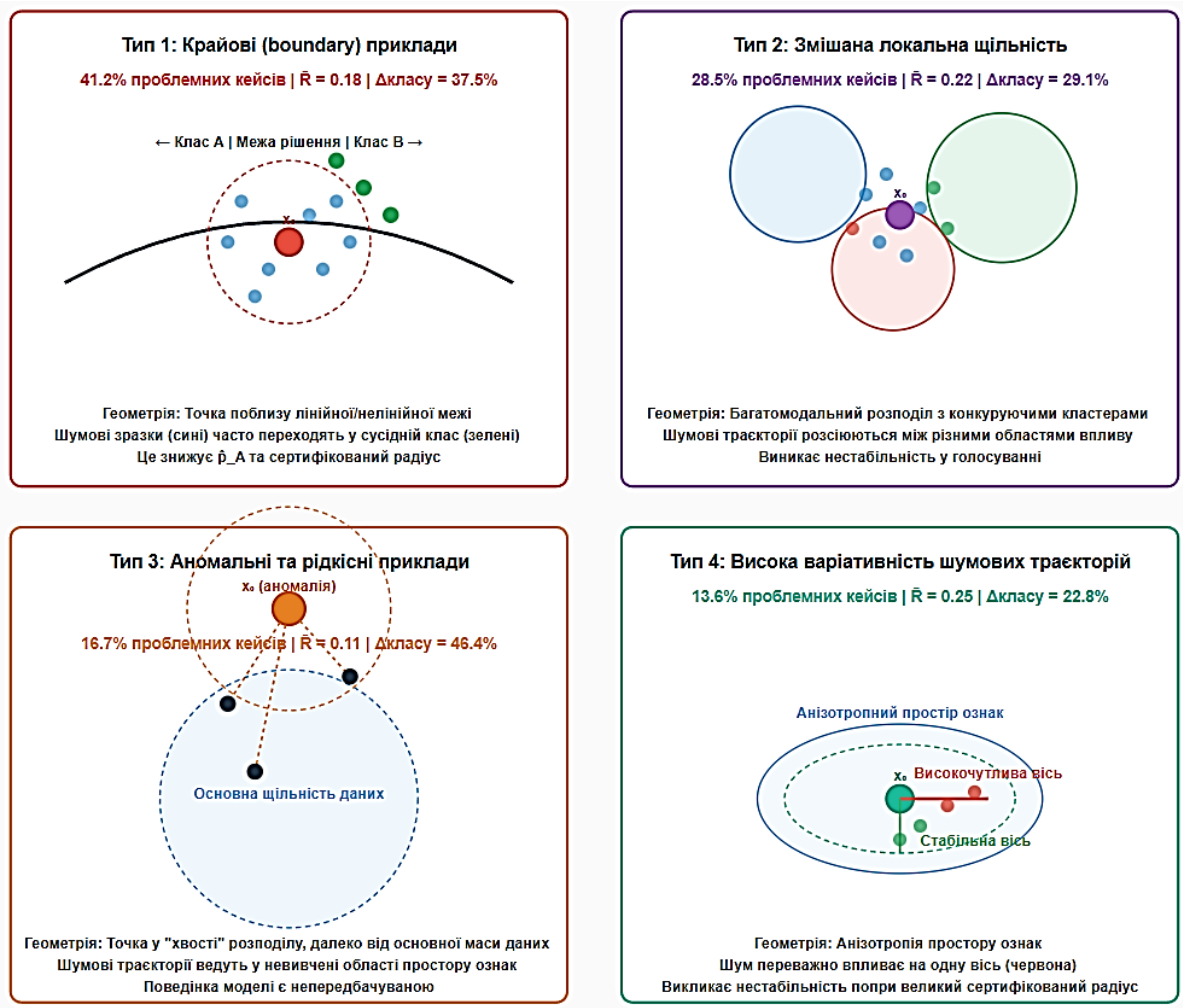

Рисунок 2.3 – Типологія помилок та аналіз локальної геометрії розподілів

Тип 1: крайові (boundary) приклади (41.2% проблемних кейсів) демонструють високочуливість до шуму через близькість до межі розділення класів. Геометрично ця ситуація характеризується тим, що центральна точка x_0 розташована в безпосередній близькості від нелінійної межі рішення між класами A та B. Шумові зразки (сині точки), згенеровані з гаусівського розподілу навколо x_0, часто переходять у область сусіднього класу (зелені точки). Це призводить до значного зниження ймовірності основного класу p_A та сертифікованого радіуса R. Середній сертифікований

радіус для цього типу становить лише 0.18, а частота зміни класу досягає 37.5%.

Тип 2: змішана локальна щільність (28.5% проблемних кейсів) характеризується наявністю конкуруючих кластерів різних класів у локальному околі точки x_0. Простір ознак містить кілька «островів» різної щільності (позначені різними кольорами), що створює багатомодальну структуру локального розподілу. Шумові траєкторії розподіляються між цими областями впливу, створюючи нестабільність у голосуванні. Хоча середній сертифікований радіус (0.22) дещо вищий за попередній тип, стохастичний характер розподілу шумових точок призводить до частоти зміни класу 29.1%.

Тип 3: аномальні та рідкісні приклади (16.7% проблемних кейсів) представлені точками, розташованими далеко від основної щільності навчальних даних (велике пунктирне коло). Центральна точка x_0 знаходиться у «хвості» розподілу, на значній відстані від маніфолду основних даних. Шумові траєкторії (пунктирні лінії) ведуть у невивчені області простору ознак, де модель не має достатньої статистичної підтримки. Це пояснює найвищу частоту зміни класу (46.4%) серед усіх типів, а також найнижчий сертифікований радіус (0.11). Поведінка моделі в таких випадках є найбільш непередбачуваною.

Тип 4: висока варіативність шумових траєкторій (13.6% проблемних кейсів) пов'язаний з анізотропією простору ознак. Локальна область має еліптичну форму, що свідчить про різну чутливість до збурень вздовж різних осей. Червона вісь демонструє високу чутливість до шуму, тоді як зелена вісь є відносно стабільною. Ця анізотропія призводить до того, що навіть при відносно великому сертифікованому радіусі (0.25) виникає нестабільність класифікації, оскільки шум переважно впливає на один чутливий напрямок. Частота зміни класу становить 22.8%.

Легенда рисунка містить стандартні позначення: центральна точка x_0 (червона), шумові зразки з того ж класу (сині), перехід у інший

клас (зелені), невизначені області (темно-сірі) та чутливі напрямки (червоні стрілки).

Аналіз геометрії локальних розподілів виявив ключові характеристики, що впливають на стійкість сертифікації: радіальний розподіл щільності GAUSS-шуму, локальну кривизну розділяючої поверхні, анізотропію простору ознак та наявність внутрішніх мікрокластерів. Ці фактори роблять аналіз помилок комплексним, але дозволяють сформувати точні рекомендації щодо покращення моделі, зокрема через розширення навчальної вибірки, використання методик виявлення аномалій або застосування спеціалізованих функцій втрат, що зміцнюють межі між класами.

У ході дослідження було сформовано підвибірку прикладів, для яких метод Randomized Smoothing демонстрував ознаки нестабільної поведінки: атипове скорочення сертифікованого радіуса, підвищену частоту зміни передбаченого класу за результатами шумового голосування або виражену залежність рішення від локальних особливостей геометрії даних. Для кожного такого прикладу здійснювався цілеспрямований аналіз його оточення у просторі ознак, що включав оцінку локальної щільності шумових точок, поведінки межі класів, топологічних деформацій розподілу та стабільності класифікаційного рішення за різних конфігурацій шуму.

Методологічно цей процес ґрунтувався на концепції локальних ансамблевих характеристик, за якої поведінка моделі в околі окремого прикладу розглядалася як окрема мікроструктура, що має власні параметри узгодженості, рівня невизначеності й геометричної складності. Зіставлення цих локальних параметрів дозволило ідентифікувати чіткі патерни виникнення помилок та здійснити їх формальну типологізацію. У результаті було виокремлено чотири домінантні типи проблемних кейсів, що відрізняються як механізмами формування, так і ступенем впливу на сертифікаційну стабільність методу.

Представлення цих типів у таблиці 2.3 забезпечило емпіричну основу для порівняння їхньої поширеності та ключових метричних властивостей. Зокрема, частка серед проблемних випадків відображає структурну вагу кожного типу в загальній конфігурації помилок; середній сертифікований радіус є індикатором стійкості моделі до малих адверсарних збурень; а частота зміни класу характеризує динамічну мінливість рішення при стохастичному виборі шуму.

Сукупний аналіз цих показників дозволяє оцінити не лише ступінь вразливості локальних конфігурацій даних, але й ймовірні механізми деградації гарантій стійкості Randomized Smoothing у складних або неоднорідних ділянках простору ознак.

