https://doi.org/10.31891/2219-9365-2025-83-4
УДК 004.89:004.8

КУЧЕРУК Володимир
Уманський національний університет
https://orcid.org/0000-0002-6422-7779
e-mail: vladimir.kucheruk@gmail.com
КУЛАКОВ Павло
Уманський національний університет
https://orcid.org/0000-0002-0167-2218
e-mail: kulakovpi@gmail.com
ЛІЩУК Роман
Уманський національний університет
https://orcid.org/0000-0002-2051-5365
e-mail: roma0lir@gmail.com
КОНЦЕБА Сергій
Уманський національний університет
https://orcid.org/0000-0003-4161-5587
e-mail: kontseba@meta.ua
МАНЬКОВСЬКА Вікторія
Уманський національний університет
https://orcid.org/0000-0003-0552-5482
e-mail: viktoriasergiivna@gmail.com

КРИТЕРІЇ ЕФЕКТИВНОСТІ ТА ЯКОСТІ НЕЙРОННИХ МЕРЕЖ

У роботі проаналізовано ключові критерії ефективності штучних нейронних мереж: точність прогнозування, швидкість навчання, витрати ресурсів та здатність до узагальнення. Мета дослідження – систематизувати показники та методи порівняльної оцінки різних архітектур. Запропоновано експериментальну методологію оцінювання на реальних даних. Практична значущість – сприяння оптимальному підбору та налаштуванню мереж для прикладних задач, підвищення якості результатів при мінімальних витратах ресурсів.

Ключові слова: нейронні мережі; критерії ефективності; метрики якості; генералізаційні властивості; обчислювальна ефективність; стійкість і надійність; оцінювання моделей.

KUCHERUK Volodymyr, KULAKOV Pavlo, LISHCHUK Roman,
KONTSEBA Serhii, MANKOVSKA Viktoriya
Uman National University

CRITERIA FOR THE EFFICIENCY AND QUALITY OF NEURAL NETWORKS

The article examines the criteria for evaluating the efficiency and quality of artificial neural networks (ANNs), emphasizing the importance of balancing predictive performance with computational feasibility. In the era of rapidly growing data volumes and increasing model complexity, the need for systematic approaches to assessing both quality and efficiency becomes critical. The study provides a comprehensive classification of metrics into three major groups: quality metrics, efficiency metrics, and generalization properties. Quality criteria such as accuracy, precision, recall, F1-score, AUC-ROC, and regression-based measures (MSE, MAE, RMSE, R^2, MAPE) are analyzed as primary indicators of predictive reliability. At the same time, efficiency is measured through computational costs (training time, inference time, FLOPs, memory footprint, energy consumption), structural parameters (number of layers and parameters, compression potential), and practical adaptability. Generalization ability is addressed through overfitting and underfitting analysis, validation and test errors, cross-validation, and the bias–variance trade-off. Furthermore, the paper highlights the importance of robustness and reliability criteria, including sensitivity to noise, adversarial resistance, reproducibility, and stability across datasets. Integral evaluation approaches, such as weighted score, quality-to-cost ratio, and sustainability indices, are proposed as tools for holistic assessment of ANN performance in real-world environments. The practical significance of the study lies in enabling informed decision-making in architecture design and hyperparameter optimization, supporting efficient deployment of neural networks in domains ranging from real-time embedded systems and IoT devices to large-scale industrial and medical applications. The findings emphasize that a balanced multi-criteria framework not only ensures predictive accuracy but also promotes resource efficiency, scalability, and long-term reliability of AI solutions, thereby contributing to sustainable and context-aware artificial intelligence development.

Keywords: neural networks; efficiency criteria; quality metrics; generalization properties; computational efficiency; robustness and reliability; model evaluation

Стаття надійшла до редакції / Received 22.07.2025
Прийнята до друку / Accepted 14.08.2025

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

У сучасну епоху цифрових технологій нейронні мережі стали невід'ємним інструментом для вирішення широкого спектра завдань: від розпізнавання зображень і обробки природної мови до медичної

діагностики та автономного управління. Зростаюча складність моделей, збільшення обсягів даних і потреба в реальному застосуванні штучного інтелекту вимагають не лише високої якості результатів, але й ефективності обчислювальних рішень. Саме тому актуальним є системне використання критеріїв якості та ефективності нейронних мереж (НМ) [1].

Критерії якості - такі як точність (accuracy), повнота (recall), F1-міра, середньоквадратична помилка (MSE) тощо - дозволяють оцінити наскільки добре модель виконує свою основну задачу. Ці метрики є важливими на етапі побудови, налаштування та валідації НМ. Однак у реальних умовах висока якість моделі не завжди є достатньою - особливо коли йдеться про використання в системах з обмеженими ресурсами, таких як мобільні пристрої, вбудовані системи або промислове обладнання [1, 2].

У таких випадках на перший план виходять критерії ефективності, які враховують обчислювальні витрати**,** енергоспоживання**,** час роботи**,** кількість параметрів моделі**,** об'єм пам'яті**,** необхідний для розгортання тощо. У багатьох сферах - таких як робототехніка, інтернет речей (IoT), автономний транспорт чи телекомунікації - НМ мають працювати в режимі реального часу та демонструвати прийнятну якість за мінімальних витрат ресурсів [1, 2]. Саме тому виникає потреба в багатокритеріальному підході до оцінювання моделей, де якість і ефективність розглядаються у взаємозв'язку.

