УДК 004.8
DOI https://doi.org/10.32782/2663-5941/2024.6.2/16

Марчук Д.М.
Національний університет «Львівська політехніка»

Фесюк І.І.
Національний університет «Львівська політехніка»

Роса Т.В.
Національний університет «Львівська політехніка»

Карпін О.О.
SRI ASR "Infineon Technologies"

Максимюк Т.А.
Національний університет «Львівська політехніка»

ДОСЛІДЖЕННЯ ІНФОРМАТИВНОСТІ ТА ІНТЕРПРЕТОВАНОСТІ АЛГОРИТМІВ ШТУЧНОГО ІНТЕЛЕКТУ

У сучасних дослідженнях штучного інтелекту проблема забезпечення інтерпретованості моделей набуває дедалі більшої ваги, особливо у критично важливих галузях, таких як медицина, фінансовий сектор та правова сфера. Значна частина моделей машинного навчання все ще функціонує як «чорні скриньки», що обмежує рівень довіри до прийнятих ними рішень та ускладнює їх упровадження на практиці. Потреба у пояснюваному штучному інтелекті (xAI) є відповіддю на ці виклики, адже зрозуміле тлумачення логіки та механізмів функціонування моделі є ключовим чинником для забезпечення прозорості та прийнятності її висновків. У статті запропоновано новий підхід до інтеграції навчання з підкріпленням із залученням зворотного зв'язку від людини (RLHF) у модель класифікації. На відміну від традиційних методів, де оновлення параметрів моделі здійснюється шляхом оптимізації функції втрат, у межах запропонованого підходу корекцію виконано безпосередньо на рівні логітів. Такий підхід базується на врахуванні реалістичного зворотного зв'язку, який моделює реакцію користувача на певний прогноз моделі. Це дозволяє не лише підвищити точність класифікації, але й надати дослідникові інструментарій для відстеження впливу зворотного зв'язку на кінцевий результат, тим самим сприяючи покращенню інтерпретованості. Експериментальні дослідження показали, що інтеграція RLHF підвищила точність класифікації з 59,20% до 88,70%. Аналіз гістограм розподілу ймовірнісних оцінок підкреслив зростання впевненості моделі та зменшення невизначеності під час прийняття рішень. Застосування метрики ROC-AUC продемонструвало здатність методу не лише покращувати точність, але й забезпечувати адекватний баланс між чутливістю і специфічністю класифікації. Запропонований підхід також вирізняється стійкістю до змінних умов та можливістю врахування зовнішніх сигналів без суттєвого збільшення обчислювальної складності. Подальші дослідження можуть бути зосереджені на динамічному налаштуванні параметрів, інтеграції складніших функцій винагороди та розширенні галузі застосування методу на багатовимірні задачі або великі реальні датасети.

Ключові слова: *xAI, RLHF, модель класифікації.*

Постановка проблеми. Сучасний розвиток штучного інтелекту (ШІ) супроводжується зростаючою потребою в методах, здатних забезпечити прозорість та інтерпретованість результатів автоматизованого аналізу даних. Попри значні успіхи у підвищенні точності моделей машинного навчання, багато з них функціонують як «чорні скриньки», позбавляючи користувачів розуміння внутрішніх механізмів прийняття рішень. Така ситуація створює суттєві перешкоди для застосування ШІ в критично важливих сферах, таких як медицина, фінанси та право, де обґрунтована довіра та здатність пояснити поведінку системи є вирішальними чинниками. Поява концепції пояснюваного штучного інтелекту (Explainable AI, xAI) покликана вирішити дану проблему, шляхом розроблення інструментів для аналізу рішень штучного інтелекту та підвищення рівня їх прийнятності та прозорості.

