УДК 681.784.7:615.849.19

ЮЛІЯ ПИЛИПЕЦЬ, ЯРОСЛАВ ЯРОСЛАВСЬКИЙ, ОЛЕКСАНДР ВОЛОСОВИЧ

ОСОБЛИВОСТІ ВИКОРИСТАННЯ EXPLAINABLE AI У БІОМЕДИЧНІЙ ОБРОБЦІ ЗОБРАЖЕНЬ: ПРОЗОРІСТЬ ТА ІНТЕРПРЕТОВАНІСТЬ МОДЕЛЕЙ

Вінницький національний технічний університет, м. Вінниця, Україна,
e-mail: 00-24-043.stud@vntu.edu.ua
Військово-медичний клінічний центр Центрального регіону, м. Вінниця, Україна
Національний університет «Одеська Політехніка»
ДП «Вінницький науково-дослідний та проектний інститут землеустрою»

Анотація. Штучний інтелект (ШІ) глибоко інтегрувався в численні наукові галузі, включаючи біомедичну обробку зображень і сигналів. Зростаючий інтерес до цієї галузі призвів до сплеску досліджень, про що свідчить різке зростання наукової активності. Використовуючи великі і різноманітні набори біомедичних даних, моделі машинного навчання і глибокого навчання трансформували різноманітні завдання - такі як моделювання, сегментація, реєстрація, класифікація і синтез - часто перевершуючи ефективність традиційних методів.

Тим не менш, основна проблема залишається: складність перекладу результатів, отриманих за допомогою ШІ, в клінічно або біологічно значущі рішення, що обмежує практичну корисність цих моделей. Пояснюваний ШІ (Explainable AI, XAI) прагне подолати цю прогалину, покращуючи інтерпретованість систем ШІ та пропонуючи прозорі пояснення їхніх рішень. Для вирішення цієї проблеми розробляється все більше підходів, і інтерес до цієї теми в науковому співтоваристві продовжує зростати.

Ключові слова: Штучний інтелект (ШІ), пояснювальний ШІ (XAI), біомедична обробка зображень, біомедична обробка сигналів, глибоке навчання, машинне навчання, інтерпретованість, прозорість моделі, підтримка прийняття клінічних рішень, сегментація, класифікація, моделі на основі даних.

Abstract. Artificial intelligence (AI) has become deeply integrated into numerous scientific fields, including biomedical image and signal processing. The growing interest in this field has led to a surge in research, as evidenced by the sharp increase in scientific activity. Using large and diverse biomedical datasets, machine learning and deep learning models have transformed a variety of tasks - such as modeling, segmentation, registration, classification, and synthesis - often outperforming traditional methods.

However, a major challenge remains: the difficulty of translating AI-derived results into clinically or biologically meaningful solutions, which limits the practical utility of these models. Explainable AI (XAI) seeks to bridge this gap by improving the interpretability of AI systems and offering transparent explanations for their decisions. More and more approaches are being developed to address this problem, and interest in the topic in the scientific community continues to grow.

Keywords: Artificial Intelligence (AI), Explanatory AI (XAI), biomedical image processing, biomedical signal processing, deep learning, machine learning, interpretability, model transparency, clinical decision support, segmentation, classification, data-driven models.

DOI: 10.31649/1681-7893-2025-50-2-210-214

АКТУАЛЬНІСТЬ ДОСЛІДЖЕННЯ

Останніми роками штучний інтелект (ШІ), зокрема, глибоке навчання (DL) і машинне навчання (ML), викликав неабиякий інтерес завдяки революційним досягненням у цій галузі. Ці технології набувають все більшого поширення в різних галузях, включаючи біомедичну обробку сигналів і зображень, де інтеграція мультимодальних, багатовимірних і багатопараметричних даних створює значні труднощі для звичайних аналітичних методів.

Незважаючи на те, що системи штучного інтелекту розробляються для підтримки користувачів у виконанні рутинних завдань, вони часто стикаються зі скептицизмом щодо їхніх результатів. Брак довіри з боку користувачів суттєво перешкоджає або затримує їхнє широке впровадження та практичне застосування. Щоб вирішити цю проблему, зростає потреба в методах, які роблять моделі ШІ більш прозорими, зрозумілими та верифікованими, узгоджуючи їхні прогнози з усталеними знаннями або асоціативними дослідженнями.