Актуальність цього питання також зумовлена розвитком таких напрямів, як:

- **AutoML та нейромережевий дизайн** (Neural Architecture Search), де критеріями вибору є не лише точність, а й продуктивність;
- **Green AI** - підхід, орієнтований на зменшення вуглецевого сліду та енергоспоживання штучного інтелекту;
- **Edge AI** - реалізація моделей штучного інтелекту на периферійних пристроях, де критичною є обмеженість обчислювальних потужностей.

Таким чином, використання критеріїв ефективності та якості є ключовим інструментом для раціональної оцінки, оптимізації та впровадження НМ у практичні застосування. Це дозволяє забезпечити баланс між точністю та ресурсною доцільністю, що є визначальним чинником у сучасних інтелектуальних системах.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Наведемо узагальнену класифікацію існуючих критеріїв ефективності та якості НМ**,** що охоплює основні аспекти оцінювання сучасних моделей штучного інтелекту є наступною.

Критерії оцінювання НМ поділяються на дві основні групи: критерії якості (якісні метрики) та критерії ефективності (ресурсні або структурні метрики). Така класифікація дає змогу комплексно аналізувати як продуктивність моделі, так і доцільність її використання в реальних умовах.

Критерії якості НМ (рис. 1) призначені для оцінювання правильності та надійності моделі виконувати своє основне завдання (класифікацію, регресію, тощо) [3].

Критерії якості НМ (рис. 1) призначені для оцінювання правильності та надійності моделі виконувати своє основне завдання (класифікацію, регресію, тощо) [3].

1. **Класифікаційні метрики** - це кількісні показники, що використовуються для оцінювання точності та якості роботи моделей машинного навчання при розв'язанні задач класифікації. До них відносяться:

1.1. **Accuracy (Точність)** - частка правильних передбачень.

Визначає частку правильно класифікованих прикладів серед усіх.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (1)$$

де TP – істинно позитивні, TN – істинно негативні, FP – хибно-позитивні, FN – хибно-негативні.

Ця метрика добре працює для збалансованих класів, але можуть бути хибні результати, якщо дані нерівномірно розподілені (наприклад, 95% одного класу).

1.2. **Precision (Точність позитивного передбачення)** - частка істинно позитивних серед усіх передбачених позитивними.

Відображає, яка частка передбачених позитивних прикладів є правильними.

$$Precision = \frac{TP}{TP+FP}. \quad (2)$$

Дана метрика використовується у випадках, коли потрібно мінімізувати кількість помилкових результатів (наприклад, у медичній діагностиці).

1.3. **Recall (Повнота)** - частка виявлених позитивних серед усіх реальних позитивних.

Показує, яка частка реальних позитивних прикладів була знайдена моделлю.

$$Recall = \frac{TP}{TP+FN}. \quad (3)$$

Дана метрика використовується у випадках, коли критично знайти всі позитивні випадки (наприклад, виявлення хвороб чи шахрайства).

