МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
КОРАБЛЕБУДУВАННЯ ІМЕНІ АДМІРАЛА МАКАРОВА
МИКОЛАЇВСЬКА ОБЛАСНА ДЕРЖАВНА АДМІНІСТРАЦІЯ
ДЕПАРТАМЕНТ ОСВІТИ І НАУКИ
МИКОЛАЇВСЬКОЇ ОБЛАСНОЇ ДЕРЖАВНОЇ АДМІНІСТРАЦІЇ

# ІННОВАЦІЇ В СУДНОБУДУВАННІ ТА ОКЕАНОТЕХНІЦІ

XVI Міжнародна науково-технічна конференція

## МАТЕРІАЛИ

**25–26 вересня 2025 рік**

*Національний університет кораблебудування*
*імені адмірала Макарова*
*просп. Героїв України, 9*

ВИДАВНИЦТВО
НАЦІОНАЛЬНОГО УНІВЕРСИТЕТУ
КОРАБЛЕБУДУВАННЯ
ІМ. АДМІРАЛА МАКАРОВА
2025

**УДК 004.8:616-079**

# ОПТИМІЗАЦІЯ МЕТОДІВ XAI У МЕДИЧНІЙ ДІАГНОСТИЦІ НА ОСНОВІ ШІ

**Вербицький О.С.**
*магістрант з комп'ютерних наук*
*Національний університет кораблебудування імені адмірала Макарова,*
*м. Миколаїв, Україна*
*soomoalex@gmail.com*
*ORCID: 0009-0000-6848-8880*
**Гайдаєнко О.В.**
*кандидат технічних наук, доцент*
*Національний університет кораблебудування імені адмірала Макарова,*
*м. Миколаїв, Україна*
*oksana.gaidaienko@nuos.edu.ua*
*ORCID: 0000-0002-6614-5443*

**Анотація**. Дослідження спрямоване на аналіз методів пояснюваності (XAI) нейронних мереж у медицині, зокрема LIME, Grad-CAM, для підвищення довіри до ШІ-рішень. Виявлено, що ключові обмеження існуючих методів – складність інтерпретації, локальність пояснень та залежність від типу даних. Запропоновано концепцію системи, що інтегрує Grad-CAM для візуалізації зображень та LIME. Результати підтверджують ефективність цих методів у клінічній практиці.

**Ключові слова**: пояснюваність нейронних мереж, штучний інтелект у медицині, медична діагностика, прозорість моделей ШІ, аналіз медичних даних.

**Вступна частина**. Сучасний штучний інтелект (ШІ) став невід'ємною частиною медичної діагностики: від виявлення пухлин на ранніх стадіях до прогнозування ризиків серцевих захворювань. Однак його «чорний ящик» залишається ключовим бар'єром для довіри лікарів. Наприклад, у 23% випадків моделі ШІ генерують галюцинації при аналізі МРТ-знімків, а 17% помилкових діагнозів у кардіології пов'язані з некоректною інтерпретацією ознак. Існуючі інженерні рішення (наприклад, патчі для окремих архітектур) не забезпечують системного розуміння причинно-наслідкових зв'язків у роботі нейронних мереж. Це обмежує їхнє використання у критичних сценаріях, де кожне рішення впливає на життя пацієнтів. Виникає потреба у переході від ситуативних модифікацій до фундаментальної теорії пояснюваності, яка б забезпечила прозорість, передбачуваність та адаптацію методів XAI до різноманітних медичних контекстів [1].

**Мета роботи** полягає у подоланні проблеми «чорного ящика» у медичному ШІ через аналіз та оптимізацію методів пояснюваності, таких як LIME і Grad-CAM. Дослідження спрямоване на систематизацію обмежень існуючих підходів, розробку рекомендацій для їх адаптації до медичних даних (DICOM, лабораторні аналізи) і клінічних вимог, а також визначення критеріїв оцінки ефективності, включаючи точність, інтуїтивність і час генерації пояснень [2]. Окрема увага приділена створенню теоретичного фундаменту, який забезпечить

передбачуваність результатів і сприятиме довірі лікарів до ШІ-діагностики, особливо у критичних сферах – онкології та кардіології.

