КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.32.1086
УДК 004.93:004.056

Лащевська Наталія Олександрівна
к.т.н., доцент, завідувач кафедри Комп'ютерної інженерії
Державний університет інформаційно-телекомунікаційних технологій, Київ, Україна

Мішкур Юрій Валентинович
аспірант кафедри Комп'ютерної інженерії
Державний університет інформаційно-телекомунікаційних технологій, Київ, Україна

ВИЯВЛЕННЯ СТЕГАНОГРАФІЇ НА ЗОБРАЖЕННІ З ВИКОРИСТАННЯМ ЛЕГКОВАЖНИХ МОДЕЛЕЙ ГЛИБОКОГО НАВЧАННЯ

Анотація. У роботі представлено комплексне експериментальне дослідження можливостей застосування легковажних згорткових нейронних мереж для задач стегоаналізу цифрових зображень в умовах обмежених обчислювальних ресурсів. Метою роботи є оцінювання ефективності компактних архітектур MobileNetV2 та EfficientNetV2S у поєднанні з високочастотним препроцесінгом зображень та визначення умов, за яких такі моделі можуть забезпечити високу точність виявлення прихованих повідомлень без використання ресурсоємних спеціалізованих мереж. У ході дослідження встановлено, що високочастотна попередня обробка є критично необхідною складовою сучасних систем стегоаналізу. За відсутності блоку високочастотних фільтрів легковажні CNN не здатні ефективно відокремлювати слабкий стегошум від семантичного контенту зображення. Показано, що недостатня кількість HPF-фільтрів або шарів препроцесінгу призводить до швидкого перенавчання та нестабільності результатів класифікації. Водночас використання структурованого та достатньо потужного блоку HPF забезпечує стійке навчання і високу узагальнювальну здатність моделей. Окрему увагу приділено порівнянню фіксованих та тренованих високочастотних фільтрів. Результати експериментів свідчать, що можливість адаптації HPF-ядер у процесі навчання дозволяє моделі підлаштовуватися під статистичні особливості стегошуму, забезпечуючи додатковий приріст точності. Такий підхід ефективно поєднує апріорні знання класичного стегоаналізу з перевагами глибокого навчання. Запропонований підхід було перевірено на зображеннях з прихованими повідомленнями, вбудованими методом LSB за різних значень обсягу вбудовування. Отримані результати підтверджують, що архітектури MobileNetV2 та EfficientNetV2S у поєднанні з HPF-препроцесінгом забезпечують високу якість виявлення за низьких ресурсних витрат. Водночас подальші дослідження необхідні для поширення підходу на складні адаптивні методи стеганографічного вбудовування.

Ключові слова: стегоаналіз, глибоке навчання, EfficientNet, MobileNet, HPF, згорткові нейронні мережі, залишкові шуми.

ВСТУП

Стеганографія – це процес приховування повідомлень у цифровому контенті (зображеннях, аудіо, відео) з метою передачі інформації так, щоб її існування було невідоме стороннім.

У сучасну цифрову епоху стеганографія – наука про приховування інформації в каналах передачі даних – стала потужним інструментом як для захисту конфіденційності, так і для зловмисних дій (передача команд ботнетам з командно-

керуючий центром, витік корпоративних даних). Протидію цьому забезпечує стегоаналіз, метою якого є виявлення факту наявності прихованого повідомлення [1-3].

Стегоаналіз потрібен для виявлення прихованої інформації в цифрових об'єктах (зображеннях, аудіо, відео, мережевому трафіку), що є критично важливим для кібербезпеки, судової експертизи та запобігання витоку даних. Його основна мета – довести факт наявності стеганографічного повідомлення (тобто, що файл був модифікований) і, за можливості, витягти це повідомлення [1]. Таким чином, стеганалітика (steganalysis) – це зворотна задача: виявлення та (іноді) локалізація вбудованих повідомлень. Сучасні адаптивні алгоритми стеганографічного вбудовування (зокрема S-UNIWARD, WOW, HUGO та MiPOD) зменшують статистичну помітність модифікацій, що робить їх детектування складним, передусім за малих коефіцієнтів вбудовування [4-6].

Широке використання алгоритмів глибокого навчання для обробки зображень привело до формування CNN-архітектур, придатних для стегоаналізу. Але архітектури CNN для стегоаналізу відрізняються від стандартних мереж (наприклад, VGG або ResNet), оскільки вони повинні виявляти надзвичайно слабкі, шум-подібні сигнали (стегошум), а не семантичний зміст зображення. Тому вони включають спеціалізовані вхідні шари та механізми придушення контенту [1-3]. Навчання найпотужніших стеганалітичних мереж (наприклад, SRNet та JYeNet) зазвичай займає багато часу. Крім того, більшість існуючих мереж спеціально розроблені для одного домену. Наприклад, YeNet призначений для просторової області, тоді як JYeNet – для області JPEG. Їхня продуктивність зазвичай незадовільна для інших областей.

Разом з тим, розгортання важких моделей (ResNet-подібних або SRNet) у прикладних або вбудованих системах обмежено через обмежені ресурси обчислення та пам'яті. Тому дослідження легковажних моделей (сімейства MobileNet або EfficientNet та їх модифікації) у стеганалітиці є актуальним науковим завданням і має практичну цінність.

Аналіз літературних даних і постановка проблеми. Методика на основі глибокого навчання значно покращила комп'ютерний зір, обробку природної мови та інші галузі. Стеганаліз на основі глибокого навчання привернув увагу дослідників, і було запропоновано багато пов'язаних досліджень [6-9].