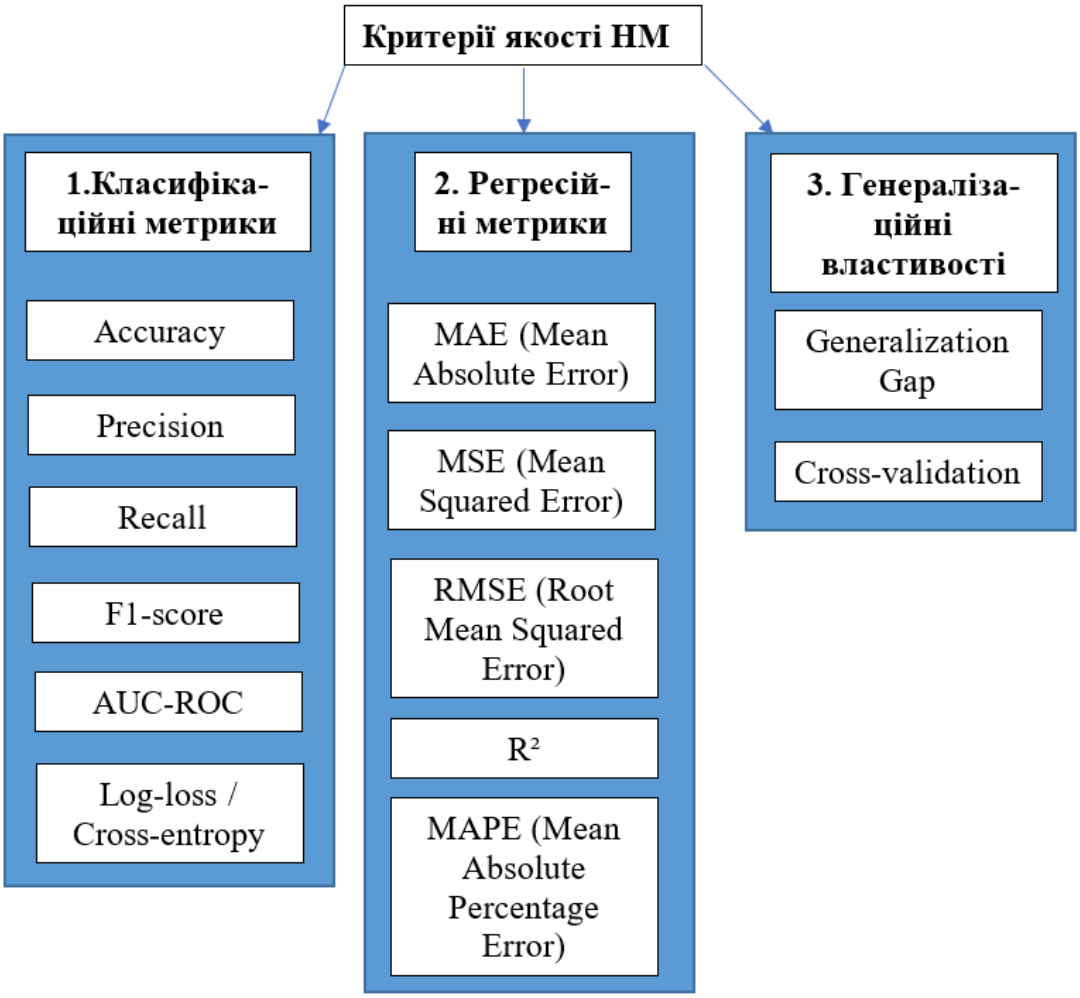

Рис. 1. Узагальнена класифікація критеріїв якості НМ

1.4. **F1-score** - гармонічне середнє між Precision та Recall.
Гармонічне середнє між Precision і Recall.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \quad (4)$$

Використовується, коли важливо знайти баланс між Precision і Recall. Добре підходить для оцінки моделей у сильно незбалансованих класах.

1.5. **AUC-ROC** - площа під ROC-кривою, що відображає здатність моделі відрізняти класи.

ROC-крива показує співвідношення True Positive Rate (Recall) та False Positive Rate при різних порогах класифікації. AUC – площа під цією кривою (від 0 до 1). Чим ближче до 1, тим краще модель розрізняє класи. При дуже незбалансованих класах краще використовувати Precision-Recall AUC.

На рис. 2 наведена графічна ілюстрація критерію AUC-ROC: крива показує співвідношення між часткою істинних позитивних спрацьовувань (TPR) та хибних позитивних спрацьовувань (FPR), а площа під нею (AUC) характеризує здатність моделі відрізняти позитивні й негативні класи.

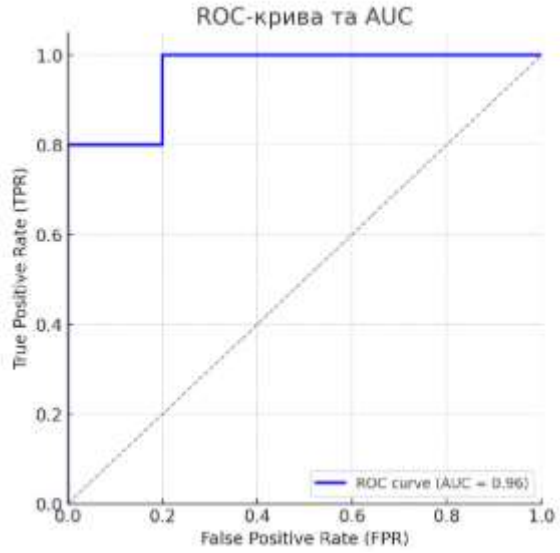

Рис. 2. Класифікаційна метрика критерію AUC-ROC

1.6. **Log-loss/Cross-entropy** (Logarithmic Loss, Cross-Entropy Loss) - функція втрат для оцінки впевненості моделі у прогнозі. Оцінює якість ймовірнісних прогнозів.

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i)], \quad (5)$$

де y_i - реальне значення класу (0 або 1), p_i - передбачена ймовірність.
Чим нижчий Log-loss, тим краще модель "впевнено" прогнозує.
В табл. 1 наведені результати порівняння класифікаційних метрик.

2. **Регресійні метрики -** кількісні показники, що використовуються для оцінювання точності та якості прогнозів моделей машинного навчання при розв'язанні задач регресії [3].
2.1. **MAE (Mean Absolute Error,)** - середня абсолютна похибка.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (6)$$

Показує середню абсолютну різницю між фактичними значеннями y_i і прогнозами $\hat{y}_i$.
Легко інтерпретується, бо має ті ж одиниці виміру, що й вихідна змінна.
2.2. **MSE (Mean Squared Error)** - середньоквадратична похибка.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \quad (7)$$

Середній квадрат різниці між реальними і прогнозованими значеннями. Краще знешкоджує великі похибки, ніж MAE.
2.3. **RMSE (Root Mean Squared Error)** - корінь середньої квадратичної похибки)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \quad (8)$$

Є квадратним коренем з MSE. Дає похибку в тих самих одиницях, що й цільова змінна. Дуже популярна метрика у практиці.
2.4. **R²** (**Coefficient of Determination)** - коефіцієнт детермінації, частка дисперсії, поясненої моделлю.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \quad (9)$$

де $\bar{y}_i$ - середнє фактичних значень.
Метрика вимірює, яку частку варіації цільової змінної пояснює модель. Значення: $R^2 = 1$ - ідеальне прогнозування; $R^2 = 0$ - модель не краща за середнє; $R^2 < 0$ - модель гірша за просте середнє.

