УДК 004.8:[316.472.4:070.16 **DOI:** 10.31673/2412-9070.2025.050411
М. С. ГНАТИШИН, аспірант;
ORCID: 0009-0009-0813-3602
О. Л. НЕДАШКІВСЬКИЙ, доктор техн. наук, професор,
ORCID: 0000-0002-1788-4434
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ

МОДЕЛЮВАННЯ СТРУКТУРИ ПОЯСНЮВАЛЬНОГО ШІ ДЛЯ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН ЗА ДОПОМОГОЮ МАШИННОГО НАВЧАННЯ

У статті досліджено необхідність підвищення зрозумілості та пояснюваності у системах виявлення фейкових новин, побудованих на основі алгоритмів машинного навчання. Пропонується нова концептуальна модель, яка забезпечує цілісне включення елементів пояснювального штучного інтелекту (ПШІ) з метою подолання проблеми "чорної скриньки".

Метою дослідження є створення умов для прозорого функціонування таких систем, підвищення рівня довіри з боку користувачів і підтримка ефективного управління інформаційним середовищем, що, в свою чергу, сприяє протидії цифровій дезінформації.

Запропонований підхід передбачає формування архітектурної основи для впровадження пояснювальних механізмів у процеси ідентифікації фейкових новин. Водночас аналізується, як методи типу LIME, SHAP та увагові механізми можуть бути адаптовані до інформаційних потреб різних цільових груп — від розробників до журналістів, модераторів і кінцевих користувачів. Також беруться до уваги специфічні виклики, пов'язані з мультимодальністю фейкового контенту, що включає текст, зображення, відео та інші формати.

Практична релевантність концепції ілюструється за допомогою змодельованих кейсів та прикладів, які демонструють її потенційні функціональні та технічні переваги.

Наукова значущість роботи полягає у розробці інтегрованої, орієнтованої на потреби різних груп користувачів структури ПШІ, яка дозволяє забезпечити не лише прозорість, а й адаптивність у контексті складних, змінюваних інформаційних потоків. На відміну від фрагментарних рішень, запропонований підхід створює уніфіковану рамку для узгодження архітектурних рішень із типами пояснень, необхідних у практиці взаємодії з системами виявлення фейкових новин.

Отримані висновки свідчать про те, що така структура здатна посилити прозорість ШІ-рішень, надати користувачам більше можливостей для критичної оцінки інформації, підвищити гнучкість моделей машинного навчання та сприяти ефективнішій взаємодії людини і ШІ. Застосування запропонованого підходу може стати основою для подальших практичних і емпіричних досліджень у сфері протидії дезінформації у цифровому середовищі.

Ключові слова: фейкові новини; пояснювальний ШІ; машинне навчання; прозорість; взаємодія людина–ШІ; архітектура ШІ-систем; дезінформація; інженерія програмного забезпечення; вебзастосунки.

Вступ

Поширення фейкових новин, що охоплює як навмисну дезінформацію, так і ненавмисне поширення неправдивої інформації, становить серйозний глобальний виклик у цифрову епоху [4]. Підсилювані онлайн-платформами, такі новини підривають довіру суспільства, дестабілізують демократичні процеси та можуть мати серйозні соціальні наслідки, що підкреслює нага-

льну потребу в ефективних засобах протидії. Машинне навчання (МН) стало основним інструментом для виявлення такого контенту і відповідні моделі (наприклад, глибинне навчання, методи на основі обробки природної мови) демонструють високу точність [4]. З огляду на це, постає потреба у використанні новітніх підходів для виявлення та класифікаці зазанченого контенту та ефективної розробки програмного забезпечення мобільних, кросплатформних і вебзастосунків [14].

Однак багато сучасних МН-систем працюють як "чорні скриньки", тобто їхні внутрішні процеси ухвалення рішень залишаються незрозумілими для користувачів [6]. Така непрозорість суттєво знижує ефективність цих інструментів у чутливому контексті виявлення фейкових новин. Вона породжує недовіру з боку користувачів і модераторів, ускладнює виявлення та усунення упередженості моделей, заважає ефективному нагляду з боку людини у складних випадках (наприклад, при виявленні сатири) та робить системи вразливими до витончених атак. Ця ключова проблема непрозорості традиційних МН-моделей у протиставленні меті створення зрозумілих та пояснюваних систем узагальнено представлена на рис. 1.