На думку [10], дослідження методи побудови розрахункових детекторів на основі ШНМ і впливу архітектурних характеристик ШНМ на точність виявлення стегозображень є актуальним та важливим завданням.

Моделі глибокого навчання пропонують динамічний та адаптивний підхід до стеганографії. Алгоритми глибокого навчання можуть автоматично вивчати складні шаблони та представлення з необроблених даних, не покладаючись на заздалегідь визначені ознаки. Ця гнучкість дозволяє моделям глибокого навчання краще узагальнюватися для різних типів стеганографії, включаючи ті, що використовують складні або нові методи. Як результат, моделі стеганологічного аналізу на основі глибокого навчання досягають вищої точності виявлення [6, 7].

Більшість сучасних моделей (наприклад, SRNet) є ресурсоємними, мають мільйони параметрів та вимагають потужних GPU для навчання і роботи. Це ускладнює їх використання в режимі реального часу на пристроях з обмеженою обчислювальною потужністю, таких як мобільні телефони. При цьому мобільні пристрої вразливі до стеганографічних загроз, що вимагає рішень, які можуть ефективно працювати з мінімальним споживанням ресурсів для виявлення прихованих даних [10, 11].

Один з шляхів вирішення проблеми продуктивності систем стегоаналізу – інтеграція до них легких архітектур глибокого навчання, зокрема MobileNet або та EfficientNet. Ці моделі мають сучасну архітектуру, яка оптимізує продуктивність, мінімізуючи обчислювальні вимоги [9].

Але стегосигнали у зображеннях зазвичай проявляються як слабкі, локальні зміни в шумовій/високочастотній компонентах зображення. Тому для підсилення цих сигналів і відсунення вкладу контенту (низькочастотної інформації) відомою практикою є використання фільтрів високих частот (HPF) або банків резидуальних фільтрів (SRM) [10,11].

Традиційний стегоаналіз, зокрема виявлення методів приховування у просторовій області (наприклад, LSB або адаптивні WOW/S-UNIWARD), ґрунтується на виявленні стего-шуму, який є побічним продуктом модифікації пікселів. Цей шум є надзвичайно слабким і має високу частоту. Звичайні інструменти обробки зображень пригнічують його як незначний артефакт.

Для ефективного виявлення цього шуму необхідний метод, який здатний придушити семантичний контент зображення-носія, залишаючи лише залишки (residuals). Саме для цього було розроблено концепцію Spatial Rich Model (SRM) [2]. SRM-фільтри – це набір лінійних фільтрів високих частот (HPF), призначених для вилучення локальних ознак кореляції та залишкових шумів із зображення. Ключовою вимогою до цих фільтрів є те, що сума їхніх коефіцієнтів має дорівнювати нулю, що гарантує пригнічення низькочастотних складових (основного контенту зображення).

Модель SRM, запропонована в [2], використовувала великий набір різноманітних ядер, які можна розділити на кілька функціональних груп: базові високочастотні фільтри, фільтри другого порядку, усереднюючі фільтри. В роботі [2] було використано набір із близько 30 ядер розміром 5x5.

При побудові мереж CNN для стегоаналізу також використовують SRM-фільтри [12, 13] в двох варіантах: фіксований вхідний шар з постійними вагами і навчальний вхідний шар з можливістю оновлювати ваги під час навчання.

Використання HPF-ядер, отриманих із SRM, є ключовим фактором, що дозволяє з моделям з CNN-архітектурами (в тому числі легковажним мережам, таким як MobileNet або EfficientNet) досягати високої точності в стегоаналізі [7, 14, 15]. Без застосування таких архітектурних рішень нейронна мережа сприймає стего-сигнал як незначний шум.

Питання вибору кількості фільтрів високих частот (HPF), особливо тих, що походять від Spatial Rich Model (SRM), на вході CNN є критичним компромісом між інформативністю та обчислювальною складністю моделей для стегоаналізу.

Збільшення кількості різних SRM-фільтрів у вхідному шарі веде до зростання кількості інформації про локальні залежності та шуми зображення.

Використання широкого набору ядер (наприклад, 30-ти або 40-ка різних фільтрів 5x5, як роботах [2, 16, 17]) дозволяє виділити залишкові шуми, і тому покращує здатність моделі розділяти природні шуми носія і додаткові шуми прихованого вкладення. Більш багатий простір ознак може сприяти кращій генералізації моделі на різні типи зображень або невідомі алгоритми вбудовування. Але кожен додатковий вхідний канал (вихід фільтра) значно збільшує кількість операцій у наступних згорткових шарах. Крім того, частина ядер SRM може бути сильно корельованою. Збільшення кількості фільтрів після певного порогу приносить незначне покращення точності, але значно збільшує обчислювальну вартість.

Потеційне використання легковажних архітектур (MobileNet, EfficientNet) прагне мінімізувати кількість вхідних фільтрів. Дослідження [1, 2, 18, 19] показали, що більшість інформації про стего-шум концентрується у фільтрах, що вимірюють горизонтальні, вертикальні та діагональні різниці. В роботі [20] встановлено, що хоча HPF-фільтри (локальні різниці) є ключовими, використання їхньої оптимізованої підмножини необхідне для уникнення надмірності, перенавчання та покращення загальної ефективності CNN.

Збільшення кількості фільтрів вимагає більше даних для навчання, щоб уникнути перенавчання моделі. Для меншої кількості фільтрів навчання може бути більш стабільним і швидшим.

У даній роботі зроблено спробу розв'язання вищеназваних проблем.

