УДК 004.932:004.89:004.7 DOI: 10.31673/2412-4338.2026.029106

Азарний Дмитро Ігорович
Аспірант факультету радіофізики, електроніки та комп'ютерних систем
Київський національний університет імені Тараса Шевченка, Київ
ORCID 0009-0000-9549-8873
dazarny@gmail.com

Давиденко Анатолій Миколайович
Доктор технічних наук, професор, професор кафедри комп'ютерної інженерії
Державний університет інформаційно-комунікаційних технологій
ORCID ID 0000-0001-6466-1690
davidenkoan@gmail.com

ВИЯВЛЕННЯ ДИПФЕЙКОВИХ ЗОБРАЖЕНЬ У СИСТЕМАХ ЕЛЕКТРОННИХ КОМУНІКАЦІЙ ЗА ДОПОМОГОЮ НЕЙРОННИХ МЕРЕЖ

Анотація. *У статті досліджується критично важлива проблема виявлення згенерованого візуального контенту (дипфейків) у сучасних системах електронних комунікацій. Зі стрімким розвитком генеративних змагальних мереж (GAN) та дифузійних моделей класичні просторові детектори швидко перенавчаються та втрачають свою ефективність при появі нових алгоритмів генерації, що створює серйозні загрози для інформаційної безпеки та довіри до цифрових каналів зв'язку. Метою даної роботи є розробка та комплексне дослідження модифікованих нейромережевих архітектур, у яких завдяки синергійному поєднанню просторових та частотних механізмів уваги досягається суттєве підвищення робастності та ефективності детекції згенерованих зображень. У ході дослідження було запропоновано та програмно реалізовано шість унікальних гібридних архітектур на базі глибокої згорткової мережі ResNet-50. Ефективність класифікації було значно підвищено завдяки цілеспрямованій інтеграції механізмів просторової багатомасштабної уваги (MSA), міжканальної уваги (CBAM) та частотної уваги, заснованої на швидкому перетворенні Фур'є (FFT), а також використанню трансформерних модулів (Transformer Decoder) для аналізу глобальних структурних взаємозв'язків. Крім того, розроблено фінальну ансамблеву модель, яка за допомогою методу м'якого голосування (soft voting) ефективно об'єднує прогнози спеціалізованих гібридів, мінімізуючи хибні спрацьовування окремих класифікаторів. Експериментальне тестування моделей проводилось на комплексних наборах даних (DFFD, FaceForensics++, HiDF) та незалежному балансованому зовнішньому датасеті (Deepfake vs Real 60K) для об'єктивної перевірки здатності до крос-доменного узагальнення. За результатами тестування на валідаційній вибірці розроблений ансамбль продемонстрував найвищу ефективність: загальна точність (Accuracy) склала 99.20 %, а F1-міра - 0.9910. Тестування на зовнішніх даних показало зниження точності до 74.7 %, що підтвердило гіпотезу про складність узагальнення ознак нових генераторів. За допомогою побудови карт активації ознак Grad-CAM було детально візуалізовано принципи прийняття рішень моделями та математично доведено критичну важливість частотної гілки (архітектура ResNetSECBAMFFT) для локалізації спектральних аномалій у випадках високоякісних підробок, які повністю ігноруються класичними просторовими детекторами.*

Ключові слова: *дипфейк, згорткові нейронні мережі, ResNet-50, механізми уваги, перетворення Фур'є, ансамблеве навчання, Grad-CAM, кібербезпека.*

Dmytro Azarnyi
PhD student, faculty of radiophysics, electronics and computer systems
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID 0009-0000-9549-8873
dazarny@gmail.com

Anatolii Davydenko
Doctor of Technical Sciences, Professor, Professor of the Department of Computer Engineering
State University of Information and Communication Technologies, Kyiv, Ukraine

ORCID ID 0000-0001-6466-1690
davidenkoan@gmail.com

DETECTION OF DEEPFAKE IMAGES IN ELECTRONIC COMMUNICATION SYSTEMS USING NEURAL NETWORKS

Abstract. *The article investigates the critically important problem of detecting generated visual content (deepfakes) in modern electronic communication systems. With the rapid development of Generative Adversarial Networks (GANs) and diffusion models, classical spatial detectors quickly overfit and lose their effectiveness when new generation algorithms emerge, which poses serious threats to information security and trust in digital communication channels. The aim of this work is to develop and comprehensively analyze modified neural network architectures in which, due to the synergistic combination of spatial and frequency attention mechanisms, a significant increase in the robustness and efficiency of detecting generated images is achieved. During the research, six unique hybrid architectures based on the deep convolutional network ResNet-50 were proposed and programmatically implemented. Classification efficiency was significantly improved through the targeted integration of multi-scale spatial attention (MSA), convolutional block attention module (CBAM), and frequency attention based on the Fast Fourier Transform (FFT), as well as the use of Transformer modules (Transformer Decoder) to analyze global structural relationships. Furthermore, a final ensemble model was developed, which effectively combines the predictions of specialized hybrids using the soft voting method, minimizing the false positive rates of individual classifiers. Experimental testing of the models was conducted on comprehensive datasets (DFFD, FaceForensics++, HiDF) and an independent balanced external dataset (Deepfake vs Real 60K) to objectively verify the capability for cross-domain generalization. According to the test results on the validation set, the developed ensemble demonstrated the highest efficiency: the overall accuracy reached 99.20%, and the F1-score was 0.9910. Testing on external data showed a decrease in accuracy to 74.7%, which confirmed the hypothesis about the difficulty of generalizing the features of new generators. By constructing Grad-CAM feature activation maps, the decision-making principles of the models were visualized in detail, and the critical importance of the frequency branch (ResNetSECBAMFFT architecture) for localizing spectral anomalies in cases of high-quality forgeries, which are completely ignored by classical spatial detectors, was mathematically proven.*

Keywords: *deepfake, convolutional neural networks, ResNet-50, attention mechanisms, Fourier transform, ensemble learning, Grad-CAM, cybersecurity.*

1. Вступ

Дипфейки - це згенеровані за допомогою глибоких нейронних мереж зображення або відео, що імітують справжні обличчя. Зі зростанням реалістичності таких підробок, проблема їх виявлення стає критичною для забезпечення безпеки та «чистоти» систем електронних комунікацій. Надійні методи верифікації візуального контенту є життєво необхідними для захисту цифрових каналів зв'язку від зловмисних маніпуляцій, фільтрації скомпрометованого трафіку та запобігання поширенню дезінформації в сучасних телекомунікаційних мережах. Враховуючи стрімкий розвиток генеративних технологій, традиційні підходи до перевірки медіаконтенту швидко втрачають свою ефективність. Сучасні системи інформаційної безпеки потребують впровадження нових, комплексних рішень, що виходять за межі простого аналізу пікселів. Актуальним напрямом у цій сфері є розробка методів детекції, які базуються на синергійному поєднанні різних типів ознак - зокрема, просторових та частотних. Використання оптимізованих нейромережевих архітектур із сучасними механізмами уваги дозволяє створити надійні інструменти для глибокого аналізу та виявлення прихованих артефактів генерації, підвищуючи загальний рівень довіри до цифрового середовища.

2. Постановка проблеми

Незважаючи на активний розвиток засобів інформаційної безпеки, класичні детектори базуються переважно на аналізі локальних просторових артефактів. Через це їхні алгоритми страждають від проблеми перенавчання (overfitting) та відносно легко компрометуються новими поколіннями генеративних моделей (сучасними GAN та дифузійними мережами). Це призводить до критичного падіння точності класифікації при крос-доменному переході на незалежні набори даних (алгоритми синтезу, ознаки яких були відсутні у навчальній вибірці). Фундаментальна проблема полягає в тому, що більшість існуючих рішень ігнорують приховані високочастотні спектральні аномалії,

фокусуючись виключно на візуальних піксельних дефектах, або ж використовують надмірно громіздкі архітектури, що ускладнює їх практичне застосування. Таким чином, гострою та досі невирішеною науково-практичною проблемою залишається розробка оптимізованих та робастних методів детекції з високою узагальнювальною здатністю. Вирішення цієї проблеми вимагає концептуально нового підходу, заснованого на розробці ансамблевих архітектур із синергійною інтеграцією механізмів просторової, міжканальної та частотної уваги, які здатні комплексно локалізувати стійкі до еволюції генераторів ознаки підробленого контенту.