У цьому контексті пояснюваний ШІ (Explainable AI, XAI) стає центральним елементом в еволюції біомедичних систем, керованих ШІ, відіграючи вирішальну роль у розвитку наступного покоління інтелектуальних технологій, орієнтованих на людину. Ця стаття пропонує огляд ключових досягнень галузі ШІ у біомедичній сфері, що охоплює як методологічні підходи, так і способи застосування. Основна увага приділяється розбору компонентів алгоритмів і з'ясуванню взаємодії між моделями і даними для поліпшення розуміння і зручності використання.

Мета дослідження. Досліждення використання Explainable AI (XAI) у біомедичній обробці зображень.

ОСНОВНІ МАТЕРІАЛИ ДОСЛІДЖЕНЬ

Пояснюваний штучний інтелект (Explainable Artificial Intelligence, XAI) останнім часом привертає значну увагу як ключовий підхід до розуміння "чому" і "як", що стоять за прогнозами, які генеруються моделями машинного навчання (ML) і глибокого навчання (DL). Кілька комплексних оглядів [1, 3, 4, 5] надають глибокий огляд галузі, окреслюючи основні концепції, пропонуючи таксономії та класифікуючи сучасні методи відповідно до них.

Перш ніж заглиблюватися в конкретні методи, важливо прояснити термінологію. Як зазначено в джерелах [3], інтерпретованість означає узгодження результатів моделі з людською інтуїцією, що допомагає зрозуміти причинно-наслідкові зв'язки між входами і виходами - поняття, тісно пов'язане з причинно-наслідковим зв'язком. На відміну від цього, пояснюваність стосується здатності розшифрувати внутрішні механізми моделі, які не завжди можуть бути інтуїтивно зрозумілими або зрозумілими з людської точки зору. Таким чином, інтерпретованість за своєю суттю не означає пояснюваність, і навпаки.

Для того, щоб система була дійсно інтерпретованою, вона повинна надавати пояснення, які є зрозумілими для людини, а характеристики, які роблять такі пояснення ефективними, повинні бути чітко визначені. Хоча консенсус щодо точних визначень все ще формується, загальна класифікація включає попередні та постфактум методи [4]. Попередні методи за своєю суттю є інтерпретованими, в той час як постфактум методи намагаються пояснити поведінку непрозорих моделей «чорної скриньки» після навчання. Ці методи можуть бути діагностичними або специфічними для конкретної моделі, а їхня сфера застосування може бути як локальною (зосередженою на окремих прогнозах), так і глобальною (спрямованою на загальну поведінку моделі).

Серед найпоширеніших підходів до інтерпретації post-hoc є методи атрибуції ознак, такі як SHAP (SHapley Additive exPlanations) [6] та LIME (Local Interpretable Model-Agnostic Explanations) [7]. Для DL-моделей зазвичай застосовують градієнтні методи. До них належать SmoothGrad, Class Activation Mapping (CAM), Grad-CAM, DeepLIFT (Deep Learning Important FeaTures) [8] і Layer-wise Relevance Propagation (LRP) [9]. Ці методи зазвичай генерують візуалізації або теплові карти, які виділяють області вхідних даних - наприклад, певні ділянки зображення - які найбільше вплинули на вихідні дані моделі.

Біомедичні дані охоплюють широкий спектр модальностей та різні просторові та часові масштаби. За своєю суттю вони неоднорідні і часто схильні до впливу шуму, пропущених значень і сильної залежності від обладнання для збору даних. Таке розмаїття вимагає спеціальних алгоритмічних рішень, здатних впоратися з такою мінливістю [2].

Ці типи даних варіюються від мікроскопічних молекулярних даних до омічних (наприклад, геномних, протеомних, транскриптомних, метаболомних), медичних зображень, клінічних звітів та електронних медичних записів. Якість і формат даних широко варіюються, включаючи аналогові, цифрові і текстові форми, зі складними структурами, такими як послідовності, дерева або графіки різного розміру. Розвиток високопродуктивних технологій, таких як секвенування наступного покоління або вдосконалена візуалізація (наприклад, функціональна та дифузійно-зважена МРТ), призвів до вибуху біомедичних даних, що вимагають спеціалізованих стратегій обробки.

