UDC 004.032.26:070.16
DOI https://doi.org/10.32689/maup.it.2025.2.4

Mykhailo HNATYSHYN
Postgraduate Student at the Department of Software Engineering in Energy,
Educational and Scientific Institute of Atomic and Thermal Power Engineering,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", gnatyshynmisha@gmail.com
***ORCID:** 0009-0009-0813-3602*

Oleksii NEDASHKIVSKIY
Doctor of Technical Sciences, Professor at the Department of Software Engineering in Energy,
Educational and Scientific Institute of Atomic and Thermal Power Engineering,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", al_1@ua.fm
***ORCID:** 0000-0002-1788-4434*

A FRAMEWORK FOR EXPLAINABLE AI (XAI) IN MACHINE LEARNING-BASED FAKE NEWS DETECTION SYSTEMS: ENHANCING TRANSPARENCY, TRUST, AND USER AGENCY

Abstract. *The subject of this article is the critical need for interpretability in machine learning (ML) based fake news detection systems and the proposal of a novel conceptual framework for the systematic integration of Explainable AI (XAI) to address this.*

*The **purpose** is to enhance transparency, user trust, and effective moderation, thereby improving the fight against the significant threat of online disinformation.*

*The proposed **methodology** involves delineating key architectural components for integrating XAI into fake news detection workflows, mapping diverse XAI techniques (e.g., LIME, SHAP, attention mechanisms) to the specific explainability needs of various stakeholders (end-users, journalists, moderators, developers), and considering the challenges of multimodal fake news. The potential benefits and operational characteristics of this framework are illustrated conceptually through mock experimental scenarios and illustrative case studies.*

*The **scientific novelty** of this work lies in its comprehensive, stakeholder-centric XAI framework specifically tailored for the complexities of fake news detection. Unlike ad-hoc applications, this framework offers a systematic approach addressing multimodal content, outlining architectural considerations for integration, and linking explanation types to differentiated user requirements, aiming for a more holistic solution to the «black-box» problem in this domain.*

***Conclusions** from this conceptual study suggest that the proposed XAI framework provides a structured pathway towards developing more trustworthy, accountable, and effective AI-driven fake news detection systems. Its implementation is projected to enhance transparency, improve user agency in information assessment, facilitate model refinement, and support robust human-AI collaboration, thereby contributing a foundational approach for future empirical validation in combating online disinformation.*

Key words: *Fake News Detection, Explainable AI (XAI), Machine Learning, Trustworthy AI, Conceptual Framework, Software Engineering, Misinformation, Disinformation.*

Михайло ГНАТИШИН, Олексій НЕДАШКІВСЬКИЙ. СТРУКТУРА ПОЯСНЮВАЛЬНОГО ШТУЧНОГО ІНТЕЛЕКТУ (ПШІ) В СИСТЕМАХ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН НА ОСНОВІ МАШИННОГО НАВЧАННЯ: ПОСИЛЕННЯ ПРОЗОРОСТІ, ДОВІРИ ТА СУБ'ЄКТНОСТІ КОРИСТУВАЧІВ

Анотація. *Предметом цієї статті є критична потреба в інтерпретованості систем виявлення фейкових новин на основі машинного навчання (МН) та пропозиція нової концептуальної структури для систематичної інтеграції Пояснювального Штучного Інтелекту (ПШІ) для її вирішення.*

***Метою** є посилення прозорості, довіри користувачів та ефективної модерації, тим самим покращуючи боротьбу зі значною загрозою онлайн-дезінформації.*

*Запропонована **методологія** передбачає визначення ключових архітектурних компонентів для інтеграції ПШІ в робочі процеси виявлення фейкових новин, співвіднесення різноманітних методів ПШІ (наприклад, LIME, SHAP, механізми уваги) зі специфічними потребами в поясненнях різних зацікавлених сторін (кінцевих користувачів, журналістів, модераторів, розробників) та врахування проблем мультимодальних фейкових новин. Потенційні переваги та операційні характеристики цієї структури концептуально ілюструються за допомогою модельованих експериментальних сценаріїв та наочних прикладів.*