3. Аналіз останніх досліджень і публікацій

Одним із найбільш ефективних математичних апаратів для вирішення задачі детекції маніпуляцій із візуальними даними є нейронні мережі. В низці наукових праць вже досліджувались шляхи вирішення даної задачі за допомогою нейронних мереж. Наприклад, у моделі SpecXNet [1] поєднано просторову та спектральну гілки для захоплення локальних текстурних аномалій та глобальних спектральних невідповідностей. У дослідженні [2], де запропоновано модель FMSI, застосовано двопоточну архітектуру з маскуванням зображень у частотній області, що дозволяє виділити загальні патерни різних технологій генерації. Архітектура EDFM [3] підсилює високочастотні шуми за допомогою глибинних розділених згорток та багатомасштабного злиття, тоді як Dual Frequency Branch Framework [4] використовує дискретне вейвлет-перетворення та фазу швидкого перетворення Фур'є для моделювання локальних взаємозв’язків. Крім того, досліджувались конволюційно-атенційні ансамблі з перетворенням вейвлетів (CAE-Net) [5], методи багатомасштабної декомпозиції для уникнення перенавчання [6] та використання мульти-уваги для ізоляції змін незалежно від особи (ID-insensitive detector) [7]. В кожному з наведених запропонованих рішень є не тільки переваги, але й суттєві недоліки. Зокрема, складні гібридні ансамблі (наприклад, поєднання кількох важких мереж) вимагають значних обчислювальних ресурсів і пам'яті, що ускладнює їх інтеграцію в системи електронних комунікацій реального часу. З іншого боку, більш легкі просторові детектори страждають від проблеми перенавчання (overfitting), демонструючи різке зниження точності при крос-доменному тестуванні на раніше небачених методах генерації підробок. Тому актуальною є задача розробки таких методів детекції, які здатні усунути ці недоліки завдяки використанню оптимізованих, обчислювально легких модулів уваги та ефективного поєднання частотних і просторових ознак.

4. Мета і задачі дослідження

З огляду на вищезазначене, метою даної роботи є: розробити модифіковані нейромережеві архітектури та їхній ансамбль, в котрих завдяки комплексній інтеграції механізмів просторової, міжканальної та частотної уваги досягається підвищення ефективності детекції дипфейкових зображень у системах електронних комунікацій.

Для досягнення поставленої мети необхідно розв’язати наступні задачі:

1. Провести аналіз сучасних методів виявлення згенерованого контенту та визначити їх ключові обмеження щодо крос-доменного узагальнення.
2. Запропонувати та програмно реалізувати низку гібридних архітектур на базі згорткової мережі ResNet-50 шляхом інтеграції модулів просторової, міжканальної та частотної уваги для підвищення точності локалізації ознак підробки.
3. Розробити ансамблеву модель на основі навчених спеціалізованих гібридів для синергійного поєднання просторових та частотних ознак з метою мінімізації хибних класифікацій та підвищення робастності системи.
4. Провести експериментальне тестування розроблених моделей на комплексних наборах даних та виконати порівняльний аналіз їхньої ефективності застосування для вирішення задачі виявлення згенерованого контенту (зокрема за допомогою побудови карт активації ознак Grad-CAM).

5. Результати дослідження

Базова архітектура: ResNet-50 та SE-блоки

На основі проведеного аналізу, для вирішення задачі ефективного виявлення дипфейкових зображень в якості базової моделі (backbone) було обрано глибоку згорткову нейронну мережу ResNet-50. Архітектура даної мережі представлена в таблиці 1. Вибір саме цієї архітектури обґрунтовується її оптимальним балансом між високою точністю вилучення ознак та помірною обчислювальною складністю. Завдяки механізму залишкових зв'язків (residual connections), ця мережа успішно долає проблему згасаючого градієнта при навчанні, що дозволяє формувати високоякісні багаторівневі карти ознак (feature maps), які є ідеальним фундаментом для подальшої інтеграції спеціалізованих модулів уваги.

Таблиця 1

Архітектура базової згорткової нейронної мережі ResNet-50, адаптованої для задачі детекції

Назва макро-блоку	Розмір вихідного тензора	Параметри шарів (ResNet-50)
Conv1	112 × 112	Згортка 7×7, 64 фільтри, крок 2. Max Pooling 3×3, крок 2
Conv2_x	56 × 56	$[Згортка\ 1 \times 1, 64\ Згортка\ 3 \times 3, 64\ Згортка\ 1 \times 1, 256] \times 3$
Conv3_x	28 × 28	$[Згортка\ 1 \times 1, 128\ Згортка\ 3 \times 3, 128\ Згортка\ 1 \times 1, 1024] \times 4$
Conv4_x	14 × 14	$[Згортка\ 1 \times 1, 256\ Згортка\ 3 \times 3, 256\ Згортка\ 1 \times 1, 1024] \times 6$
Conv5_x	7 × 7	$[Згортка\ 1 \times 1, 512\ Згортка\ 3 \times 3, 512\ Згортка\ 1 \times 1, 2048] \times 3$
Вихідний класифікаційний блок	1 × 1	Global Average Pooling Fully Connected (1 нейрон) Активація Sigmoid

Характеристики вхідних даних. Для вирішення задачі розпізнавання дипфейків на вхідний шар базової мережі ResNet-50 та запропонованих гібридних нейромережевих архітектур, котрі будуть розглянуті далі, подаються два типи просторово-частотних характеристик зображень (залежно від конфігурації моделі): 1.Просторові характеристики (Колірний RGB-тензор): Для всіх досліджуваних у даній роботі моделей базовим входом є нормалізований триканальний матричний тензор (Red, Green, Blue) розмірністю 224×224 пікселі. На цьому рівні аналізуються інтенсивності пікселів, що дозволяє першим згортковим шарам мережі виділяти такі низькорівневі ознаки, як колірні невідповідності, мікротекстура шкіри, різкі градієнти освітлення та характерні візуальні артефакти на межах блендінгу (злиття згенерованого обличчя з реальним фоном). 2.Частотні характеристики (Амплітудний спектр): Для двопотокових та гібридних архітектур (DualFreqNet, ResNetSEFFT, ResNetSECBAMFFT на вхід додаткової гілки аналізу подається двовимірний масив амплітуд, обчислений за допомогою швидкого перетворення Фур'є (FFT), застосованого до зображення, попередньо переведеного у градації сірого. Ця математична характеристика позбавлена просторової інформації, натомість вона репрезентує частотний розподіл сигналу. Її аналіз дозволяє нейромережі фіксувати невидимі для людського ока високочастотні шуми, періодичні патерни («шахові» артефакти) та спектральні аномалії, які неминуче залишають сучасні генеративні алгоритми (GAN та дифузійні моделі) під час синтезу зображення.

ResNet-50 та залишкові зв'язки. Як вказано в базовій науковій праці Kaiming He та співавторів [8], у якій було вперше представлено архітектуру глибокого залишкового навчання, залишкові функції дозволяють мережі вивчати відхилення від ідентичності та суттєво полегшують тренування навіть дуже глибоких мереж. Структурно архітектура ResNet-50 є глибокою згортковою нейронною мережею, що складається з 50 шарів. Її структуру можна детально описати за допомогою наступних послідовних макро-блоків

Вхідний згортковий блок (Conv1): Призначення: Початкове виділення низькорівневих ознак (контурів, градієнтів, базових текстур зображення обличчя) та первинне зменшення просторової розмірності тензора. Структура та вагові коефіцієнти: Шар містить одну згортку з великим розміром ядра 7 ×7 і кроком (stride) 2. Кількість вихідних фільтрів - 64. Цей шар містить порівняно невелику кількість параметрів (близько 9,4 тис. вагових коефіцієнтів), але виконує найважливішу роботу з первинної обробки пікселів. Активація та нормалізація: Для стабілізації навчання після згортки обов'язково застосовується пакетна нормалізація (Batch Normalization), після чого сигнал проходить через нелінійну функцію активації ReLU. Завершується етап операцією максимального пулінгу (Max Pooling) з вікном 3 ×3 і кроком 2.

Залишковий блок 1 (Conv2_x): Призначення: Формування базових текстурних карт ознак. Структура та вагові коефіцієнти: Складається з 3 послідовних «вузьких» (bottleneck) блоків (рис. 1). Кожен такий блок містить три згорткових шари: 1 ×1 (64 фільтри - стиснення), 3 ×3 (64 фільтри - обробка) та 1 ×1 (256 фільтрів - відновлення розмірності). Загальна кількість параметрів у цьому макро-блоці становить близько 0,2 млн. Активація: Усі згорткові операції всередині блоку супроводжуються Batch Normalization та функцією ReLU.

Залишковий блок 2 (Conv3_x):Призначення: Виділення складніших середньорівневих ознак (наприклад, елементів обличчя, очей, меж шкіри, локальних артефактів блендінгу).
Структура та вагові коефіцієнти: Складається з 4 bottleneck-блоків. Кількість фільтрів зростає вдвічі: 1 ×1 (128), 3 ×3 (128) та 1 ×1 (512). Перший блок використовує крок 2 для зменшення просторової розмірності (downsampling). Блок містить близько 1,2 млн параметрів. Активація: Batch Normalization та ReLU після кожного згорткового шару.

