УДК 004.056.55:004.8.032.26:004.891 **DOI:** 10.31673/2412-9070.2026.027616

Ю. В. МІШКУР[1], аспірант;
ORCID: 0009-0004-6807-6914
О. С. ЗАХАРЧЕНКО[2], ст. викл.,
ORCID 0000-0002-1604-0416
[1]Державний університет інформаційно-комунікаційних технологій, Київ
[2]Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ГІБРИДНИЙ ПІДХІД ДО СТЕГОАНАЛІЗУ НА ОСНОВІ МУЛЬТИМОДАЛЬНИХ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ТА ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ

У статті розглянуто гібридний підхід до стегоаналізу цифрових зображень, що поєднує можливості спеціалізованих згорткових нейронних мереж (ЗНМ) для низькорівневого виявлення статистичних аномалій із семантичним аналізом мультимодальних великих мовних моделей (МВММ). Запропонований підхід спрямований на подолання трьох фундаментальних обмежень існуючих монолітних ЗНМ-детекторів: низької здатності до узагальнення на невідомі стеганографічні алгоритми, відсутності контекстного аналізу мультимодальних метаданих та непрозорості процесів прийняття рішень. Архітектура гібридної системи реалізована у середовищі TensorFlow/Keras із використанням трьох ЗНМ-архітектур - MobileNetV2, ResNet50 та EfficientNetB0 – модифікованих спеціалізованими вхідними шарами фільтрації на основі ядра Лапласа та банку фільтрів SRM для виділення стеганографічно значущих залишкових сигналів. Інтеграція із мовними моделями реалізована через локальне розгортання Ollama у середовищі Google Colab із використанням моделей Gemma 3:4b, Gemma 3:12b та Llama 3.2 Vision 11B. Остаточне рішення формується через механізм «м'якого» злиття (Decision Fusion) зважених виходів ЗНМ- та ВММ-компонент, де вагові коефіцієнти динамічно коригуються залежно від виявленого семантичного контексту зображення.

Експериментальна перевірка виконана на синтезованому наборі даних на базі CIFAR-10 (LSB-вбудовування) та на еталонному наборі ALASKA2. Найвищу точність виявлення забезпечила конфігурація ResNet50 + Gemma 3:12b: 95.8% на CIFAR-10 і 91.7% на ALASKA2. Отримані результати свідчать про перспективність гібридного підходу для підвищення точності, узагальнюваності та інтерпретованості систем стегоаналізу.

Ключові слова: стегоаналіз, стеганографія, згорткові нейронні мережі, великі мовні моделі, MobileNetV2, ResNet50, EfficientNet, Gemma3, Llama, Ollama, гібридна архітектура.

Вступ

Стеганографія - мистецтво прихованої передачі інформації шляхом вбудовування секретних даних у цифрові медіафайли - набула надзвичайної актуальності в епоху стрімкого розвитку інформаційно-комунікаційних технологій. Здатність приховувати факт самої передачі даних, на відміну від криптографії, яка лише захищає їх зміст, робить стеганографію потужним інструментом як для легітимного захисту авторських прав і конфіденційних комунікацій, так і для кіберзлочинної діяльності, зокрема поширення шкідливого програмного забезпечення, обходу систем фільтрації контенту та організації прихованих каналів зв'язку між зловмисниками [1]. Останніми роками дедалі частіше фіксується використання стеганографічних технік для маскування загроз і проведення складних операцій, що зумовлює потребу в частковому переосмисленні концепції шкідливого програмного забезпечення. За даними [2] стеганографічні методи активно використовуються в складних цільових атаках (APT) та операціях кіберрозвідки, що зумовлює нагальну потребу у розробці ефективних засобів стегоаналізу.

Стегоаналіз - наукова дисципліна, спрямована на виявлення прихованої інформації у цифрових носіях - пройшов значний еволюційний шлях від класичних статистичних методів до сучасних підходів на основі глибокого навчання. Традиційні алгоритми, засновані на аналізі гістограм, моментів вищих порядків та текстурних характеристик зображень, демонструють суттєві обмеження при виявленні адаптивних стеганографічних методів, зокрема WOW (Wavelet Obtained Weights), HUGO (Highly Undetectable steGO) та S-UNIWARD, які мінімізують статистично помітні збурення в областях з багатою текстурою [3]. Ці методи здатні вбудовувати дані з надзвичайно малою ймовірністю виявлення, що є серйозним викликом для існуючих систем стегоаналізу.

Революційний прогрес у галузі згорткових нейронних мереж (ЗНМ) відкрив нові перспективи для стегоаналізу зображень. Архітектури на зразок SRNet, Ye-Net та XuNet продемонстрували вражаючі результати у виявленні цифрової стеганографії завдяки здатності до автоматичного вилучення ієрархічних ознак з просторового домену зображень [4]. Ключовою особливістю цих мереж є застосування спеціалізованих шарів попередньої обробки з фіксованими фільтрами виявлення різниць суміжних пікселів, що дозволяє підсилювати слабкі стеганографічні сигнали перед подачею їх до класифікаційних шарів. Незважаючи на суттєві досягнення, ЗНМ-підходи мають певні обмеження: вони потребують великих обсягів розмічених навчальних даних, схильні до перенавчання на конкретних стеганографічних алгоритмах і не здатні ефективно аналізувати мультимодальний контекст - аудіо, текстові метадані та послідовності зображень одночасно.

Поява великих мовних моделей (ВММ) із розширеними мультимодальними можливостями - таких як GPT-4V, LLaVA та Gemini - відкриває принципово нові можливості для стегоаналізу. Ці моделі, навчені на колосальних обсягах різнорідних даних, здатні розуміти семантичний контекст зображень, ідентифікувати аномалії у відповідності між візуальним вмістом та метаданими, а також виконувати міркування вищого порядку щодо природи медіафайлів [5]. Інтеграція мультимодальних ВММ із спеціалізованими ЗНМ-детекторами стеганографії утворює синергетичну систему, де нейронна мережа забезпечує точний попіксельний аналіз статистичних аномалій, а мовна модель - семантичне осмислення контексту та пояснення прийнятих рішень. Такий гібридний підхід є перспективним напрямком подолання обмежень монолітних архітектур.