**Основна частина**. Прозорість рішень штучного інтелекту (ШІ) є критичною для медицини, де кожна помилка може мати незворотні наслідки. Існуючі моделі ШІ, попри високу точність, залишаються «чорними ящиками», що обмежує їхнє використання у клінічній практиці [3]. Наприклад, 30% лікарів відмовляються від алгоритмічної підтримки через недовіру до незрозумілих результатів. Пояснюваність стає ключовим інструментом для ранньої діагностики, де інтерпретація рішень дозволяє не лише виявити патології, але й обґрунтувати лікувальні протоколи.

**Об'єкт і предмет дослідження.** Дослідження зосереджено на аналізі та вдосконаленні двох ключових компонентів, необхідних для подолання проблеми «чорного ящика» в медичному ШІ.

Об'єкт – рішення нейронних мереж у медичній діагностиці (аналіз зображень, прогнозування ризиків).

Предмет – методи пояснюваності (xai), зокрема grad-cam для візуалізації аномалій на dicom-знімках та lime для інтерпретації табличних даних (лабораторні аналізи, діагностичні шкали).

**Порівняльний аналіз методів XAI.** В ході дослідження отримано статистично значимі результати, що підтверджують ефективність запропонованих методів пояснюваності у медичній діагностиці.

Grad-CAM досягає 95% точності у виявленні пухлин на МРТ завдяки локалізації аномалій у реальному часі. Його інтеграція з PACS дозволяє автоматизувати анотацію звітів, зменшуючи час діагностики на 25%.

LIME демонструє 87% ефективності у прогнозуванні серцевих захворювань, але потребує оптимізації для роботи зі складними архітектурами.

Інші методи (SHAP, Integrated Gradients) мають високу технічну точність, але їхні результати важко інтерпретувати без спеціальних знань, що робить їх менш прийнятними для лікарів.

**Концепція системи**.Запропоновано архітектуру, яка поєднує:

– Grad-CAM для аналізу медичних зображень (МРТ, КТ) з накладанням heatmap на DICOM-файли;

– LIME для пояснення прогнозів на основі лабораторних даних, з генерацією звітів у форматі FHIR для EHR.

Система враховує критерії оцінки: інтуїтивність візуалізацій, час відгуку (<2 сек) та відповідність клінічним стандартам.

**Практичні результати.** У онкології використання Grad-CAM зменшило кількість помилкових позитивних результатів на 18% завдяки точній локалізації пухлин. У кардіології LIME покращив точність прогнозу інфаркту на 22%, виділяючи ключові фактори ризику (рівень тропоніну, анамнез). У пульмонології комбінація методів дозволила скоротити час аналізу рентгенівських знімків на 35%.

**Рекомендації.** На основі результатів порівняльного аналізу та виявлених обмежень запропоновано шляхи вдосконалення методів пояснюваності для медичної практики.

Адаптація методів:

– використання grad-cam для 3d-зображень;
– розширення lime для роботи з мультимодальними даними.

Теоретичні принципи:

– універсальність (відокремлення пояснень від архітектури моделі);
– валідація через клінічні критерії.

Це демонструє, що Grad-CAM та LIME є оптимальними методами для медицини, поєднуючи точність, прозорість та інтеграцію з клінічними стандартами. Їхнє впровадження не лише підвищує довіру до ШІ, але й закладає основу для системної теорії пояснюваності, необхідної для майбутніх досліджень.

**Висновки.** Дослідження довело, що комбінація методів Grad-CAM та LIME є найперспективнішими методами пояснюваності для медичної діагностики. Grad-CAM забезпечує точну візуалізацію аномалій на зображеннях (МРТ, КТ) з інтуїтивною інтеграцією в клінічні процеси, а LIME дозволяє інтерпретувати табличні дані, такі як лабораторні аналізи, зрозуміло для лікарів без глибоких знань AI. Однак ключові обмеження цих методів – необхідність адаптації Grad-CAM до 3D-зображень та обмежена сумісність LIME з EHR – вимагають подальшої оптимізації.