***Наукова новизна** цієї роботи полягає в її комплексній, орієнтованій на зацікавлених сторін структурі ПШІ, спеціально адаптованій до складнощів виявлення фейкових новин. На відміну від ситуативних застосувань, ця структура пропонує систематичний підхід, що охоплює мультимодальний контент, визначає архітектурні міркування для інтеграції та пов'язує типи пояснень з диференційованими вимогами користувачів, маючи на меті більш цілісне вирішення проблеми «чорної скриньки» в цій галузі.*

Висновки *з цього концептуального дослідження свідчать, що запропонована структура ПШІ забезпечує структурований шлях до розробки більш надійних, підзвітних та ефективних систем виявлення фейкових новин на основі ШІ. Прогнозується, що її впровадження посилить прозорість, покращить суб'єктність користувачів в оцінці інформації, полегшить вдосконалення моделей та підтримає надійну співпрацю людини та ШІ, тим самим роблячи внесок у фундаментальний підхід для майбутньої емпіричної валідації у боротьбі з онлайн-дезінформацією.*

Ключові слова: *виявлення фейкових новин, пояснювальний штучний інтелект (ПШІ), машинне навчання, довірений ШІ, концептуальна структура, програмна інженерія, дезінформація.*

Introduction. The proliferation of fake news, encompassing both deliberate disinformation and unintentional misinformation, presents a formidable global challenge in the digital era [4]. Amplified by online platforms, it erodes public trust, disrupts democratic processes, and can have severe societal consequences, underscoring the urgent need for effective countermeasures. Machine learning (ML) has become a primary tool in detecting such content, with models (e.g., deep learning, NLP-based approaches) achieving notable accuracy [4].

However, many advanced ML detectors operate as "black boxes", their internal decision-making processes opaque to users [6]. This lack of transparency critically undermines their utility in the sensitive context of fake news. It breeds mistrust among users and moderators, complicates the identification and mitigation of model biases, hinders effective human oversight in nuanced cases (like satire), and leaves systems vulnerable to sophisticated adversarial attacks. This fundamental challenge of opacity in conventional ML systems versus the goal of transparent, explainable systems is visually summarized in (Fig. 1).

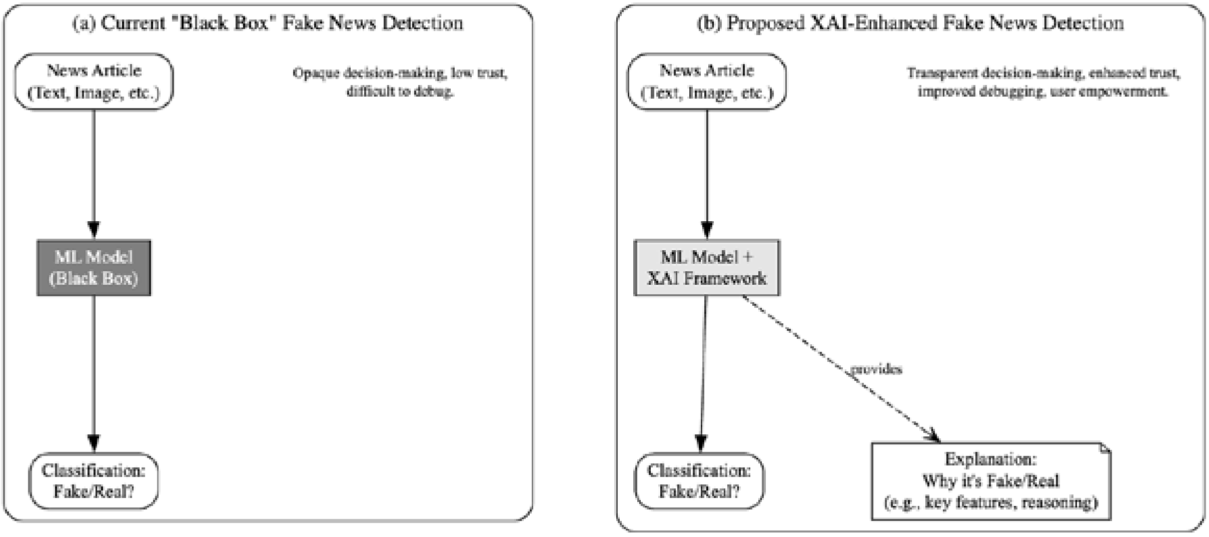
(a) Current "Black Box" Fake News Detection
News Article (Text, Image, etc.)
Opaque decision-making, low trust, difficult to debug.
ML Model (Black Box)
Classification: Fake/Real?
(b) Proposed XAI-Enhanced Fake News Detection
News Article (Text, Image, etc.)
Transparent decision-making, enhanced trust, improved debugging, user empowerment.
ML Model + XAI Framework
provides
Classification: Fake/Real?
Explanation: Why it's Fake/Real (e.g., key features, reasoning)