2.5. **MAPE (Mean Absolute Percentage Error)** – це метрика регресії, яка показує середню відносну похибку моделі у відсотках. Вона обчислюється за формулою:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|, \quad (10)$$

де: y_i - фактичні значення; $\hat{y}_i$ - прогнозовані значення; n - кількість спостережень.

Інтерпретація наступна. Якщо *MAPE* = 5%, то це означає, що в середньому модель помиляється на 5% від реального значення. Чим менше *MAPE*, тим краща якість прогнозу.

В табл. 2 наведені результати порівняння регресійних метрик.

Таблиця 1.

Порівняння класифікаційних метрик

Метрика	*Що показує*	*Переваги*	*Недоліки*
Accuracy	Частка правильно класифікованих прикладів	Легка для розуміння, універсальна	Неефективна при незбалансованих класах
Precision	Частка коректних серед передбачених позитивів	Важлива, коли важлива «чистота» передбачень	Ігнорує пропущені позитивні приклади (FN)
Recall	Частка знайдених позитивних серед усіх позитивних	Корисна, коли важливо не пропустити позитиви (наприклад, хворих)	Може бути високою при великій кількості FP
F1-Score	Баланс між Precision і Recall	Добре підходить для незбалансованих даних	Менш інтерпретована, ніж Accuracy
AUC-ROC	Здатність моделі відрізняти позитивні та негативні	Не залежить від порогу класифікації	Може вводити в оману при дуже незбалансованих класах
Log-loss	Якість ймовірнісних прогнозів	Враховує ступінь впевненості передбачення	Складніший для інтерпретації, чутливий до неправильних «впевнених» прогнозів

3. Генералізаційні властивості - здатність моделі машинного навчання (зокрема НМ) не лише запам’ятати навчальні дані, а й робити коректні передбачення на нових прикладах, що не входили до навчальної вибірки. Іншими словами: добра генералізація - модель прогнозує нові дані з високою точністю (точність ~90% на тренувальних даних і ~88% на тестових; погана генералізація - модель або перенавчена (overfitting) (модель показує точність ~99% на тренувальних даних і лише ~70% на тестових), або недонавчена (underfitting) (модель має ~65% точність і на тренувальних, і на тестових даних - вона занадто проста) [3].

Таблиця 2.

Порівняння регресійних метрик

Метрика	*Що показує*	*Переваги*	*Недоліки*
MAE	Середнє відхилення прогнозу від фактичних значень	Легка для інтерпретації; не чутлива до викидів	Не враховує вагу великих помилок
MSE	Середній квадрат відхилень прогнозу від істини	Показує значимість великих похибок; добре для оптимізації	Чутлива до викидів; важко інтерпретувати одиниці
RMSE	Середня похибка у тих самих одиницях, що й цільова змінна	Інтуїтивно зрозуміла; сильніше карає великі помилки, ніж MAE	Так само чутлива до викидів, як MSE
R^2	Частку варіації цільової змінної, пояснену моделлю	Дає узагальнену оцінку якості; зрозумілий для порівняння моделей	Може бути від’ємним; не завжди показує якість прогнозу окремих значень
MAPE	Середню відносну похибку у відсотках	Інтуїтивно зрозуміла у %; зручна для бізнес-аналітики	Некоректна при $y \approx 0$; переоцінює маленькі значення

Фактори, що впливають на генералізацію:

- **Обсяг і якість даних.** Більший та різноманітніший датасет зазвичай підвищує здатність моделі узагальнювати. Шумні або нерелевантні дані погіршують генералізацію.

- **Складність моделі.** Занадто складна модель (дуже глибока НМ) може «запам’ятати» тренувальні дані - overfitting. Занадто проста модель може не вловити залежності - underfitting.

- **Регуляризація.** Використання L1/L2-регуляризації, Dropout, Batch Normalization, Data Augmentation допомагає уникнути перенавчання.

- **Розбиття даних.** Генералізаційні властивості оцінюють на валідаційній та тестовій вибірках. Якщо модель добре працює лише на тренувальних даних, але гірше на тестових, то генералізація слабка.

3.1. **Training Error (помилка на навчальній вибірці)** - середня помилка моделі на тих даних, які використовувалися під час навчання. Вона показує наскільки добре модель запам'ятала та відтворює закономірності з навчального набору.

3.2. **Validation Error (помилка на валідаційній вибірці)** - помилка моделі на валідаційних даних, які не використовувалися під час навчання, але застосовуються для контролю процесу тренування. Вона показує здатність моделі узагальнювати закономірності на нових даних, подібних до тренувальних.