Незважаючи на значні досягнення у кожній із зазначених галузей, комплексних досліджень, присвячених інтеграції мультимодальних ВММ та ЗНМ для потреб стегоаналізу, залишається вкрай мало. Більшість існуючих робіт розглядають ці підходи ізольовано, не використовуючи їх взаємодоповнюючий потенціал. Крім того, відкритими залишаються питання узагальнюваності запропонованих рішень на невідомі стеганографічні алгоритми, ефективності роботи у реальних умовах з різноякісними носіями, а також інтерпретованості рішень, що є критично важливим для застосування у правоохоронній та криміналістичній практиці [6].

Метою даної статті є розробка та дослідження гібридної архітектури стегоаналізу, що поєднує переваги спеціалізованих згорткових нейронних мереж для низькорівневого виявлення стеганографічних аномалій із можливостями мультимодальних великих мовних моделей для семантичного аналізу та контекстного осмислення медіаконтенту. Запропонований підхід спрямований на підвищення точності виявлення прихованої інформації у зображеннях, поліпшення узагальнюваності на нові стеганографічні схеми та забезпечення інтерпретованості результатів аналізу для практичного застосування у задачах кібербезпеки та цифрової криміналістики.

Постановка проблеми

Сучасні адаптивні стеганографічні алгоритми, зокрема S-UNIWARD (Spatial Universal Wavelet Relative Distortion), WOW (Wavelet Obtained Weights) та HILL (High-pass, Low-pass, Low-pass), мінімізують статистично помітні збурення шляхом концентрації вбудовуваних змін у текстурно-складних областях зображень. Це призводить до того, що відношення сигнал/шум для стеганографічних артефактів стає надзвичайно малим - типова відносна ємність сучасних

схем вбудовування становить 0.1–0.4 біт/піксель при практично нульовій візуальній помітності [7].

Незважаючи на значний прогрес архітектур на основі глибокого навчання - SRNet, YeNet, XuNet та їх модифікацій - вони мають ряд принципових обмежень, що перешкоджають їх широкому практичному застосуванню. По-перше, ці архітектури навчаються та оптимізуються для конкретного стеганографічного алгоритму або вузького класу алгоритмів. При застосуванні до невідомого методу вбудовування точність виявлення суттєво знижується - до рівня, що порівнянний з випадковим вгадуванням. Це явище, відоме як проблема узагальнюваності (generalization problem), є одним із ключових нерозв'язаних питань сучасного стегоаналізу [4].

По-друге, існуючі архітектури згорткових нейронних мереж оперують виключно у просторовому домені пікселів окремого зображення, не враховуючи ширшого контексту: супровідних метаданих файлу, послідовності графічних об'єктів, текстових анотацій або інших мультимодальних сигналів, що можуть свідчити про стеганографічну активність. У реальних сценаріях кіберрозвідки зловмисники нерідко застосовують специфічні патерни передачі даних, характерні службові параметри або поведінкові особливості, які виходять за межі аналізу одиничного файлу. Ігнорування цього контексту призводить до системного недовикористання доступної інформації.

По-третє, сучасні архітектури згорткових нейронних мереж є закритими системами, які видають бінарне рішення без надання жодних пояснень чи обґрунтувань. Це є критичним недоліком для застосування у цифровій криміналістиці та правоохоронній практиці, де висновок детектора має бути зрозумілим, відтворюваним і юридично обґрунтованим. Відсутність механізмів прозорості та логічного підтвердження результатів суттєво обмежує практичну цінність наявних систем.

Мультимодальні великі мовні моделі (ВММ) класу GPT-4V, LLaVA та Gemini демонструють вражаючі можливості у розумінні візуального контенту, міркуванні вищого порядку та інтеграції різнорідних джерел інформації. Завдяки попередньому навчанню на колосальних обсягах різнорідних даних ці моделі набувають широкого семантичного розуміння зображень, здатності виявляти артефакти обробки і аномалії між візуальним вмістом та контекстуальними метаданими [5]. Це відкриває принципово нову можливість для стегоаналізу: замість низькорівневого аналізу статистичних відхилень у розподілі пікселів ВММ можуть здійснювати семантичне осмислення зображення у контексті всіх доступних метаданих.

Водночас ВММ мають суттєві обмеження для задачі стеганоаналізу. Їхня просторова роздільна здатність при обробці зображень значно нижча порівняно зі спеціалізованими ЗНМ (згортковими нейронними мережами), а тонкі статистичні аномалії на рівні окремих пікселів, характерні для адаптивної стеганографії, практично недоступні для сприйняття поточними архітектурами ВММ.

Крім того, значна обчислювальна вартість процесу виконання та отримання результатів у ВММ робить їх непридатними для самостійного застосування у системах реального часу. Нарешті, наразі відсутні методології ефективного поєднання низькорівневих сигналів від ЗНМ із семантичними висновками ВММ в єдину узгоджену систему прийняття рішень. Таким чином, на основі аналізу сучасного стану галузі можна виокремити три фундаментальних протиріччя, які обумовлюють наукову проблему даного дослідження.

Перше протиріччя полягає у невідповідності між високою складністю адаптивних стеганографічних схем, що мінімізують статистичну помітність, та обмеженою здатністю монолітних архітектур виявляти контент, прихований за алгоритми, які виходять за межі їхньої навчальної вибірки.