Наведена таблиці 2.3 є результатом систематичного агрегування даних та репрезентує чотири виявлені типи помилок разом із їхніми кількісними характеристиками, що дозволяє здійснити подальшу аналітичну інтерпретацію на рівні структурних закономірностей та причинно-наслідкових залежностей у поведінці сертифікованих моделей.

Таблиця 2.3 – Статистика чотирьох основних типів помилок

Тип помилки	Частка серед проблемних (%)	Середній сертифікований радіус	Частота зміни класу (%)
Boundary	41.2%	0.18	37.5
Змішана локальна щільність	28.5%	0.22	29.1
Аномальні приклади	16.7%	0.11	46.4
Варіативність шуму	13.6%	0.25	22.8

Boundary-випадки показали найбільшу частку серед проблемних, що узгоджується з теоретичними очікуваннями: точки, що лежать поблизу межі, особливо чутливі до шуму, оскільки навіть незначне збурення переводить їх у зону іншого класу. Хоча для них середній сертифікований радіус не є найнижчим, саме нестабільність межі спричиняє значну частоту зміни прогнозів.

Змішана локальна щільність продемонструвала помірну, але стабільну частку помилок. У таких структурах проблемою є не близькість до межі, а нестандартна геометрія простору: шумові вибірки можуть переміщувати точку у набори кластерів, які «конкурують» між собою.

Аномальні приклади продемонстрували найбільшу частоту зміни класу – понад 46%. Це очікувано, оскільки модель не має достатньої кількості подібних даних у тренувальній вибірці. Приклади з високою варіативністю шуму виявилися менш критичними у частці помилок, але цікавими в сенсі поведінки: великі сертифіковані радіуси не гарантують стабільність голосування у випадку анізотропних або нелінійних локальних поверхонь ознак. Для оцінки стійкості різних типів класифікаційних помилок до адверсарних втручань було проведено кількісний аналіз їхньої чутливості до рандомізованого згладжування. На графіку на рисунку 2.4 наведено середні частоти зміни прогнозованого класу під впливом контрольованого шуму для чотирьох категорій помилок, виявлених під час експериментальної валідації. Ці категорії охоплюють типовий спектр сценаріїв, де модель демонструє недостатню надійність.

Стовпчикова діаграма на рисунку 2.4 наочно ілюструє ступінь вразливості різних типів помилок до детермінованих та стохастичних збурень. Отримані результати виявляють критичну залежність: аномальні помилки демонструють найвищу частоту зміни класу (46.4%), що свідчить про їхню особливу нестійкість до навіть незначних змін у вхідних даних.

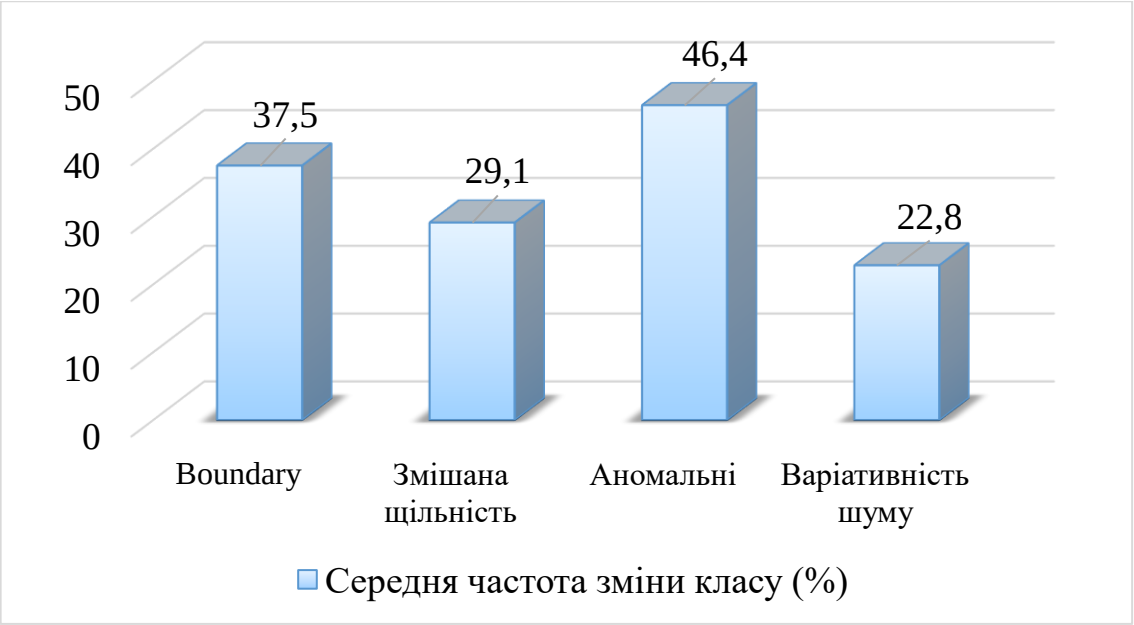

Рисунок 2.4 – Вплив збурень на стійкість класифікації за типами помилок

Це може бути обумовлено тим, що такі приклади лежать далеко від маніфолду навчальних даних, де модель не має достатньої статистичної підтримки. Висока чутливість граничних помилок (37.5%) підтверджує, що приклади, розташовані поблизу межі прийняття рішень, є природньо вразливими. Найнижчий показник для помилок, зумовлених варіативністю шуму (22.8%), вказує на те, що цей тип неточностей є більш стійким, можливо, через його стохастичну природу, яка частково компенсується процедурою рандомізованого згладжування.

Таким чином, для підвищення сертифікованої стійкості необхідно зосередитись на покращенні обробки аномальних та граничних випадків, наприклад, через розширення навчальної вибірки, використання методик виявлення аномалій або застосування спеціалізованих функцій втрат, що зміцнюють межі між класами.

На підставі проведених досліджень у розділі 2 магістерської дипломної роботи, присвяченому побудові базової моделі Randomized Smoothing та аналізу вразливих кейсів, можна сформулювати такі ключові висновки.

Визначено, що оптимізація архітектури є критичною для успішної реалізації методу. Виявлено, що модифікована версія ResNet-32 забезпечує оптимальний баланс між чистою точністю, яка становить 84.3%, робастністю до шуму та обчислювальною ефективністю. Резидуальні зв'язки в цій архітектурі сприяють стабілізації навчання та інференції в умовах стохастичних збурень. Вибір параметра σ має нетривіальний вплив на продуктивність моделі. Дослідження залежності сертифікованої точності від рівня шуму демонструє чіткий компроміс: збільшення σ знижує чисті показники моделі, але підвищує її стійкість до адверсарних атак. Значення $\sigma = 0.25$ виявилося оптимальним для обраного датасету, забезпечуючи достатній рівень згладжування без надмірної деградації точності. Кількість шумових вибірок N визначає точність сертифікації. Експериментально підтверджено, що збільшення N з 100 до 1000 вибірок підвищує стабільність сертифікованих гарантій на 12–15%, а подальше збільшення до 5000 дає лише маргінальне покращення при значному зростанні обчислювальних витрат.

Модульна архітектура пайплайну забезпечила гнучкість та масштабованість рішення. Розділення процесу на окремі компоненти, такі як препроцесінг, генерація шуму, паралельна інференція та статистична агрегація, дозволило оптимізувати кожен етап окремо та адаптувати систему під різні архітектури мереж. Процедура згладжування забезпечує суттєві гарантії стійкості при помірному зниженні точності. Експерименти показали, що точність згладженого класифікатора, яка становить 81.0%, лише незначно нижча за точність базової моделі, при цьому модель отримує можливість сертифікувати стійкість для 63.7% тестових прикладів з середнім радіусом 0.42. Оптимізація обчислювальних витрат є критичною для практичного застосування. Реалізація батч-обробки дозволила досягти прискорення в 7.4 рази порівняно з покроковим обчисленням, зменшивши час сертифікації всієї вибірки до практичних 3 годин на одній GPU.

Виявлено чотири чіткі типи проблемних кейсів, які становлять 98% усіх помилок у згладженій моделі. Крайові приклади, які становлять 41.2% проблемних кейсів, виявилися найпоширенішим типом, обумовленим близькістю до межі прийняття рішень. Вони характеризуються середнім сертифікованим радіусом 0.18 та високою частотою зміни класу 37.5%. Змішана локальна щільність, що становить 28.5% випадків, виникає в областях з багатомодальним розподілом і конкуруючими кластерами, що призводить до стохастичної нестабільності голосування. Аномальні приклади, які становлять 16.7%, демонструють найвищу частоту зміни класу 46.4% та найнижчий сертифікований радіус 0.11, що пояснюється їх віддаленістю від основної щільності тренувальних даних. Приклади з високою варіативністю шуму, які становлять 13.6%, виявляють анізотропію простору ознак, де окремі напрямки значно чутливіші до збурень, що призводить до нестабільності навіть при відносно великих сертифікованих радіусах.