3.3. **Test Error (помилка на тестовій вибірці)** - помилка на абсолютно незалежних даних, які не використовувались ні в навчанні, ні у валідації. Вона показує реальну узагальнюючу здатність моделі, тобто як вона працюватиме у практичному застосуванні.

3.4. **Bias-Variance Trade-off (компроміс між зміщенням та дисперсією)** - концепція, яка описує баланс між двома джерелами помилок: зміщення (Bias): помилка, що виникає через надмірне спрощення моделі (недонавчання); Дисперсія (Variance): помилка, що виникає через занадто сильне пристосування до тренувальних даних (перенавчання).

Вона показує чи модель схильна більше до недонавчання чи до перенавчання.

3.5. **Cross-Validation Score (оцінка за перехресною перевіркою)** - метод, коли дані багаторазово розбиваються на навчальні та валідаційні підмножини, і середня помилка обчислюється за всіма ітераціями. Показує наскільки стійкою і стабільною є модель при різних розбиттях даних.

В табл. 3 наведені результати порівняння генералізаційних властивостей.

Таблиця 3.

Порівняння генералізаційних властивостей

Метрика	*Що показує*	*Переваги*	*Недоліки*
Training Error	Наскільки добре модель вивчила дані, на яких тренувалася	Дозволяє виявити недонавчання	Не відображає здатності до узагальнення
Validation Error	Якість прогнозу на нових даних, що не використовувались у тренуванні	Використовується для налаштування гіперпараметрів	Може змінюватись залежно від вибору валідаційної вибірки
Test Error	Рівень узагальнення на незалежних даних	Найбільш об'єктивна оцінка ефективності	Потребує якісної та репрезентативної тестової вибірки
Bias-Variance Trade-off	Баланс між зміщенням (спрощенням моделі) і розкидом (перепідгонкою)	Пояснює причини помилок моделі; корисний для аналізу	Важко кількісно оцінити; потребує додаткових методів аналізу
Cross-Validation Score	Стійкість і стабільність результатів моделі при різних розбиттях даних	Зменшує ризик перенавчання; об'єктивна оцінка	Обчислювально затратна, особливо для великих моделей

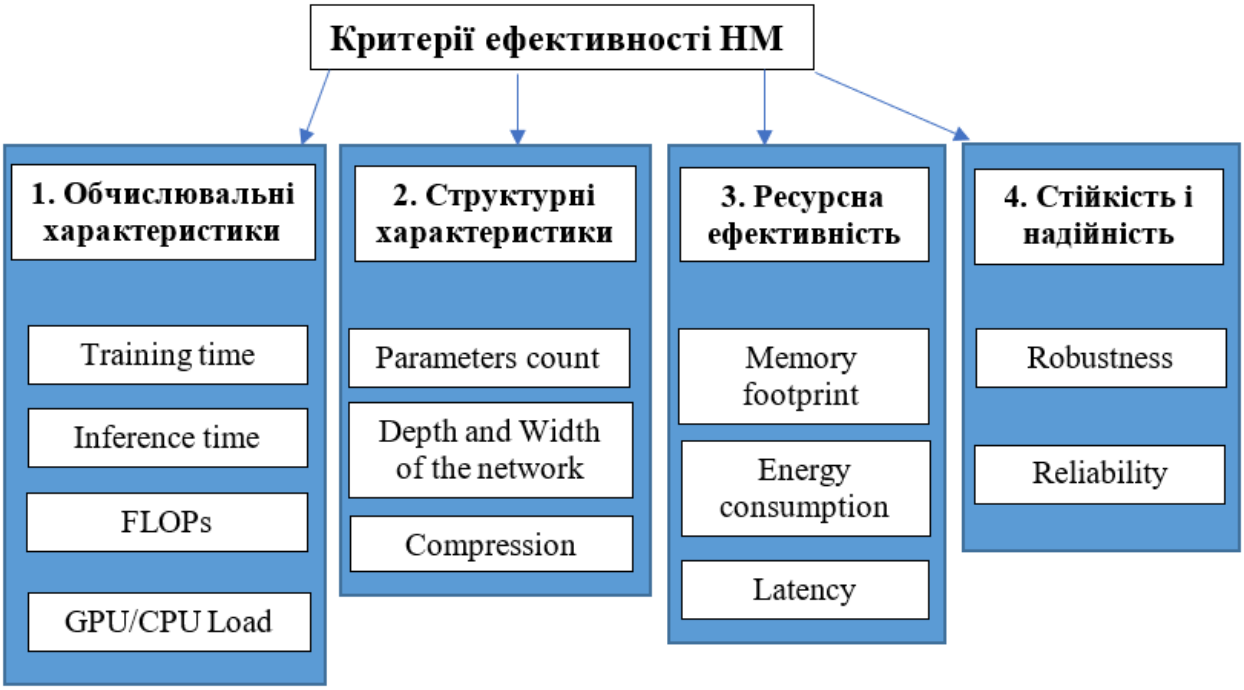

Рис. 3. Узагальнена класифікація критеріїв ефективності НМ

2. Критерії ефективності нейронної мережі

Мета критеріїв ефективності (рис. 3) - оцінити, наскільки раціонально модель використовує ресурси для досягнення бажаної якості [2, 3].