Метою дослідження є адаптація легковажних архітектур конволюційних нейронних мереж (MobileNet, EfficientNet) для стеганалітики без втрати якості летектування у порівнянні з більш важкими спеціалізованими мережами (Xu-Net, SRNet), із одночасною суттєвою економією ресурсів.

Для досягнення поставленої мети вирішено такі **завдання**:

– проаналізувати вплив HPF-препроцесінгу (фіксованих SRM-фільтрів) на продуктивність легковажних CNN;

– проаналізувати взаємозв'язок між точністю виявлення та ресурсними вимогами для легковажних CNN;

– обрати оптимальну структуру вхідного фільтра для покращення продуктивності легковажних CNN.

Методика дослідження.

1. Побудова набору даних для навчання і перевірки моделі.

Для вбудовування текстових повідомлень було використано відомий набір даних CIFAR10 [21].

CIFAR-10 – це широко використовуваний датасет у комп'ютерному зорі, який містить 60 000 кольорових зображень розміром 32x32 пікселів у 10 класах.

Ключові особливості CIFAR-10 для стеганографії:

1) Обмежена кількість пікселів зображення (32x32) означає, що відносний обсяг вбудованого повідомлення (payload capacity) швидко стає високим.

2) Набір даних містить кольорові зображення (RGB), що забезпечує 3 байти або 24 біти для кожного пікселя. Це збільшує ємність вбудовування.

3) Зображення зберігаються у форматі PNG або raw-форматі без втрат, що унеможливлює внесення артефактів, які зазвичай ускладнюють LSB-вбудовування.

На думку [22, 23, 24] CIFAR-10 є хорошим джерелом для зображень для побудови великого навчального набору (120 000 фотографій), який було використано для порівняння ефективності різних AI/ML моделей у виявленні прихованих повідомлень.

Для додавання прихованого напису було використано техніку LSB. Стеганографія LSB – це підхід до приховування повідомлень, який безпосередньо змінює біти, найменш значущі для кольору пікселя: останній(і) біт(и) [25, 26]. Точніше, він замінює значення існуючих бітів двійковим значенням повідомлення. Підхід LSB є найбільш традиційним та найпростішим у реалізації стеганографічним підходом. Хоча простий LSB-метод є легко виявним, він має і деяки переваги для створення датасетів і перевірки роботи моделей.

Ключові переваги LSB-методів при генерації датасетів [1,12]:

1) Ізоляція стего-шуму: LSB-заміна (або її адаптовані варіанти) в просторовій області на нестиснутих носіях (наприклад, файли PNG або PPM з BOSSBase) дозволяє

створити синтетичний датасет, на якому єдиною суттєвою відмінністю між зображенням-носієм та стего-зображенням є дисторсія, внесена LSB. Це дозволяє нейронній мережі сфокусуватися виключно на вивченні залишків LSB-шуму.

2) LSB є найпростішим для виявлення методом. Навчання CNN на LSB-зразках гарантує, що модель спочатку засвоїть основні принципи виділення шумів за допомогою HPF-фільтрів, перш ніж переходити до більш складних, адаптивних алгоритмів (наприклад, WOW або S-UNIWARD).

3) LSB-вбудовування може забезпечити високу щільність (наприклад, 1.0 біт на піксель або 0.5 bpp), яка необхідна для забезпечення чіткого сигналу помилки для CNN.

2. Легковажні архітектури для побудови моделі стегоаналізу.

За думкою [27], легковажні CNN – це життєздатний і часто бажаний вибір для стегоаналізу. Менша кількість параметрів і використання ефективних операцій (наприклад, глибинно-роздільні згортки) дозволяють швидше навчати моделі та отримувати результати висновків.

Але ці архітектури потрибують інтеграції з фіксованими HPF-фільтрами (наприклад, SRM-ядрами). Фактично, блок фільтрів виділяє шумоподібні стего-сигнали, а легковажна мережа ефективно агрегує ці ознаки, мінімізуючи обчислювальні витрати.

MobileNetV2 – легка CNN-архітектура, яка використовує інвертовані резидуальні блоки та глибинні згортки [27]. Її вибір у роботі обумовлений достатньою глибиною для виділення складних патернів, помірною кількістю параметрів, придатністю до розгортання в обмежених середовищах.

EfficientNetV2 ґрунтується на оптимізації як архітектури, так і процесу навчання, використовуючи автоматичне масштабування (Compound Scaling) та оптимізовані згорткові блоки [28]. Ця архітектура забезпечує найкращий баланс між точністю та обчислювальною вартістю. Для стегоаналізу це дозволяє використовувати більш глибокі варіанти (наприклад, B0 або B3), ніж можна було б дозволити зі старими мережами.

3. Побудова блоку фільтрів.

Високочастотні фільтри (HPF) відіграють центральну роль у більшості сучасних систем стегоаналізу, заснованих на глибокому навчанні (CNN). Вони є необхідною передумовою для успішного виявлення стегосигналу. Основна мета HPF – виділення залишкового шуму зображення, в якому містяться слабкі статистичні аномалії, внесені стеганографічним вбудовуванням (наприклад, LSB).

Ключова вимога до HPF-ядра K у стегоаналізі полягає в тому, що сума його коефіцієнтів має дорівнювати нулю $\left(\sum_{i,j} K_{i,j} = 0\right)$. Це гарантує, що фільтр усуває низькочастотні, контентні компоненти (плавні області), максимізуючи при цьому контрастність між природним шумом та стего-шумом [1, 17, 30].

В перших роботах, присвячених побудові систем стегоаналізу, було використано фіксовані (нетреновані) HPF-шари, що передують основній CNN-архітектурі. Це забезпечує стабільний і стандартизований вхід у вигляді залишкового зображення [1, 2, 17, 30].