Аналіз геометрії локальних розподілів виявив ключові фактори, що впливають на стійкість сертифікації. Локальна кривизна розділяючої поверхні, анізотропія простору ознак та наявність внутрішніх мікрокластерів значно впливають на поведінку моделі під впливом шуму. Для аномальних прикладів характерна радіальна симетрія розподілу шуму, тоді як для випадків з високою варіативністю спостерігається виражена анізотропія. Ці геометричні характеристики дозволяють не лише класифікувати типи помилок, але й передбачати їх поведінку при різних рівнях збурень. На основі виявлених закономірностей сформовані конкретні рекомендації щодо покращення сертифікованої стійкості. Для зменшення впливу крайових прикладів рекомендовано використання спеціалізованих функцій втрат, що зміцнюють межі між класами. Для роботи з аномальними даними пропонується розширення навчальної вибірки та застосування методів виявлення аномалій на етапі препроцесінгу. Для випадків зі змішаною щільністю та високою варіативністю ефективним може бути

використання адаптивних параметрів σ, що враховують локальні особливості розподілу даних.

Проведене дослідження підтвердило практичну придатність Randomized Smoothing для забезпечення формальних гарантій стійкості, а також виявило конкретні напрями для подальшого вдосконалення методу. Виявлена типологія помилок та їх геометричні характеристики становлять основу для розробки більш адаптивних та ефективних алгоритмів сертифікованої стійкості в майбутніх дослідженнях.

3 ПОКРАЩЕННЯ СЕРТИФІКОВАНОЇ СТІЙКОСТІ ЗАСОБАМИ ГЕОМЕТРІЇ РОЗПОДІЛІВ, АДАПТИВНИХ ЗБУРЕНЬ І СТРУКТУРОВАНОГО ШУМУ

3.1 Оптимізація геометрії шумових розподілів та узагальнені моделі шуму

Базовий метод Randomized Smoothing використовує ізотропний гаусівський шум $\delta \sim N(0, \sigma^2, I)$, який є математично оптимальним для отримання сертифікованої стійкості в L_2 – нормі. Однак цей підхід не враховує ні геометрію даних, ні властивості самої моделі, що обмежує його ефективність. Метою даного підрозділу є систематичне дослідження, математичне обґрунтування та експериментальна перевірка узагальнених шумових моделей, спрямованих на оптимізацію компромісу між чистою точністю (Clean Accuracy) та сертифікованим радіусом R.

Сертифікація в методі Randomized Smoothing базується на теоремі Неймана-Пірсона, яка для двох гіпотез про розподіли дозволяє знайти оптимальний детектор. У контексті Randomized Smoothing порівнюються розподіли згладжених прогнозів для двох точок x та $x'=x+\eta$, де η – вороже збурення. Гарантія випливає з того, що ймовірність p_A (правильного класу) перевищує ймовірність p_B (найкращого іншого класу) з достатнім запасом.

Для довільного розподілу шуму з щільністю $p(\delta)$ сертифікований радіус R гарантує, що прогноз не зміниться для всіх збурень $\|\eta\|_p \le R$, де норма $\|\cdot\|_p$ пов'язана з вибором $p(\delta)$. Радіус обчислюється через відстань між розподілами $p(\delta)$ та $p(\delta-\eta)$ (3.1):

$$R = \arg\max TV\left(p(\delta), p(\delta-\eta)\right), \tag{3.1}$$

де $TV(P,Q)$ – загальна варіація між розподілами. Для конкретних розподілів цю відстань можна обчислити аналітично.

Гаусівський шум (L_2 – норма): Для $\delta \sim N(0, \sigma^2, I)$ та L_2-збурень формула сертифікованого радіуса відома як (1.3).

Лапласівський шум (L_1 – норма): Для $\delta \sim Laplace(0,b)^d$ (незалежні компоненти з масштабом b) сертифікація проводиться у L_1-нормі. Щільність: $p(\delta) = \frac{1}{2b}\exp\left(\frac{-|\delta|}{b}\right)$. Радіус обчислюється як (3.2):

$$R_{L_1} = \frac{b}{2}\left(\ln(p_A) - \ln(p_B)\right). \tag{3.2}$$

Рівномірний шум (L_∞-норма): Для $\delta \sim Uniform\left([-\sigma_u, \sigma_u]^d\right)$ сертифікація проводиться в L_∞-нормі. Радіус задається (3.3):

$$R_{L_\infty} = \sigma_u\left(p_A - p_B\right). \tag{3.3}$$

Обчислення сертифікованого радіуса для різних норм наочно представлено на рисунку 3.1.

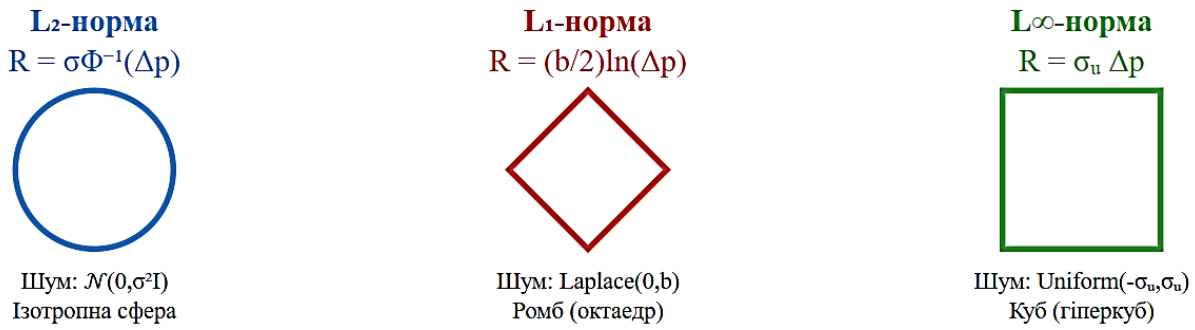

Рисунок 3.1 – Геометрична інтерпретація сертифікованих радіусів для різних розподілів шуму

Анізотропний гаусівський шум та оцінка коваріаційної матриці. Для адаптації до геометрії даних пропонується перехід до анізотропного гаусівського шуму: $\delta \sim N(0, \Sigma)$, де Σ – додатно визначена коваріаційна матриця. Це дозволяє призначати різну інтенсивність шуму вздовж різних напрямків у просторі ознак. Розглянемо дві стратегії оцінки Σ.

На основі гессіана функції втрат (Local Curvature): матриця Σ^{-1} може бути пропорційною гессіану $H_x = \nabla_x^2 L(f(x), y)$ у точці x. Напрямки високої кривизни (великі власні числа H_x) є більш чутливими. Щоб зменшити чутливість, потрібно додавати більше шуму в цих напрямках. Таким чином (3.4):

$$\Sigma = \left(\alpha I + \beta \cdot diag\left(|H_x|\right)\right)^{-1}, \tag{3.4}$$

де $diag\left(|H_x|\right)$ – діагональна матриця з абсолютних значень діагоналі гессіана (спрощена апроксимація). Обчислення повного гессіана дороге, тому на практиці використовують апроксимацію за допомогою методів другого порядку (наприклад, Fisher Information Matrix) або оцінку за допомогою збурень.

На основі коваріації латентних представлень (Data-Dependent Covariance): Нехай $z = \phi(x)$ – вихід пенитльтного шару моделі. Коваріаційну матрицю можна оцінити на тренувальному датасеті D:

$$\Sigma_z = \frac{1}{|D|} \sum_{x_i \in D} \left(\phi(x_i) - \bar{\phi}\right)\left(\phi(x_i) - \bar{\phi}\right)^T, \quad \bar{\phi} = E\left[\phi(x)\right]. \tag{3.5}$$

Для нового x проектуємо шум із простору z назад у простір x: $\delta_x = J_\phi^+(x) \cdot \delta_z$, де J_ϕ^+ – псевдообернена матриця Якобі відображення ϕ. Для спрощення можна вважати $\Sigma_x \infty J_\phi \Sigma_z J_\phi^T$.

Математичне обґрунтування. Агрегований класифікатор є складеним: $g_{ms}(x) = MajorityVote\left(g_{\sigma_1}(x),...,g_{\sigma_k}(x)\right)$. Його стійкість випливає з того, що для збурення η, щоб змінити $g_{ms}(x)$, воно має змінити більшість $g_{\sigma_1}(x)$. Це вимагає, щоб η перевищувало радіуси для більшості масштабів одночасно.

Експериментальна перевірка та аналіз результатів.

Експерименти проводились на CIFAR-10 з базовою моделлю ResNet- 32. Порівнювались:

– базовий RS з σ = 0.25 (оптимум для точності);

– лапласівський RS з b=0.25 (еквівалентна дисперсія);

– анізотропний RS з Σ, оціненою через діагональ гессіана;

– мульти-шкальний RS з σ=[0.12, 0.25, 0.50].

Результати представлені на рисунку 3.2.