1. Обчислювальні характеристики - це сукупність показників, що відображають витрати обчислювальних ресурсів (часу, пам'яті, енергії) під час навчання та використання моделі.

1.1. **Training time (Час навчання)** - загальний час, необхідний для тренування моделі.

1.2. **Inference time (Час передбачення)** - час, потрібний для обробки одного або кількох прикладів.

1.3. **FLOPs (Floating Point Operations)** - кількість операцій з плаваючою комою при інференсі.

1.4. **GPU/CPU Load** - рівень завантаження апаратних ресурсів.

2. Структурні характеристики - це параметри, що описують її архітектуру, зокрема кількість шарів (глибину), кількість нейронів у шарах (ширину), типи зв'язків та організацію мережі [5].

2.1. **Кількість параметрів (Parameters count)** - розмір моделі, що впливає на швидкість, пам'ять і потребу в зберіганні.

2.2. **Глибина і ширина мережі** - кількість шарів і нейронів.

2.3. **Можливість стискання (Compression)** - потенціал для зменшення обсягу без втрати якості (наприклад, pruning, quantization).

3. Ресурсна ефективність - це здатність моделі досягати заданої якості прогнозування при мінімальному використанні обчислювальних, пам'яттєвих та енергетичних ресурсів [5].

3.1. **Memory footprint** - обсяг оперативної пам'яті, що використовується моделлю.

3.2. **Energy consumption** - кількість енергії, витраченої під час інференсу або навчання.

3.3. **Latency (затримка)** - час реакції системи в реальному часі.

4. Критерії стійкості та надійності - характеризують поведінку моделі в нестандартних або критичних умовах [4, 5].

Критерії стійкості - це показники, що характеризують здатність НМ зберігати ефективність при впливі шуму, викидів чи атак.

До параметрів стійкості (Robustness) відносяться:

- **Чутливість до шуму у вхідних даних** – наскільки змінюється результат при наявності випадкових перешкод.
- **Стійкість до викидів (outliers)** – збереження адекватних прогнозів при аномальних даних.
- **Adversarial robustness** – здатність протистояти спеціально створеним вхідним прикладам для обману моделі.
- **Загальна варіативність результатів** – ступінь зміни вихідних значень при малих варіаціях вхідних.

Критерії надійності - це показники, що відображають здатність НМ відтворювати стабільні та правильні результати в різних умовах і при повторних запусках.

До параметрів надійності (Reliability) відносяться:

- **Відтворюваність результатів** – стабільність якості при повторному навчанні на тих самих даних.
- **Стабільність на різних вибірках** – збереження продуктивності на тренувальних, валідаційних і тестових даних.
- **Середній час безвідмовної роботи (uptime, fault tolerance)** – здатність працювати без збоїв у реальних системах.
- **Ймовірність відмови** – ризик некоректних або непередбачуваних вихідних результатів.

5. Інтегральні (комплексні) критерії

Інтегральні (комплексні) критерії ефективності НМ використовуються тоді, коли потрібно оцінити модель **не за одним показником**, а за **сукупністю різних характеристик**, що відображають як якість прогнозування, так і витрати ресурсів, стійкість, узагальнювальні властивості тощо [6-10].

До них відносяться:

5.1. **Зважена сума метрик (Weighted Score).**

Мета цього показника - об'єднання кількох показників (точність, швидкість, використання пам'яті) у єдиний критерій шляхом призначення ваг кожному параметру. Дозволяє врахувати пріоритети задачі (наприклад, у мобільному застосунку важливіша швидкодія, ніж надточність).

5.2. **Критерій "якість/ресурси" (Quality-to-Cost Ratio)** - відношення якості моделі (наприклад, точності чи RMSE) до витрат ресурсів (часу навчання, FLOPs, обсягу пам'яті). Показує, наскільки ефективно модель досягає результатів при заданих обмеженнях.

5.3. **Індекс узагальнювальної здатності (Generalization Index)** - співвідношення продуктивності моделі на тренувальних і тестових даних. Чим менша різниця - тим краща генералізація.

5.4. **Критерій збалансованості (Balanced Performance Criterion) -** оцінка моделі одночасно за кількома властивостями: точністю, стабільністю, стійкістю до шуму. Часто використовується у вигляді середньої гармонічної метрик (аналог F1-score для багатьох показників).

5.5. **Інтегральний показник надійності (Reliability Index).** Включає в себе стабільність результатів на різних вибірках, відтворюваність при повторному навчанні, чутливість до випадкових факторів. Використовується в критичних системах (автопілот, медичні застосування).

5.6. **Енергетично-ефективний критерій (Energy-aware Efficiency Criterion).** Враховує якість прогнозу відносно енергоспоживання системи (наприклад, у вбудованих пристроях чи мобільних роботах). Дає змогу оцінювати «економність» моделі.

5.7. **Комплексний показник життєздатності моделі (Model Sustainability Score).** Враховує довготривале використання: простоту перенавчання, масштабованість, адаптивність до нових даних. Використовується у виробничих системах із динамічними даними.