Сучасні архітектури для посилення виділення стегосигналу використовують глибинний HPF-стем [31-33], що складається з кількох послідовних HPF-шарів. Перший шар може містити декілька спрямованих ядер (наприклад, горизонтальні, вертикальні, діагональні SRM-фільтри). Наступні шари можуть бути як фіксованими, так і навчальними.

В цьому дослідження було використано декілька варіантів архітектури вхідного блоку фільтрів, а саме:

1) Перший варіант містив декілька однакових фільтрів 5x5 з фіксованим ядром. Кількість використаних фільтрів змінювалась від 1 до 60.
2) Другий варіант також містив декілька однакових фільтрів 5x5 з початковим фіксованим ядром, але з можливістю тренування фільтруючих шарів.
3) Третий варіант містив декілька груп спрямованих ядер 3x3 (горизонтальні, вертикальні, діагональні SRM-фільтри) з можливістю тренування фільтруючих шарів.
4) Четвертий варіант містив декілька груп спрямованих ядер 5x5 (горизонтальні, вертикальні, діагональні SRM-фільтри) з можливістю тренування фільтруючих шарів.
5) П’ятий варіант містив декілька груп спрямованих ядер 7x7 (горизонтальні, вертикальні, діагональні SRM-фільтри) з можливістю тренування фільтруючих шарів.

4. Проведення обчислювальних експерниментів

Для проведення експериментів були використані моделі наступної структури (рис. 1):

- Блок попередньої обробки (один з 4 варіантів);
- Легковажна конволюційна мережа (MobileNetv2 або EfficientNetV2S);
- Шар GlobalAveragePooling і щільний шар з активацією «сігмоїд»;
- Блок виводу ілюстрацій і перевірки відновлення вбудованого тексту.

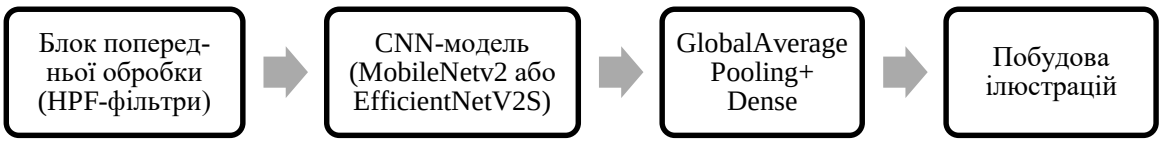

Рис. 1 Архітектура моделі стегоаналізу з використанням легковажних CNN

Всі експерименти виконувались в середовищі Google Collaboratory з використанням графічного прискорювача T4. Використовувалась мова програмування python, для побудови нейромережевих моделей був використаний пакет tensorflow з інтерфейсом keras. Також було використано деякі модулі пакету scikit-learn.

Зображення з датасету Cifar10 32x32x3 перед вбудовуванням прихованого тексту перетворювались на зображення 96x96x3. Для забезпечення контрольованих умов дослідження було сформовано декілька варіантів штучного датасета, які містили від 3000 до 60000 зображень, з яких: 50% – cover-зображення (без змін), 50% – stego-зображення (з вбудованим повідомленням). Розглядалися різні значення обсягу вбудовування (payload), що вимірювався в бітах на піксель (bpp).

Для вбудовування використовувались текстові повідомлення як англійською, так і українською мовою, використовувалось кодування utf-8 (це було враховано при побудові послідовності бітів).

Навчання проводилося з використанням оптимізатора Adam з регульованою початковою швидкістю навчання (в більшості експериментів 0.0001)та функції втрат binary cross-entropy.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Результати експериментів показали, що без використання HPF-препроцесінгу навіть при високих значеннях обсягу вбудовування виявлення прихованого тексту не досягається.

Для забезпечення можливості виявлення вбудованих артефактів треба використати хоча б один HPF-фільтр. Але навчання моделі з зовсім низькою кількістю фільтрів дуже швидко призводить до помітного перенавчання моделі (рис.2).

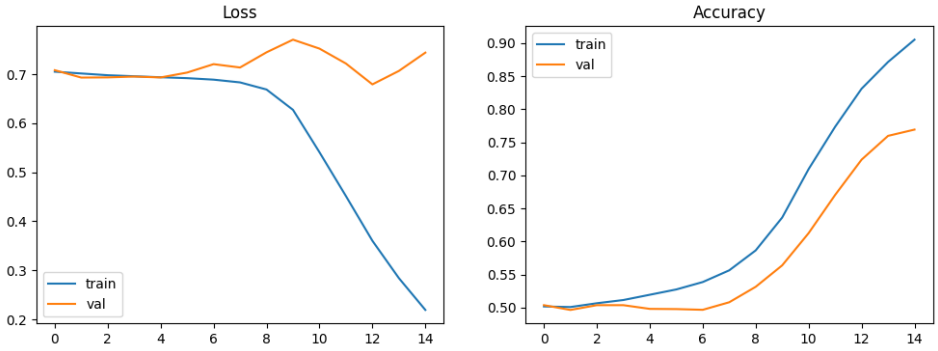

Рисунок 2. Процес навчання моделі MobileNetV2 з одним фільтром 5x5 у початковому блоці

Збільшення кількості фільтрів до 5 і більше значно покращує можливості навчання моделі і стійкість результату. Приклад кривих навчання моделі з 5 фільтрами на вході наведено на рис.3. Але цей результат не стійкий і дуже залежить від кількості зображень в навчальному датасеті і може змінитись при додаванні або видаленні лише одного фільтра.