Друге протиріччя обумовлене наявністю потужних мультимодальних ВММ, здатних до семантичного аналізу, за відсутності обґрунтованих методів їхньої інтеграції із спеціалізованими ЗНМ-детекторами для задач стегоаналізу.

Третє протиріччя виникає між потребою у прозорих та інтерпретованих рішеннях для цифрової криміналістики та закритою природою існуючих нейромережевих моделей, що не надають обґрунтувань для прийнятих рішень.

Аналіз останніх досліджень і публікацій

Сучасний стегоаналіз пройшов довгий шлях від класичних статистичних методів до архітектур глибокого навчання. Визначальним етапом стала розробка моделі просторово-насиченого стеганоаналізу [1], яка запропонувала розгалужений набір ознак, отриманих за допомогою лінійних фільтрів залишків сусідніх пікселів. Це на тривалий час закріпило парадигму ручного проєктування характерних рис зображення. Проте етап самостійного конструювання ознак суттєво обмежував адаптивність до нових стеганографічних схем і потребував значних зусиль при переході між різними алгоритмами приховування даних.

Революційний внесок у стегоаналіз глибоким навчанням було здійснено у роботі [8], в який представлено Ye-Net - архітектуру з попередньо визначеними фільтрами виявлення залишкових сигналів у першому шарі та активаційними функціями укорочений лінійний блок (TLU - Truncated Linear Unit). Цей підхід дозволив суттєво підвищити чутливість до слабких стеганографічних змін і встановив новий орієнтир точності на наборах даних BOSSBase та BOWS2 для алгоритмів WOW та S-UNIWARD.

Водночас було представлено архітектуру XuNet, яка запропонувала спрощену, проте ефективну п'ятишарову структуру із вбудованим модулем попередньої обробки на основі фіксованого фільтра високих частот. Таке рішення забезпечило конкурентоспроможні результати у виявленні прихованих даних за рахунок акцентування на високочастотних складових зображення [9].

Важливим кроком у розвитку ЗНМ-стегоаналізу стало створення архітектури SRNet [4], яка вперше об'єднала два паралельні потоки обробки: перший виконує низькорівневе вилучення стеганографічних артефактів через залишкові фільтри, другий - традиційне ієрархічне кодування семантики зображення. Механізм пізнього злиття обох потоків забезпечив SRNet найвищу точність серед тогочасних рішень і заклав основу для подальших досліджень у сфері багатопотокової обробки. Але водночас SRNet потребує значних обчислювальних ресурсів та великих навчальних наборів, що обмежує її практичне застосування.

Серйозним викликом для ЗНМ-підходів є проблема узагальнюваності: моделі, навчені на конкретному стеганографічному алгоритмі, суттєво деградують при тестуванні на невідомих схемах. Змагання ALASKA2 Challenge [10] стимулювало розробку реалістичніших протоколів оцінювання, що враховують різноманіття умов зйомки, форматів стиснення та типів камер. Провідні рішення змагання ALASKA2 базувалися на використанні архітектури EfficientNet як основної моделі, яку поєднували з механізмами вибіркової уваги та методами об'єднання прогнозів декількох мереж для досягнення найвищих показників точності у виявленні прихованих даних [10].

На думку авторів [11], архітектура MobileNetV2 часто використовувалася в ансамблях для забезпечення різноманітності ознак та прискоренні попереднього скринінгу зображень, хоча важчі моделі (EfficientNet) показали кращу точність. Архітектура MobileNetV2 [11] базується на концепції інвертованих залишкових блоків (inverted residuals) з лінійними bottleneck-шарами та depth-wise separable convolutions. Ключова перевага MobileNetV2 - радикальне скорочення обчислювальної складності (~300M FLOPS для вхідного зображення 224x224) при збереженні прийнятної точності класифікації. Параметрична ефективність моделі (~3.4M параметрів) робить її привабливою для розгортання в ресурсообмежених середовищах, зокрема в умовах Google Colab з обмеженим часом GPU-сесії. Реалізація через tf.keras.applications.MobileNetV2 надає зручний доступ до ImageNet-попередньо навчених ваг.

Архітектура EfficientNet [12] запропонувала принципово новий підхід до масштабування ЗНМ через одночасне пропорційне збільшення глибини, ширини та роздільної здатності вхідних зображень (compound scaling coefficient). Сімейство моделей EfficientNet-B0 через B7 охоплює широкий спектр співвідношення точність/ефективність. У задачах виявлення маніпуляцій із зображеннями EfficientNet демонструє значно кращу точність порівняно з MobileNetV2 при помірному збільшенні обчислювальних витрат. Дослідження [13] підтвер-дили, що архітектури EfficientNet-B4/B5 досягають конкурентних результатів у змаганні ALASKA2 при відповідному підборі параметрів навчання. У тій самій роботі наведено й досвід використання

модифікованих мереж ResNet-50 для виявлення вбудовувань у колірних каналах JPEG-зображень.

Глибока архітектура ResNet-50 [14] залишається еталонною базовою моделлю в задачах комп'ютерного зору завдяки механізму skip connections, що вирішує проблему затухання градієнтів у глибоких мережах. Численні дослідження в галузі цифрової криміналістики демонструють, що ResNet-50 з відповідно модифікованим першим шаром (заміна стандартного згорткового шару на high-pass фільтри виявлення залишкових сигналів) є ефективним інструментом стегоаналізу. Реалізація у TensorFlow/Keras через tf.keras.applications.ResNet50 надає зручний доступ до попередньо навчених ваг з можливістю гнучкого fine-tuning для специфічних задач детекції стеганографії. Механізм Grad-CAM для ResNet-50 забезпечує генерацію просторових карт активацій, критично важливих для інтеграції з ВММ-компонентою гібридної системи.