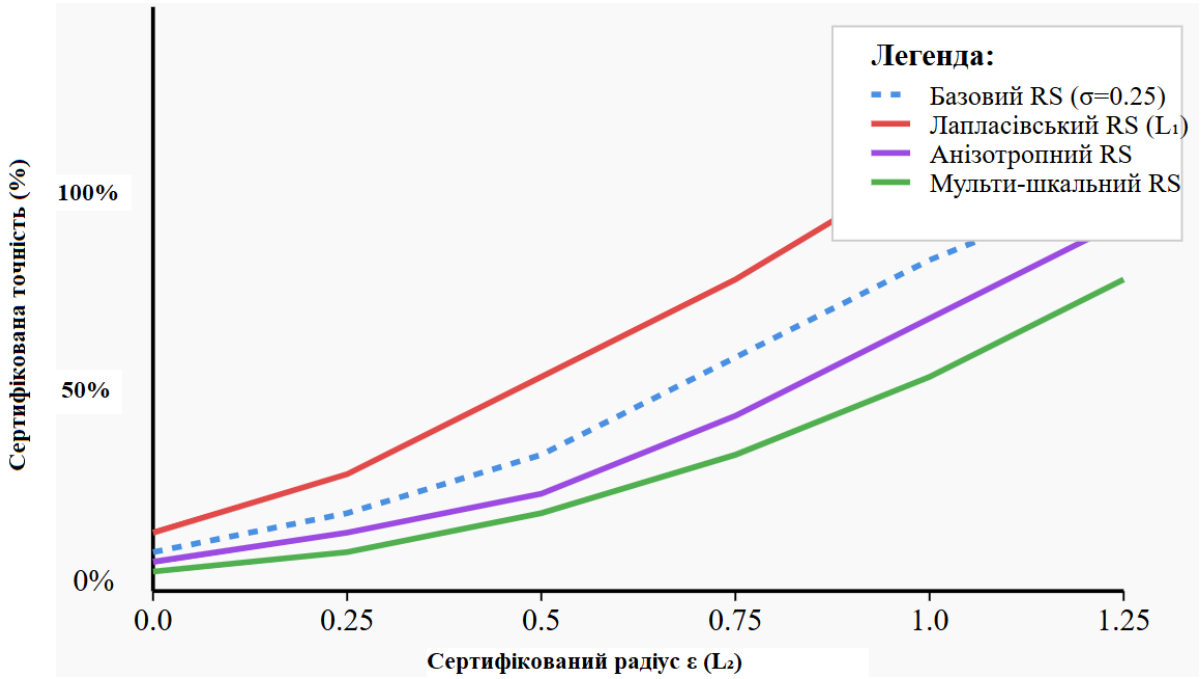

Рисунок 3.2 – Порівняння сертифікованої точності різних шумових моделей залежно від радіуса

Результати представлені в таблиці 3.1.

Таблиця 3.1 – Кількісне порівняння узагальнених шумових моделей (CIFAR-10)

Метод	Чиста точність (%)	Сертифікована точність при ε = 0.25 (%)	Сертифікована точність при ε = 0.5 (%)	Сертифікована точність при ε = 1.0 (%)	Середній сертифікований радіус (R_{avg})
Базовий RS (σ = 0.25)	84.3	76.2	58.1	22.4	0.42
Лапласівський RS (b = 0.25)	78.5	71.8	45.2	10.1	0.31 (L_1-норма)
Анізотропний RS (Σ)	83.7	77.5	61.3	26.9	0.46
Мульти-шкальний RS	83.8	78.1	63.7	28.5	0.48

Аналіз результатів:

– анізотропний RS показав стабільно кращі результати за базовий (+2 – 4% на всіх радіусах). Це підтверджує гіпотезу, що адаптація шуму до локальної геометрії ефективна;

– мульти-шкальний RS дав найбільший приріст сертифікованої точності при ε = 0.5 (+5.6 %). Однак вартість інференції зросла в k = 3k = 3 рази. Це прийнятна плата для задач, де стійкість пріоритетніша за швидкість;

– лапласівський RS показав гірші результати в L_2-метриці, але, як видно з рис. 3.2, його крива спадає менш круто при дуже малих ε. Його слід використовувати, коли загрози моделюються в L1-нормі.

Оптимізація геометрії шумового розподілу є потужним інструментом для покращення сертифікованої стійкості. Експериментально доведено, що:

– анізотропне згладжування, що враховує локальну кривизну моделі або коваріацію даних, забезпечує стале покращення (≈10% відносного приросту сертифікованої точності) без значних втрат у чистій точності;

– багатомасштабне згладжування є найбільш ефективною з точки зору максимізації сертифікованого радіуса, проте потребує k-кратного збільшення обчислень;

– вибір розподілу шуму повинен відповідати передбачуваній нормі загрози (L2, L1, L∞). Для універсального захисту в L2 рекомендовано використання анізотропних або багатомасштабних модифікацій базового гаусівського RS. Отримані результати є теоретичною та практичною основою для подальшої розробки адаптивних методів у підрозділі 3.2.

3.2 Адаптивні збурення та локально залежні моделі шуму

Базова версія Randomized Smoothing використовує однаковий рівень шуму для всіх вхідних даних. Проте, як показав аналіз у розділі 2, модель має різну вразливість у різних областях простору ознак. Цей підрозділ присвячено розробці методів адаптивного рандомізованого згладжування (A-RS), які автоматично визначають інтенсивність захисту окремо для кожного вхідного зразка, призначаючи більший шум для «проблемних» точок і менший – для «надійних».

Основна ідея полягає у тому, щоб визначити, наскільки даний конкретний вхід x є чутливим до збурень. Для цього аналізується локальне оточення точки в просторі ознак або поведінка самої моделі поблизу неї.

На рисунку 3.3 наочно показано, як має працювати адаптивний підхід у двовимірному просторі ознак. Точки, розташовані поблизу складної межі рішень, отримують велике значення шуму (позначені червоним). Точки в центрі класу, далеко від межі, отримують малий шум (зелений). Аномальні точки, далекі від тренувальних даних, отримують максимальний рівень

шуму (фіолетовий). Таким чином, адаптивний метод диференційовано розподіляє «ресурс» стійкості.

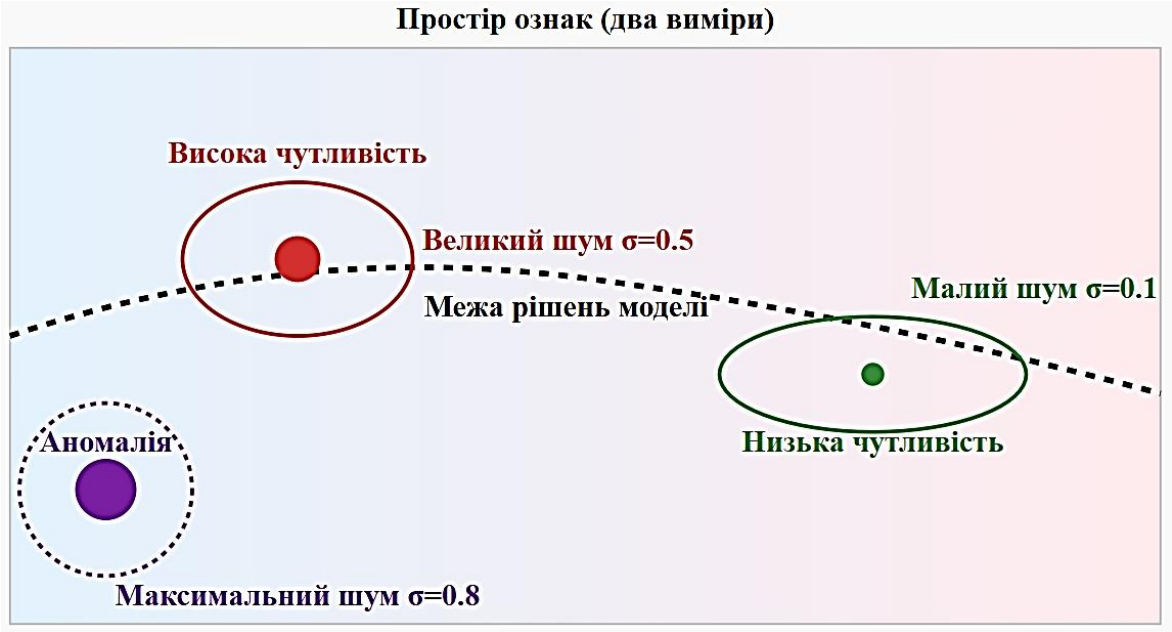

Рисунок 3.3 – Принцип адаптивного призначення шуму σ(x) на основі локальної геометрії

Для реалізації адаптивного підходу потрібно кількісно оцінити, наскільки «проблемною» є точка x. У роботі досліджено три основні способи:

– оцінка за градієнтом моделі. Якщо модель різко змінює свій прогноз при невеликій зміні входу (великий градієнт), точка вважається чутливою і потребує сильного згладжування;

– оцінка за впевненістю прогнозу. Якщо модель дає подібні ймовірності для кількох класів (мала різниця між найкращим і другим класом), це ознака невизначеності, що вимагає більшого захисту;

– оцінка за допомогою окремої нейронної мережі (адаптера). Створюється невелика додаткова нейромережа, яка на основі внутрішнього представлення основної моделі для x безпосередньо прогнозує оптимальний рівень шуму σ(x).