Залишковий блок 3 (Conv4_x): Призначення: Формування високорівневих семантичних ознак. На цьому етапі мережа "розуміє" глобальну структуру обличчя. Структура та вагові коефіцієнти: Це найглибший макро-блок мережі, що складається з 6 bottleneck-блоків. Параметри фільтрів знову подвоюються: 1 ×1 (256), 3 ×3 (256) та 1 ×1 (1024). Містить понад 7 млн вагових коефіцієнтів. Перший блок також має крок 2.
Активація: Аналогічно до попередніх етапів, застосовується пакетна нормалізація (Batch Normalization) та функція ReLU.

Залишковий блок 4 (Conv5_x): Призначення: Фінальна агрегація найбільш абстрактних ознак перед етапом класифікації. Структура та вагові коефіцієнти: Складається з 3 bottleneck-блоків із фільтрами: 1 ×1 (512), 3 ×3 (512) та 1 ×1 (2048). Це найважчий за кількістю параметрів етап виділення ознак (близько 15 млн вагових коефіцієнтів). Активація: Batch Normalization та ReLU після кожного згорткового шару всередині блоку.

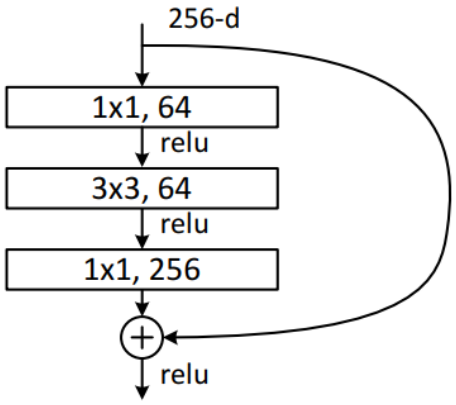

Рис. 1. Архітектура bottleneck-блоку

Вихідний класифікаційний блок (Fully Connected): Призначення: Перетворення багатовимірного тензора ознак у кінцеву ймовірність приналежності до класу. Структура та активація: Спочатку застосовується глобальний середній пулінг (Global Average Pooling), який стискає просторовий тензор 7 ×7 у плоский вектор-дескриптор розмірності 2048. Оскільки вирішується задача бінарної класифікації (реальне зображення чи дипфейк), повнозв'язний шар адаптовано: він має лише один вихідний нейрон (замість стандартних 1000) з активаційною функцією Sigmoid, що видає ймовірність підробки від 0 до 1. Кількість параметрів - 2049 (2048 ваг + 1 bias).

Ключовою особливістю кожного описаного залишкового блоку є наявність обхідних з'єднань (shortcut connections). Математично функціонування базового залишкового блоку описується формулою (1):

$$Y = F(X) + X \qquad (1)$$

де X та Y - вхідний та вихідний вектори відповідних шарів. Функція F(X) представляє залишкове відображення (residual mapping), яке мережа повинна вивчити. Операція додавання виконується за допомогою залишкового зв'язку (skip connection) та поелементного додавання тензорів, після чого застосовується нелінійна функція активації ReLU. Необхідно зазначити, що незважаючи на високу ефективність базової архітектури ResNet-50, для задачі виявлення дипфейків важливо фокусувати увагу мережі на конкретних каналах ознак, де артефакти генерації є найбільш помітними. Для вирішення цієї задачі, в рамках даної роботи, базову архітектуру було розширено за допомогою інтеграції спеціального модуля - Squeeze-and-Excitation (SE) блоку [9].

Squeeze-and-Excitation (SE) блок. SE-блоки виконують адаптивну перенормалізацію каналів: просторову інформацію стискають до вектора ознак через глобальне згортання, пропускають через невелику двошарову мережу та використовують для масштабування вихідних каналів. Такий механізм забезпечує значне підвищення точності з мінімальними обчислювальними витратами [9]. Операція стиснення (Squeeze) агрегує просторову інформацію (розмірності H x W) для кожного каналу c за допомогою глобального середнього пулінгу (Global Average Pooling), формуючи дескриптор каналу z_c за формулою (2):

$$z_c = F_{sq}(X_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \qquad (2)$$

Для захоплення залежностей між каналами застосовується операція збудження (Excitation) - механізм шлюзування з двома повнозв'язними шарами (зменшення та збільшення розмірності), де δ позначає функцію активації ReLU, а σ - сигмоїду (3):

$$s = F_{ex}(z, W) = \sigma\big(g(z, W)\big) = \sigma\big(W_2 \delta(W_1 z)\big) \qquad (3)$$

Нарешті, вихідний блок виконує адаптивне перемасштабування, перемножуючи початкові карти ознак X_c на отримані ваги каналів s_c , що описується виразом (4):

$$X_c = F_{scale}(X_c, s_c) = s_c \cdot X_c \qquad (4)$$

На основі аналізу функціональних можливостей механізмів уваги, архітектуру базової мережі ResNet-50 було посилено інтегрованими SE-блоками. Таке поєднання дозволяє моделі не лише ефективно вилучати просторову інформацію, але й динамічно фокусуватися на найважливіших каналах виділення артефактів підробки [8], [9].

Модифіковані архітектури детекції. Для усунення раніше виявлених недоліків базових конволюційних мереж, зокрема проблеми слабкого узагальнення та різкого падіння точності при крос-доменному тестуванні, в рамках даної роботи було запропоновано та програмно реалізовано низку

унікальних гібридних архітектур шляхом авторської інтеграції та послідовного комбінування існуючих модулів уваги. Головна ідея полягає в розширенні базової моделі ResNet-50 додатковими спеціалізованими модулями (просторової, міжканальної та частотної уваги), що дозволяє мережі комплексніше аналізувати специфічні артефакти генерації облич. Нижче наведено детальний опис розроблених конфігурацій.

ResNetSEMSA. Архітектура запропонованої гібридної моделі представлена на рис. 2. Модель використовує базову ResNet-50 з SE-блоками та модуль багатомасштабної уваги (MSA). MSA виконує паралельні згортки 1×1, 3×3 та 5×5, об'єднуючи їх через сигмоїдну функцію. Така обробка дозволяє виявляти дипфейкові артефакти різних масштабів: маленькі локальні дефекти та великі глобальні аномалії. Подібні багатомасштабні механізми успішно застосовуються для детекції тонких структур та ефективного уникнення перенавчання безпосередньо в сучасних задачах розпізнавання дипфейків [6].

ResNetSEMSA накладає MSA поверх базової мережі, що дозволяє враховувати дрібні та великі артефакти одночасно. У нашому випадку MSA покращує здатність мережі до узагальнення без значного збільшення числа параметрів. Математично застосування модуля багатомасштабної уваги до вхідного тензора X записується формулою (5), де M_{MSA}- згенерована карта уваги, а $\odot$ позначає поелементне множення.

$$X_{msa} = M_{MSA} \odot X \tag{5}$$

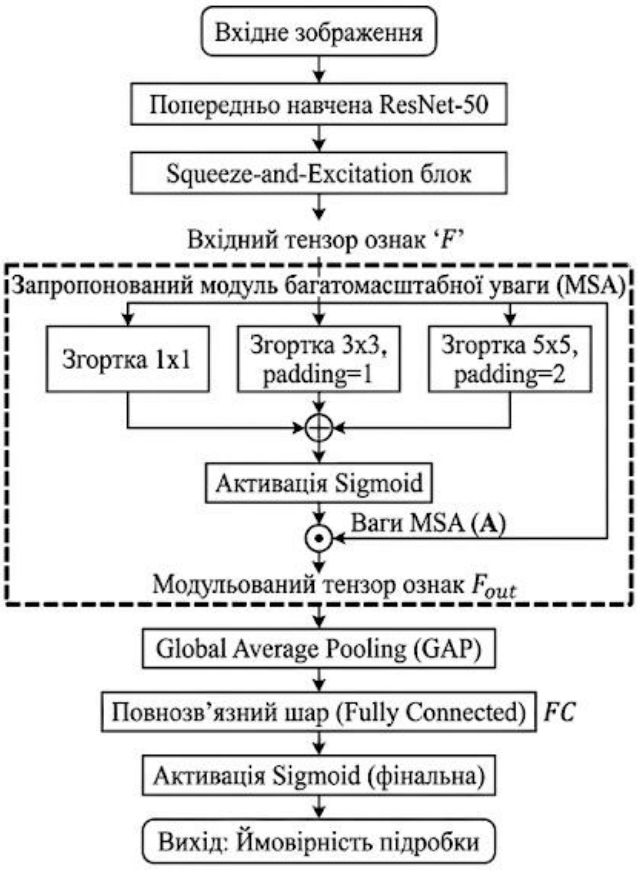

Рис. 2. Архітектура запропонованої гібридної моделі ResNetSEMSA

ResNetSETransformer. У цій архітектурі (рис. 3) після базової мережі додається трансформерний декодер. Transformer використовує багатоголову увагу (Multi-Head Attention): запити Q, ключі K та значення V формуються з ознак ResNet-50, після чого застосовується механізм

масштабованого скалярного добутку для оцінки залежностей між різними регіонами обличчя. Обчислення уваги (Scaled Dot-Product Attention) відбувається за класичною формулою (6):

$$z(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (6)$$

Застосування такого математичного механізму обчислення уваги дозволяє мережі ефективно знаходити глобальні взаємозв'язки у зображенні. Механізми багатоголової уваги та трансформерні архітектури вже продемонстрували високі результати у багатьох задачах комп'ютерного зору, зокрема при виявленні маніпуляцій з обличчями, де вони забезпечують високу точність та ізоляцію ознак на різних наборах даних [7]. Однак, через велику кількість параметрів, трансформерні модулі вимагають більше пам'яті та обчислювального часу.