Порівняльний аналіз трьох архітектур у контексті задачі стегоаналізу виявляє наступну закономірність: MobileNetV2 є оптимальним для обмеженого обчислювального бюджету та швидкого прототипування; EfficientNet-B4 забезпечує найкращий компроміс між точністю та обчислювальними витратами в умовах Google Colab; ResNet-50 є переважним вибором при необхідності інтерпретованості через Grad-CAM та широкому доступі до бібліотечних реалізацій пояснювального ШІ.

Архітектура LLaVA (Large Language and Vision Assistant) [15], реалізує ефективний підхід до побудови мультимодальних систем: візуальний енкодер CLIP ViT-L/14 генерує векторні представлення зображень, які через навчений проекційний шар перетворюються на токени, сумісні з простором текстового декодера на базі Vicuna/LLaMA. Версія LLaVA 1.6 (LLaVA-NeXT) суттєво покращила попередника завдяки збільшенню роздільної здатності вхідних зображень до 672x672 через механізм Dynamic High Resolution, кращим навчальним даним та вдосконаленому instruction tuning. Модель LLaVA-1.6-Mistral-7B (~4.1 ГБ у форматі Q4_K_M) доступна для локального розгортання через Ollama, що забезпечує конфіденційність аналізованих даних у прикладних дослідженнях.

Сімейство моделей Llama 3 від Meta AI [16] є одним з найпотужніших відкритих мовних моделей станом на 2024 рік. Llama 3.1 включає варіанти з 8B, 70B та 405B параметрів, навчені на корпусі понад 15 трлн. токенів. Для мультимодальних задач особливо релевантна модель Llama 3.2-Vision (11B), що підтримує безпосередній аналіз зображень через вбудований візуальний енкодер. Ключова перевага Llama 3 для наукових досліджень - відкрита ліцензія Meta Llama 3, що дозволяє комерційне використання та модифікацію моделі. Розгортання через Ollama забезпечує стандартизований REST API, сумісний з OpenAI Chat Completions, що спрощує інтеграцію з наявним кодом Python/TensorFlow. Llama-3.2-Vision-11B у форматі Q4_K_M займає близько 6.5 ГБ VRAM, що вкладається в обмеження T4 GPU в Google Colab Pro.

Серія моделей Gemma 3 від Google DeepMind [17] представляє сімейство відкритих моделей, оптимізованих для ефективного розгортання. Gemma 3 використовує архітектуру decoder-only transformer з груповим multi-head attention (GQA), RoPE positional embeddings та розширеним словником токенів. Моделі Gemma 3 доступні у варіантах 1B, 4B, 12B та 27B параметрів; версія 4B-IT (instruction-tuned) демонструє оптимальне співвідношення якості та обчислювальних вимог для середовища Google Colab T4. Gemma 3 4B-IT у форматі Q4_K_M займає близько 3.3 ГБ VRAM, що дозволяє одночасно завантажувати декілька моделей для порівняльних досліджень. Мультимодальні можливості Gemma 3 реалізовані через інтеграцію з SigLIP-подібним візуальним енкодером, оптимізованим для аналізу зображень.

Платформа Ollama [18] є відкритою системою для локального розгортання та керування великими мовними моделями. Ключова архітектурна особливість Ollama - підтримка квантизованих моделей у форматі GGUF (Q4_K_M, Q5_K_M, Q8_0), що дозволяє запускати моделі з мільярдами параметрів на обладнанні з обмеженою VRAM, зокрема на NVIDIA T4 (16 ГБ) в Google Colab. Встановлення Ollama в Google Colab виконується командою curl -fsSL https://ollama.ai/install.sh | sh та подальшим запуском сервера як фонового процесу. API доступний через localhost:11434 та повністю сумісний з OpenAI REST API, що дозволяє замінити хмарний GPT-4V на локальний LLaVA або Llama 3.2-Vision без зміни коду клієнта.

Дослідження в галузі edge-розгортання ВММ, зокрема робота [19] щодо QLoRA-квантизації, демонструють, що 4-бітна квантизація зберігає понад 95% продуктивності повноточних моделей для більшості задач розуміння тексту та зображень. Для задачі стегоаналізу це означає практичну застосовність підходу: Llama-3.2-Vision-11B у форматі Q4_K_M (~6.5 ГБ) та Gemma-3-4B-IT (~3.3 ГБ) вкладаються в ліміт VRAM Google Colab T4, тоді як LLaVA-1.6-Mistral-7B (~4.1 ГБ) є ще компактнішим варіантом. Вибір формату квантизації Q4_K_M (замість Q4_0) обумовлений кращою збалансованістю точності та швидкості інференсу, що задокументовано в порівняльних дослідженнях спільноти llama.cpp.

Інтеграція Ollama з конвеєром TensorFlow/Keras реалізується через офіційний Python-клієнт ollama або стандартні HTTP-запити. Після класифікації ЗНМ-компонентою (MobileNetV2/ResNet-50/EfficientNet) результати аналізу - ймовірність стегоносія, карта активацій Grad-CAM у форматі base64, числові ознаки - передаються ВММ у складі структурованого промпту. Відповідь ВММ містить природно-мовне пояснення рішення, виявлені аномалії та оцінку достовірності. Такий підхід реалізує принцип Human-in-the-Loop для криміналістичних застосувань, де інтерпретованість рішення є юридичною вимогою.

Концепція гібридних систем стегоаналізу, що поєднують різнорідні аналітичні компоненти, набуває дедалі більшого поширення. Дослідження [20] запропонувало статистично обгрунтований критерій для поєднання класифікаційних рішень від кількох незалежних детекторів. Більш сучасні ансамблеві підходи об'єднують декілька ЗНМ-детекторів через механізм зваженого голосування для підвищення узагальнюваності. Однак всі ці підходи залишаються в межах одного модального простору (пікселі зображення) та не використовують семантичні знання ВММ.