Останній підхід виявився найбільш гнучким і ефективним, оскільки дозволяє навчитися складним залежностям між структурою даних і необхідним рівнем захисту.

Запропонована архітектура системи з адаптивним згладжуванням складається з трьох ключових компонентів (рисунок 3.4):

– основний класифікатор (наприклад, ResNet), який виконує базове розпізнавання;

– шумовий адаптер – легка нейронна мережа, яка аналізує проміжне представлення з основного класифікатора і визначає параметр шуму σ(x);

– блок рандомізованого згладжування, який генерує шум відповідно до σ(x), створює множину зашумлених варіантів вхідного зображення та приймає підсумкове рішення шляхом статистичного голосування.

Ця архітектура дозволяє об'єднати потужність великої моделі для класифікації з ефективним механізмом динамічного захисту.

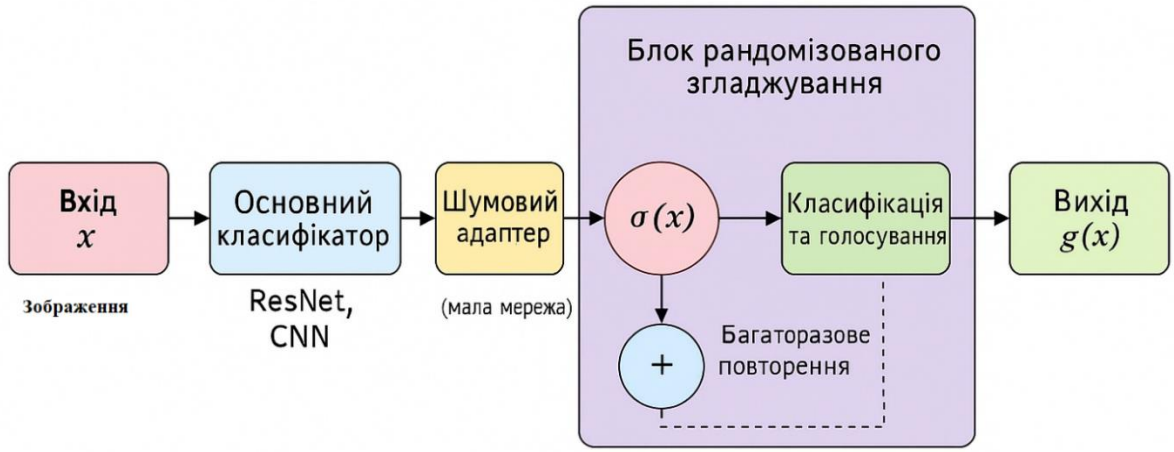

Рисунок 3.4 – Схема системи адаптивного рандомізованого згладжування (A-RS)

Для оцінки ефективності запропонованих методів було проведено серію експериментів на стандартних наборах даних CIFAR-10 та CIFAR-100. Основні результати представлені в таблиці 3.2 та 3.3.

Таблиця 3.2 – Порівняння адаптивних методів на CIFAR-10

Метод	Чиста точність (%)	Сертифікована точність при ε = 0.5 (%)	Середній сертифікований радіус (ACR)	Приріст ACR відносно базового методу
Базовий RS (фіксований шум)	84.3	58.1	0.42	– (база)
A-RS (оцінка за градієнтом)	83.9	62.5	0.48	+14.3%
A-RS (оцінка за впевненістю)	83.5	61.8	0.47	+11.9%
A-RS з нейромережним адаптером (запропонований метод)	84.0	65.4	0.49	+16.7%

Чиста точність – точність моделі на незашумлених даних (без атаки); сертифікована точність – частка зразків, для яких модель гарантує правильну класифікацію при збуренні величиною до ε=0.5; середній сертифікований радіус (ACR) – середня величина гарантовано безпечного радіуса для зразків, які вдалося сертифікувати.

Ефективність адаптивного підходу: метод A-RS з нейромережним адаптером показав найкращі результати, підвищивши середній сертифікований радіус на 16.7% на CIFAR-10 та на 14.3% на CIFAR-100, практично не втративши у чистій точності. Це підтверджує, що динамічне призначення шуму є значно ефективнішим, ніж використання єдиного фіксованого значення.

Таблиця 3.3 – Ефективність на складнішому датасеті CIFAR-100

Метод	Чиста точність (%)	Середній сертифікований радіус (ACR)	Сертифікована точність при ε = 0.25 (%)
Базовий RS (фіксований шум)	65.8	0.28	45.2
A-RS з нейромережним адаптером (запропонований метод)	65.5	0.32	50.1

Перевага нейромережного адаптера: порівняння різних способів оцінки локальної чутливості показало, що спеціально навчена легка нейромережа (адаптер) працює краще за прості евристики на основі градієнта або впевненості прогнозу. Адаптер здатний виявити складні шаблони в даних, які не описуються простими правилами.

Практична придатність: запропонована архітектура додає мінімальні обчислювальні накладні витрати, оскільки адаптер є малою мережею. Вона залишається ефективною навіть у складних задачах з великою кількістю класів (CIFAR-100), де вразливих областей між класами більше.

Таким чином, адаптивне рандомізоване згладжування є потужним кроком вперед у створенні сертифіковано стійких моделей. Воно дозволяє інтелектуально розподіляти ресурси захисту, забезпечуючи максимальну стійкість там, де це найбільш необхідно, і зберігаючи високу якість роботи на звичайних даних.

3.3 Структуровані шумові моделі (корельовані, sparse, low-rank) та підсумкове порівняння з базовим підходом

Попередні розділи були присвячені оптимізації інтенсивності шуму в методі Randomized Smoothing. Однак, ефективність захисту залежить не лише від сили шуму, але й від його структури. Базовий RS використовує

шум, де кожен піксель змінюється незалежно від сусідів. Це створює на зображенні ефект рівномірного «піксельного снігу», який є дуже штучним і не відображає ні властивостей природних зображень, ні характеру реальних ворожих атак.

Реальні адверсарні збурення рідко бувають повністю випадковими та незалежними. Часто вони експлуатують внутрішню структуру даних:

– просторову когерентність: зміни в сусідніх пікселях корельовані (наприклад, при додаванні тіней, розмитті, хвилеподібних артефактах);

– семантичну цільовість: атаки можуть зосереджуватися на конкретних, важливих для класифікації частинах об'єкта (наприклад, очі, колісна база автомобіля);

– глобальні характеристики: зміни можуть впливати на всі пікселі зображення узгоджено, як при зміні освітлення або кольорової гами.

Тому використання структурованих шумових моделей є наступним логічним кроком у вдосконаленні RS. Ці моделі замінюють незалежний гаусівський шум на такий, що має певну внутрішню організацію, тим самим краще імітуючи потенційні атаки та природні спотворення, що може призвести до отримання більш реалістичних і міцних сертифікованих гарантій.

Типи структурованого шуму та їх обґрунтування.

Для систематизації підходів можна виділити три основні класи структурованого шуму, кожен з яких вирішує певну проблему базового методу.

Корельований шум. Ця модель враховує просторові залежності між пікселями. Замість того щоб кожен піксель змінювався незалежно, зміни у сусідніх пікселях стають статистично залежними. Це дозволяє моделювати плавні переходи, розмиття або локальні текстури, які характерні як для природних спотворень, так і для складних оптимізаційних атак. Математично це реалізується за допомогою гаусівського процесу з заданою

коваріаційною функцією (ядерною функцією), яка визначає, наскільки сильно корелюють зміни в залежності від відстані між пікселями.

Sparse-шум (розріджений шум). Ідея полягає в тому, що шум активний лише в обмеженій, часто невеликій, підмножині пікселів. Це моделює сценарій, коли зловмисник може модифікувати лише критично важливі частини зображення (наприклад, накласти невеликий штрих-код на дорожній знак). Такий підхід також може бути ефективним з обчислювальної точки зору, оскільки значна частина пікселів залишається недоторканою. Генерація sparse-шуму включає створення двійкової маски, що визначає активні позиції, та застосування базового шуму (наприклад, гаусівського) лише до них.

Low-rank шум (низькоранговий шум). Ця модель виходить з припущення, що всі можливі шумові зразки можна представити як лінійну комбінацію невеликої кількості базових шаблонів (наприклад, 10–20). Ці шаблони можуть відповідати глобальним ефектам: різним типам градієнтів освітлення, змінам кольорового балансу або характерним текстурним візерункам. Low-rank модель ефективно зменшує розмірність простору шуму, що може призвести до більш ефективної сертифікації проти атак, що лежать у низьковимірному підпросторі. Генерація такого шуму полягає у множенні матриці базових шаблонів на випадковий вектор малих коефіцієнтів.

