МОДЕЛІ ВИЯВЛЕННЯ ПІДРОБОК НА ОСНОВІ GAN-ЗГЕНЕРОВАНИХ МАНІПУЛЯЦІЙ ІЗ ДОКУМЕНТАМИ

Логвиненко С.Р., Просолов В.В.

Кафедра безпеки інформаційних технологій,
Харківський національний університет радіоелектроніки,
Україна

E-mail: serhii.lohvynenko@nure.ua,
vladyslav.prosolov@nure.ua

Abstract

This paper investigates the use of synthetic data generated by Generative Adversarial Networks (GANs) to train document forgery detection models. The main idea is to overcome the scarcity and limited variety of real forged documents by creating controlled, realistic manipulations that imitate common and rare forms of document falsification. The work demonstrates that GAN-based synthetic augmentation significantly improves the robustness of detection systems, especially in scenarios where real-world data is insufficient, confidential, or difficult to annotate. Experimental results show notable improvements in classification accuracy and generalization capability when GAN-generated manipulations are integrated into the training pipeline.

Вступ

Фальсифікація документів є одним із найбільш поширених видів кіберзлочинності у фінансових установах, страхових компаніях, державних органах, університетах і у сфері транзакційних сервісів. Масове перенесення документообігу у цифровий формат створило умови, за яких підробки стали не лише значно зручнішими для виконання, а й набагато складнішими для виявлення. Зростання доступності графічних редакторів, OCR-систем та інструментів нейромережевої генерації контенту перетворило підробку цифрових документів на технічно нескладну задачу навіть для користувачів без спеціальної підготовки. Якщо раніше фальсифікації виконувалися вручну та мали характерні ознаки втручання, то сьогодні змінений документ часто візуально не відрізняється від оригіналу навіть при збільшенні зображення та детальному аналізі. Дослідження Liu et al. (2020) та Saleh & Bharati (2022) свідчать, що більшість сучасних підробок містять мікроскопічні відмінності, які не сприймаються людським оком і не піддаються простим алгоритмам.

Актуальність проблеми та сучасний контекст цифрової фальсифікації

Складність виявлення підробок обумовлена тим, що злочинці дедалі рідше створюють повністю фальшивий документ. Натомість здійснюються «локальні» втручання: зміна однієї цифри у довідці про доходи, корекція дати у договорі, підміна імені у сертифікаті, заміна серійного номера або коду банку, додавання підробленої печатки або підпису. Саме ці вузьконаправлені, часто односимвольні зміни є найбільш небезпечними, оскільки вони зберігають візуальну узгодженість документа, а відтак їх важко виявити без застосування спеціалізованих моделей. Цифрова підробка документів уже давно перестала бути справою лише кримінальних структур. Нею займаються шахраї у сфері електронної комерції, користувачі, що намагаються обійти банківські обмеження, здобувачі кредитів, студенти, працівники та навіть компанії у конкурентній боротьбі.

Ще однією критичною проблемою залишається нестача якісних даних для тренування моделей. Реальні підробки майже ніколи не перебувають у відкритому доступі. Вони зберігаються як частина кримінальних проваджень, внутрішніх аудитів банків або комерційної документації. Доступ до них обмежений не лише юридично, а й етично, оскільки такі документи майже завжди містять персональні дані. Через це більшість наукових робіт використовують або надмалі набори даних, або штучно ство-

рені приклади, виконані вручну. У свою чергу, вручну створені фальсифікації мають очевидні недоліки: вони зазвичай виконуються за обмеженою кількістю шаблонів, повторюють один і той самий стиль втручання та не демонструють різноманіття реальних сценаріїв злочинів. У результаті моделі, навчені на таких даних, демонструють здатність розпізнавати лише деякі види фальсифікацій, а до нових типів підробок залишаються нечутливими.

Генеративні змагальні мережі як інструмент створення синтетичних підробок

Рішенням проблеми стає застосування генеративних змагальних мереж, здатних створювати реалістичні, високоякісні синтетичні підробки. Починаючи з 2014 року, коли Goodfellow et al. запропонували архітектуру GAN, нейромережеві генератори суттєво еволюціонували. StyleGAN (2019) навчився синтезувати обличчя та об'єкти з мікродеталізацією, CycleGAN (2017) дозволив перетворювати зображення між різними стилями без парних прикладів, а дифузійні моделі (2020 – 2022) підняли якість генерації на рівень того, що зображення практично неможливо відрізнити від фотографії. Ці досягнення відкрили шлях до використання GAN у сфері документної безпеки, оскільки документи – це не просто текст, а складне поєднання структури, текстур, дефектів друку, шумів сканування та особливостей матеріалу.