Застосування ВММ для задач цифрової безпеки та криміналістики зображень - відносно новий напрям. У роботі [21] та в дослідженнях в рамках проєкту SIMARGL (EU Horizon 2020) було вивчено ЗНМ-підходи для виявлення stegomalware у реальних умовах. Ці роботи підкреслюють важливість семантичного контексту при аналізі підозрілих медіафайлів. Зокрема, аналіз метаданих EXIF, історії модифікацій файлу та статистики розподілу пікселів у поєднанні з результатами ЗНМ-класифікації може суттєво підвищити впевненість детектора та зменшити кількість хибних позитивних результатів у реальних умовах застосування.

Ключова прогалина в існуючих дослідженнях полягає у відсутності науково обгрунтованої архітектури інтеграції ЗНМ та ВММ для задачі стегоаналізу. Поточні роботи або розглядають ЗНМ ізольовано (без семантичного контексту ВММ), або використовують ВММ лише для опису зображень загального характеру (без спеціалізованого стегоаналітичного контексту). Запропонований гібридний підхід, що використовує ЗНМ на базі MobileNetV2/ResNet-50/EfficientNet (TensorFlow/Keras) у поєднанні з ВММ Gemma 3/Llama 3 (Ollama, Google Colab), заповнює цю прогалину через формалізовану дворівневу архітектуру прийняття рішень з інтерпретованим виходом.

Проведений аналіз останніх досліджень і публікацій дозволяє сформулювати такі висновки:

По-перше, спеціалізовані ЗНМ-архітектури SRNet та Ye-Net досягли значних успіхів у виявленні адаптивної стеганографії, проте їх узагальнюваність на невідомі алгоритми залишається незадовільною. Ефективні архітектури загального призначення MobileNetV2, ResNet-50 та EfficientNet, реалізовані через TensorFlow/Keras, забезпечують практичний компроміс між точністю, обчислювальною вартістю та інтерпретованістю - три ключові характеристики для гібридної системи.

По-друге, мультимодальні ВММ Llama 3 або Gemma 3, доступні для локального розгортання через Ollama у середовищі Google Colab, відкривають нову можливість семантичного аналізу та генерації інтерпретованих пояснень. Квантизовані версії (Q4_K_M) цих моделей вкладаються в обмеження оперативної пам’яти GPU T4 Google Colab, що робить підхід відтворюваним без спеціалізованого обладнання. Інтеграція через стандартизований Ollama REST API забезпечує модульність та взаємозамінність ВММ-компоненти.

По-третє, відсутність систематичних досліджень інтеграції ЗНМ та ВММ для потреб стегоаналізу визначає наукову новизну запропонованого підходу. Гібридна архітектура, де ЗНМ виявляє слабкі статистичні аномалії на рівні пікселів, а ВММ інтерпретує ці результати в семантичному контексті метаданих та формує природно-мовне обгрунтування, відповідає реальним вимогам застосувань у кібербезпеці. Детальний опис реалізації гібридної архітектури наведено нижче.

Метою дослідження є розробка та верифікація гібридної архітектури стеганоаналізу, що поєднує згорткові нейронні мережі (ЗНМ) для виявлення низькорівневих аномалій із великими мультимодальними моделями (ВММ) для семантичного аналізу контексту. Це дозволить підвищити точність, універсальність та інтерпретованість рішень автоматизованих систем.

Основна частина

Гібридна архітектура в цьому дослідженні базується на поєднанні двох технологічних стеків: низькорівневої екстракції ознак за допомогою згорткових нейронних мереж та високорівневого семантичного арбітражу за допомогою великих мультимодальних моделей (ВММ).

Реалізація запропонованого гібридного конвеєра базувалася на чотирьох послідовних етапах обробки даних. Процес аналізу зображення починався з виділення шумової складової зображення: за допомогою фільтрів високих частот (HPF) усувається основний візуальний контент, що дозволяє зосередитися на мікроскопічних аномаліях пікселів. На наступному етапі здійснювалося автоматичне вилучення ознак за допомогою спеціалізованих нейронних архітектур, таких як MobileNetV2, ResNet50 або EfficientNet, які виконували роль аналітичних сенсорів. Отримані технічні показники були основою для формування комплексного мультимодального запиту, який було спрямовано до локальної великої мовної моделі. Фінальна стадія передбачає використання потужностей Gemma 3, Llama 3 або LLaVA для інтелектуального синтезу висновків, де статистичні дані об’єднувались з семантичним контекстом зображення для формування вичерпного експертного звіту.

Багатошаровий HPF-фільтр (High-Pass Filter) будувався з банку фільтрів, який містив 12 базових ядер розміром 5x5. Цей банк фільтрів було застосовано до кожного з трьох RGB-каналів незалежно. Він містив направлені фільтри SRM (8 ядер - горизонтальні, вертикальні, діагональні та антидіагональні). Направлені ядра були призначені для виявлення розривів у статистичних зв'язках між сусідніми пікселями в різних напрямках. Додатково було використано 4 ядра з оператором Лапласа, але з різними наборами вагових коефіцієнтів для аналізу зон з різною інтенсивністю природного шуму.

Для того, щоб отримати 36 каналів на виході препроцесора, використовується глибинна згортка (Depthwise Convolution), що дозволило обраній ЗНМ-моделі (наприклад, MobileNetV2) аналізувати 36 різних «шумових портретів» одного й того самого зображення.

Крім того, замість функції активації Relu було використано власну функцію активації ABS (обчислення за модулем). Застосування операції обчислення модуля до результатів фільтрації високих частот дозволяє акумулювати повний обсяг амплітудних відхилень шуму, що виникають як у позитивному, так і в негативному діапазонах. Такий підхід забезпечує нейронну мережу вдвічі більшою кількістю інформативних ознак для подальшого аналізу. Оскільки стандартного об'єкта layers.ABS не існує, це реалізується через шар Lambda:

```
from tensorflow.keras import layers
x = layers.Lambda(lambda t: tf.abs(t), name="abs_activation")(x)
```

Такий шар було розміщено одразу після блоку HPF-фільтрів перед основними згортковими блоками мережі.