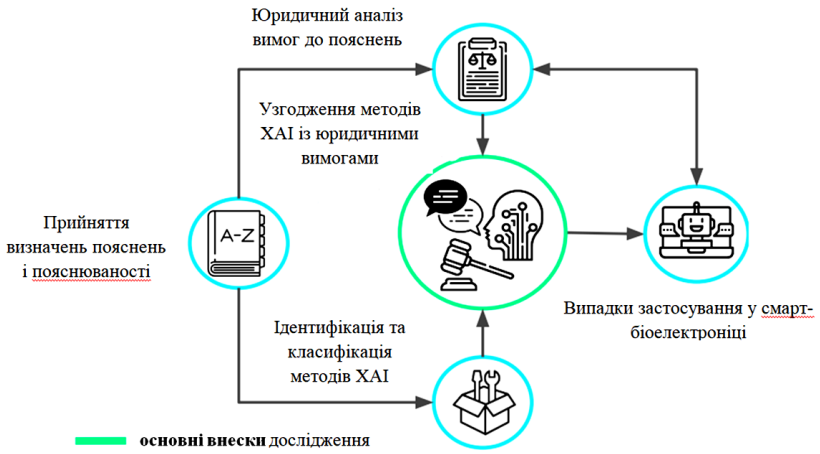

Рисунок 1 – Схематичний огляд методології дослідження щодо інтеграції правових вимог та інструментів XAI

Незважаючи на поширеність великих біомедичних даних, доступ до їх високоякісних наборів залишається непослідовним. Методи глибокого навчання (DL) є природним вибором для управління такими складними великими обсягами даних. Однак загальною проблемою в біомедичному контексті є обмежена доступність, особливо на місцевому або інституційному рівнях. Щоб пом'якшити цю проблему, було впроваджено різноманітні підходи, включаючи методи регуляризації (наприклад, відсіювання), ранню зупинку, напівконтрольоване навчання, яке використовує як марковані, так і немарковані дані, а також доповнення даних - розширення наборів даних за допомогою штучних перетворень або введення шуму.

Однак ці стратегії маніпулювання даними додають ще один рівень складності, особливо в біомедичній галузі, де будь-яке перетворення має залишатися біологічно та клінічно обґрунтованим. Хоча цеі методи не намагаються вирішити всі проблеми, притаманні обробці біомедичних даних, вони зосереджується саме на інтерпретованості. Біомедичні дані становлять величезний виклик для пояснювального ШІ (XAI), вимагаючи моделей, здатних впоратися з невизначеністю, і суворою перевіркою їхніх результатів - області, яка залишається недостатньо вивченою.

У біомедичній галузі інтерпретованість є ключовою вимогою, тоді як валідація залишається основною проблемою [10]. Як мінімум, моделі ШІ повинні бути зрозумілими для людини, надійними та надійними. Після того, як модель генерує пояснення, вони повинні пройти належну валідацію, що особливо важливо в біомедичних контекстах і включає в себе кілька вимірів.

Про пояснюваність в XAI. Щоб вважатися надійними, пояснення повинні не тільки виглядати правдоподібними з біомедичної точки зору, але й демонструвати узгодженість між різними реалізаціями, включаючи різні методи і архітектури. Крім того, вони повинні бути стійкими до заплутаних змінних і демонструвати статистичну значущість. Одна з найпоширеніших методик перевірки інтерпретованості передбачає використання теплових карт, які візуально показують області вхідних даних, найбільш відповідальні за вплив на рішення моделі.

Комплексний приклад такої валідації [11], де для пояснення результатів моделі в задачі стратифікації пацієнтів, що розрізняє прогресуючу і рецидивуючо-ремітуючу форми розсіяного склерозу, було застосовано пошарове поширення релевантності (Layer-wise Relevance Propagation, LRP). Процес валідації не тільки оцінював потенційний вплив чинників, але також включав кореляційний аналіз між картами LRP і мікроструктурними маркерами, які, як відомо, відрізняються між двома підтипами захворювання.

В іншому прикладі [12-16] порівнюються методи XAI на основі ознак і на основі прикладів з використанням набору даних UCI Heart Disease Dataset (посилання Kaggle) для порівняння їхніх переваг і обмежень.