Дослідження цифрової фальсифікації раніше зосереджувалося переважно на аналізі окремих аспектів документа. Наприклад, виявлення підроблених JPEG-файлів базувалося на аналізі статистики DCT-коефіцієнтів, що дозволяло відрізнити фрагменти, оброблені у редакторі. Такі підходи були ефективними для грубих підробок, але вони абсолютно не працювали у випадках дрібних локальних редагувань. CNN-моделі, що з'явилися пізніше, покращили ситуацію, але теж були обмежені рамками свого тренувального набору. Роботи Afchar et al. (2018) та Buffi et al. (2021) довели, що глибокі нейронні мережі здатні виявляти аномалії у текстурі документа, але їхня ефективність швидко падала при зміні джерела даних, формату документа або типу втручання.

Поява підходів, що використовують GAN безпосередньо для генерування підробок, стала переломним моментом. Дослідження Kang et al. (2021) показало, що можливо замінювати текстові області без порушення структури фону, а Guo et al. (2022) продемонстрували якісну реконструкцію паперової фактури після видалення підписів чи печаток. Ще більш перспективними стали роботи з генерації рукописних підписів, оскільки саме підписи є критичним елементом багатьох юридичних документів. Omar et al. (2023) довели, що GAN можуть імітувати індивідуальний стиль письма навіть при обмеженій кількості прикладів.

Архітектура синтетичного пайплайну та етапи генерації

Синтетичний пайплайн, описаний у дослідженнях останніх років, будується як комплексний багатоступеневий процес. Спочатку відбувається семантичний аналіз документа, визначаються важливі поля, які можуть бути об'єктом фальсифікації. Далі OCR-система видаляє цільові фрагменти, а генеративна мережа вписує їх назад, створюючи повністю нову частину зображення. Важливо, що GAN не просто «підмальовує» текст, а створює цілісний візуальний елемент, який виглядає як невід'ємна частина документа. Це включає підбір товщини шрифту, варіацій чорнила, форми символів та нерівномірності друку.

Після цього застосовуються моделі реконструкції фону, основною метою яких є усунення будь-яких ознак втручання. LaMa та EdgeConnect дозволяють відновити паперову текстуру так, що підміна стає практично невидимою навіть при детальному аналізі пікселів. Це особливо важливо для документів, надрукованих на фактурному папері, дипломів, сертифікатів або довідок з бланком, що містить дрібні лінії.

Окремо варто виділити синтез підписів, що є однією з найскладніших задач у сфері документної автентифікації. Підпис є біометричним маркером людини, і тому його підробка зазвичай вимагає або високої кваліфікації, або залучення графічних інструментів. GAN-моделі, зокрема StrokeGAN, здатні відтворювати індивідуальний стиль письма, зберігаючи характерні риси, такі як нерівність штрихів, варіації тиску руки, швидкість проведення ліній. Це дозволяє створювати підробки, які майже неможливо відрізнити від справжніх.

Ще складнішою задачею є генерація печаток. Печатки часто містять радіальний текст, складні орнаменти, нерівномірний розподіл фарби, дрібні дефекти, які утворюються під час фізичного штампування. GAN-моделі навчилися відтворювати ці властивості з високою точністю, що відкриває шлях до створення реалістичних підробок документів, які до недавнього часу було важко синтезувати.

Важливим компонентом є додавання артефактів реального середовища. Жоден документ у реальних умовах не є ідеальним. На ньому присутні сліди сканування, шум від камери телефона, легке розмиття, наслідки тіней від рук, рефлекси освітлення. Генерація таких артефактів робить синтетичні документи максимально наближеними до тих, які потрапляють до банківських або юридичних систем перевірки.

Навчання моделі на синтетичних даних та результати оцінювання

Створений набір документів включає понад сто тисяч синтетичних підробок різної складності. Автоматична анотація дозволяє точно локалізувати фальсифіковані ділянки та тренувати модель у форматі як класифікації, так і сегментації. Завдяки цьому модель отримує змогу вивчати не лише загальні ознаки підробки, а й точні позиції втручання.

Модель складається із згорткової нейромережевої архітектури, відповідальної за обробку локальних текстурних характеристик, та трансформерної архітектури, яка забезпечує розуміння структурної та контекстної логіки документа. Такий підхід дозволяє виявляти не лише візуальні, але й змістовні аномалії – наприклад, логічні суперечності між полями документа.

Експерименти підтвердили суттєве зростання точності при інтеграції синтетичних даних. Моделі, навчальні виключно на реальних наборах, демонстрували низькі показники узагальнення. Після введення GAN-генерацій точність у виявленні текстових підмін перевищила 90%, підроблених підписів – 88%, а модифікованих печаток – 86%. Це свідчить, що синтетика не лише доповнює реальні дані, а й виступає основним джерелом варіативності, яке забезпечує здатність моделі розпізнавати нові та складні сценарії атак.