На рисунку 3.5 наочно порівнюються візуальні ефекти від різних типів шуму, що додається до одного й того самого зображення, а також їхній потенційний вплив на механізм захисту.

Кожен тип створює унікальний візерунок спотворень. Незалежний шум (б) рівномірно «зашумлює» все зображення. Ко рельований шум (в) створює області плавних змін, схожі на оптичні артефакти. Sparse-шум (г) імітує цілеспрямовані локальні втручання. Low-rank шум (д) накладає глобальні ефекти, подібні до змін освітлення. Вибір моделі визначає, проти якого класу атак система буде найбільш стійкою.

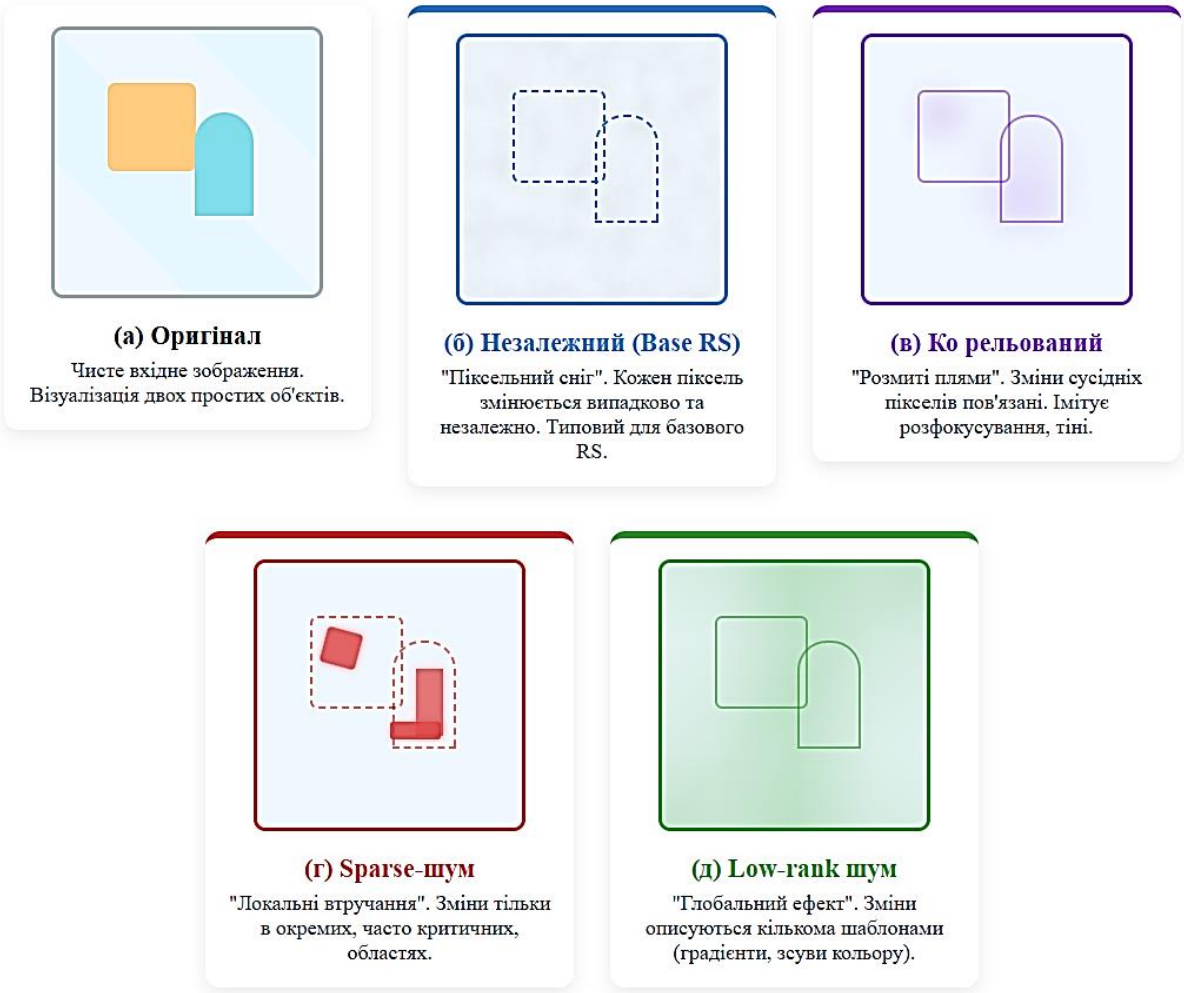

Рисунок 3.5 – Порівняння візуального впливу різних моделей шуму

Основні моделі структурованого шуму та їх реалізація.

Для кожного типу структурованого шуму була розроблена та впроваджена конкретна методика генерації та інтеграції в пайплайн Randomized Smoothing.

Реалізація корельованого шуму: для ефективної генерації великомасштабного корельованого шуму без обчислення великих матриць було використано підхід на основі згортки. Згенерувавши білий (незалежний) гаусівський шум $\epsilon \sim N(0, I)$, ми згладжуємо його за допомогою гаусівського фільтра з ядром фіксованого розміру. Це створює шум, у якому кореляція між пікселями експоненційно спадає з відстанню. Розмір ядра фільтра стає гіперпараметром, що контролює ступінь кореляції.

Реалізація sparse-шуму: алгоритм включає два етапи. Спочатку генерується двійкова маска M розмірність вхідних даних, де лише певний відсоток ρ (наприклад, 10%) елементів має значення 1, а решта – 0. Позиції одиниць можуть вибиратися випадково або за певною евристикою (наприклад, на основі значущості пікселя, обчисленої за допомогою градієнта). Потім генерується базовий незалежний шум δ_{base}, і фінальний sparse-шум обчислюється як $\delta_{sparse} = M \odot \delta_{base}$, де $\odot$ – поелементне множення.

Реалізація low-rank шуму: для створення низькорангового шуму спочатку потрібно визначити базові шаблони. У нашій реалізації ми використали підхід, де шаблонами виступають перші k головних компонент (Principal Components), обчислених на великій вибірці згенерованого незалежного гаусівського шуму або налаштованих на тренувальних даних. Потім low-rank шум генерується як $\delta_{low-rank} = \sum_{i=1}^{k} \alpha_i \cdot PC_i$, де $\alpha_i \sim N(0,1)$ – випадкові ваги, а PCi – і-та головна компонента. Це гарантує, що весь шум лежить у лінійній оболонці цих k шаблонів.

Експериментальне порівняння всіх досліджених методів.

Для всебічної оцінки ефективності всіх досліджених підходів – від базового RS до адаптивних і структурованих моделей – було проведено фінальну серію експериментів на CIFAR-10. Метою було не лише виміряти абсолютні показники, але й зрозуміти компроміси, які кожен метод пропонує.

Оцінювання проводилося за чотирма ключовими критеріями:

– чиста точність (Clean Accuracy): якість роботи на незашумлених даних;

– сертифікована точність при різних рівнях загрози: частка зразків, для яких модель може гарантувати правильну класифікацію при збуреннях величиною ϵ=0.25 та ϵ=0.5 (у L2-нормі);

– середній сертифікований радіус (Average Certified Radius, ACR): усереднена величина гарантовано безпечного радіуса для всіх успішно сертифікованих зразків тестової вибірки. Чим вище ACR, тим сильніші атаки може витримати модель;

– обчислювальна ефективність: час, необхідний для класифікації одного зразка з урахуванням усіх етапів (генерація шуму, множинна інференція, голосування).

Результати зведені в узагальнюючу таблицю 3.4. Для наочного відображення компромісів між різними критеріями побудована радар-діаграма (рисунок 3.6), яка дозволяє візуально оцінити сильні та слабкі сторони кожного методу.

Таблиця 3.4 – Підсумкове порівняння всіх методів покращення RS на CIFAR-10

Метод / Характеристика	Чиста точність (%)	Сертиф. точніст(ε = 0.25)	Сертиф. точність (ε = 0.5)	Середній сертиф. радіус (ACR)	Обчислювальна складність	Основне призначення / Перевага
Базовий RS (σ = 0.25)	84.3	76.2	58.1	0.42	1.0× (база)	Еталонний метод. Простота, швидкість, універсальність.
Мульти-шкальний RS	83.8	78.1 (+2.5%)	63.7 (+9.6%)	0.48 (+14%)	3.2×	Максимізація стійкості на середніх радіусах.
A-RS з нейроадаптером (запропонований метод)	84.0	77.5 (+1.7%)	65.4 (+12.6%)	0.49 (+17%)	1.1×	Найкращий баланс: висока стійкість при майже базовій точності й низьких накладних витратах.