Fig. 1. Conceptual Illustration of (a) The "Black Box" Challenge in Traditional ML-Based Fake News Detection versus (b) The Transparency Offered by an XAI-Enhanced Approach

Consequently, there is a pressing need for Explainable AI (XAI)–methods that can, as illustrated in Figure 1(b), render ML decisions understandable to humans by providing insights into their reasoning [3]. Integrating XAI is essential for transforming these detection tools into more accountable, robust, and ultimately trustworthy systems [10] in the ongoing effort to combat online disinformation.

Analysis of Recent Research and Publications. Research in automated fake news detection has rapidly advanced [4], alongside a growing interest in Explainable AI (XAI) to address the "black-box" nature of complex models [6]. ML approaches have evolved from traditional models (e.g., SVMs, Naive Bayes) relying on feature engineering to sophisticated deep learning architectures, particularly Transformer-based models (e.g., BERT), which excel at understanding textual nuances. Recognizing the multifaceted nature of fake news, current research increasingly focuses on multimodal systems that analyze text, images/videos (often with CNNs or Vision Transformers), and network propagation patterns (using GNNs) [2]. Despite progress, challenges persist in dataset quality, model adaptability to evolving tactics, and handling nuanced or culturally specific content [4]. The trend is towards more complex models, which often exacerbates the explainability challenge.

Prominent Explainable AI (XAI) Techniques. XAI offers methods to interpret ML model decisions [7]. Key categories include:

● Model-Agnostic Methods: LIME [9] and SHAP [8] are widely used to explain individual predictions by identifying critical input features, offering both local and global perspectives.

● Model-Specific Methods**:** Techniques like attention mechanism visualizations (for Transformers) or gradient-based methods (for neural networks) provide insights into model-specific internal workings. Other approaches include counterfactual explanations [11] (showing what input changes would alter a decision) and example-based explanations [5, 6]. The goal is to provide human-understandable justifications for model outputs [3].

Existing Applications of XAI in Fake News Detection**:** The application of XAI in fake news is emerging [4]. Studies have utilized LIME and SHAP to identify influential textual features in classifications and to interpret deepfake detection models [2]. Initial findings suggest XAI can improve users' ability to assess news trustworthiness. However, these applications are often focused on specific XAI techniques or data modalities rather than integrated systems.

Highlighting Unresolved Parts of the General Problem (Identifying the Gap)**:** Despite progress, critical gaps hinder the effective use of XAI in combating fake news:

● Lack of Comprehensive Frameworks**:** Most XAI applications are ad-hoc, lacking holistic frameworks for systematic integration across diverse data types and detection stages [4].

● Challenges in Explaining Multimodal Fake News**:** Generating coherent explanations for sophisticated, multimodal fake news remains difficult [2, 4].

● Inadequate Evaluation of Explanation Effectiveness**:** Robust, context-specific methods for evaluating how XAI impacts user understanding and decision-making in the fake news domain are underdeveloped [3, 5].

● Operationalization Hurdles**:** Integrating XAI into real-world, scalable fake news detection systems presents significant software engineering and ethical challenges [10].

This paper seeks to bridge these gaps by proposing a structured, stakeholder-aware XAI framework tailored for fake news detection.

Formulation of the Purpose of the Article (Statement of the Task). Addressing the identified gaps in current research, particularly the need for a more integrated and stakeholder-focused approach to explainability in fake news detection, this article sets out to propose a novel conceptual framework. The **primary purpose** is to detail a systematic approach for integrating Explainable AI (XAI) techniques within machine learning-based fake news detection systems, with the overarching goal of enhancing transparency, user trust, and overall system effectiveness.