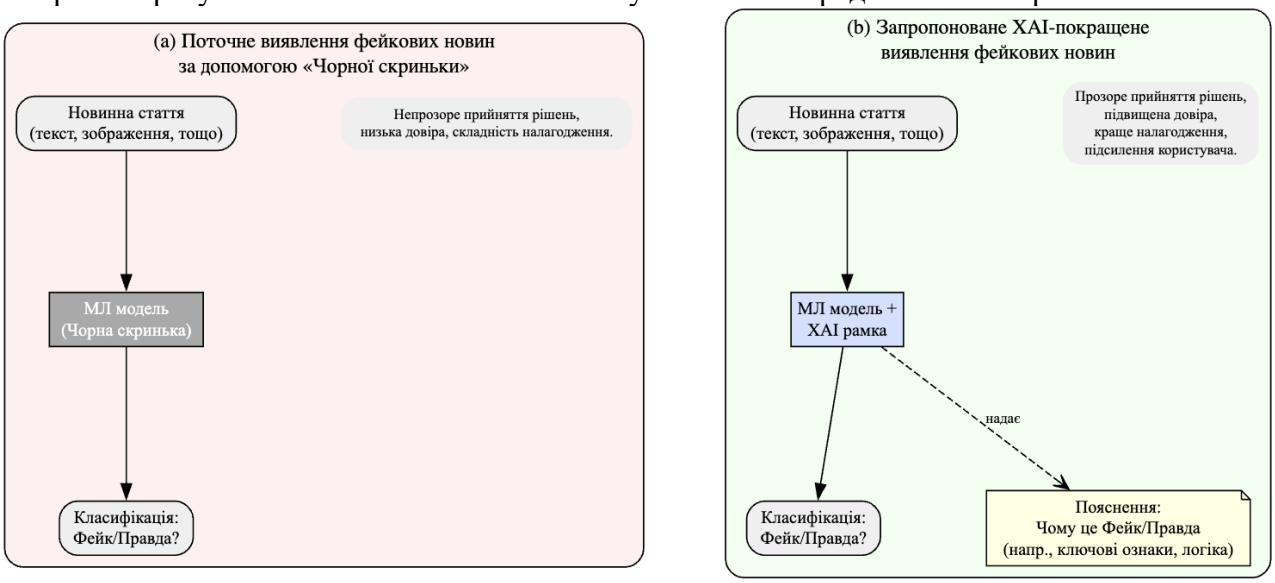

Рис. 1. Концептуальна ілюстрація: (а) проблема "чорної скриньки" у традиційному виявленні фейкових новин на основі машинного навчання порівняно з (б) прозорістю, яку забезпечує підхід із використанням пояснюваного ШІ (XAI)

Відповідно, виникає нагальна потреба у використанні пояснюваного ШІ (XAI) — методів, які, як показано на рис. 1(б), дозволяють зробити рішення моделей машинного навчання зрозумілими для людини, надаючи уявлення про логіку їхнього функціонування [3]. Інтеграція XAI є ключовою для перетворення таких інструментів виявлення на більш підзвітні, надійні та, зрештою, довірені системи [10] у тривалій боротьбі з онлайн-дезінформацією.

Результати проведених досліджень отримані під час роботи над проєктами з використанням технологій штучного інтелекту для боротьби з дезінформацією та розробки інформаційних технологій визначення тональності та класифікації текстового контексту інформації на основі нейромережних методів відповідно до Наказу МОН України №1202 від 04.10.2023р. [13] про пріоритетну тематику (п. 84 та п. 11 відповідно).