Іншим важливим аспектом є те, що традиційні методи оптимізації моделей штучного інтелекту

переважно не враховують складність динамічних середовищ, де моделі повинні оперативно адаптуватися до змінних умов та інтегрувати зовнішні сигнали, включно з поведінкою та оцінками користувачів. У такому контексті інтеграція навчання з підкріпленням із залученням людського зворотного зв'язку (Reinforcement Learning from Human Feedback, RLHF) постає перспективним напрямком. RLHF розширює можливості системи, дозволяючи їй навчатися не лише за рахунок формальних функцій втрат, а й на основі реалістичних, суб'єктивних оцінок користувачів. Це, у свою чергу, надає можливість не просто підвищити адаптивність моделей, а й глибше зрозуміти вплив зовнішніх сигналів на формування кінцевих висновків, тим самим сприяючи зміцненню довіри до ШІ та підвищенню інтерпретованості.

Аналіз останніх досліджень і публікацій. Інтерпретованість моделей штучного інтелекту стала одним із ключових напрямків сучасних досліджень у галузі машинного навчання [1]. Зі зростанням складності моделей, особливо нейронних мереж глибокого навчання, виникає потреба у методах, які дозволяють пояснити та зрозуміти процес прийняття рішень цими моделями. Це важливо не лише для довіри користувачів, але й для виявлення потенційних упереджень та забезпечення етичності систем ШІ [2].

Одним із головних напрямків є розвиток поясненого штучного інтелекту (xAI). Методи xAI спрямовані на розкриття внутрішніх механізмів моделей та надання зрозумілих пояснень їхніх рішень [3]. Серед популярних підходів варто зазначити метод, що базується на теорії ігор для оцінки внеску кожної ознаки у передбачення моделі (SHAP – SHapley Additive exPlanations), а також механізми уваги (Attention mechanisms), які використовуються в нейронних мережах для фокусування на важливих частинах вхідних даних, що полегшує інтерпретацію моделей [4].

В останні роки набуває популярності підхід навчання з підкріпленням із залученням зворотного зв'язку від людини (RLHF – Reinforcement Learning from Human Feedback), який дозволяє створювати інтерпретовані моделі [5]. Перевагою даного методу, є те, що він дозволяє моделям навчатися не лише на основі алгоритмічних функцій втрат, але й враховувати оцінки та побажання користувачів. Це особливо важливо для забезпечення відповідності моделей етичним нормам та очікуванням суспільства.

Зокрема, компанія OpenAI активно досліджує RLHF у контексті великих мовних моделей, таких як GPT-3 та GPT-4 [6]. Вони показали, що RLHF може суттєво покращити якість згенерованого тексту, роблячи його більш релевантним та етичним [7]. Завдяки інтеграції зворотного зв'язку від людини, моделі стали краще розуміти контекст та потреби користувачів, що підвищило їхню практичну цінність. Останні дослідження вказують на те, що поєднання RLHF та методів XAI може сприяти підвищенню інтерпретованості моделей [8]. Наприклад, інтерактивні системи, де користувачі можуть надавати зворотний зв'язок щодо роботи моделі, дозволяють краще зрозуміти її поведінку та вплив окремих факторів на рішення. Залучення зворотного зв'язку може призвести до формування більш зрозумілих та інтерпретованих внутрішніх станів моделі, що полегшує їхнє пояснення та аналіз [9]. Це особливо актуально в медичній діагностиці, фінансовому аналізі та інших сферах, де точність та інтерпретованість рішень є критичними [10]. Аналіз останніх досліджень показує, що інтеграція RLHF є перспективним шляхом для підвищення інтерпретованості та адаптивності моделей ШІ. Поєднання методів пояснюваного штучного інтелекту з навчанням на основі зворотного зв'язку від людини може привести до створення більш довірених та ефективних систем [11]. Подальші дослідження у цьому напрямку є важливими для розвитку ШІ, який відповідає потребам суспільства та етичним стандартам [12].