При побудові гібридного варіанта знаходження стеганографічних вкладень для ініціалізації ЗНМ-частини було використано ваги для набору даних ImageNet. Додаткове навчання мережі не проводилося.

Конволюційна нейронна мережа була використана для формування багатоканальних карт ознак, що містять статистичні девіації пікселів.

МВММ отримувала ознаки від ЗНМ-частини та поєднувала їх із текстовим запитом (промптом) і наявним контекстом задачі.

Вихідні тензори ЗНМ перетворювались у візуальні токени, зрозумілі для мовної моделі, за допомогою шару-адаптера. Основний блок МВММ обробляв послідовність токенів і формував логічний висновок про наявність/відсутність вбудовування. Крім того, мовна модель надавала текстові пояснення до цього висновку.

Для реалізації технічного аналізу в середовищі TensorFlow/Keras обрано три архітектури, що представляють різні підходи до масштабування:

- MobileNetV2: використовує інвертовані залишкові блоки (Inverted Residuals), що критично для мінімізації обчислювальної складності;
- ResNet50: базується на концепції залишкового навчання (Residual Learning), що дозволяє уникнути згасання градієнта при аналізі дрібних артефактів;
- EfficientNet (B0): застосовує метод складеного масштабування (Compound Scaling), забезпечуючи баланс між глибиною та роздільною здатністю.

Кожна модель модифікована через додавання шару фільтрації на вході, що реалізує ядро Лапласа для виділення високочастотних ознак, притаманних LSB-стеганографії.

Для забезпечення конфіденційності та можливості тонкого налаштування, розгортання мовних моделей здійснюється локально за допомогою локального серверу Ollama в середовищі Google Colab. Було використано такий стек моделей: Gemma 3 (4B/12B), Llama 3.2 Vision (11B). Було виконано порівняння поведінки версій Gemma, а саме здатність моделі 4B до швидкої класифікації порівняно з глибшою здатністю моделі 12B до логічного обґрунтування виявлених аномалій.

Для первинної верифікації використано набір даних CIFAR-10, що містить 60 000 зображень розміром 32x32 пікселі. На його основі сформовано пари оригінальних та стеганографічних об'єктів із впровадженням тексту методом заміни найменшого значущого біта. Це забезпечило можливість точного контролю щільності вбудовування та перевірки чутливості гібридної системи до мінімальних змін цифрового сигналу. Для оцінки здатності моделі до узагальнення в реальних сценаріях застосовано еталонний набір ALASKA2, який містить різноманітні методи приховування інформації як у просторовому домені, так і у форматі JPEG.

Остаточне рішення приймається на основі «м’якого» злиття:

$$D_{final} = w_1 \cdot P_{CNN} + w_2 \cdot P_{VMM}\ , \qquad (1)$$

де w_1, w_2 - вагові коефіцієнти, що динамічно коригуються ВММ залежно від виявленого семантичного контексту зображення (наприклад, складні текстури знижують вагу CNN через ризик False Positive).

Результати дослідження

На першому етапі тестування було оцінено здатність трьох обраних архітектур (MobileNetV2, ResNet50, EfficientNetB0) ідентифікувати LSB-вбудовування у зображеннях низької роздільної здатності.

Найбільш надійний результат було отримано з використанням архітектури ResNet50. Використання архітектур MobileNetV2 і EfficientNetB0 було ускладнено виникненням перенавчання при збільшенні кількості блоків фільтрації.

Самі по собі архітектури нейронних мереж загального призначення не здатні знайти ознаки присутності стеганографічного вкладення. Без додавання вхідних шарів, які забезпечують виділення ознак стеганографічного приховування, показник Accuracy незалежно від варіанту навчання моделі і обраної архітектури буз близьким до 0.5.

Експериментально підтверджено пряму залежність між кількістю застосованих ядер фільтрації та якістю вихідного сигналу. Використання банку фільтрів SRM із 12 ядрами забезпечило достатню чутливість до невеличких обсягів вбудованих даних (до 5% ємності контейнера), що дозволило ідентифікувати структурні розриви цілісності пікселів на однорідних ділянках зображення. Використання ядер Лапласа виявилося найбільш ефективним методом детекції прихованих даних у зонах із високою деталізацією текстур, зокрема на зображеннях трави або піску, де стандартні алгоритми обробки зазвичай демонструють високу похибку.

Проведений порівняльний аналіз архітектур на етапі обробки даних визначив ієрархію ефективності моделей залежно від пріоритетів дослідження. Найвищий рівень надійності та точності виявлення прихованої інформації забезпечив використання ResNet50, оскільки ця модель сформувала найбільш стабільні вектори ознак високої розмірності для класифікації складних алгоритмів стеганографії. Проміжний результат продемонстрував впровадження EfficientNetB0, яка виділила релевантні ознаки при збереженні балансу між обчислювальною складністю та якістю даних. Водночас застосування MobileNetV2 дозволило досягти найменших вимог до системних ресурсів і найвищої швидкості опрацювання, хоча це супроводжувалося зниженням загальної точності через недостатню роздільну здатність сформованих масивів ознак.

Аналіз фінального етапу генерації звітів виявив суттєві розбіжності в аналітичних можливостях використаних моделей. Застосування Gemma 3:12b забезпечило найвищий рівень точності завдяки здатності моделі диференціювати природні артефакти стиснення від навмисного втручання, що дозволило збільшити частку виявлених на заключному етапі випадків наявності прихованого тексту. Використання Gemma 3:4b дозволило досягти високої швидкості формування висновків, проте її звіти мали лаконічний характер і містили менше деталей у складних сценаріях із високим рівнем цифрового шуму. Водночас впровадження використання Llama 3 (Vision) продемонструвало достатню ефективність у формуванні структурованих аналітичних висновків, оскільки ця модель забезпечила чітке розмежування між обчисленими статистичними показниками та семантичними аспектами виявлених даних.