Стійкий результат з точки зору відсутності перенавчання був отриманий лише для 45 фільтрів.

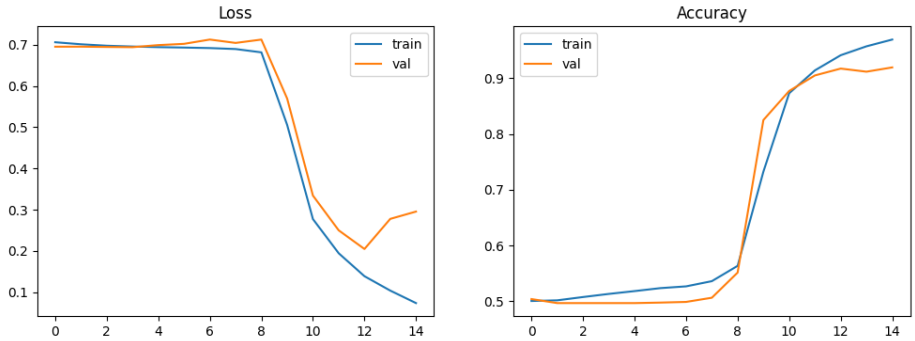

Рисунок 3. Процес навчання моделі MobileNetV2 з 5 фільтрами 5x5 у початковому блоці (набір даних з 60000 зображень)

Якщо групувати спрямовані ядра 3x3 (горизонтальні, вертикальні, діагональні SRM-фільтри), або 5x5 чи 7x7, стійкий результат з точки зору перенавчання моделі досягається при наявності не менш 5 груп фільтрів.

Порівняння результатів для моделей з фіксованими фільтрами та моделей з можливостю навчання фільтрів показало, що дозвіл на адаптацію HPF-ядер забезпечує

приріст якості на 8-12%, що свідчить про доцільність поєднання апріорних знань із глибоким навчанням.

ROC-криві продемонстрували стабільне зростання показника AUC при використанні тренованих HPF-шарів, досягаючи значень AUC ≈ 0.998, що вказує на високу роздільну здатність запропонованого детектора (рис. 4).

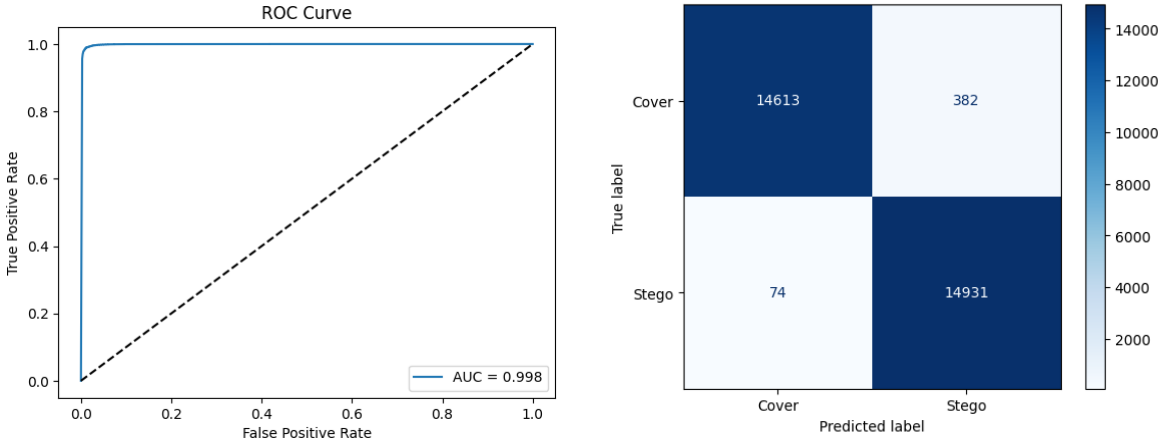

Рис.4 Оцінка якості навчання моделі стегоаналізу на базі архітектури MobileNetV2 з 15 групами спрямованих ядер 3x3 (набір даних з 60000 зображень)

Таким чином, експериментальні результати демонструють, що використання високочастотного препроцесінгу у поєднанні з MobileNetV2 суттєво покращує ефективність стегоаналізу. Зокрема, HPF-фільтри з можливостю тренування дозволяють моделі адаптуватися до характеру стего-шуму, що забезпечує приріст точності класифікації до 95-97% та значення AUC = 0.998. Отримані результати підтверджують доцільність поєднання апріорних знань (SRM-фільтри) з можливостями глибокого навчання.

При зміні архітектури класифікатору на модель EfficientNetV2S висновки про архітектуру блоку попередньої обробки зображень залишились без змін.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У результаті експериментальних досліджень було встановлено, що:

1. Високочастотний препроцесінг є критично важливим для ефективного стегоаналізу. Для стійкого виявлення стеганографії необхідно використання досить потужного блока структурованих HPF-фільтрів. Потреба в потужніших методах вилучення ознак, можливо, шляхом розширеної попередньої обробки або більш досконалих архітектур моделей, є очевидною. Крім того, більші, різноманітніші набори даних та включення ансамблевих методів можуть покращити точність моделей без значного збільшення обчислювальних вимог.
2. Недостатня кількість шарів HPF-фільтрів препроцесінгу призводить до помітного перенавчання класифікатора і суттєво знижує точність виявлення прихованих повідомлень на зображеннях.
3. HPF-шари з можливістю тренування забезпечують додаткову адаптацію до характеру стего-шуму;

4. Архітектури MobileNetV2 і EfficientNetV2S є доцільним вибором для побудови ефективного стегоаналітичного детектора. Але ця перевірка виконана для досить простого завдання – виявлення даних, які приховано за допомогою LSB-методу. Для уточнення умов і можливості побудови детектора більш складних методів приховування даних необхідні додаткові дослідження.