Практичне значення роботи полягає в зменшенні часу діагностики (на 25-35% залежно від спеціалізації) та підвищенні точності прогнозів (до 22% у кардіології). Запропоновані критерії оцінки – інтуїтивність, час відгуку, відповідність клінічним стандартам – створюють основу для стандартизації використання XAI у медицині.

Перспективи подальших досліджень включають:

– розробку програмних модулів для автоматизації звітів на базі FHIR;
– адаптацію методів для мультимодальних даних (наприклад, поєднання зображень і текстових записів);
– валідацію системи в реальних клінічних умовах, зокрема в онкологічних та кардіологічних відділеннях.

Ця робота стала важливим етапом у визначенні структурних елементів для побудови теорії пояснюваності нейронних мереж у медицині, де прозорість і довіра формують основу інноваційних клінічних рішень.

## Література

[1] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*. 25. DOI: 10.1038/s41591-018-0300-7

[2] Arrieta, A. B. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*. 58. P. 82–115. DOI: 10.1016/j.inffus.2019.12.012

[3] Вербицький, О. (2024). Аналіз методів поясності нейронних мереж для подолання проблеми чорного ящика. *Збірник матеріалів конференції ITMAS – 2024: Інформаційні технології: моделі, алгоритми, системи* / Національний університет кораблебудування імені адмірала Макарова, м. Миколаїв, Україна, 30–31 жовтня 2024. С. 48–52.

**OPTIMIZING XAI METHODS FOR AI-BASED MEDICAL DIAGNOSTICS**

**Verbytskyi O.S., Gaidaienko O.V.**
Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine.

**Abstract**. The study analyzes explainable AI (XAI) methods in medical diagnostics, focusing on LIME, Grad-CAM, and SHAP. It reveals that existing approaches face limitations in interpretability and integration with clinical standards. Grad-CAM (for imaging) and LIME (for tabular data) are identified as optimal due to their transparency and compatibility with DICOM/EHR. The proposed system architecture and evaluation criteria aim to enhance trust in AI-driven decisions among medical professionals.

**Keywords**: explainability of neural networks; artificial intelligence in medicine; medical diagnostics; transparency of AI models; medical data analysis.

**УДК 004.032.26:004.89**

# LOCALIZATION AND GENERATIVE CONTROL OF CONCEPT-SPECIFIC NEURONS IN DEEP NEURAL NETWORKS

**Verbytskyi O.S.**
*M.Sc. Student in Computer Science*
*Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine*
*soomoalex@gmail.com*
*ORCID: 0009-0000-6848-8880*
**Gaidaienko O.V.**
*Ph.D. in Technology, Associate Professor,*
*Admiral Makarov National University of Shipbuilding, Mykolayiv, Ukraine*
*oksana.gaidaienko@nuos.edu.ua*
*ORCID: 0000-0002-6614-5443*

**Abstract**. Deep neural networks (DNNs) remain "black boxes" due to the lack of interpretable mechanisms for concept representation. This study proposes a method to localize minimal sets of concept-specific neurons (e.g., "cat") using L0-optimization heuristics and activation masking, enabling generative control and pareidolia analysis. Experiments on VGG16 (ImageNet) confirm the approach's efficacy: ablation of identified neurons reduces class accuracy by >75%, while their activation on noisy images reveals human-like illusory pattern recognition.

**Keywords**: explainable ai, concept neurons, activation masking, neuron ablation, pareidolia.

**Introduction**. Deep neural networks (DNNs) have achieved breakthrough performance in computer vision tasks, yet their internal decision-making mechanisms remain opaque. This "black box" nature impedes trust and reliability in critical applications, from medical diagnostics to autonomous systems [1]. A core challenge lies in identifying how high-level concepts (e.g., "cats")

**СЕКЦІЯ 7. ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА УПРАВЛІННЯ ПРОЕКТАМИ**