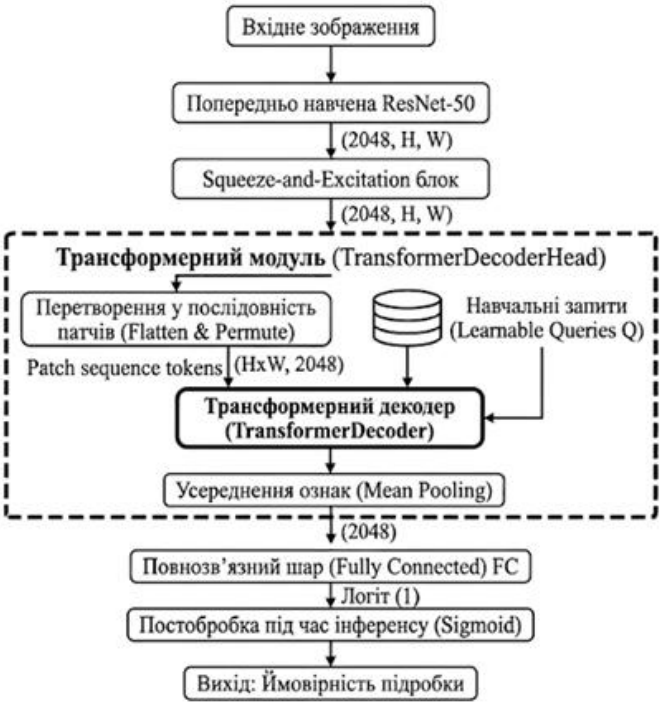

Рис. 3. Архітектура запропонованої гібридної моделі ResNetSETransformer

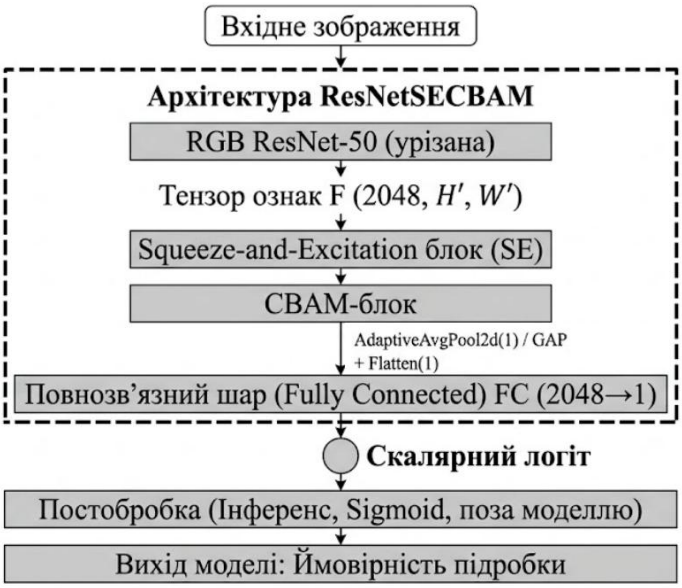

Рис. 4. Архітектура запропонованої гібридної моделі ResNetSECBAM

DualFreqNet. Ця модель (рис.5) складається з двох паралельних гілок: одна обробляє оригінальне зображення у просторовій області, а інша - його спектр амплітуд, отриманий за допомогою швидкого перетворення Фур'є (FFT). На частотну гілку застосовується увага до амплітуди, що дозволяє підсилити найінформативніші частоти.

Перехід у частотну область та застосування механізму уваги описується наступними кроками. Спочатку виконується перетворення Фур'є для отримання спектра згідно з виразом (10). Далі генерується карта частотної уваги M_{freq} на основі амплітуди спектра за формулою (11). Після чого отримані ваги застосовуються до вихідних ознак згідно з формулою (12).

$$X_{freq} = FFT(X) \quad (10)$$

$$M_{freq} = \sigma\left(W_f \cdot \left|X_{freq}\right|\right) \quad (11)$$

$$X_{fa} = M_{freq} \odot X \quad (12)$$

Подібна двопоточна комбінація була використана у роботах FMSI [2] та SpecXNet [1], де поєднання просторових і частотних ознак дозволило досягти state-of-the-art результатів і кращого переносу між генераторами дипфейків. Багатошарова увага дозволяє моделі ураховувати залежності на великих відстанях; схожі механізми мульти-уваги застосовує ID-insensitive multi-attention detector, що показує високу точність на різних наборах дипфейків [7].

ResNetSEFFT. Архітектура даної моделі представлена на рис. 6. У ній SE-блоки доповнюються блоком частотної уваги, який працює з амплітудним представленням FFT. Спочатку обчислюється спектр зображення, після чого його високочастотні компоненти пропускають через глибоку нейронну мережу та комбінують із просторовими ознаками за допомогою механізму уваги. Формально генерацію карти частотної уваги (Mfreq) на основі амплітудного спектра та її застосування до просторових ознак після SE-блоку (Xse) можна описати наступним чином. Спочатку обчислюється спектр і нормалізується його амплітуда за формулою (13). Після цього отримана мапа уваги поелементно множиться на вихідні просторові ознаки (14).

$$M_{freq} = \sigma\left(W_{conv} \cdot Norm(|FFT(X_{se})|)\right) \quad (13)$$

$$X_{out} = X_{se} \odot M_{freq} \quad (14)$$

Такі підходи були продемонстровані у FMSI та CAE-Net, де частотні ознаки допомагають відокремити глобальні структурні невідповідності від локальних текстурних артефактів [2], [5].

ResNetSECBAMFFT. Це найкомплексніша архітектура (рис.7), що одночасно містить SE-блоки, CBAM та частотну увагу. Комбінація міжканальної, просторової та частотної уваги дозволяє мережі захоплювати як дрібні, так і глобальні артефакти.

Математично повний цикл перетворення вхідного тензора у цій гібридній архітектурі записується як послідовне накладання всіх згенерованих мап уваги у вигляді формули (15):

$$X_{final} = M_{freq} \odot \left(M_s \odot (M_c \odot X_{se})\right) \qquad (15)$$

Такий багатоступеневий підхід дозволяє кожному наступному модулю уточнювати ознаки, вже відфільтровані попереднім механізмом. Ідея об'єднання кількох модулів уваги була реалізована у SpecXNet та EDFM, де багатомодальна фільтрація значно підвищувала робастність і узагальнення [1], [3].