Порівняльна характеристика різних варіантів побудови гібридної системи виявлення стеганографії наведена у таблицях 1 та 2.

Найвищий показник точності в обох випадках зафіксовано у конфігурації ResNet50 + Gemma 3:12b. Порівняння з результатами для версії Gemma 3:4b свідчить, що збільшення кількості параметрів мовної моделі безпосередньо впливає на здатність системи до правильної класифікації складних випадків.

Але спостерігається суттєве зростання часу надання висновку при використанні повнорозмірних моделей. Використання Gemma 3:12b збільшує загальний час обробки приблизно у 2-3 рази порівняно з версією 4b.

Таблиця 1

Порівняльна характеристика варіантів архітектури гібридної системи виявлення стеганографії за результатами аналізу синтезованого набору даних на базі CIFAR10

Конфігурація системи (Модель аналізу + Модель інтерпретації)	Точність виявлення (Accuracy)	Час обробки одного зображення (мс)
MobileNetV2 + Gemma 3:4b	78.4%	65
MobileNetV2 + Gemma 3:12b	83.9%	225
ResNet50 + Gemma 3:4b	88.9%	125
ResNet50 + Gemma 3:12b	95.8%	255
EfficientNetB0 + Gemma 3:4b	88.5%	115
EfficientNetB0 + Gemma 3:12b	91.2%	245

Таблиця 2

Порівняльна характеристика варіантів архітектури гібридної системи виявлення стеганографії за результатами аналізу набору даних Alaska2

Конфігурація системи (Модель аналізу + Модель інтерпретації)	Точність виявлення (Accuracy)	Час обробки одного зображення (мс)
MobileNetV2 + Gemma 3:4b	72.1%	85
MobileNetV2 + Gemma 3:12b	78.5%	245
ResNet50 + Gemma 3:4b	84.1%	155
ResNet50 + Gemma 3:12b	91.7%	285
EfficientNetB0 + Gemma 3:4b	80.3%	140
EfficientNetB0 + Gemma 3:12b	86.4%	275

Найшвидша комбінація (MobileNetV2 + Gemma 3:4b) демонструє найменший час відгуку, що дозволяє використовувати її у потокових системах моніторингу, попри нижчу точність. Час обробки для ResNet50 є найбільшим, що пояснюється складністю формування глибинних векторів ознак для зображень високої роздільної здатності.

Наведені дані підтверджують, що архітектура ResNet50 + Gemma 3:12b є найбільш стабільною для еталонних наборів типу ALASKA2. Її здатність до формування стабільних векторів ознак забезпечує надійну роботу з різними типами стиснення JPEG, які є характерними для цього набору даних.

Висновки

За результатами проведеного дослідження та аналізу варіантів архітектури системи виявлення стеганографії встановлено, що спільне використання згорткових мереж для первинної обробки сигналу та мультимодальних великих мовних моделей для фінальної інтерпретації забезпечує вищу достовірність результатів порівняно з ізольованим застосуванням кожного з підходів. Нейронні мережі ефективно фіксують статистичні відхилення у структурі сукупності пікселів зображення, тоді як мовні моделі успішно відсікають помилкові спрацювання, спричинені природними шумами або артефактами стиснення.

Експериментально доведено, що серед досліджених архітектур модель ResNet50 забезпечила найвищу точність і стабільність завдяки використанню залишкових зв'язків, що дозволяє формувати стійкі вектори ознак для класифікації складних алгоритмів приховування даних. Модель EfficientNetB0 посідає проміжну позицію, демонструючи прийнятний компроміс між точністю виявлення та обчислювальними витратами. Архітектура MobileNetV2 визначена як оптимальна для потокових систем моніторингу з обмеженими ресурсами через мінімальні вимоги до обсягу відеопам'яті та найвищу швидкість опрацювання інформації.

Експериментально виявлено вплив потужності ВММ на якість семантичного аналізу. Порівняння конфігурацій із Gemma 3:4b та Gemma 3:12b підтвердило пряму залежність між кількістю параметрів мовної моделі та здатністю системи диференціювати природні артефакти стиснення від навмисних стеганографічних вкладень. Gemma 3:12b формує розгорнуті аналітичні висновки і краще справляється із складними сценаріями при високому рівні цифрового шуму, натомість Gemma 3:4b забезпечує значно вищу швидкість інференсу при лаконічніших поясненнях. Llama 3.2 Vision демонструє ефективне розмежування між обчисленими статистичними показниками та семантичними аспектами виявлених аномалій.

Виявлено оптимальні конфігурації гібридної системи стеганографії за різними критеріями. Конфігурація ResNet50 + Gemma 3:12b досягає найвищої точності (95.8% на CIFAR-10; 91.7% на ALASKA2) і рекомендується для задач, де пріоритетом є максимальна надійність виявлення. Конфігурація MobileNetV2 + Gemma 3:4b є оптимальною для потокового моніторингу завдяки найменшому часу відгуку (65 мс/85 мс відповідно), незважаючи на нижчу точність (78.4% / 72.1%). Збільшення розміру ВММ збільшує загальний час обробки у 2–3 рази.

Подальші дослідження доцільно спрямувати на: розширення тестування на невідомих стеганографічних алгоритмах для верифікації узагальнюваності; вдосконалення механізму динамічного злиття рішень за рахунок навченого мета-класифікатора; дослідження можливостей тонкого налаштування ВММ на стегоаналітичних наборах даних через підходи LoRA/QLoRA; оцінку практичної ефективності системи у реальних умовах кіберрозвідки з урахуванням різноманітності форматів медіаконтенту та каналів передачі.