Аналіз останніх досліджень і публікацій

Дослідження в галузі автоматизованого виявлення фейкових новин за останні роки демонструють стрімкий розвиток [4], паралельно з чим зростає інтерес до пояснюваного штучного інтелекту (XAI) як засобу подолання проблеми "чорної скриньки" у складних моделях [6]. Методи машинного навчання еволюціонували від традиційних підходів (наприклад, метод опорних векторів (SVM), наївний байєсівський класифікатор), що покладалися на ручну побудову ознак, до сучасних глибинних архітектур, зокрема моделей на базі трансформерів (наприклад, BERT), які демонструють високі результати в обробці текстової інформації.

Усвідомлюючи багатогранну природу фейкових новин, сучасні дослідження дедалі частіше зосереджуються на мультимодальних системах, що аналізують текст, зображення та відео (з використанням згорткових нейронних мереж (CNN) або віжн-трансформерів), а також моделі поширення в мережах (наприклад, із застосуванням графових нейронних мереж, GNN) [2]. Незважаючи на досягнення, залишаються проблеми якості датасетів, адаптивності моделей до змін у тактиці поширення фейків, а також складнощі у виявленні культурно специфічного або нюансованого контенту [4]. Загальна тенденція до ускладнення моделей ще більше загострює проблему їх пояснюваності.

1. Провідні методи пояснюваного ШІ (XAI)

XAI пропонує набір методів, що дозволяють інтерпретувати рішення моделей машинного навчання [7]. Основні категорії включають:

● Агностичні до моделі методи: LIME [9] та SHAP [8] широко використовуються для пояснення індивідуальних передбачень шляхом виявлення найбільш значущих вхідних ознак, забезпечуючи як локальні, так і глобальні пояснення.

● Специфічні для моделі методи: До них належать візуалізації механізмів уваги (для трансформерів) або градієнтні методи (для нейронних мереж), що дають уявлення про внутрішні процеси моделі. Інші підходи включають контрфактичні пояснення [11] (які демонструють, які зміни у вхідних даних змінять рішення моделі) та прикладні пояснення [5, 6]. Мета цих методів — надати людині зрозуміле обґрунтування результатів моделі [3].

2. Існуючі застосування XAI у виявленні фейкових новин

Використання XAI у контексті виявлення фейкових новин наразі лише формується [4]. У деяких дослідженнях застосовували LIME і SHAP для ідентифікації ключових текстових ознак у класифікації або для інтерпретації моделей виявлення дипфейків [2]. Первинні результати свідчать, що XAI здатен підвищити здатність користувачів оцінювати достовірність новин. Однак більшість таких застосувань зосереджені на окремих методах XAI або типах даних, а не на цілісних, інтегрованих системах.

3. Виявлення нерозв'язаних аспектів загальної проблеми (ідентифікація прогалин)

Попри досягнення, низка критичних обмежень заважає ефективному використанню XAI у боротьбі з фейковими новинами:

● Відсутність комплексних структур: Більшість прикладів застосування XAI є ситуативними та не мають цілісних фреймворків для систематичної інтеграції на різних етапах виявлення та для різних типів даних [4].

● Складність пояснення мультимодальних фейкових новин: Створення зрозумілих пояснень для складного мультимодального контенту залишається відкритою проблемою [2, 4].

● Недостатність методів оцінки ефективності пояснень: Відсутні надійні, контекстно чутливі підходи до оцінки того, як XAI впливає на розуміння та прийняття рішень користувачами у сфері фейкових новин [3, 5].

● Проблеми операціоналізації: Впровадження XAI у масштабовані системи виявлення фейкових новин у реальному середовищі супроводжується суттєвими викликами в галузі програмної інженерії та етики [10].

Ця стаття має на меті подолати вказані прогалини шляхом розробки структурованої, орієнтованої на зацікавлені сторони концептуальної структури XAI, спеціально адаптованої до завдань виявлення фейкових новин.

Мета та завдання дослідження

Мета дослідження полягає у розробці нової концептуальної структури для інтеграції методів пояснюваного штучного інтелекту (ПШІ) у системи виявлення фейкових новин на основі

машинного навчання з метою підвищення прозорості, довіри користувачів та загальної ефективності таких систем.