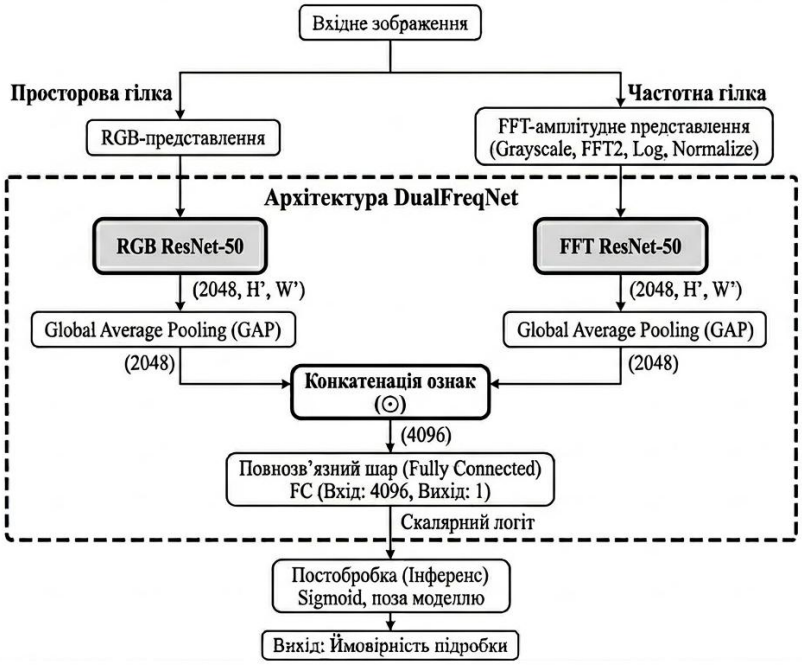

Рис. 5. Архітектура запропонованої двопотокової моделі DualFreqNet

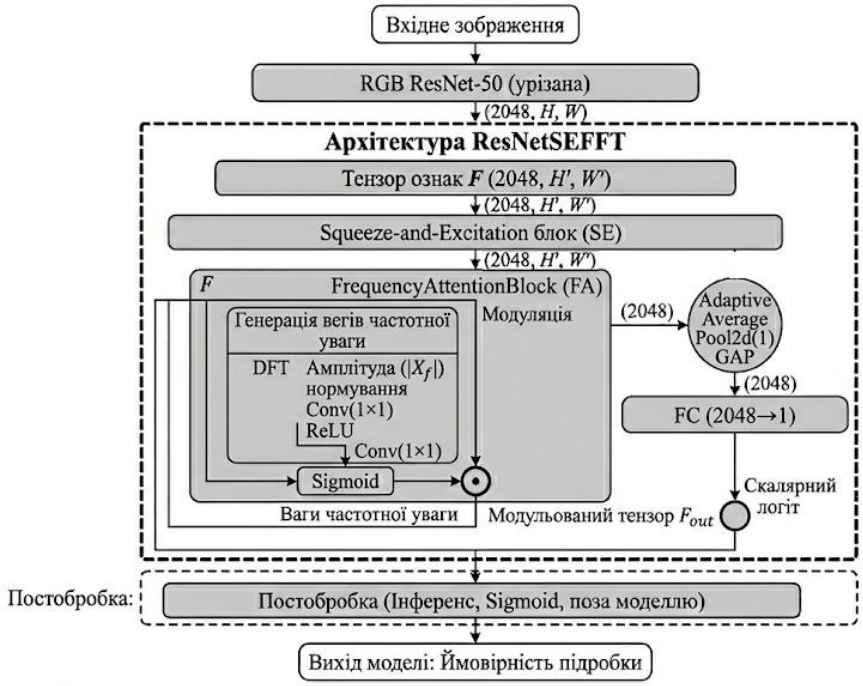

Рис. 6. Архітектура запропонованої гібридної моделі ResNetSEFFT

Методика експериментів. Для дослідження ефективності застосування розроблених модифікованих архітектур для вирішення задачі виявлення дипфейкових зображень було проведено низку експериментів, у яких визначалися ключові метрики класифікації. Для об'єктивної оцінки продуктивності моделей використовувалися наступні параметри, розраховані на основі матриці помилок (де TP - істинно позитивні, TN - істинно негативні, FP - хибно позитивні, FN - хибно негативні результати).

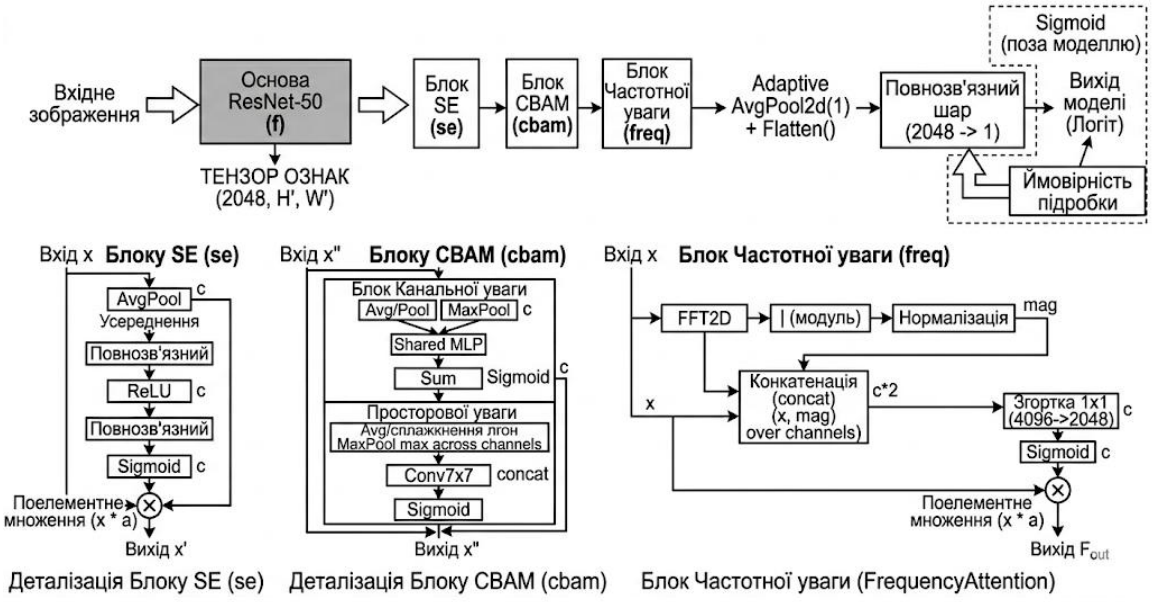

Рис. 7. Архітектура найкомплекснішої гібридної моделі ResNetSECBAMFFT

1. Точність (Accuracy): загальна частка правильно класифікованих зображень, $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$).

2. Влучність (Precision): частка дійсних дипфейків серед усіх зображень, які модель класифікувала як підробки, $Precision = \frac{TP}{TP+FP}$. Цей параметр дозволяє оцінити рівень хибних спрацьовувань.
3. Повнота (Recall): здатність моделі знаходити всі наявні дипфейки у вибірці, $Recall = \frac{TP}{TP+FN}$. Для систем безпеки це критичний параметр, оскільки пропуск підробки часто є гіршим сценарієм, ніж хибна тривога.
4. F1-міра (F1-Score): гармонійне середнє між влучністю та повнотою $F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$. Вона слугує головною комплексною метрикою для порівняння моделей в умовах можливого дисбалансу класів.

Площа під ROC-кривою (AUC): агрегований показник, що відображає здатність моделі коректно ранжувати справжні та підроблені зображення незалежно від обраного порогу класифікації. Математично показник AUC обчислюється як інтеграл функції частки істинно позитивних результатів (TPR) за часткою хибно позитивних результатів (FPR):

$$AUC = \int_0^1 TPR(FPR)dFPR \qquad (16)$$

де $TPR = \frac{TP}{TP+FN}$, а $FPR = \frac{FP}{FP+FN}$. Даний показник дозволяє оцінити якість класифікації в цілому, де значення 1.0 відповідає ідеальній моделі, а 0.5 - випадковому вгадуванню. Оцінка за вказаними метриками проводилась для наступних розроблених моделей: ResNetSEMSA, ResNetSETransformer, ResNetSECBAM, DualFreqNet, ResNetSEFFT, ResNetSECBAMFFT, а також для їхнього ансамблю.

Загальна методика проведення експериментальних досліджень складалася з наступних послідовних етапів:

1. Формування цільових наборів даних та їх розділення на валідаційну і тестову вибірки з урахуванням унікальності осіб.
2. Попередня обробка зображень (кадрування, нормалізація, частотне перетворення) та застосування методів аугментації.
3. Двоетапне тренування запропонованих нейромережевих архітектур (базове навчання на комбінованому датасеті та fine-tuning).
4. Формування ансамблю моделей для підвищення загальної точності та робастності детекції.
5. Тестування моделей на зовнішньому наборі даних із подальшим розрахунком та порівнянням метрик ефективності.
6. Побудова та аналіз карт активації ознак (Grad-CAM) для візуальної інтерпретації принципів прийняття рішень як окремими моделями, так і їхнім ансамблем, а також поглиблене дослідження причин помилкових класифікацій (False Negatives).

Набори даних. Для проведення експериментів усі використані набори зображень було логічно розділено на дві цільові групи: основний комбінований набір (для навчання та первинного тестування) та незалежний зовнішній набір (виключно для перевірки крос-доменного узагальнення). До основного комбінованого набору увійшли бази DFFD (≈299 тис. зображень, 58 703 реальні та 240 336 підроблені, 4 типи маніпуляції) [11], FaceForensics++ (понад 1.8 млн маніпульованих зображень, 4 методи маніпуляції) [12] та HiDF (з якого для навчання було використано лише статичний набір із 62 тис. зображень високої реалістичності) [13]. Результати тестування моделей, наведені у таблиці 2, отримані саме на валідаційній частині цього об'єднаного набору. В якості незалежного зовнішнього набору, який не використовувався на етапі навчання, було

обрано датасет Deepfake vs Real 60K (балансований набір з 30 тис. реальних та 30 тис. підроблених зображень) [14].