Отримані результати підтверджують ефективність запропонованого методу та його перспективність для практичного застосування.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Fridrich, J. (2010). *Steganography in digital media: Principles, algorithms, and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9781139192903
2. Fridrich, J., & Kodovský, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, *7*(3), 868–882. https://doi.org/10.1109/TIFS.2012.2190402
3. Ma, Y., Wang, J., Zhang, X., Wang, G., Xin, X., & Zhang, Q. (2025). LGS-Net: A lightweight convolutional neural network based on global feature capture for spatial image steganalysis. *IET Image Processing*, *19*. https://doi.org/10.1049/ipr2.70005
4. Karamanji, A., Ahmed, A., & Fadhil, A. (2024). Comparative deep learning models in applications of steganography detection. *Journal of Image and Graphics*, *12*(3), 312–319. https://doi.org/10.18178/joig.12.3.312-319
5. Zhang, R., Zhu, F., Liu, J., & Liu, G. (2018). Efficient feature learning and multi-size image steganalysis based on CNN. *arXiv Preprint*. https://arxiv.org/abs/1807.11428
6. Agarwal, S., & Jung, K.-H. (2022). Identification of content-adaptive image steganography using convolutional neural network guided by high-pass kernel. *Applied Sciences*, *12*(22), Article 11869. https://doi.org/10.3390/app122211869
7. Deng, X.-Q., Chen, B.-L., Luo, W.-Q., et al. (2022). Universal image steganalysis based on convolutional neural network with global covariance pooling. *Journal of Computer Science and Technology*, *37*, 1134–1145. https://doi.org/10.1007/s11390-021-0572-0
8. Yatsura, P., & Progonov, D. (2025). A review of modern methods for steganalysis and localization of embedded data in digital images. *Theoretical and Applied Cybersecurity*, *7*. https://doi.org/10.20535/tacs.2664-29132025.1.328265
9. Hao, L., Yi, Z., Jinwei, W., Weiming, Z., & Xiangyang, L. (2024). Lightweight steganography detection method based on multiple residual structures and transformer. *Chinese Journal of Electronics*, *33*(4), 965–978. https://doi.org/10.23919/CJE.2022.00.452
10. Mazurczyk, W., & Caviglione, L. (2014). Steganography in modern smartphones and mitigation techniques. *arXiv Preprint*. https://arxiv.org/abs/1410.6796
11. Chen, N., & Chen, B. (2022). Defending against OS-level malware in mobile devices via real-time malware detection and storage restoration. *Journal of Cybersecurity and Privacy*, *2*. https://doi.org/10.3390/jcp2020017
12. Boroumand, M., Chen, M., & Fridrich, J. (2019). Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, *14*(5), 1181–1193. https://doi.org/10.1109/TIFS.2018.2871749
13. Ye, J., Ni, J., & Yi, Y. (2017). Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, *12*(11), 2545–2557. https://doi.org/10.1109/TIFS.2017.2710946
14. Hu, F., Xu, R., & Cheng, Z. (2021). A fast and effective image steganalysis model based on convolutional neural network. *Journal of Physics: Conference Series*, *1861*, Article 012074. https://doi.org/10.1088/1742-6596/1861/1/012074
15. Agarwal, S., Kim, H., & Jung, K.-H. (2023). High-pass-kernel-driven content-adaptive image steganalysis using deep learning. *Mathematics*, *11*(20), Article 4322. https://doi.org/10.3390/math11204322
16. Qian, Y., Dong, J., Wang, W., & Tan, T. (2015). Deep learning for steganalysis via convolutional neural networks. In *Media watermarking, security, and forensics 2015* (Vol. 9409, Article 94090J). SPIE. https://doi.org/10.1117/12.2083479