Постановка завдання. У сучасних задачах машинного навчання традиційні методи оптимізації часто не враховують складність реальних сценаріїв, де потрібна адаптивність і врахування зовнішніх сигналів. Метою даної статті є інтеграція підходу RLHF у типову модель класифікації, яка часто використовується у критичних сферах, для підвищення її ефективності та інтерпретованості. Замість традиційного оновлення ваг моделі через оптимізацію функції втрат, корекція здійснюється безпосередньо на рівні логітів шляхом додавання реалістичного зворотного зв'язку. Такий підхід не лише підвищує точність моделі, але й надає можливість простежити, як зворотний зв'язок впливає на кінцеві передбачення, що є ключовим для інтерпретованого ШІ. Це, в свою чергу дасть змогу підвищити інтерпретованість моделей ШІ без значного збільшення обчислювальної складності.

Виклад основного матеріалу. Для задачі класифікації модель $\mathcal{M}$, визначається через обчислення ймовірності приналежності входу x до класу $y - p(y \mid x)$, з використанням набору логітів $\mathbf{z} \in \mathbb{R}^C$, де C – кількість класів. Для прикладу, роз-

глянемо модель класифікації, яка поєднує стандартну схему обчислення ймовірностей через функцію softmax з інтеграцією реалістичного зворотного зв'язку, який моделюється як реакція користувача на результати передбачень [13]. Основою для моделі без RLHF є класичний алгоритм softmax, який перетворює логіти на ймовірності, відображаючи рівень впевненості моделі у приналежності вхідного зразка до певного класу:

$$p(y = i \mid x) = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \quad (1)$$

Однак у стандартному підході ця функція не враховує додаткові зовнішні сигнали або зворотний зв'язок, що обмежує її ефективність у реалістичних сценаріях. Для розв'язання цієї проблеми в класичний алгоритм інтегровано RLHF, основна ідея якого полягає в корекції логітів моделі перед застосуванням softmax шляхом додавання зворотного зв'язку, що генерується на основі правильності передбачень. У цьому випадку логіти модифікуються з урахуванням зворотного зв'язку через функцію:

$$\mathbf{z}' = \mathbf{z} + \mathbf{f}_{\text{RLHF}}(\mathbf{z}, \mathbf{r}), \quad (2)$$

де $\mathbf{f}_{\text{RLHF}}$ – функція адаптації логітів залежно від зворотного зв'язку $\mathbf{r}$. Вектор $\mathbf{r} = [r_1, \ldots, r_C]$ представляє реакцію системи на результати передбачень, причому:

$$r_i = \begin{cases} +\delta & \textit{для правильних передбачень,} \\ -\delta & \textit{для неправильних передбачень.} \end{cases}$$

Застосування RLHF змінює обчислення ймовірностей у функції softmax, оновлюючи (1) до вигляду:

$$p(y = i \mid x, \mathbf{r}) = \frac{\exp(z_i + r_i)}{\sum_{j=1}^{C} \exp(z_j + r_j)} \quad (3)$$

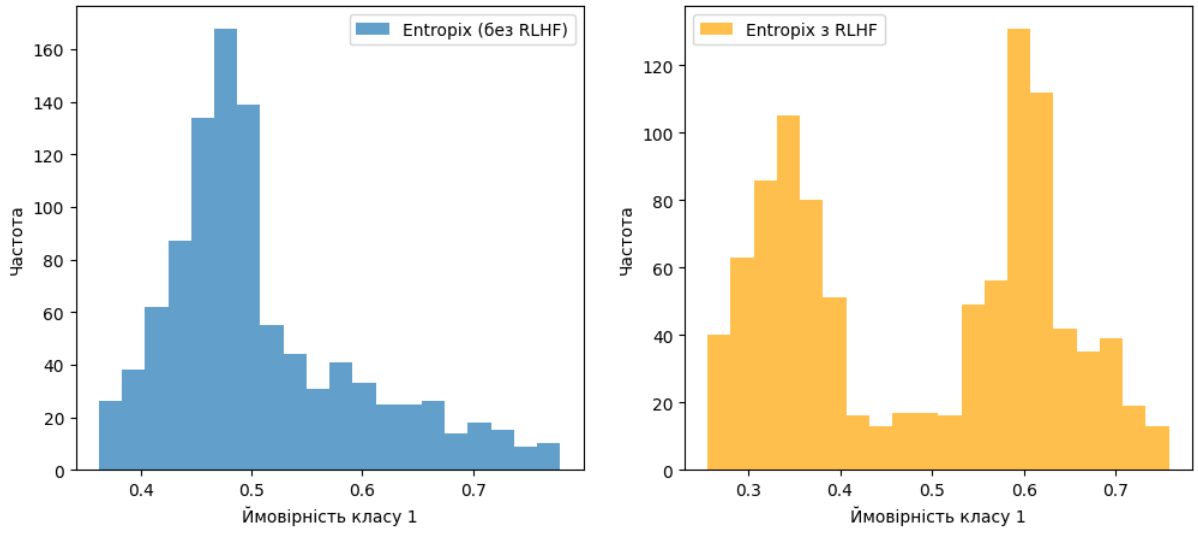
Це дозволяє моделі враховувати не лише базові логіти, але й коригувати свої передбачення на основі наданого зворотного зв'язку. Під час навчання така модель оптимізується з використанням наступної функції втрат:

$$\mathcal{L} = -\sum_{i=1}^{N} \log p(y^{(i)} \mid x^{(i)}, \mathbf{r}^{(i)}) \quad (4)$$

де N – кількість прикладів у навчальній вибірці. Функція втрат включає штраф за неправильні передбачення, що враховує внесок RLHF.

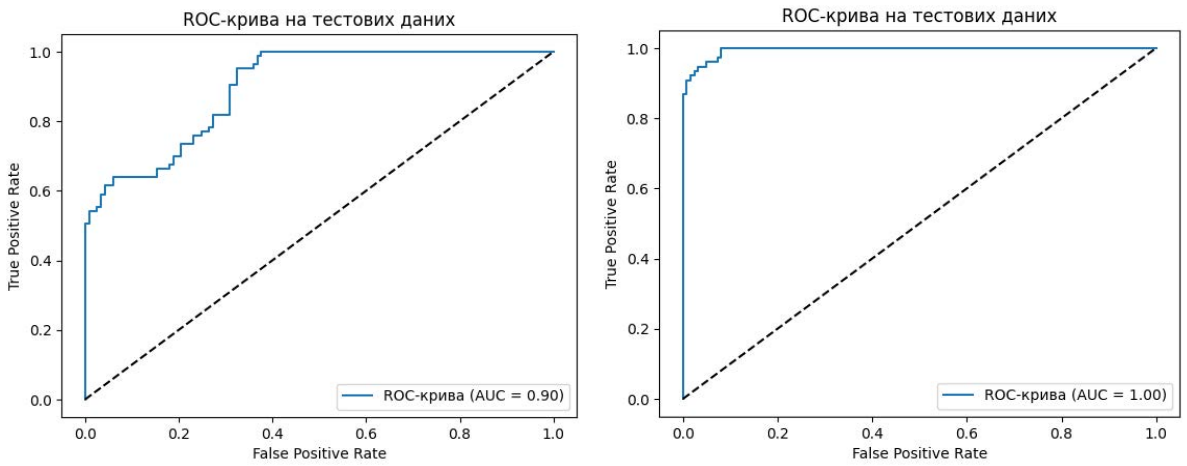
Додатково RLHF забезпечує адаптивність моделі через параметризацію зворотного зв'язку $\mathbf{r}$ за допомогою спеціальної моделі $\mathcal{F}$, яка залежить від логітів і історії передбачень $\mathbf{h}$. Формально це виражається як $r_i = \mathcal{F}(z_i, h_i)$. Такий підхід дозволяє динамічно коригувати впевненість моделі відповідно до реалістичного зворотного зв'язку користувача, що значно підвищує точність, стійкість і адаптивність системи в умовах складних задач класифікації.