Внесок авторів

Юрій МІШКУР – концептуалізація, розробка методології дослідження, програмна реалізація гібридної моделі, проведення обчислювальних експериментів, збір і обробка даних, ана-

ліз результатів, підготовка первинної версії рукопису; Оксана ЗАХАРЧЕНКО – наукове консультування, формулювання наукової проблеми, методологічна верифікація, інтерпретація результатів, критичне рецензування змісту, наукове редагування тексту.

Декларація про використання штучного інтелекту

Під час підготовки статті інструменти штучного інтелекту не використовувалися ні для отримання наукових результатів, ні для аналізу даних, написання основного змісту статті чи формулювання висновків.

Конфлікт інтересів

Автори заявляють про відсутність конфлікту інтересів. Під час підготовки цієї роботи не існувало жодних комерційних, фінансових чи інших взаємовідносин, які могли б вплинути на результати дослідження або їх інтерпретацію. Робота виконана з дотриманням принципів ака-демічної доброчесності, етичних норм проведення наукових досліджень і вимог редакційної політики щодо запобігання конфлікту інтересів.

Список використаної літератури

1. 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

2. Liguori, A., Zuppelli, M., Gallo, D., Guarascio, M., & Caviglione, L. (2025). A deep learning-based approach for stegomalware sanitization in digital images. Journal of Intelligent Information Systems, 64, 121–144. https://doi.org/10.1007/s10844-025-00936-6

3. Holub, V., Fridrich, J., & Denemark, T. (2014). Universal distortion function for steganography in an arbitrary domain. EURASIP Journal on Information Security, 2014(1), 1–13. https://doi.org/10.1186/1687-417X-2014-1

4. 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

5. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2304.08485

6. Yousfi, Y., Butora, J., Khvostikov, A., & Fridrich, J. (2021). Breaking ALASKA: Color separation is not enough. Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 97–108. https://doi.org/10.1145/3437880.3460395

7. Holub, V., & Fridrich, J. (2013). Digital image steganography using universal distortion. Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. https://doi.org/10.1145/2482513.2482514

8. 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

9. Xu, G., Wu, H. Z., & Shi, Y. Q. (2016). Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 23(5), 708–712. https://doi.org/10.1109/LSP.2016.2548421

10. Cogranne, R., Giboulot, Q., & Bas, P. (2020). The ALASKA#2 image steganalysis challenge: A first step towards steganalysis into the wild. Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 1–11. https://doi.org/10.1145/3369412.3395075

11. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

12. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning (PMLR 97), 6105–6114. https://doi.org/10.48550/arXiv.1905.11946

13. Wei, K., Luo, W., Tan, S., & Huang, J. (2022). Universal deep network for steganalysis of color images based on channel representation. IEEE Transactions on Information Forensics and Security, 17, 3022–3036. https://doi.org/10.48550/arXiv.2111.12231

14. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 770–778. https://doi.org/10.1109/CVPR.2016.90

15. Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.2310.03744

16. Meta AI. (2024). The Llama 3 herd of models. arXiv preprint arXiv:2407.21783. https://doi.org/10.48550/arXiv.2407.21783

17. Gemma Team, Google DeepMind. (2024). Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295. https://doi.org/10.48550/arXiv.2403.08295

18. Ollama. (2024). Ollama: Get up and running with large language models locally. GitHub. https://github.com/ollama/ollama

19. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2305.14314

20. Cogranne, R., Sedighi, V., & Fridrich, J. (2017). Practical strategies for content-adaptive batch steganography and pooled steganalysis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2122–2126. https://doi.org/10.1109/ICASSP.2017.7952530

21. Caviglione, L., & Mazurczyk, W. (2022). Never mind the malware, here's the stegomalware. IEEE Security & Privacy, 20(5), 101–106. https://doi.org/10.1109/MSEC.2022.3178205

Yu. Mishkur, O. Zakharchenko

A HYBRID APPROACH TO STEGOANALYSIS BASED ON MULTIMODAL LARGE LANGUAGE MODELS AND CONVOLUTIONAL NEURAL NETWORKS

The article considers a hybrid approach to stegoanalysis of digital images, which combines the capabilities of specialized convolutional neural networks (CNNs) for low-level detection of statistical anomalies with the semantic analysis of multimodal large language models (MLLMs). The proposed approach is aimed at overcoming three fundamental limitations of existing monolithic CNN detectors: low generalization ability to unknown steganographic algorithms, lack of contextual analysis of multimodal metadata, and opacity of decision-making processes. The architecture of the hybrid system is implemented in the TensorFlow/Keras environment using three CNN architectures – MobileNetV2, ResNet50 and EfficientNetB0 - modified with specialized input filtering layers based on the Laplace kernel and the SRM filter bank to extract steganographically significant residual signals. Integration with language models is implemented through a local deployment of Ollama server in the Google Colab environment using the Gemma 3:4b, Gemma 3:12b and Llama 3.2 Vision 11B models. The final solution is formed through a "soft" fusion mechanism (Decision Fusion) of the

weighted outputs of the CNN and MLLM components, where the weights are dynamically adjusted depending on the detected semantic context of the image.

Experimental verification is performed on a synthesized dataset based on CIFAR-10 (LSB embedding) and on the ALASKA2 reference set. The highest detection accuracy was provided by the configuration ResNet50 + Gemma 3:12b: 95.8% on CIFAR-10 and 91.7% on ALASKA2. The obtained results indicate the promise of the hybrid approach for increasing the accuracy, generalizability and interpretability of stegoanalysis systems.

Keywords: stegoanalysis, steganography, convolutional neural networks, large language models, MobileNetV2, ResNet50, EfficientNet, Gemma3, Llama, Ollama, hybrid architecture.

Надійшла до редакції: 10.03.2026
Прийнята до друку: 21.04.2026
Опубліковано: 27.04.2026