17. Holub, V., & Fridrich, J. (2012). Designing steganographic distortion using directional filters. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*. https://doi.org/10.1109/WIFS.2012.6412655
18. Hussain, I., Zeng, J., & Tan, S. (2020). A survey on deep convolutional neural networks for image steganography and steganalysis. *KSII Transactions on Internet and Information Systems, 14*, 1228–1248. https://doi.org/10.3837/tiis.2020.03.017
19. Wu, L., Han, X., Wen, C., & Li, B. (2020). A steganalysis framework based on CNN using the filter subset selection method. *Multimedia Tools and Applications, 79*. https://doi.org/10.1007/s11042-020-08831-8
20. Krizhevsky, A. (2012). *Learning multiple layers of features from tiny images* [Technical report]. University of Toronto.
21. DiSalvo, N. (2025). Steganographic embeddings as an effective data augmentation. *arXiv Preprint*. https://arxiv.org/abs/2502.15245
22. Stefanek, G., Gulbransen, L., Spink, G., Morawski, J., Filla, D., & Rabello De Castro, R. (2024). A comparison of AI models to detect hidden messages in images. *Issues in Information Systems*, 119–132. https://doi.org/10.48009/3_iis_2024_110
23. K, V., Annem, P., Devarakonda, M., Jyothi, A., & Rayudu, N. (2025). Image steganography with CNN-based encoder-decoder. *International Journal for Modern Trends in Science and Technology*, *11*(4), 60–64. https://doi.org/10.5281/zenodo.15108976
24. Fridrich, J., Goljan, M., & Du, R. (2001). Detecting LSB steganography in color and gray-scale images. *IEEE Multimedia*, *8*, 22–28. https://doi.org/10.1109/93.959097
25. Kombrink, M. H., Geradts, Z. J. M. H., & Worring, M. (2025). Image steganography approaches and their detection strategies: A survey. *ACM Computing Surveys*, *57*(2), Article 33. https://doi.org/10.1145/3694965
26. Bauravindah, A., & Fudholi, D. (2024). Lightweight models for real-time steganalysis: A comparison of MobileNet, ShuffleNet, and EfficientNet. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 8*, 737–747. https://doi.org/10.29207/resti.v8i6.6091
27. Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4510–4520).
28. Zhou, A., Ma, Y., Ji, W., Zong, M., Yang, P., Wu, M., & Liu, M. (2022). Multi-head attention-based two-stream EfficientNet for action recognition. *Multimedia Systems*, *29*, 1–12. https://doi.org/10.1007/s00530-022-00961-3
29. Farooq, N., & Mir, R. (2024). Image steganalysis using deep convolution neural networks: A literature survey. *International Journal of Sensors, Wireless Communications and Control*, *14*(4), 247–264. https://doi.org/10.2174/0122103279296370240529075507
30. Chaumont, M. (2020). Deep learning in steganography and steganalysis. In *Digital media steganography*. https://doi.org/10.1016/B978-0-12-819438-6.00022-0
31. Wu, S., Zhong, S.-H., & Liu, Y. (2020). A novel convolutional neural network for image steganalysis with shared normalization. *IEEE Transactions on Multimedia*, *22*(1), 256–270. https://doi.org/10.1109/TMM.2019.2920605
32. Dwaik, A., & Belkhouche, Y. (2024). Enhancing the performance of convolutional neural network image-based steganalysis in spatial domain using spatial rich model and 2D Gabor filters. *Journal of Information Security and Applications, 85*, Article 103864. https://doi.org/10.1016/j.jisa.2024.103864

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023

DETECTION OF STEGANOGRAPHY IN IMAGES USING LIGHTWEIGHT DEEP LEARNING MODELS

Abstract. This paper presents a comprehensive experimental study on the applicability of lightweight convolutional neural networks (CNNs) for the task of digital image steganalysis under resource-constrained conditions. The main objective of the research is to evaluate whether compact architectures such as MobileNetV2 and EfficientNetV2S can achieve high detection accuracy when combined with appropriate high-frequency preprocessing, while significantly reducing computational and memory requirements compared to heavy specialized steganalytic networks. The study demonstrates that high-pass filtering (HPF) is a critical prerequisite for effective steganography detection. Without explicit suppression of image content, lightweight CNNs fail to distinguish weak stego-noise from natural image structures. Experimental results show that an insufficient number of HPF layers or filters leads to rapid overfitting and unstable classification performance. In contrast, the use of a sufficiently expressive and structured HPF preprocessing block enables stable convergence and high generalization ability. Special attention is paid to the influence of fixed versus trainable HPF filters. It is shown that allowing HPF kernels to adapt during training provides an additional performance gain by adjusting to the statistical characteristics of stego-noise. This hybrid approach effectively combines prior knowledge from classical steganalysis (SRM-based filters) with the adaptive capabilities of deep learning. The proposed detection framework was evaluated using images with data embedded by the least significant bit (LSB) method under various payload conditions. Both MobileNetV2 and EfficientNetV2S demonstrated high detection accuracy and near-optimal ROC characteristics while maintaining low computational complexity. These results confirm that lightweight CNN architectures, when properly augmented with high-frequency preprocessing, represent a viable solution for practical steganalysis, including deployment on embedded or mobile platforms. However, the current evaluation is limited to LSB-based embedding. Future work will focus on extending the proposed approach to more sophisticated adaptive steganographic algorithms and on improving robustness across different image domains.

Keywords: stegoanalysis, deep learning, EfficientNet, MobileNet, HPF, convolutional neural networks, residual noise.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Fridrich, J. (2010). *Steganography in digital media: Principles, algorithms, and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9781139192903
2. Fridrich, J., & Kodovský, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, *7*(3), 868–882. https://doi.org/10.1109/TIFS.2012.2190402
3. Ma, Y., Wang, J., Zhang, X., Wang, G., Xin, X., & Zhang, Q. (2025). LGS-Net: A lightweight convolutional neural network based on global feature capture for spatial image steganalysis. *IET Image Processing*, *19*. https://doi.org/10.1049/ipr2.70005