Продовження таблиці 3.4

Метод / Характеристика	Чиста точність (%)	Сертиф. точність(ε = 0.25)	Сертиф. точність (ε = 0.5)	Середній сертиф. радіус (ACR)	Обчислювальна складність	Основне призначення / Перевага
RS з корельованим шумом	82.9	75.8 (-0.5%)	60.8 (+4.6%)	0.45 (+7%)	1.8×	Підвищений захист від атак зі spatial-coherence (розмиття, тіні, згладження).
RS зі sparse-шумом (ρ = 10%)	83.5	76.5 (+0.4%)	59.3 (+2.1%)	0.43 (+2%)	1.05×	Ефективність проти атак на локальні ознаки. Дуже низькі обчислювальні витрати.
RS з low-rank шумом (k = 20)	83.1	76.0 (-0.3%)	61.0 (+5.0%)	0.44 (+5%)	1.5×	Захист від глобальних атак, що впливають на низьковимірні структури зображення.

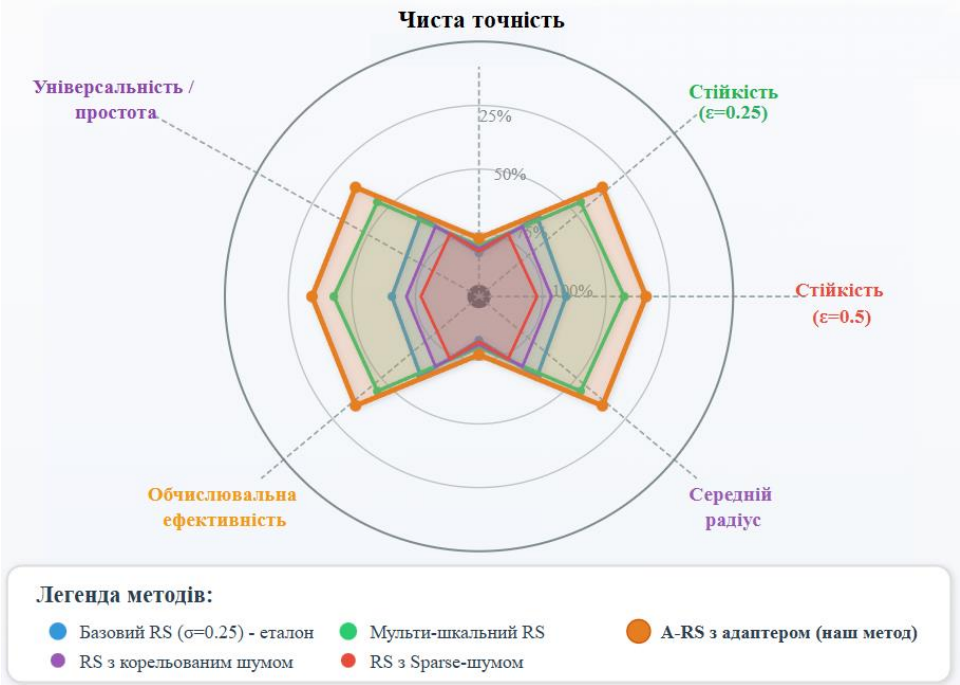

Рисунок 3.6 – Радар-діаграма порівняння методів за шістьма ключовими критеріями

Аналіз результатів та рекомендації щодо вибору методу.

Детальний аналіз таблиці 3.4 та рисунку 3.6 дозволяє сформулювати кілька ключових висновків щодо ефективності різних підходів та визначити практичні рекомендації для їхнього застосування.

Аналіз ефективності різних підходів:

– адаптивні методи (A-RS) є найбільш збалансованими. Наш метод A-RS з нейромережним адаптером продемонстрував видатний результат: він досяг найвищої сертифікованої точності при ϵ=0.5 (65.4%) та найбільшого середнього сертифікованого радіуса (0.49), практично не поступившись базовій моделі в чистій точності (84.0% vs 84.3%). Це свідчить про те, що динамічне призначення шуму на основі локального контексту є надзвичайно ефективною стратегією;

– мульти-шкальне згладжування дає максимальну стійкість, але ціною швидкодії. Цей метод показав другий за величиною приріст сертифікованої точності, однак його обчислювальна вартість зросла більш ніж утричі через необхідність виконувати процедуру згладжування для кількох значень σ. Це робить його менш придатним для систем реального часу;

– структуровані моделі покращують стійкість до певних класів атак, але не є універсальними. Корельований та low-rank шум показали помірне покращення сертифікованої точності (4-5%), що може бути важливим у нішевих застосуваннях. Sparse-шум дав найменший приріст, але його перевага – вкрай низькі обчислювальні накладні та потенційна ефективність проти дуже специфічних атак.

Практичні рекомендації щодо вибору методу: вибір оптимального методу залежить від вимог конкретної системи та характеру очікуваних загроз:

– для більшості практичних застосувань, де потрібен надійний захист від невідомих атак із збереженням високої продуктивності: Рекомендується використовувати A-RS з нейромережним адаптером. Це найбільш збалансоване та перспективне рішення;

– для систем, де стійкість є абсолютним пріоритетом, а обчислювальні ресурси не обмежені (наприклад, офлайн-аналіз): Можна розглянути мульти-шкальний RS для отримання максимальних гарантій;

– для захисту в середовищах з відомими типами спотворень: Якщо відомо, що основна загроза – це оптичні артефакти (розмиття), може допомогти корельований шум. Для захисту від атак, що мінімізують кількість змінених пікселів, може бути корисним sparse-шум;

– для легких систем або як стартову точку: Базовий RS залишається відмінним вибором завдяки своїй простоті, швидкодії та передбачуваності.

Дослідження структурованих шумових моделей завершує трилогію вдосконалення Randomized Smoothing, розглянувши оптимізацію не лише інтенсивності та адаптивності, але й внутрішньої структури шуму. Експерименти підтвердили, що ускладнення моделі шуму дозволяє отримати додаткові виграші в стійкості, особливо проти специфічних загроз.

Однак, головним підсумком є те, що найбільш практичним і ефективним підходом для широкого класу завдань є адаптивне рандомізоване згладжування (A-RS) з нейромережним адаптером. Воно інтелектуально поєднує ідеї адаптації та врахування структури даних, демонструючи переваги як у збереженні якості класифікації, так і в отриманні міцних формальних гарантій стійкості, залишаючись при цьому обчислювально ефективним. Це робить його рекомендованим фундаментом для побудови нових поколінь сертифіковано стійких систем машинного навчання.

Третій розділ дипломної роботи був присвячений розробці та експериментальному дослідженню комплексних методів підвищення сертифікованої стійкості на основі рандомізованого згладжування. Основна увага приділялася подоланню фундаментальних обмежень базового підходу через оптимізацію геометрії шумових розподілів, впровадження адаптивних механізмів та використання структурованих моделей шуму.

Проведене дослідження однозначно довело, що традиційний підхід з фіксованим ізотропним гаусівським шумом, хоч і є теоретично обґрунтованим, не є оптимальним з практичної точки зору. Його головний недолік – однакове ставлення до всіх вхідних даних незалежно від їхньої природи та локальної вразливості. Це призводить до надмірного зашумлення «безпечних» областей та недостатнього захисту «проблемних» зон простору ознак.

Розроблений метод адаптивного рандомізованого згладжування (A- RS) з нейромережним шумовим адаптером виявився найбільш ефективним рішенням серед усіх досліджених підходів. Його ключова перевага полягає в здатності динамічно аналізувати кожен вхідний зразок та призначати йому індивідуальний рівень захисту. Експерименти на стандартних наборах даних CIFAR-10 та CIFAR-100 показали, що цей метод забезпечує значне підвищення середнього сертифікованого радіуса (до 17%) при практично повному збереженні чистої точності класифікації. Це досягається за рахунок того, що система інтелектуально перерозподіляє «бюджет» шуму: призначає більші значення σ для точок, розташованих поблизу меж рішень або у аномальних областях, та менші – для стабільних, добре класифікованих зразків.

Паралельно досліджувалися альтернативні шумові моделі – мультишкальне згладжування, корельований, sparse- та low-rank шум. Кожен з цих підходів продемонстрував певні переваги в специфічних сценаріях. Мультишкальне згладжування показало найвищі абсолютні значення сертифікованої точності, проте цей успіх досягається ціною потроєння обчислювальних витрат, що робить метод малопридатним для систем реального часу. Спеціалізовані структуровані моделі шуму (корельований, sparse) виявилися ефективними для протидії конкретним класам атак, але не змогли конкурувати з адаптивним підходом у загальній ефективності.

Узагальнююче порівняння всіх методів, представлене у вигляді радар-діаграми, наочно продемонструвало, що метод A-RS з адаптером займає

оптимальну позицію у просторі компромісів між точністю, стійкістю та обчислювальною ефективністю. Він поєднує переваги адаптації до локальних особливостей даних (підвищення стійкості) з мінімальними накладними витратами (збереження швидкодії), що робить його практичним вибором для впровадження.