Інтегральні критерії (табл. 4) – це комбінація різних метрик, яка дозволяє об'єктивніше оцінити модель у реальних умовах, коли важливе не лише «наскільки точна мережа», а й «наскільки вона швидка, стійка, надійна і ресурсоефективна».

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Дослідження критеріїв ефективності та якості НМ є надзвичайно актуальним з огляду на швидкий розвиток штучного інтелекту та широке застосування НМ у різноманітних сферах (класифікація, прогнозування, моделювання та ін.). НМ відомі здатністю навчатися на даних та забезпечувати високу точність результатів, однак вони також вимагають значних обчислювальних ресурсів та великої кількості тренувальних даних. Таким чином, оцінка їх ефективності через систему критеріїв дозволяє збалансувати потребу в якості прогнозування з обмеженнями на час обчислення, пам'ять та енергію. У роботі висвітлено ключові аспекти теми: важливість комплексної оцінки НМ, виділено основні типи метрик, що охоплюють як якісні, так і кількісні показники роботи моделей, а також розглянуто вплив цих критеріїв на практичні рішення при побудові архітектури.

Таблиця 4.

Порівняння інтегральних критеріїв

Критерій	*Що показує*	*Переваги*	*Недоліки*
Зважена сума метрик (Weighted Score)	Загальну ефективність моделі з урахуванням важливості різних метрик	Гнучкість, можна налаштовувати ваги під задачу	Суб'єктивність вибору ваг; складність калібрування
Критерій "якість / ресурси" (Quality-to-Cost Ratio)	Наскільки добре модель працює відносно витрат ресурсів	Баланс між точністю і продуктивністю	Може недооцінювати складні моделі з високою якістю
Індекс узагальнювальної здатності (Generalization Index)	Різницю між навчанням і тестуванням	Дає оцінку перенавчання	Не враховує швидкодію та інші аспекти
Критерій збалансованості (Balanced Performance)	Поєднання точності, стабільності, стійкості до шуму	Комплексна оцінка, подібна до F1-score	Складність у виборі метрик для поєднання
Інтегральний показник надійності (Reliability Index)	Стійкість результатів при різних умовах	Важливий для критичних застосувань	Важко формалізувати і перевіряти
Енергетично-ефективний критерій (Energy-aware Efficiency)	Співвідношення якості і енергоспоживання	Ключовий для мобільних та IoT-систем	Не завжди актуальний для серверних моделей
Показник життєздатності моделі (Sustainability Score)	Масштабованість, простота перенавчання, адаптивність	Орієнтація на довготривале використання	Важко кількісно виміряти

Основними результатами роботи є систематизація критеріїв оцінки продуктивності НМ за двома напрямками. По-перше, розглянуто метрики якості. Для завдань класифікації найчастіше використовують показники точності (accuracy), чутливості (recall), специфічності (specificity) та влучності (precision), які вимірюють частку правильно класифікованих об'єктів серед усіх, позитивних і негативних прикладів відповідно. Важливим доповненням є F1-міра – гармонійне середнє precision і recall, що дає збалансовану оцінку у разі нерівномірних класів. Для багатокласових завдань аналогічно застосовують середню по класах точність або інші похідні метрики. Окремо слід відзначити, що в деяких випадках використовують ROC-криві та площу під ними (AUC) – оцінка, незалежна від конкретного порогу рішення. У разі регресії критеріями є похибки апроксимації (наприклад, середньоквадратична помилка MSE або середня абсолютна помилка MAE): чим менше ці значення, тим точніше модель повторює реальні дані. Таким чином, кожен із цих критеріїв враховує різні аспекти якості роботи моделі: наприклад, точність визначає загальну успішність класифікації, recall – здатність знайти всі позитивні випадки, а F1-міра – компроміс між ними. Ігнорування цих показників може призвести до упереджень (наприклад, висока точність при дуже нерівномірних класах)

або занижених висновків про реальні можливості мережі. У підсумку робота узагальнила, що поєднана оцінка через набір таких метрик забезпечує всебічний аналіз ефективності моделі.

Висвітлено критерії обчислювальної ефективності та складності мережі, що відіграють вирішальну роль при практичному застосуванні. З одного боку, це показники, пов'язані з ресурсами: передусім загальна кількість параметрів моделі (що визначає її розмір і обсяг пам'яті) та обчислювальна складність (кількість операцій, FLOPs). Рівень FLOPs прямо корелює з часом інференсу (швидкістю роботи моделі) – менший «літній» вказує на швидше виконання та менші енергозатрати. Водночас кількість параметрів вказує на обсяг пам'яті, потрібний для збереження ваг мережі. Обидва критерії мають враховуватися одночасно, оскільки компактна модель (з малою кількістю параметрів) може мати велику кількість обчислень і навпаки. Крім того, до критеріїв ефективності відносять швидкість збіжності (темпи навчання), стійкість алгоритму до збоїв та потребу в даних, що впливають на витрати часу та праці при налаштуванні.