Підготовка даних. Для забезпечення коректного навчання моделей та уникнення перенавчання, була проведена попередня обробка зображень, яка складалась з наступних кроків:

Виявлення та кадрування облич: Зображення облич вирізалися з оригінальних кадрів за допомогою каскадного детектора MTCNN.

Масштабування та нормалізація: Отримані кадри приводилися до єдиного просторового стандарту шляхом масштабування до розміру 224×224 пікселів, після чого нормалізувались їхні значення.

Частотне перетворення: Для моделей, що використовують частотну увагу (DualFreqNet, ResNetSEFFT, ResNetSECBAMFFT), додатково розраховувалося швидке перетворення Фур'є (FFT) для отримання амплітудного спектра.

Аугментація даних: З метою розширення навчальної вибірки застосовувались методи просторової та кольорової аугментації, такі як горизонтальне віддзеркалення, випадкові повороти та зміна яскравості.

Розділення за ідентифікаторами осіб (Subject-aware splitting): Набори розділялись на тренувальну (train) та валідаційну (validation) вибірки суворо за унікальними особами людей на фото. Це є критично важливим кроком для запобігання «витоку даних» (data leakage), коли зображення однієї й тієї ж людини опиняються в обох вибірках. Таке суворе розділення гарантує, що нейромережа не просто запам'ятовує знайомі обличчя, а дійсно вчиться виявляти загальні артефакти генерації на абсолютно нових людях, забезпечуючи об'єктивність тестування.

Стратегія тренування. Навчання моделей на основному комбінованому наборі здійснювалось у два етапи: спочатку проводилось базове тренування на об'єднаних даних DFFD та FaceForensics++, після чого виконувалось донавчання (fine-tuning) на датасеті HiDF для підвищення чутливості мереж до високореалістичних маніпуляцій. Використовувався оптимізатор Adam зі швидкістю 10^{-4}, розміром пакету (batch size) 32, крос-ентропійною функцією втрат (cross-entropy loss) та механізмом ранньої зупинки (early stopping). Для трансформерної моделі застосовувалась маска для підключення патчів.

Ансамбль. Рішення щодо застосування ансамблевого підходу обґрунтовується тим, що кожна з запропонованих гібридних архітектур є спеціалізованою та фокусується на виділенні конкретного типу ознак: модуль MSA захоплює різномасштабні просторові артефакти, Transformer аналізує глобальні структурні взаємозв'язки, а FFT-моделі ефективно виявляють аномалії генерації в частотній області. Оскільки різні алгоритми створення дипфейків залишають принципово різні сліди, жодна окрема мережа не здатна забезпечити абсолютну робастність до всіх видів маніпуляцій. Тому шість навчених моделей було об’єднано в єдиний ансамбль шляхом усереднення їхніх прогнозних ймовірностей (soft voting). В розробленому програмному модулі здійснюється паралельне завантаження збережених ваг кожної мережі, обчислення ймовірностей та визначення фінального класу за встановленим порогом. Така стратегія дозволяє ефективно синергувати сильні сторони різних архітектур, мінімізувати дисперсію помилок індивідуальних класифікаторів та суттєво підвищити загальну надійність детекції.

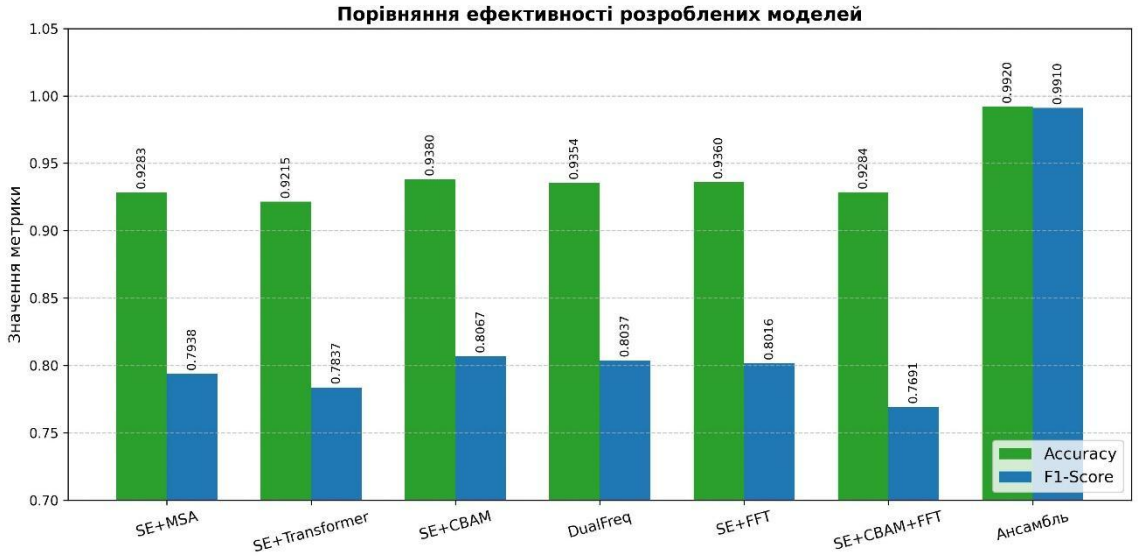

Рис. 8. Порівняльна гістограма метрик Accuracy та F1-міри для досліджуваних моделей

Результати проведених експериментів щодо оцінки ефективності досліджуваних нейромережевих архітектур наведено на Рисунку 8. Для наочного порівняння моделей на гістограму винесено дві ключові інтегральні метрики: загальну точність (Accuracy) та F1-міру, які є найбільш показовими для оцінки загальної продуктивності та збалансованості класифікаторів. Розгорнутий покомпонентний аналіз за всіма п'ятьма розрахованими параметрами зведено у таблицю 2. Як видно з графіка, індивідуальні гібридні моделі демонструють високі базові показники, проте саме їх ансамблювання дозволяє досягти максимальної продуктивності.

Таблиця 2

Результати порівняльного тестування моделей детекції дипфейків

Модель	Accuracy	Precision	Recall	F1	AUC
ResNetSECBAM	0.9380	0.8095	0.8039	0.8067	0.9707
ResNetSEMSA	0.9283	0.7387	0.8577	0.7938	0.9705
ResNetSEFFT	0.9360	0.7991	0.8041	0.8016	0.9713
ResNetSETransformer	0.9215	0.7039	0.8838	0.7837	0.9698
DualFreqNet	0.9354	0.7862	0.8219	0.8037	0.9736
ResNetSECBAMFFT	0.9284	0.7996	0.7407	0.7691	0.9635
Ансамбль	0.9920	0.9871	0.9953	0.9910	0.9960

Детальні кількісні значення всіх виміряних метрик (Accuracy, Precision, Recall, F1-міра та AUC) для кожної архітектури зведено у Таблицю 2.

Як видно з таблиці 2, ансамбль демонструє найкращі результати на внутрішній валідаційній вибірці (точність 99.20 % та F1-міра 0.9910). Водночас, для об'єктивної оцінки крос-доменного узагальнення, було проведено додаткове тестування ансамблевої моделі на незалежному зовнішньому наборі Deepfake vs Real 60K. За результатами цього експерименту ключові інтегральні показники склали: загальна точність (Accuracy) - 74.7 % та F1-міра - 79.2 %. Таке зниження загальної точності (приблизно до 75 %) є очікуваним і підтверджує наявність проблеми обмеженої узагальнювальної здатності нейромереж при обробці контенту від абсолютно нових типів генераторів, ознаки яких були відсутні у навчальних даних. Найсильніші окремі моделі - ResNetSECBAM та ResNetSEFFT. ResNetSETransformer забезпечує високу точність на складних прикладах завдяки здатності моделювати глобальний контекст, але має значну кількість параметрів та вимагає більше обчислювальних ресурсів. ResNetSEMSA покращує роботу на дрібних артефактах, але його ефективність залежить від якості багатомасштабної інформації. Для поглибленого розуміння принципів прийняття рішень

розробленими гібридними моделями та для наочної візуалізації локалізованих ними артефактів генерації, було побудовано карти активації ознак (Grad-CAM) для усіх компонентів ансамблю. На Рисунку 9 наведено приклад успішної детекції (True Positive), де ансамбль впевнено класифікував зображення як підробку з імовірністю 67.2 %.

Отримані теплові карти наочно демонструють синергію різних архітектур. Наприклад, модель із просторовою та канальною увагою (ResNetSECBAM) сфокусувалася на області шиї та ключиць, виявляючи характерні артефакти блендінгу (злиття згенерованого обличчя з реальним тілом). Водночас інші моделі (ResNetSEMSA, ResNetSETransformer) виявили просторові аномалії в освітленні та текстурі шкіри на правій частині обличчя. Завдяки такому комплексному аналізу, ансамбль зміг компенсувати невпевненість окремої двопотокової моделі (DualFreqNet, 48.1 %) та прийняти правильне фінальне рішення.