Для досягнення цієї мети поставлено такі завдання:

1. Визначити потреби у пояснюваності, характерні для різних зацікавлених сторін екосистеми фейкових новин (користувачів, журналістів, модераторів, розробників).

2. Провести аналіз відповідних методів ПШІ та зіставити їх із компонентами моделей виявлення фейкових новин та потребами користувачів.

3. Розробити модульну архітектуру програмного забезпечення, що забезпечує гнучку інтеграцію компонентів ПШІ у процес виявлення фейкових новин.

4. Проілюструвати потенційні переваги та практичну користь запропонованої структури за допомогою тематичних кейсів і змодельованих експериментальних сценаріїв.

5. Розглянути основні етичні аспекти та обмеження, пов'язані з впровадженням розробленої структури, для всебічного оцінювання її ефективності.

Концептуальні основи та керівні принципи пояснюваного штучного інтелекту (ПШІ) для виявлення фейкових новин

Розробка структури базується на усвідомленні того, що ефективна пояснюваність у складній і чутливій сфері виявлення фейкових новин має бути надійною та орієнтованою на користувача [3, 5]. Пояснення вважається цінним, якщо воно:

- Відповідає істині (faithful): точно відображає процес ухвалення рішення моделлю [5, 7].
- Зрозуміле: є доступним і зрозумілим для цільового користувача, враховуючи різний рівень технічної підготовки та специфічні інформаційні потреби [3].
- Практичне: дає можливість користувачу робити обґрунтовані висновки або приймати відповідні рішення.
- Своєчасне і контекстно-орієнтоване: надається ефективно з урахуванням релевантної контекстної інформації для максимізації користі.

Дизайн структури спирається на кілька ключових принципів:

- Модульність та розширюваність: архітектура передбачає модульність, що спрощує інтеграцію різних методів ПШІ [7] та дозволяє адаптуватися до нових моделей машинного навчання або змін у характеристиках фейкових новин.
- Підтримка мультимодальності: оскільки фейкові новини часто поєднують текст, зображення, відео та схеми поширення у мережах, структура розроблена з урахуванням можливості пояснень, що враховують ці мультимодальні сигнали [2, 4].
- Прозорість пояснень: структура має на меті чітко комунікувати не лише логіку моделі, але й можливості та обмеження застосовуваних методів ПШІ [1, 5].
- Сприяння співпраці людини і ШІ: позиціонування ПШІ як моста для ефективнішої взаємодії між людською експертизою і можливостями штучного інтелекту у боротьбі з дезінформацією [10].

Запропонована структура ПШІ: архітектура та функціонал

Запропонована структура ПШІ, ілюстрована на рис. 2, передбачає багаторівневу систему. Вона інтегрується з основною моделлю (або ансамблем моделей) виявлення фейкових новин і систематично генерує, синтезує та доставляє змістовні пояснення [1].