Практична цінність наведених критеріїв полягає у тому, що вони безпосередньо впливають на вибір архітектури, налаштування та впровадження мережі під конкретне завдання. Зокрема, якщо пріоритетом є висока точність прогнозування (наприклад, у діагностиці захворювань), обирають моделі з глибшою архітектурою і підвищеною потужністю розпізнавання, навіть за рахунок більших обчислювальних витрат. Натомість у задачах реального часу (наприклад, автономна навігація) чи на вбудованих пристроях критичним стає зменшення FLOPs і параметрів. Приклади з практики показують: порівнюючи FLOPs та точність різних архітектур, розробники можуть вибрати оптимальну модель для обмежених ресурсів – наприклад, легку мережу з достатньою точністю для мобільного застосунку. Аналогічно, критерії впливають на гіперпараметричну оптимізацію: наприклад, розмір батчів, темп навчання або регуляризація налаштовуються з урахуванням балансу між швидкістю збіжності і точністю. Висновком можна вважати, що ґрунтовний аналіз поєднання якісних і ресурсних критеріїв дає змогу проєктувати НМ, що відповідають особливим вимогам конкретних завдань, та ефективно їх реалізовувати на обчислювальних платформах.

Література

1. Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge, MA: MIT Press, 2016. 775 p.
2. Rohlfs C. Generalization in Neural Networks: A Broad Survey [Electronic resource]. arXiv:2209.01610, 2022. Available at: https://arxiv.org/abs/2209.01610
3. Wang Y., Han Y. et al. Computation-efficient Deep Learning for Computer Vision: A Survey [Electronic resource]. arXiv:2308.13998, 2023. 47 p. Available at: https://arxiv.org/pdf/2308.13998
4. Wang J. A Survey of Neural Network Robustness Assessment in Image Recognition [Electronic resource]. arXiv:2404.08285, 2024. 34 p. Available at: https://arxiv.org/pdf/2404.08285
5. Sze V., Chen Y.-H., Yang T.-J., Emer J. S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proceedings of the IEEE. 2017. Vol. 105, No. 12. P. 2295–2329. Available at: https://semiwiki.com/wp-content/uploads/2018/03/2017_pieee_dnn.pdf
6. Rao S. Machine Learning, Illustrated: Evaluation Metrics for Classification [Electronic resource]. Towards Data Science (Medium). Apr. 2023. Available at: https://medium.com/data-science/machine-learning-illustrated-classification-evaluation-metrics-dfc33b373c43
7. Zuccarelli E. Performance Metrics in Machine Learning – Part 1: Classification [Electronic resource]. Towards Data Science. 29 Dec. 2020. Available at: https://medium.com/data-science/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92
8. Zuccarelli E. Performance Metrics in Machine Learning – Part 2: Regression [Electronic resource]. Towards Data Science. 4 Jan. 2021. Available at: https://medium.com/data-science/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92
9. Rainio O., Teuho J., Klén R. Evaluation metrics and statistical tests for machine learning. Scientific Reports. 2024. Vol. 14. Art. 6086. Available at: https://www.nature.com/articles/s41598-024-56706-x
10. TensorFlow Developers. tf.keras.metrics – TensorFlow 2.16.1 Documentation [Electronic resource]. 2023. Available at: https://www.tensorflow.org/api_docs/python/tf/keras/metrics

References

1. Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge, MA: MIT Press, 2016. 775 p.
2. Rohlfs C. Generalization in Neural Networks: A Broad Survey [Electronic resource]. arXiv:2209.01610, 2022. Available at: https://arxiv.org/abs/2209.01610
3. Wang Y., Han Y. et al. Computation-efficient Deep Learning for Computer Vision: A Survey [Electronic resource]. arXiv:2308.13998, 2023. 47 p. Available at: https://arxiv.org/pdf/2308.13998
4. Wang J. A Survey of Neural Network Robustness Assessment in Image Recognition [Electronic resource]. arXiv:2404.08285, 2024. 34 p. Available at: https://arxiv.org/pdf/2404.08285
5. Sze V., Chen Y.-H., Yang T.-J., Emer J. S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proceedings of the IEEE. 2017. Vol. 105, No. 12. P. 2295–2329. Available at: https://semiwiki.com/wp-content/uploads/2018/03/2017_pieee_dnn.pdf
6. Rao S. Machine Learning, Illustrated: Evaluation Metrics for Classification [Electronic resource]. Towards Data Science (Medium). Apr. 2023. Available at: https://medium.com/data-science/machine-learning-illustrated-classification-evaluation-metrics-dfc33b373c43

7. Zuccarelli E. Performance Metrics in Machine Learning – Part 1: Classification [Electronic resource]. Towards Data Science. 29 Dec. 2020. Available at: https://medium.com/data-science/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92
8. Zuccarelli E. Performance Metrics in Machine Learning – Part 2: Regression [Electronic resource]. Towards Data Science. 4 Jan. 2021. Available at: https://medium.com/data-science/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92
9. Rainio O., Teuho J., Klén R. Evaluation metrics and statistical tests for machine learning. Scientific Reports. 2024. Vol. 14. Art. 6086. Available at: https://www.nature.com/articles/s41598-024-56706-x
10. TensorFlow Developers. tf.keras.metrics – TensorFlow 2.16.1 Documentation [Electronic resource]. 2023. Available at: https://www.tensorflow.org/api_docs/python/tf/keras/metrics