Подальша перспектива[13], де стверджується, що пояснюваність слід перефразувати як "ефективну оспорюваність". З точки зору пацієнтоцентризму, автори вважають, що люди повинні мати право і можливість оскаржувати діагнози, поставлені штучним інтелектом. Для того, щоб це мало сенс, системи повинні надавати доступ до різних форм інформації, зокрема про те, як використовуються дані, потенційні алгоритмічні упередження і загальну продуктивність системи.

СПОСОБИ ЗАСТОСУВАННЯ

Нейровізуалізація: Старіння мозку

Оцінка віку мозку за допомогою нейровізуалізації привернула увагу завдяки своєму потенціалу відрізняти старіння від специфічних для захворювання змін. Для інтерпретації модельних прогнозів використовуються такі методи, як лінійні моделі, PCA/ICA (ante-hoc) і новіші методи post-hoc, такі як карти важливості ознак і значущості. Останні підходи до глибокого навчання, включаючи SmoothGrad і Grad-CAM, виділяють ділянки мозку, пов'язані з прогнозованим віком. Хоча моделі, що піддаються поясненню, є поширеними, пояснення не завжди вдається отримати. Поступово візуалізація в латентному просторі та методи на основі перестановок стають ключовими інструментами для розуміння поведінки моделі.

XAI в імунотерапії раку (mIF Images)

Імунотерапія раку спирається на просторове профілювання тканин для аналізу оточення пухлини. Для інтерпретації зображень мультиплексної імунофлуоресценції була розроблена модель глибокого навчання під назвою CGAT (Cell Graph Attention Network). Вона класифікує ступінь тяжкості захворювання та визначає ключові клітинні структури, що впливають на прогнози, особливо на межі пухлини. Цей підхід підвищує точність і пояснюваність, підтримуючи експертну інтерпретацію в діагностиці раку. У біомедичній галузі інтерпретованість є ключовою вимогою, в той час як валідація залишається основним викликом [10]. Як мінімум, моделі ШІ повинні бути зрозумілими для людини, надійними та надійними. Після того, як модель генерує пояснення, вони повинні пройти належну валідацію, що особливо важливо в біомедичному контексті і включає в себе кілька вимірів.

ВИСНОВКИ

Пояснюваний штучний інтелект (Explainable Artificial Intelligence, XAI) охоплює набір методів, спрямованих на те, щоб допомогти людям інтерпретувати, довіряти та оцінювати рішення, прийняті моделями машинного навчання. У біомедичній галузі XAI відіграє життєво важливу роль у забезпеченні точності, прозорості та надійності діагнозів і прогнозів за допомогою ШІ. Однак впровадження XAI в біомедичну обробку сигналів і зображень залишається складним завданням.

Одна з головних проблем полягає у відсутності узгоджених визначень і концептуальної ясності - проблема, що корениться у філософських, психологічних та етичних міркуваннях. Після того, як ці основоположні концепції будуть встановлені, виникає необхідність визначення вимірюваних атрибутів і рамок валідації. Існують різні підходи, включаючи типи валідації, засновані на людському вкладі (прикладні, засновані на людині) і ті, що використовують об'єктивні (функціональні, засновані на функціональності).

Валідація залишається критично важливим кроком: пояснення повинні бути надійними, узгодженими між різними впровадженнями та узгодженими з попередніми клінічними знаннями. Відсутність фундаментальної істини додає ще один рівень складності, що вимагає рішень, специфічних для конкретної галузі. Невизначеність також повинна бути кількісно оцінена, оскільки вона є наслідком дизайну моделі, якості даних і методологічних відмінностей.

Крім того, часто виникає компроміс між продуктивністю і інтерпретованістю. Включення знань з біомедичної галузі в моделі ML/DL може допомогти збалансувати це протиріччя і покращити пояснюваність. Однак такі проблеми, як упередженість навчання, коли результати моделі спотворюються через незбалансовані дані, неправильні конфігурації або невідповідності між моделлю і даними, залишаються важливими. Ці упередження можуть призвести до недостовірних результатів і викликати занепокоєння, пов'язані з «безпекою навчання» ШІ.

Для вирішення цих проблем досліджуються такі методи, як навчання на основі правил, візуалізації процесів і оцінки моделей, орієнтованих на людину. Незважаючи на досягнутий прогрес, необхідні додаткові дослідження та валідація, щоб забезпечити інтерпретацію, надійність і клінічну значущість методів ШІ в біомедичних науках.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ / REFERENCES

1. Almpani S., Kiouvrekis Y., Stefaneas P., Frangos P. Computational argumentation for medical device regulatory classification // *International Journal on Artificial Intelligence Tools*. – 2022.
2. Aydemir B., Hoffstetter L., та ін. Tempsal-uncovering temporal information for deep saliency prediction // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. – 2023.