To achieve this, the article will first delineate stakeholder-specific explainability needs pertinent to the diverse actors in the fake news ecosystem. Following this, it will map suitable XAI techniques to various components of fake news detection models and these distinct user requirements. A key aspect of this work involves outlining a modular software architecture conducive to the flexible integration of these XAI components. Furthermore, the potential benefits and operational utility of the proposed framework will be illustrated through a combination of illustrative case studies and mock experimental scenarios. Finally, the discussion will extend to key ethical considerations and inherent limitations associated with such a framework, thereby providing a comprehensive exposition of the proposed solution.

The Proposed XAI Framework for Fake News Detection. This section details the proposed conceptual framework for integrating Explainable AI (XAI) into machine learning-based fake news detection systems. The framework aims to transform opaque detection models into transparent tools, enhancing user trust [10] and providing actionable insights to combat the pervasive challenge of online disinformation.

Conceptual Foundations and Guiding Principles. The development of this framework is rooted in the understanding that effective explainability in the complex and sensitive domain of fake news detection must be robust and user-focused [3, 5]. An explanation is deemed valuable if it is:

● **Faithful:** Accurately reflects the underlying model's decision-making process [5, 7].

● **Understandable:** Comprehensible to the target stakeholder, considering their varying levels of technical expertise and specific informational needs [3].

● **Actionable:** Empowers the stakeholder to make informed judgments or take appropriate next steps.

● **Timely and Context-Aware:** Provided efficiently and incorporating relevant contextual information to maximize its utility.

The framework's design adheres to several core principles:

1. **Modularity and Extensibility:** The architecture is designed to be modular, facilitating the integration of various XAI techniques [7] and allowing for future adaptation to new ML models or evolving fake news characteristics.

2. **Multimodality Support:** Given that fake news often leverages a combination of text, images, video, and network propagation patterns, the framework is conceptualized to support explanations that can account for these multimodal signals [2, 4].

3. **Transparency of Explanations:** The framework aims to clearly communicate not only the model's reasoning but also the capabilities and limitations of the XAI methods employed [1, 5].

4. **Facilitation of Human-AI Collaboration:** Positioning XAI as a bridge that enables more effective partnership between human expertise and AI capabilities in addressing disinformation [10].

The Proposed XAI Framework: Architecture and Functionality. The proposed XAI framework, illustrated in **Figure 2**, is envisioned as a multi-layered system. It integrates with an underlying fake news detection model (or an ensemble of such models) to systematically generate, synthesize, and deliver insightful explanations [1].