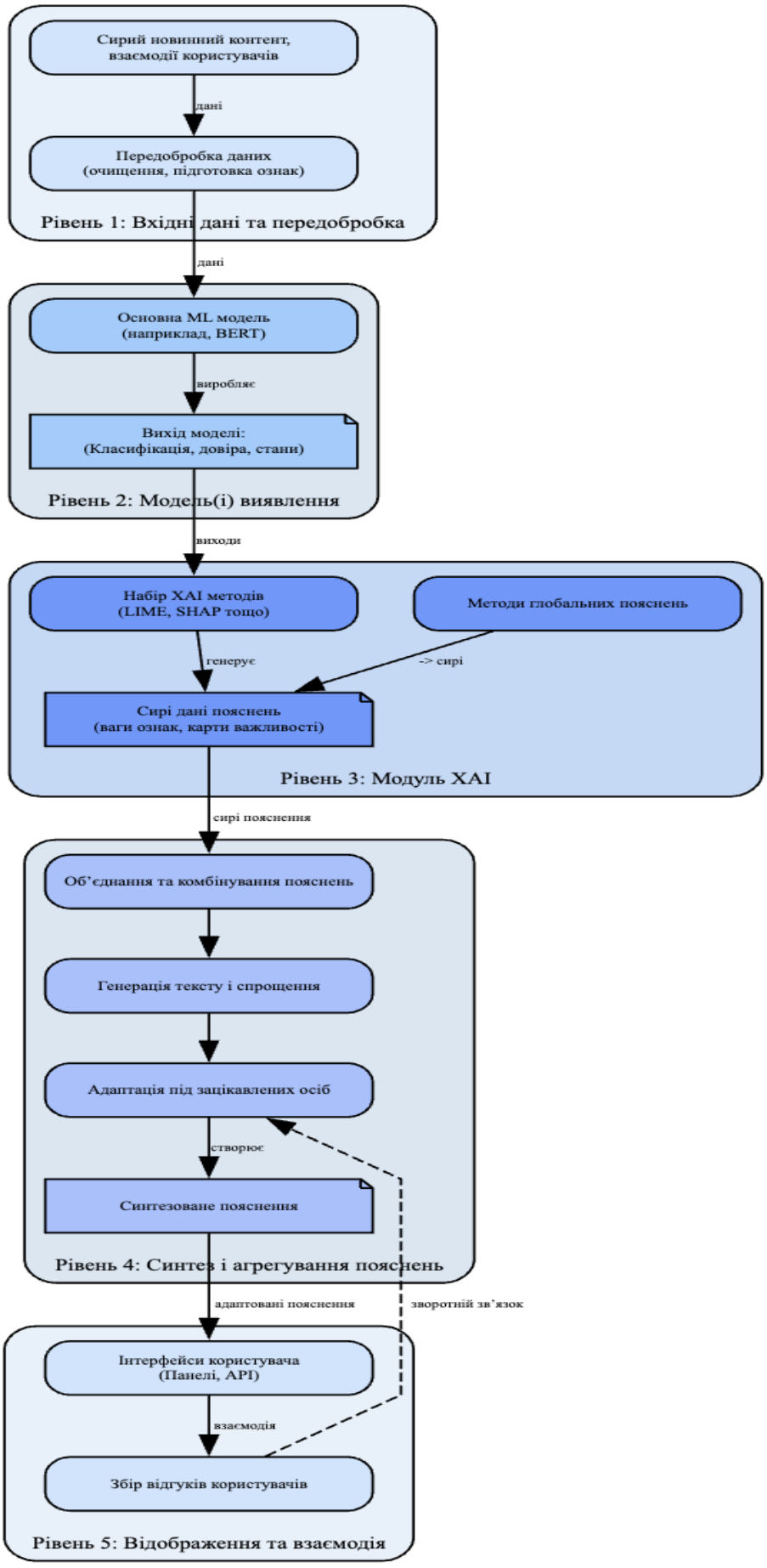

Рис. 2. Компактна архітектурна діаграма запропонованої структури пояснюваного штучного інтелекту (ПШІ)

Основні шари та компоненти запропонованої структури

Шар 1: Вхідні дані та їх попередня обробка.

Перший шар відповідає за прийом широкого спектру вхідних даних, включаючи необроблений контент новин (текст, зображення, посилання на відео, а також супровідні метадані — заголовки, авторство, дати публікації), дані про взаємодію користувачів (наприклад, поширення, коментарі, скарги, якщо доступно) та інформацію про джерела. На цьому етапі виконується необхідна попередня обробка, адаптована як для подальшої роботи моделі виявлення фейкових новин, так і для специфічних вимог методів пояснюваного ШІ (ПШІ).

Шар 2: Модель(і) виявлення фейкових новин.

Цей шар містить основну модель або ансамбль моделей машинного навчання, наприклад, сучасні трансформерні архітектури для аналізу тексту, Vision Transformers (ViTs) або згорткові нейронні мережі (CNN) для роботи з зображеннями і відео, а також графові нейронні мережі (GNN) для аналізу патернів поширення інформації та взаємозв'язків між джерелами. Моделі генерують основний класифікаційний результат (наприклад, «фейк», «легітимна», «маніпулятивна») з відповідною оцінкою впевненості. Важливо, що цей шар може також надавати внутрішні стани моделі (наприклад, ваги уваги або проміжні вектори ознак), які є цінними для деяких методів ПШІ.