3. Bach S., Binder A., Montavon G., Klauschen F., Müller K.-R., Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation // *PloS one*. – 2015.
4. Bastani O., Kim C., Bastani H. Interpretability via model extraction // *Fairness, Accountability, and Transparency in Machine Learning Workshop*. – 2017.
5. Bernal J., Mazo C. Transparency of artificial intelligence in healthcare: insights from professionals in computing and healthcare worldwide // *Applied Sciences*. – 2022.
6. Bibal A., Lognoul M., De Streel A., Frénay B. Legal requirements on explainability in machine learning // *Artificial Intelligence and Law*. – 2021.
7. European Data Protection Board. Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679 (WP251rev.01). – 2018. – Режим доступу: https://ec.europa.eu/newsroom/article29/items/612053/en.
8. Bondarenko A., Aleksejeva L., Jumutc V., Borisov A. Classification tree extraction from trained artificial neural networks // *Procedia Computer Science*. – 2017.
9. European Commission. General Data Protection Regulation, 2016. – Режим доступу: https://eur-lex.europa.eu/eli/reg/2016/679/oj.
10. European Commission. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices. – 2017. – Режим доступу: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32017R0745.
11. European Commission. Artificial Intelligence Act. – 2024. – Режим доступу: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689.
12. Dash S., Gunluk O., Wei D. Boolean decision rules via column generation // *Advances in Neural Information Processing Systems*. – 2018.
13. Dhurandhar A., Chen P.-Y., Luss R., Tu C.-C., Ting P., Shanmugam K., Das P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives // *Advances in Neural Information Processing Systems*. – 2018. – Vol. 31 (NIPS).
14. Pavlov S. V. Information Technology in Medical Diagnostics //Waldemar Wójcik, Andrzej Smolarz, July 11, 2017 by CRC Press - 210 Pages.
15. Wójcik W., Pavlov S., Kalimoldayev M. Information Technology in Medical Diagnostics II. London: (2019). Taylor & Francis Group, CRC Press, Balkema book. – 336 Pages.
16. Y. Pylypets, S. Pavlov, Y. Yaroslavsky, S. Kostyuk, and M. Ursan, “Features of the application of telemedical technologies based on artificial intelligence in disaster medicine,” Opt-el. inf-energ. tech., vol. 48, no. 2, pp. 190–195, Nov 2024.

Надійшла до редакції 10.09.2025 р.

ПИЛИПЕЦЬ ЮЛІЯ ОЛЕКСАНДРІВНА – аспірант кафедри біомедичної інженерії та оптикоелектронних систем, Вінницький національний технічний університет, Вінниця, Україна, офіцер відділення зв’язку, Військово-медичний клінічний центр Центрального регіону, м. Вінниця, Україна,
e-mail: y.hopanchuk@med.mil.ua

ЯРОСЛАВСЬКИЙ ЯРОСЛАВ ІВАНОВИЧ – к.т.н., старший викладач кафедри біомедичної інженерії, Національний університет «Одеська Політехніка, директор, ДП «Вінницький науково-дослідний та проектний інститут землеустрою», Вінниця, Україна,
e-mail: yaroslavskyidzk@gmail.com

ВОЛОСОВИЧ ОЛЕКСАНДР СЕРГІЙОВИЧ – аспірант кафедри біомедичної інженерії та оптико-електронних систем, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, Україна, 21021; ***e-mail: sashka.v0@gmail.com***

YULIIA PYLYPETS, YAROSLAV YAROSLAVSKYY, OLEKSANDR VOLOSOVYCH

FEATURES OF USING EXPLAINABLE AI IN BIOMEDICAL IMAGE PROCESSING: TRANSPARENCY AND INTERPRETABILITY OF MODELS

Military Medical Clinical Center of the Central Region, Vinnytsia, Ukraine
Vinnytsia National Technical University, Vinnytsia, Ukraine
Odessa Polytechnic National University,
SE "Vinnytsia Research and Design Institute of Land Management", Vinnytsia, Ukraine