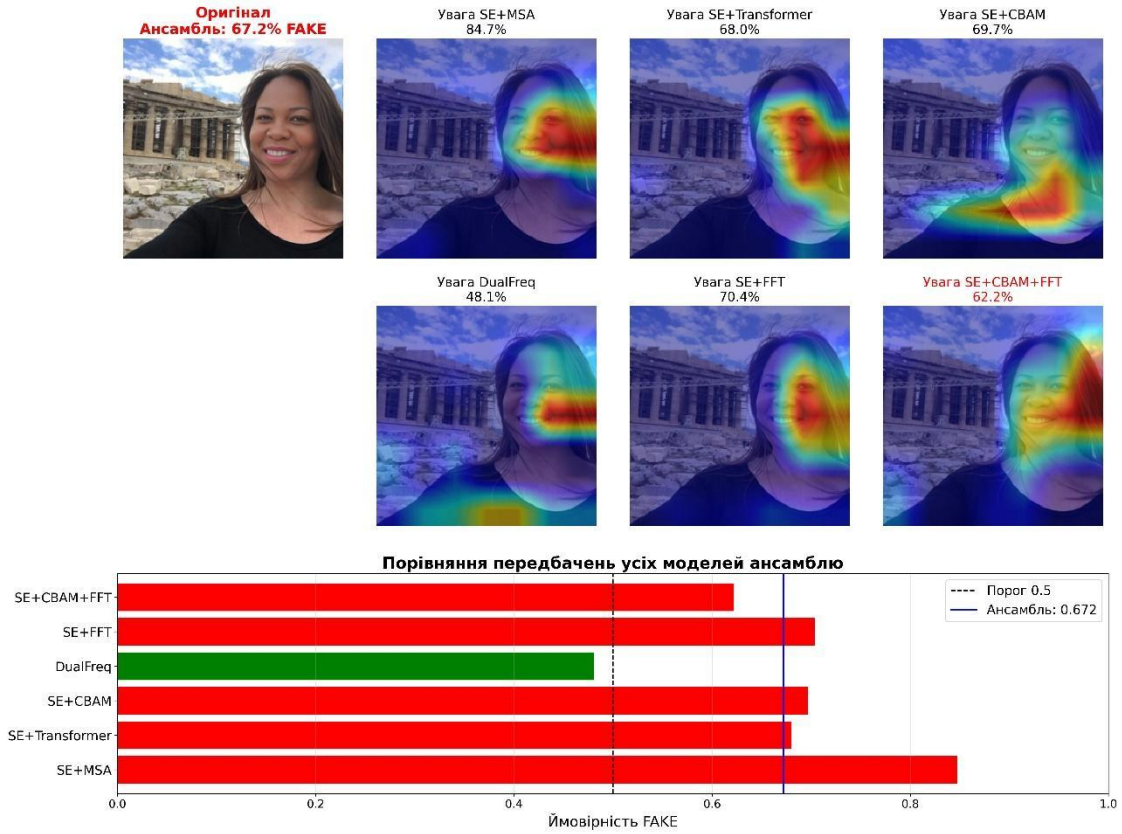

Рис. 9. Візуалізація карт уваги моделей ансамблю: приклад успішного виявлення дипфейку

Однак, окрім успішних випадків, особливий науковий інтерес становить аналіз помилок першого роду (False Negatives). На Рисунку 10 наведено показовий випадок, коли загальний ансамбль хибно класифікував високоякісний дипфейк як справжнє зображення (фінальна ймовірність підробки склала лише 18.7 %).

Аналіз цього прикладу свідчить, що більшість просторових моделей (зокрема ResNetSEMSA та ResNetSECBAM) сфокусували свою увагу на ідеально згенерованих ділянках шкіри, де відсутні візуальні дефекти, що призвело до їхньої хибної впевненості у реалістичності кадру. Натомість найскладніша гібридна архітектура ResNetSECBAMFFT виявилася єдиною моделлю, здатною розпізнати підробку (впевненість 61.4 %). Її теплова карта показує, що завдяки модулю частотної уваги фокус мережі змістився на фонові структури та межі волосся, де спектральний аналіз зафіксував невидимі для ока високочастотні шуми генерації. Цей факт підтверджує, що правильне рішення частотної моделі було нівельоване впевненими помилками просторових моделей під час простого усереднення ймовірностей (soft voting), що вказує на необхідність розробки методів адаптивного зважування в майбутніх дослідженнях.

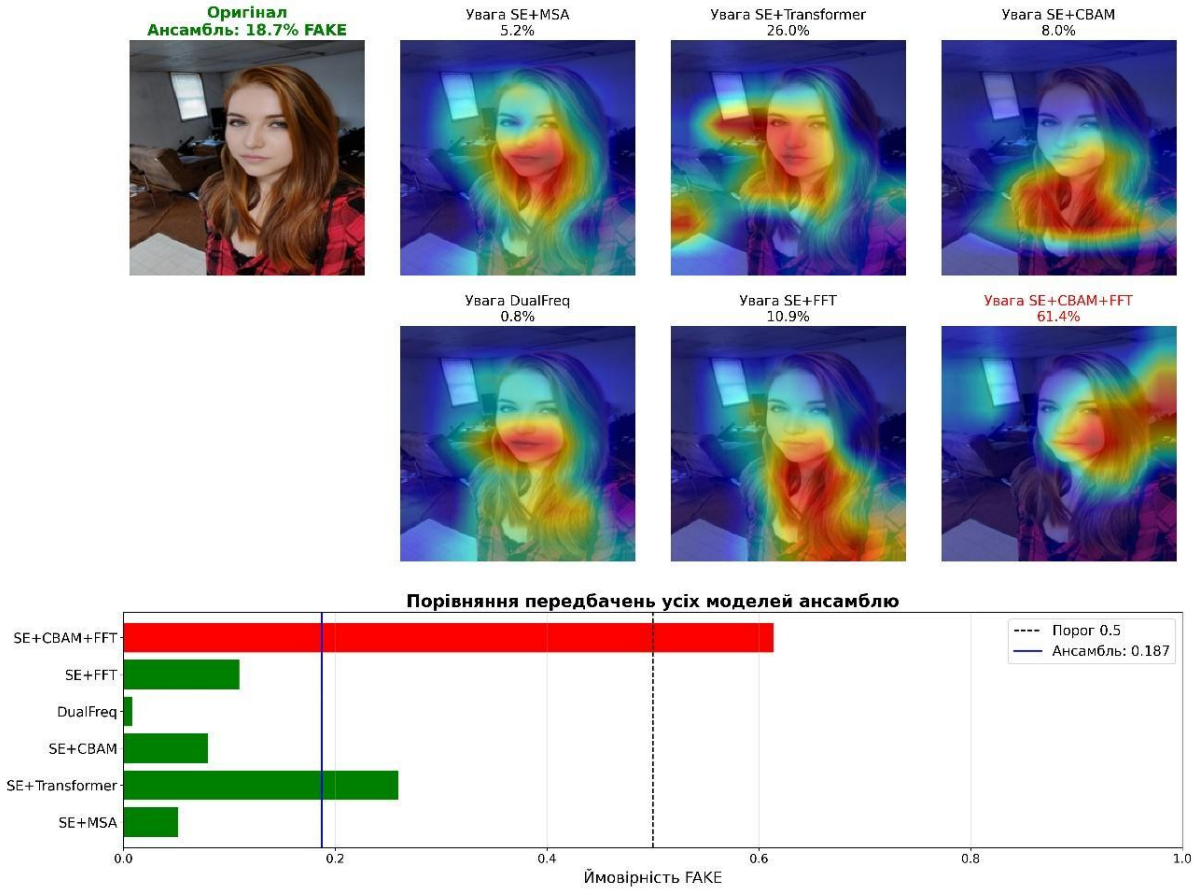

Рис. 10. Візуалізація карт уваги моделей ансамблю: аналіз помилки класифікації (False Negative)

6. Висновки та перспективи подальших досліджень

У ході виконання даної роботи було успішно розв'язано всі поставлені задачі та отримано наступні результати:

1. Проведено аналіз сучасних методів виявлення згенерованого візуального контенту . Встановлено, що класичні просторові детектори швидко перенавчаються та втрачають ефективність при зміні алгоритмів генерації, що обумовлює необхідність переходу до комплексних просторово-частотних систем.

2. Запропоновано та програмно реалізовано шість гібридних архітектур на базі мережі ResNet-50 . Їхня ефективність забезпечується цілеспрямованою інтеграцією паралельних та послідовних модулів просторової, міжканальної та частотної уваги (MSA, CBAM, Transformer, FFT). Це дозволяє моделям відійти від простого аналізу пікселів і зосередитись одночасно на локальних дефектах блендінгу та невидимих людському оку високочастотних спектральних аномаліях.