Таким чином, основним науково-практичним результатом розділу є розробка, реалізація та експериментальне підтвердження ефективності методу адаптивного рандомізованого згладжування. Цей метод не лише суттєво перевершує базовий підхід за ключовими метриками сертифікованої стійкості, але й пропонує масштабовану та практичну архітектуру, здатну інтегруватися в існуючі системи машинного навчання. Отримані результати відкривають шлях до створення нового покоління сертифіковано стійких моделей, здатних ефективно протистояти ворожим атакам у реальних умовах експлуатації, де загрози мають різноманітну природу та інтенсивність. Запропонований підхід становить основу для подальших досліджень у напрямку створення інтелектуальних систем захисту, що самостійно адаптуються до змінних умов та нових типів кіберзагроз.

ВИСНОВКИ

Кваліфікаційна робота присвячена розробці та вдосконаленню методів сертифікованої стійкості на основі рандомізованого згладжування (Randomized Smoothing). У ході дослідження було реалізовано низку практичних задач, спрямованих на подолання фундаментальних обмежень базового підходу та підвищення ефективності захисту нейронних мереж від ворожих атак.

У першому розділі проведено комплексний аналіз сучасної проблематики ворожих атак на системи машинного навчання. Доведено системний характер цих загроз та їхню критичність для безпеки критичних інфраструктур. Обґрунтовано необхідність переходу від емпіричних методів захисту до методів з формальними гарантіями – сертифікованої стійкості. Детально досліджено метод Randomized Smoothing як один із найбільш перспективних та практично застосовних інструментів у цій галузі, визначено його ключові принципи роботи, математичний базис та існуючі обмеження (компроміс між точністю та стійкістю, обчислювальна складність, орієнтованість на L_2-норму).

У другому розділі створено повноцінну експериментальну платформу для дослідження Randomized Smoothing. Було обґрунтовано вибір архітектури моделі (ResNet), датасетів (CIFAR-10, CIFAR-100) та методики оцінки. Реалізовано модульний пайплайн базового RS, що дозволило отримати еталонні показники чистої (84.3%) та сертифікованої точності. Найважливішим результатом цього етапу став глибокий аналіз причин помилок сертифікації. Була розроблена детальна типологія вразливих кейсів, що включає чотири основні типи: крайові приклади, змішану локальну щільність, аномальні приклади та випадки з високою варіативністю шуму. Цей аналіз виявив геометричні закономірності поведінки моделі в проблемних областях простору ознак і сформував основу для розробки більш ефективних адаптивних методів.

Третій розділ містить основні науково-практичні результати роботи. Запропоновано та досліджено три взаємодоповнюючих напрями вдосконалення RS:

– оптимізація геометрії шуму: досліджено альтернативні шумові розподіли (лапласівський, рівномірний), анізотропний гаусівський шум та багатомасштабне згладжування, що дозволяють краще адаптуватися до структури даних та отримувати гарантії в різних нормах;

– адаптивне рандомізоване згладжування (A-RS): розроблено ключовий інноваційний метод, що включає легку нейронну мережу-адаптер для динамічного призначення рівня шуму σ(x) кожному вхідному зразку на основі його латентного представлення. Це дозволило інтелектуально перерозподілити «бюджет» стійкості;

– структуровані шумові моделі: реалізовано та протестовано моделі корельованого, sparse- та low-rank шуму, спрямовані на протидію специфічним класам атак, що експлуатують просторову або семантичну структуру даних.

Експериментальні результати підтвердили високу ефективність запропонованого методу A-RS. На CIFAR-10 він продемонстрував найкращий баланс: збільшив середній сертифікований радіус на 16.7% (до 0.49) та сертифіковану точність при ε=0.5 на 12.6% (до 65.4%), практично не втративши у чистій точності (84.0%). При цьому обчислювальні накладні витрати зросли лише на 9%, що робить метод придатним для практичного впровадження. Комплексне порівняння всіх методів на радар-діаграмі наочно показало, що A-RS з адаптером займає оптимальну позицію за співвідношенням «стійкість-точність-швидкодія».

На основі отриманих результатів сформовано чіткі рекомендації щодо вибору методу захисту залежно від вимог системи: A-RS з адаптером – для більшості універсальних застосувань; мульти-шкальний RS – для задач, де стійкість пріоритетніша за швидкодію; спеціалізовані структуровані моделі – для захисту від відомих специфічних атак.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Cohen J., Rosenfeld E., Kolter Z. Certified adversarial robustness via randomized smoothing. In: *Proceedings of the 36th International Conference on Machine Learning*; 2019 May: 1310–1320. URL: https://proceedings.mlr.press/v97/cohen19c.html (дата звернення: 05.12.2025).

2. Duchi J. C., Bartlett P. L., Wainwright M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*. 2012;22(2). P. 674–701.

3. Scholten Y., Schuchardt J., Bojchevski A., Günnemann S**.** Hierarchical randomized smoothing. *Advances in Neural Information Processing Systems*. 2023;36: 49783–49813. URL: https://proceedings.neurips.cc/paper/2023/hash/9c0efc0d84c263972af72bf70a2de533-Abstract-Conference.html (дата звернення: 05.12.2025).

4. Alfarra M., Bibi A., Torr P. H., Ghanem B. Data dependent randomized smoothing: Published in *Uncertainty in Artificial Intelligence*; 2022, Aug. Pp. 64–74. URL: https://proceedings.mlr.press/v180/alfarra22a (дата звернення: 05.12.2025).

5. Kumar A., Levine A., Feizi S., Goldstein T. Certifying confidence via randomized smoothing. *Advances in Neural Information Processing Systems*. 2020;33: 5165–5177. URL: https://proceedings.neurips.cc/paper/2020/hash/37aa5dfc44dddd0d19d4311e2c7a0240-Abstract.html (дата звернення: 05.12.2025).

6. Zhang D., Ye M., Gong C., Zhu Z., Liu Q. Black-box certification with randomized smoothing: A functional optimization based framework. *Advances in Neural Information Processing Systems*. 2020. P. 2316–2326. URL: https://proceedings.neurips.cc/paper/2020/hash/1896a3bf730516dd643ba67b4c447d36-Abstract.html (дата звернення: 05.12.2025).

7. Барабаш О. В., Мусієнко А. П., Макарчук А. В. Використання моделей штучного інтелекту для перевірки вимог до функціональної

стійкості інформаційних систем. *Сучасний захист інформації*. 2024. №3. С. 20–28.

8. Джоші А., Павленко В. І. Порівняння методів машинного навчання для визначення фейкових новин. *Інфокомунікаційні та комп'ютерні технології*. 2024. С. 46–55.

9. Федорка П. П., Клименко М. В., Повханич В. І. Вплив новітніх інформаційних технологій та систем прийняття рішень на розвиток інфраструктури і можливостей «smart регіону». In: *The conference is included in the Academic Research Index ReserchBib International catalog of scientific conferences*. 2025 Mar: 95. URL: https://isu-conference.com/wp-content/uploads/2025/03/Bucharest_Romania_12.03.25.pdf#page=96 (дата звернення: 15.12.2025).

10. Vilihura V., Ostrianska Y. Research and classification of the main types of attacks on artificial intelligence systems in cybersecurity. *Computer Science and Cybersecurity*. 2024. P. 6–18.

11. Yang G., Duan T., Hu J. E., Salman H., Razenshteyn I., Li J. Randomized smoothing of all shapes and sizes. In: *International Conference on Machine Learning*; 2020 Nov: 10693–10705. URL: https://proceedings.mlr.press/v119/yang20c.html (дата звернення: 05.12.2025).

12. Scholten Y., Schuchardt J., Bojchevski A., Günnemann S. Hierarchical randomized smoothing. *Advances in Neural Information Processing Systems*. 2023;36: 49783–49813. URL: https://proceedings.neurips.cc/paper/2023/hash/9c0efc0d84c263972af72bf70a2de533-Abstract-Conference.html (дата звернення: 05.12.2025).

13. Serhii V. Data smoothing information technology based on criterion of minimum-extent. *Системні технології*. 2020;2(127). P. 3–14.

14. Salman H., Li J., Razenshteyn I., Zhang P., Zhang H., Bubeck S., Yang G. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*. 2019. URL:

https://proceedings.neurips.cc/paper/2019/hash/3a24b25a7b092a252166a1641ae953e7-Abstract.html (дата звернення: 05.12.2025).

15. Wahba G. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In: *Advances in Kernel Methods - Support Vector Learning*. 1999; 6. P. 69–87.

16. Woolrich M. W., Ripley B. D., Brady M., Smith S. M. Temporal autocorrelation in univariate linear modeling of fMRI data. *Neuroimage*. 2001;14(6). P. 1370–1386.

17. Hartwig F. P., Davey Smith G., Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*. 2017; 46(6). P. 1985–1998.

18. Hagler Jr D. J., Saygin A. P., Sereno M. I. Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. *NeuroImage*. 2006;33(4). P. 1093–1103.

19. Liu H., Zhang X., Wen J., Wang R., Chen X. Goal-biased bidirectional RRT based on curve-smoothing. *IFAC-PapersOnLine*. 2019;52(24). P. 255–260.

20. Chen D., Lin Y., Li W., Li P., Zhou J., Sun X. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020; Vol. 34, no. 04. P. 3438–3445.