Результати експериментів чітко демонструють ефективність RLHF. Точність моделі без застосування RLHF становила 59,20%, що відображає базовий рівень для порівняння. Після інтеграції RLHF точність зросла до 88,70%, оскільки модель навчилася коригувати свої помилки, використовуючи додаткову інформацію про правильність передбачень. Це підвищило її здатність адаптуватися до невизначеностей у даних та враховувати контекст під час прийняття рішень. Гістограми на рис. 1 демонструють розподіл впевненості



а) б)

Рис. 1. Гістограми розподілу впевненості моделі Entropix з RLHF і без RLHF



а) б)

Рис. 2. ROC характеристики моделі Entropix без використання RLHF а) та з використанням RLHF – б)

передбачень моделі у двох сценаріях: без RLHF та з його інтеграцією.

На першій гістограмі зображено розподіл ймовірностей класу 1 для моделі без використання RLHF. У цьому випадку модель базується на традиційному обчисленні ймовірностей через softmax без корекції логітів, що призводить до рівномірного розподілу ймовірностей із значним розкидом у значеннях. Такий розподіл свідчить про невпевненість моделі у своїх передбаченнях, часто оцінюючи класи з низькими або середніми ймовірностями. Подібний підхід може бути прийнятним для стандартних задач, але є недостатньо адаптивним для складних сценаріїв, де точність передбачень має вирішальне значення.

Друга гістограма демонструє розподіл ймовірностей для моделі із застосуванням RLHF. У цьому випадку корекція логітів на основі реалістичного зворотного зв'язку значно впливає на кінцеві ймовірності, що проявляється у зростанні ймовірностей у вищому діапазоні значень. Такий результат свідчить про підвищення впевненості моделі у своїх передбаченнях, зумовлене підсиленням правильних передбачень за допомогою позитивного зворотного зв'язку та зменшенням впливу помилкових передбачень через негативний зворотний зв'язок.

Порівняння двох гістограм підтверджує, що використання RLHF сприяє покращенню чіткості передбачень моделі, знижуючи невпевненість. Це особливо важливо у задачах із високою вартістю помилок, таких як медична діагностика, фінансові прогнози або критичні сценарії подвійного використання. Практична значимість отриманих результатів полягає у можливості швидкої адаптації моделі до змін умов або врахування зовнішніх сигналів. Наприклад, у медичній діагностиці RLHF може враховувати оцінки лікарів для підвищення точності, а у фінансових додатках – адаптуватися до ринкових змін. Подальше вдосконалення алгоритму може включати дослідження впливу динамічної зміни параметра alpha або інтеграцію більш складних функцій винагороди для врахування додаткових аспектів даних, а також оцінку ефективності підходу у багатовимірних задачах чи на великих наборах реальних даних.

Окрім того, проведено оцінку метрики ROC-AUC (рис. 2), що дозволило отримати більш повну характеристику ефективності моделі. Традиційна метрика точності відображає лише частку правильно класифікованих зразків, проте в задачах із дисбалансом класів вона може бути недостатньою. ROC-крива відображає здатність моделі розрізняти позитивні та негативні класи при зміні співвідношення між часткою правильно класифікованих позитивних прикладів (TPR) та часткою хибнопозитивних спрацьовувань (FPR).

Висновки. У даному дослідженні продемонстровано ефективність інтеграції навчання з підкріпленням із залученням зворотного зв'язку від людини (RLHF) у модель класифікації для підвищення її точності та інтерпретованості. Запропонований інноваційний підхід передбачає корекцію логітів моделі безпосередньо шляхом додавання реалістичного зворотного зв'язку, що дозволяє моделі враховувати додаткові зовнішні сигнали без значного збільшення обчислювальної складності. Експериментальні результати показали, що після інтеграції RLHF точність моделі зросла з 59,20% до 88,70%. Аналіз гістограм розподілу ймовірнос-

тей передбачень вказує на підвищення впевненості моделі та зменшення невизначеності у прийнятті рішень. Результати ROC-AUC підтвердили покращення здатності моделі розрізняти класи та балансувати між точністю і повнотою, що є особливо важливим у задачах з дисбалансом класів.