4. Karamanji, A., Ahmed, A., & Fadhil, A. (2024). Comparative deep learning models in applications of steganography detection. *Journal of Image and Graphics, 12*(3), 312–319. https://doi.org/10.18178/joig.12.3.312-319
5. Zhang, R., Zhu, F., Liu, J., & Liu, G. (2018). Efficient feature learning and multi-size image steganalysis based on CNN. *arXiv Preprint*. https://arxiv.org/abs/1807.11428
6. Agarwal, S., & Jung, K.-H. (2022). Identification of content-adaptive image steganography using convolutional neural network guided by high-pass kernel. *Applied Sciences, 12*(22), Article 11869. https://doi.org/10.3390/app122211869
7. Deng, X.-Q., Chen, B.-L., Luo, W.-Q., et al. (2022). Universal image steganalysis based on convolutional neural network with global covariance pooling. *Journal of Computer Science and Technology, 37*, 1134–1145. https://doi.org/10.1007/s11390-021-0572-0
8. Yatsura, P., & Progonov, D. (2025). A review of modern methods for steganalysis and localization of embedded data in digital images. *Theoretical and Applied Cybersecurity, 7*. https://doi.org/10.20535/tacs.2664-29132025.1.328265
9. Hao, L., Yi, Z., Jinwei, W., Weiming, Z., & Xiangyang, L. (2024). Lightweight steganography detection method based on multiple residual structures and transformer. *Chinese Journal of Electronics, 33*(4), 965–978. https://doi.org/10.23919/CJE.2022.00.452
10. Mazurczyk, W., & Caviglione, L. (2014). Steganography in modern smartphones and mitigation techniques. *arXiv Preprint*. https://arxiv.org/abs/1410.6796
11. Chen, N., & Chen, B. (2022). Defending against OS-level malware in mobile devices via real-time malware detection and storage restoration. *Journal of Cybersecurity and Privacy, 2*. https://doi.org/10.3390/jcp2020017
12. Boroumand, M., Chen, M., & Fridrich, J. (2019). Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security, 14*(5), 1181–1193. https://doi.org/10.1109/TIFS.2018.2871749
13. Ye, J., Ni, J., & Yi, Y. (2017). Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security, 12*(11), 2545–2557. https://doi.org/10.1109/TIFS.2017.2710946
14. Hu, F., Xu, R., & Cheng, Z. (2021). A fast and effective image steganalysis model based on convolutional neural network. *Journal of Physics: Conference Series, 1861*, Article 012074. https://doi.org/10.1088/1742-6596/1861/1/012074
15. Agarwal, S., Kim, H., & Jung, K.-H. (2023). High-pass-kernel-driven content-adaptive image steganalysis using deep learning. *Mathematics, 11*(20), Article 4322. https://doi.org/10.3390/math11204322
16. Qian, Y., Dong, J., Wang, W., & Tan, T. (2015). Deep learning for steganalysis via convolutional neural networks. In *Media watermarking, security, and forensics 2015* (Vol. 9409, Article 94090J). SPIE. https://doi.org/10.1117/12.2083479
17. Holub, V., & Fridrich, J. (2012). Designing steganographic distortion using directional filters. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*. https://doi.org/10.1109/WIFS.2012.6412655
18. Hussain, I., Zeng, J., & Tan, S. (2020). A survey on deep convolutional neural networks for image steganography and steganalysis. *KSII Transactions on Internet and Information Systems, 14*, 1228–1248. https://doi.org/10.3837/tiis.2020.03.017
19. Wu, L., Han, X., Wen, C., & Li, B. (2020). A steganalysis framework based on CNN using the filter subset selection method. *Multimedia Tools and Applications, 79*. https://doi.org/10.1007/s11042-020-08831-8
20. Krizhevsky, A. (2012). *Learning multiple layers of features from tiny images* [Technical report]. University of Toronto.
21. DiSalvo, N. (2025). Steganographic embeddings as an effective data augmentation. *arXiv Preprint*. https://arxiv.org/abs/2502.15245
22. Stefanek, G., Gulbransen, L., Spink, G., Morawski, J., Filla, D., & Rabello De Castro, R. (2024). A comparison of AI models to detect hidden messages in images. *Issues in Information Systems*, 119–132. https://doi.org/10.48009/3_iis_2024_110
23. K, V., Annem, P., Devarakonda, M., Jyothi, A., & Rayudu, N. (2025). Image steganography with CNN-based encoder-decoder. *International Journal for Modern Trends in Science and Technology, 11*(4), 60–64. https://doi.org/10.5281/zenodo.15108976
24. Fridrich, J., Goljan, M., & Du, R. (2001). Detecting LSB steganography in color and gray-scale images. *IEEE Multimedia, 8*, 22–28. https://doi.org/10.1109/93.959097

25. Kombrink, M. H., Geradts, Z. J. M. H., & Worring, M. (2025). Image steganography approaches and their detection strategies: A survey. *ACM Computing Surveys, 57*(2), Article 33. https://doi.org/10.1145/3694965
26. Bauravindah, A., & Fudholi, D. (2024). Lightweight models for real-time steganalysis: A comparison of MobileNet, ShuffleNet, and EfficientNet. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 8*, 737–747. https://doi.org/10.29207/resti.v8i6.6091
27. Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4510–4520).
28. Zhou, A., Ma, Y., Ji, W., Zong, M., Yang, P., Wu, M., & Liu, M. (2022). Multi-head attention-based two-stream EfficientNet for action recognition. *Multimedia Systems, 29*, 1–12. https://doi.org/10.1007/s00530-022-00961-3
29. Farooq, N., & Mir, R. (2024). Image steganalysis using deep convolution neural networks: A literature survey. *International Journal of Sensors, Wireless Communications and Control, 14*(4), 247–264. https://doi.org/10.2174/0122103279296370240529075507
30. Chaumont, M. (2020). Deep learning in steganography and steganalysis. In *Digital media steganography*. https://doi.org/10.1016/B978-0-12-819438-6.00022-0
31. Wu, S., Zhong, S.-H., & Liu, Y. (2020). A novel convolutional neural network for image steganalysis with shared normalization. *IEEE Transactions on Multimedia, 22*(1), 256–270. https://doi.org/10.1109/TMM.2019.2920605
32. Dwaik, A., & Belkhouche, Y. (2024). Enhancing the performance of convolutional neural network image-based steganalysis in spatial domain using spatial rich model and 2D Gabor filters. *Journal of Information Security and Applications, 85*, Article 103864. https://doi.org/10.1016/j.jisa.2024.103864

Отримано редакцією журналу / Received: 17.01.26
Прорецензовано / Revised: 30.01.26
Схвалено до друку / Accepted: 26.03.26