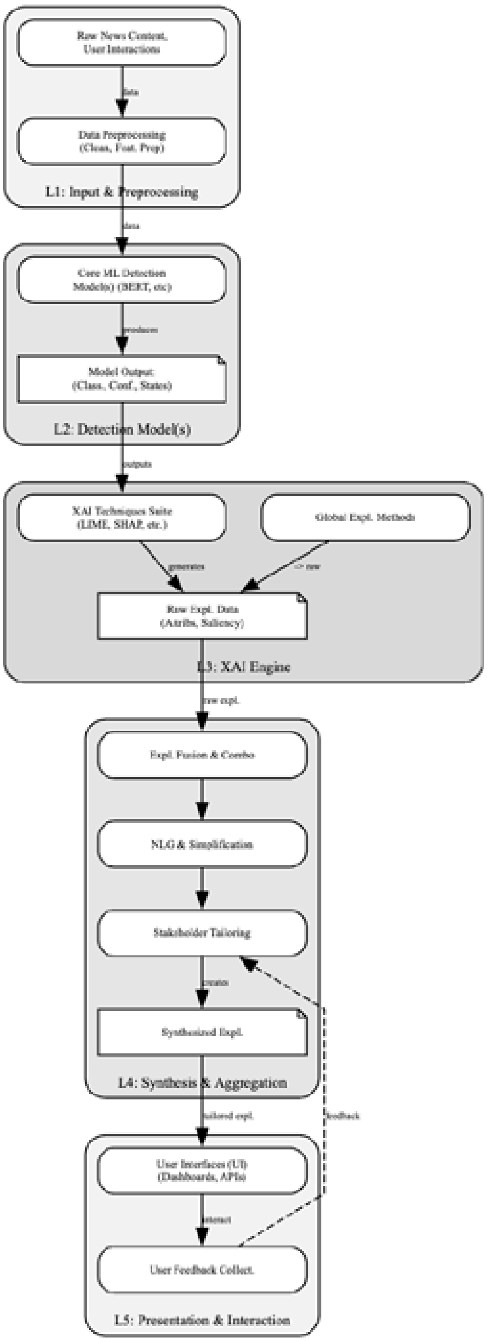

Fig. 2. Compact Architectural Diagram of Proposed XAI Framework

The primary layers and components are:

● **Layer 1: Data Input & Preprocessing:** This initial layer is responsible for ingesting a wide array of inputs, including raw news content (text, images, video URLs, associated metadata such as headlines, authorship, and publication dates), user interaction data (e.g., shares, comments, reports, if accessible), and source information. It then performs necessary preprocessing tailored for both the subsequent fake news detection model and the specific requirements of the XAI analysis techniques.

● **Layer 2: Fake News Detection Model(s):** This layer houses the core ML model(s) – for instance, advanced Transformer-based architectures for textual analysis, Vision Transformers (ViTs) or CNNs for image/video content, and Graph Neural Networks (GNNs) for analyzing propagation patterns or source relationships. These models provide the primary classification (e.g., "fake," "legitimate," "misleading") along with a confidence score. Crucially, this layer can also expose internal model states (like attention weights or intermediate feature vectors) that are valuable inputs for certain XAI techniques.

● **Layer 3: XAI Engine:** As the central nervous system for explainability, this layer contains a diverse toolkit of XAI methods. These include model-agnostic techniques like LIME and SHAP (for local and global feature attributions), model-specific methods such as attention visualization (for Transformer models) or gradient-based explanations (for deep neural networks), counterfactual explanation generators [11], and techniques like GNNExplainer for graph-based models. The engine selectively applies these techniques based on the input data modality, the type of ML model used, and the specific nature of the explanation being sought.

● **Layer 4: Explanation Synthesis & Aggregation:** Raw outputs from the XAI Engine (e.g., feature importance scores, saliency maps, rule sets) are often too complex or fragmented for direct consumption [6]. This layer's critical function is to process, refine, and synthesize these raw explanations into forms that are coherent, understandable, and directly relevant to different stakeholders. Key processes include fusing explanations from multiple XAI methods or across different data modalities, employing Natural Language Generation (NLG) to create human-readable summaries, and adapting the level of detail and technical complexity based on the target user's profile.

● **Layer 5: Presentation & Interaction:** The final layer is responsible for delivering the tailored explanations to users or other systems. This can occur through various user interfaces (UIs) such as interactive dashboards, browser extensions, or embedded notices within news platforms. It also includes provisions for API-based delivery to integrate with external content moderation tools. Conceptually, this layer incorporates a feedback mechanism, allowing users to provide input on the clarity and utility of explanations, which can inform iterative improvements to the XAI system.

Conclusions and Prospects for Further Exploration. This article addressed the critical challenge of "black-box" machine learning models in fake news detection by proposing a novel conceptual framework for integrating Explainable AI (XAI). The core contribution is a stakeholder-centric, modular, and multimodal XAI architecture designed to enhance transparency, user trust, and decision-making in combating online disinformation. The illustrative scenarios presented highlight its potential to provide clearer, more actionable insights than current non-explainable systems.

Future work must prioritize the empirical validation of this framework through prototype development and rigorous testing with real-world datasets and diverse user groups. Key research directions also include advancing XAI techniquestailored for sophisticated and multimodal fake news (including generative AI content), addressing the scalability and real-time performance challenges for practical deployment, and conducting user-centric studies to assess the real-world impact and usability of tailored explanations. Continued investigation into adaptive explanations and the ethical implications of XAI in this domain will also be crucial for evolving this conceptual blueprint into a robust and responsible tool.

Bibliography:

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.

Дата надходження статті: 01.06.2025
Дата прийняття статті: 09.06.2025
Опубліковано: 23.09.2025