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.

2. Gunning D. Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*. 2017. URL: https://www.darpa.mil/program/explainable-artificial-intelligence

3. Lundberg S.M., Erion G., Lee S.I. Consistent Individualized Feature Attribution for Tree Ensembles. *IEEE International Conference on Machine Learning (ICML):* conference proceedings. (14-17 July 2020), 2020. P. 1–17.

4. Shrikumar A., Greenside P., Kundaje A. Learning Important Features Through Propagating Activation Differences. *IEEE International Conference on Machine Learning (ICML):* conference proceedings. (Sydney, Australia, 6-11 August 2020). Sydney. 2017. P. 3145–3153.

5. Knox W.B., Stone P. TAMER: Training an Agent Manually via Evaluative Reinforcement. *IEEE 7th International Conference on Development and Learning:* conference proceedings. (USA, 2008). USA. 2008. P. 292–297.

6. Radford A., Wu J., Child R. et al. Language Models are Unsupervised Multitask Learners. *OpenAI Technical Report*. 2019. URL: https://openai.com/blog/better-language-models

7. OpenAI. GPT-3: Language Models are Few-Shot Learners. *OpenAI Blog*. 2020. URL: https://openai.com/blog/gpt-3-apps

8. Milani S., Topin N., Veloso M., Fang F. Explainable Reinforcement Learning: A Survey and Comparative Review. *ACM Computing Surveys*. 2024. Vol. 56. No. 7. Article 168.

9. Heuillet A. et al. Explainability in Deep Reinforcement Learning. *Knowledge-Based Systems*. 2021. Vol. 214. Article 106685.

10. Ali S., Akhlaq F., Imran A. S., Kastrati Z., Daudpota S. M., Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Computers in Biology and Medicine*. 2023. Vol. 166. Article 107555.

11. Gilpin L. H. e al. Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA):* conference proceedings. (Turin, Italy, 1-4 October 2018). Turin. 2018. P. 80-89.

12. Samek W., Müller K.R. Towards Explainable Artificial Intelligence. *Lecture Notes in Computer Science*. 2019. Vol. 11700. P. 5–22.

13. Entropix. *GitHub*. 2023. URL: https://github.com/xjdr-alt/entropix.

Marchuk D.M., Fesiuk I.I., Rosa T.V., Karpin O.O., Maksymiuk T.A. **RESEARCH ON THE INFORMATIVENESS AND EXPLAINABILITY OF ARTIFICIAL INTELLIGENCE ALGORITHMS**

In current artificial intelligence (AI) research, the issue of ensuring model interpretability is gaining increasing attention, especially in critical fields such as medicine, finance, and law. Many machine learning models still function as "black boxes," which limits trust in their decisions and complicates their practical deployment. The need for explainable AI (xAI) arises as a direct response to these challenges, since a clear understanding of a model's logic and operational mechanisms is crucial for ensuring transparency and acceptance of its outputs. This paper introduces a new approach for integrating reinforcement learning from human feedback (RLHF) into a classification model. Unlike conventional methods, which update model parameters by optimizing loss functions, the proposed technique applies corrections directly at the logit level. This approach incorporates realistic user feedback, simulating human reactions to the model's predictions. As a result, it not only increases classification accuracy but also allows researchers to track how the feedback affects the final outcome, thereby improving interpretability. Experimental results show that integrating RLHF raises classification accuracy from 59.20% to 88.70%. Analyzing the histograms of probability distributions shows that the model's confidence increases while uncertainty decreases during decision-making. The use of the ROC-AUC metric confirms the method's ability not only to improve accuracy but also to maintain an appropriate balance between sensitivity and specificity. Moreover, the proposed approach is robust under changing conditions and can incorporate external signals without substantially increasing computational complexity. Future studies may focus on dynamically adjusting parameters, integrating more complex reward functions, and evaluating the method's performance on high-dimensional tasks or large real-world datasets.

Key words: *xAI, RLHF, classification model.*