3. Розроблено ансамблеву модель на основі навчених спеціалізованих гібридів. Завдяки синергії різних механізмів уваги, ансамбль ефективно об'єднує просторові та частотні ознаки, що дозволяє мінімізувати хибні спрацьовування та суттєво підвищити загальну надійність системи детекції порівняно з індивідуальними класифікаторами.

4. Проведено експериментальне тестування розроблених моделей на комплексних наборах даних та виконано порівняльний аналіз їхньої ефективності. За результатами дослідження встановлено наступне:

1 - Індивідуальні архітектури (зокрема ResNetSECBAM та ResNetSEFFT) демонструють високу загальну точність (Accuracy до 93.80 %) та збалансованість (F1-міра до 0.8067) на валідаційній вибірці.

2 - Сформована ансамблева модель дозволила досягти максимальної ефективності детекції на внутрішньому наборі даних: точність склала 99.20 %, а F1-міра - 0.9910.

3 - Перевірка на незалежному зовнішньому наборі (Deepfake vs Real 60K) зафіксувала зниження загальної точності до 74.7 %, що кількісно підтверджує наявність проблеми обмеженої узагальнювальної здатності нейромереж при обробці контенту від абсолютно нових типів генераторів.

4 - Здійснений візуальний аналіз за допомогою карт Grad-CAM довів, що у випадках високоякісних підробок просторові моделі можуть хибно класифікувати зображення як справжнє, тоді як частотна гілка (ResNetSECBAMFFT) продовжує успішно локалізувати спектральні шуми генерації.

Перспективи подальших досліджень. Враховуючи виявлену проблему зниження загальної точності при крос-доменному тестуванні (до 74.7 %), пріоритетним напрямом подальших досліджень є розробка алгоритмів динамічного (адаптивного) зважування прогнозів ансамблю. Це дозволить системі автоматично надавати вищий пріоритет частотним гілкам у випадках, коли просторові ознаки не містять видимих дефектів. Крім того, перспективним є застосування методів контрастивного навчання (contrastive learning) для формування у нейромереж інваріантних до генератора ознак підробок.

Внесок авторів Дмитро Азарний - програмне забезпечення; збір і перевірка емпіричних даних; емпіричне дослідження; розробка архітектури; аналіз джерел та підготовка початкового проєкту статті. Анатолій Давиденко - концептуалізація; розробка архітектури; наукове керівництво під час проведенням дослідження.

Декларація про штучний інтелект

Під час підготовки цієї статті було використано інструмент штучного інтелекту Gemini для перекладу іншомовних наукових першоджерел українською мовою з метою їх подальшого аналізу та формування огляду літератури. Після використання цього інструменту автори ретельно перевірили отримані матеріали, самостійно підготували текст дослідження та несуть повну відповідальність за зміст публікації.

Конфлікт інтересів

Автори заявляють про відсутність конфлікту інтересів та підтверджують, що під час підготовки цієї роботи не існувало жодних комерційних, фінансових чи інших взаємовідносин, які могли б бути розцінені як такі, що здатні вплинути на результати дослідження або їх інтерпретацію. Робота виконана відповідно до принципів академічної доброчесності, етичних норм проведення наукових досліджень та вимог редакційної політики щодо запобігання конфлікту інтересів.

Список використаної літератури

1. Alam, I., Islam, M. T., & Woo, S. S. (2025). SpecXNet: A dual-domain convolutional network for robust deepfake detection. In Proceedings of the 33rd ACM International Conference on Multimedia (MM '25). ACM. https://doi.org/10.48550/arXiv.2509.22070

2. Luo, X., & Wang, Y. (2025). Frequency-domain masking and spatial interaction for generalizable deepfake detection. Electronics, 14(7), Article 1302. https://doi.org/10.3390/electronics14071302

3. Zhou, Z., Sun, J., & Li, X. (2025). EDFM: An enhanced dual-branch fusion model for face deepfake detection. Journal of Applied and Numerical Optimization, 7, 97–111. https://jano.biemdas.com/archives/1773

4. Yan, J., Li, Z., Wang, F., He, Z., & Fu, Z. (2025). Dual frequency branch framework with reconstructed sliding windows attention for AI-generated image detection. IEEE Transactions on Information Forensics and Security. https://doi.org/10.48550/arXiv.2501.15253

5. Bhattacharjee, A., Islam, K., Anan, K., Intesher, A., Fuad, A. A., Saha, U., & Imtiaz, H. (2025). CAE-Net: Generalized deepfake image detection using convolution and attention mechanisms with spatial and frequency domain features [arXiv:2502.10682]. arXiv. https://doi.org/10.48550/arXiv.2502.10682

6. Li, K., Ren, W., Li, J., Wang, W., & Cao, X. (2025). Critical forgetting-based multi-scale disentanglement for deepfake detection. Proceedings of the AAAI Conference on Artificial Intelligence, 39(1). https://doi.org/10.1609/aaai.v39i1.32021

7. Sheng, Y., Zou, Z., Yu, Z., Pang, M., Ou, W., & Han, W. (2025). ID-insensitive deepfake detection model based on multi-attention mechanism. Scientific Reports, 15, Article 11168. https://doi.org/10.1038/s41598-025-96254-6

8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778). https://doi.org/10.48550/arXiv.1512.03385

9. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7132–7141). https://doi.org/10.48550/arXiv.1709.01507

10. Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3–19). https://doi.org/10.48550/arXiv.1807.06521

11. Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5781–5790). https://doi.org/10.48550/arXiv.1910.01717

12. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1–11). https://doi.org/10.48550/arXiv.1901.08971

13. DSAIL-SKKU. (2023). HiDF: High-quality human-indistinguishable deepfake dataset [Dataset]. GitHub. https://github.com/DSAIL-SKKU/HiDF

14. prithivMLmods. (2024). Deepfake vs real 60K [Dataset]. Hugging Face. https://huggingface.co/datasets/prithivMLmods/Deepfake-vs-Real-60K

References

1. Alam, I., Islam, M. T., & Woo, S. S. (2025). SpecXNet: A dual-domain convolutional network for robust deepfake detection. In Proceedings of the 33rd ACM International Conference on Multimedia (MM '25). ACM. https://doi.org/10.48550/arXiv.2509.22070

2. Luo, X., & Wang, Y. (2025). Frequency-domain masking and spatial interaction for generalizable deepfake detection. Electronics, 14(7), Article 1302. https://doi.org/10.3390/electronics14071302

3. Zhou, Z., Sun, J., & Li, X. (2025). EDFM: An enhanced dual-branch fusion model for face deepfake detection. Journal of Applied and Numerical Optimization, 7, 97–111. https://jano.biemdas.com/archives/1773

4. Yan, J., Li, Z., Wang, F., He, Z., & Fu, Z. (2025). Dual frequency branch framework with reconstructed sliding windows attention for AI-generated image detection. IEEE Transactions on Information Forensics and Security. https://doi.org/10.48550/arXiv.2501.15253

5. Bhattacharjee, A., Islam, K., Anan, K., Intesher, A., Fuad, A. A., Saha, U., & Imtiaz, H. (2025). CAE-Net: Generalized deepfake image detection using convolution and attention mechanisms with spatial and frequency domain features [arXiv:2502.10682]. arXiv. https://doi.org/10.48550/arXiv.2502.10682

6. Li, K., Ren, W., Li, J., Wang, W., & Cao, X. (2025). Critical forgetting-based multi-scale disentanglement for deepfake detection. Proceedings of the AAAI Conference on Artificial Intelligence, 39(1). https://doi.org/10.1609/aaai.v39i1.32021

7. Sheng, Y., Zou, Z., Yu, Z., Pang, M., Ou, W., & Han, W. (2025). ID-insensitive deepfake detection model based on multi-attention mechanism. Scientific Reports, 15, Article 11168. https://doi.org/10.1038/s41598-025-96254-6

8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778). https://doi.org/10.48550/arXiv.1512.03385

9. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7132–7141). https://doi.org/10.48550/arXiv.1709.01507

10. Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3–19). https://doi.org/10.48550/arXiv.1807.06521

11. Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5781–5790). https://doi.org/10.48550/arXiv.1910.01717

12. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1–11). https://doi.org/10.48550/arXiv.1901.08971

13. DSAIL-SKKU. (2023). HiDF: High-quality human-indistinguishable deepfake dataset [Dataset]. GitHub. https://github.com/DSAIL-SKKU/HiDF

14. prithivMLmods. (2024). Deepfake vs real 60K [Dataset]. Hugging Face. https://huggingface.co/datasets/prithivMLmods/Deepfake-vs-Real-60K

Надійшла до редакції: 20.03.26
Прийнята до друку:12.06.26
Опубліковано:30.06.26