Шар 3: Модуль Пояснюваного ШІ (XAI Engine).

Цей шар виконує роль центральної системи пояснюваності, містить різноманітний набір інструментів ПШІ. Сюди входять методи, незалежні від моделі, такі як LIME і SHAP (для локального та глобального визначення важливості ознак), модель-специфічні методи — візуалізація механізму уваги (для трансформерів), градієнтні методи (для глибоких нейронних мереж), генератори контрфактичних пояснень [11], а також методи, як GNNExplainer, для графових моделей. Обробка відповідних методів здійснюється залежно від типу вхідних даних, використовуваної моделі, яка використовується, та характеру необхідного пояснення.

Шар 4: Синтез та агрегування пояснень.

Сирі результати роботи модуля ПШІ (наприклад, оцінки важливості ознак, карти уваги, набори правил) часто є надто складними або фрагментованими для прямого сприйняття [6]. Основним завданням цього шару є обробка, уточнення та інтеграція цих пояснень у форми, які будуть зрозумілими, послідовними і релевантними для різних категорій користувачів. Серед ключових процесів — об'єднання пояснень з кількох методів ПШІ або різних типів даних, використання генерації природною мовою (NLP) для створення зручних для сприйняття текстових резюме, а також адаптація рівня деталізації і технічної складності відповідно до профілю кінцевого користувача.

Шар 5: Презентація та взаємодія.

Останній шар відповідає за доставку адаптованих пояснень користувачам або іншим системам. Це може здійснюватися через різноманітні інтерфейси користувача, такі як інтерактивні панелі управління, розширення для браузерів або вбудовані повідомлення на платформах новин. Також передбачена можливість API-інтеграції для підключення зовнішніх інструментів модерації контенту. Концептуально цей шар включає механізм зворотного зв'язку, який дає користувачам змогу оцінювати зрозумілість і корисність пояснень, що сприятиме поступовому вдосконаленню системи ПШІ.

Висновки

У цій статті розглянуто критичну проблему «чорних скриньок» моделей машинного навчання у виявленні фейкових новин шляхом запропонування нової концептуальної структури для інтеграції пояснюваного штучного інтелекту (ПШІ). Головний внесок полягає у розробці орієнтованої на зацікавлених сторін, модульної та мультимодальної архітектури ПШІ, спрямованої на підвищення прозорості, довіри користувачів та якості прийняття рішень у боротьбі з онлайн-дезінформацією. Ілюстративні сценарії демонструють потенціал цієї структури забезпечувати більш зрозумілі та практичні інсайти у порівнянні з існуючими системами без пояснювальних механізмів.

Подальші дослідження повинні зосередитися на емпіричній валідації цієї структури через розробку прототипів та їх всебічне тестування на реальних наборах даних і серед різноманітних груп користувачів. Основні напрями наукової роботи включають вдосконалення методів ПШІ, адаптованих до складних і мультимодальних фейкових новин (зокрема, з урахуванням контенту, створеного генеративним ШІ), вирішення питань масштабованості та роботи у режимі реального часу для практичного застосування, а також проведення користувацьких досліджень для оцінки впливу і зручності адаптованих пояснень у реальному середовищі. Також важливим буде подальше вивчення адаптивних пояснень та етичних аспектів застосування ПШІ в цій сфері, що є ключовим для перетворення концептуального проєкту у надійний і відповідальний інструмент.

Список літератури

1. Adadi A., Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. 2018. Vol. 6. P. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

2. Alaskar R., KP S. Explainable AI for deepfake detection: A review. MDPI Applied Sciences. 2023. Vol. 13, No. 5. Art. 3021. DOI: 10.3390/app13053021.

3. Arrieta A. B. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020. Vol. 58. P. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

4. Barredo Arrieta A. et al. On the explainability of \mbox{Artifcial} Intelligence in fake news detection: \mbox{Challenges} and future directions. TechRxiv. Preprint. 2022. DOI: 10.36227/techrxiv.19309205.v1.