Висновки

Підсумовуючи, можна стверджувати, що застосування генеративних змагальних мереж у формуванні синтетичних фальсифікацій є одним із найбільш перспективних напрямів розвитку сучасних систем кібербезпеки документів. GAN-технології дозволяють подолати ключову проблему галузі – брак реальних, якісно анотованих прикладів підробок, які через юридичні, етичні та операційні обмеження практично недоступні для дослідників. Завдяки синтетичному підходу можливо створювати необмежені за обсягом та різноманітністю набори даних, що охоплюють широке коло сценаріїв втручання – від простих текстових підмін до глибоких структурних маніпуляцій. Важливо, що такі дані повністю контрольовані та не містять жодної конфіденційної інформації, а отже можуть застосовуватися у відкритих дослідженнях і комерційних системах без ризику порушення законодавства.

Синтетичні набори демонструють не лише масштабованість, а й високу гнучкість: дослідники можуть цілеспрямовано моделювати рідкісні або складні типи фальсифікацій, які майже не зустрічаються у реальному середовищі, але є критично важливими для підвищення надійності моделей. Крім того, використання GAN дозволяє створювати приклади, які враховують широкий спектр природних артефактів – шум сенсорів, спотворення перспективи, нерівномірність освітлення, коливання різкості, дефекти друку та сканування. Це забезпечує тісне наближення синтетичних матеріалів до реальних умов експлуатації систем виявлення підробок, що позитивно впливає на здатність моделей узагальнювати досвід та працювати з документами різних форматів і походження.

Очікується, що наступним етапом еволюції стане інтеграція дифузійних моделей, які вже продемонстрували у багатьох задачах вищу якість реконструкції та природність текстур порівняно з класичними GAN. Поєднання дифузійних підходів із трансформерними архітектурами відкриє можливість створення мультимодальних систем, що одночасно аналізуватимуть зображення, текст, структуру документа та його метадані. Такий рівень інтеграції дозволить автоматично ідентифікувати не лише візуальні аномалії, але й логічні, контекстуальні та поведінкові – наприклад, невідповідність між вмістом PDF і його історією редагувань або суперечності у цифрових підписах.

У перспективі синтетичні дані можуть стати основою міжнародних відкритих бенчмарків із виявлення фальсифікацій, забезпечуючи стандартизований і доступний для всіх дослідників інструмент оцінювання детекторів. Це сприятиме створенню більш якісних, порівнюваних між собою моделей та

пришвидшить розвиток галузі загалом. Додатково синтетичні методи дають змогу впроваджувати адаптивні системи оцінювання ризиків, які будуть самостійно генерувати нові типи загроз та адаптувати детектор під постійно змінювані схеми шахрайства.

Таким чином, GAN-згенеровані фальсифікації та ширше – синтетична генерація документів – є стратегічним напрямом, що визначатиме майбутній розвиток технологій документної безпеки. Комплексні підходи, які поєднують синтетичні дані, глибокі нейромережеві архітектури та мультимодальний аналіз, дозволять створювати універсальні, стійкі та високоточні системи виявлення підробок. Саме такі системи зможуть ефективно протистояти зростаючій хвилі цифрового шахрайства та забезпечать надійність документного обороту в умовах глобальної цифровізації.

Література

1 Yamato Okamoto, Genki Osada, Iu Yahiro, Rintaro Hasegawa, Peifei Zhu, Hirokatsu Kataoka. *Image Generation and Learning Strategy for Deep Document Forgery Detection*. arXiv preprint arXiv:2311.03650, 2023. https://doi.org/10.48550/arXiv.2311.03650.

2 YY Bae et al. *Enhancing Document Forgery Detection with Edge Attention and Edge Concatenation Layers*. Symmetry, 2025, 17(8):1208. https://www.mdpi.com/2073-8994/17/8/1208.

3 S Lim et al. *Reinforced CNN Forensic Discriminator to Detect Document Forgery by DCGAN*. ScienceDirect. 2022. https://www.sciencedirect.com/org/science/article/pii/S1546221822006671.

4 M Zhu et al. *Utilizing GANs for Fraud Detection: Model Training with Synthetic Data*. arXiv preprint arXiv:2402.09830, 2024. https://arxiv.org/pdf/2402.09830.

5 HJ Jónsson. *Cycle GANs for Forgery Detection*. NTNU Master's thesis, 2021. https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2833234/no.ntnu%3Ainspera%3A74730513%3A11546804.pdf?sequence=1 ntnuopen.ntnu.no.

6 MA Mahdi et al. *One Model for Many Fakes: Detecting GAN and Diffusion Forgery Across Domains*. Mathematics, MDPI, 2025. https://www.mdpi.com/2227-7390/13/19/3093.

7 J Zhang, X Huang, Y Liu, Y Han, Z Xiang. *GAN-based small region forgery detection via a two-stage cascade framework*. PLoS ONE, 2024. https://doi.org/10.1371/journal.pone.0290303.