5. Doshi-Velez F., Kim B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv. Preprint arXiv:1702.08608. 2017. URL: https://arxiv.org/abs/1702.08608 (Last accessed: 17.05.2025).

6. Guidotti R. et al. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys. 2018. Vol. 51, No. 5. Art. 93. P. 1–42. DOI: 10.1145/3236009.

7. Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 2021. Vol. 23, No. 1. Art. 18. DOI: 10.3390/e23010018.

8. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. P. 4765–4774. URL: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (Last accessed: 17.05.2025).

9. Ribeiro M. T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). 2016. P. 1135–1144. DOI: 10.1145/2939672.2939778.

10. Shneiderman B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy Human-Centered AI systems. ACM Transactions on Interactive Intelligent Systems. 2020. Vol. 10, No. 4. Art. 26. P. 1–31. DOI: 10.1145/3419764.

11. Verma S., Dickerson J., Pruthi G. Counterfactual Explanations for Machine Learning: A Review. MLR : Workshop on Human Interpretability in Machine Learning (WHI 2020). 2020. URL: http://proceedings.mlr.press/v119/verma20a.html (Last accessed: 17.05.2025).

12. Zhang Y., Chen X. Explainable Recommendation: A Survey and New Perspectives. ACM Transactions on Intelligent Systems and Technology. 2020. Vol. 11, No. 5. Art. 50. P. 1–37. DOI: 10.1145/3383581.

13. Про внесення змін до наказу Міністерства освіти і науки України від 07.09.2023 №1104 : НАКАЗ від 04.10.2023 № 1202. URL: https://mon.gov.ua/static-objects/mon/sites/1/nauka/Konkurs.vidbir.proektiv.nauk.robit-molodykh.vchenykh-2023/2023/10/04/Nakaz.MON.vid-04.10.2023-1202.pdf (дата звернення: 17.05.2025).

14. Hnatyshyn M.S., Nedashkivskiy O. L. Methods and software for the detection of false information on the Internet based on neural networks. Науково-практичний журнал «Зв'язок», 2024, № 4 (170). С. 52 – 57. URL: https://doi.org/10.31673/2412-9070.2024.045257 (дата звернення: 17.05.2025).

M. Hnatyshyn, O. Nedashkivskiy

MODELING THE STRUCTURE OF EXPLAINABLE AI FOR FAKE NEWS DETECTION USING MACHINE LEARNING

This article explores the pressing need for enhancing interpretability and transparency in fake news detection systems based on machine learning techniques. It introduces a novel conceptual framework designed to systematically integrate Explainable Artificial Intelligence (XAI) components to address the "black-box" nature of such models.

The primary goal is to strengthen system transparency, foster user trust, and support effective moderation, thereby improving efforts to counteract digital disinformation. The proposed approach outlines the architectural foundations required for embedding XAI mechanisms into fake news detection workflows. It also examines how established explanation methods — such as LIME, SHAP, and attention mechanisms — can be aligned with the specific informational needs of different stakeholder groups, including developers, moderators, journalists, and end users. Special consideration is given to the challenges posed by multimodal disinformation, which includes text, images, video, and other content types.

The framework's practical relevance is illustrated through simulated scenarios and example use cases that demonstrate its potential functional and operational benefits.

The scientific contribution of this work lies in the development of an integrated, stakeholder-centered XAI architecture, specifically tailored to the complex task of detecting fake news. Unlike ad hoc applications, the framework offers a systematic approach that encompasses multimodal content, defines integration-focused architectural considerations, and matches explanation types to differentiated user demands.

Findings from this conceptual study suggest that the proposed XAI structure offers a coherent pathway toward building more transparent, accountable, and effective fake news detection systems. Its adoption is expected to enhance users' ability to critically assess information, improve model adaptability, and support human–AI collaboration — providing a foundation for further empirical research in combating online disinformation.

Keywords: fake news detection; explainable AI (XAI); machine learning; AI transparency; conceptual framework; software engineering; disinformation; software engineering; web applications.