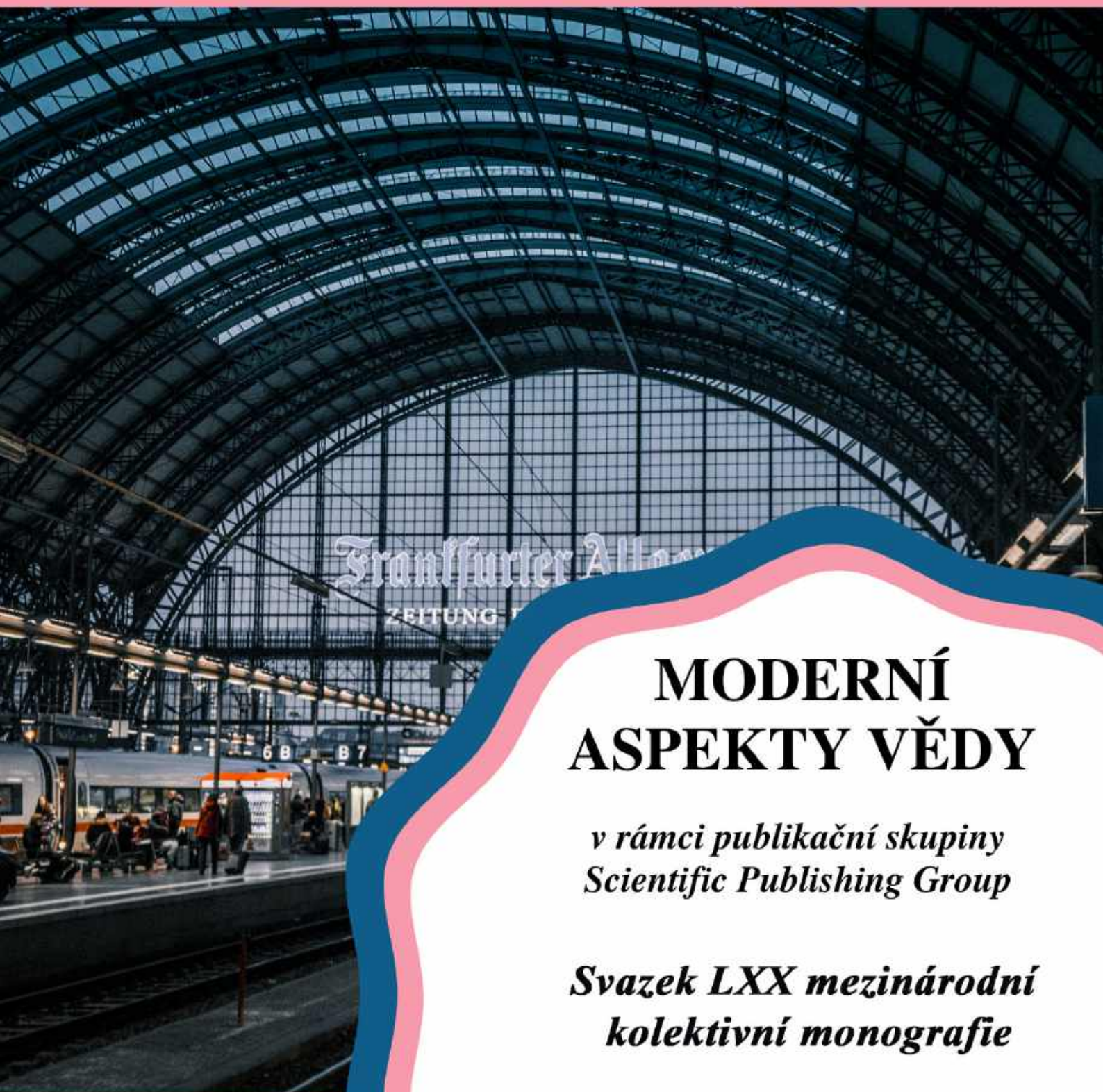




Наукові перспективи
Видавнича група



MODERNÍ ASPEKTY VĚDY

*v rámci publikační skupiny
Scientific Publishing Group*

*Svazek LXX mezinárodní
kolektivní monografie*



Česká republika
2026



Mezinárodní Ekonomický Institut s.r.o. (Česká republika)
Středoevropský vzdělávací institut (Bratislava, Slovensko)
Národní institut pro ekonomický výzkum (Batumi, Gruzie)
Al-Farabi Kazakh National University (Kazachstán)
Institut filozofie a sociologie Ázerbájdžánu Národní akademie věd
(Baku, Ázerbájdžán)
Institut vzdělávání Ázerbájdžánské republiky (Baku, Ázerbájdžán)
Regionální akademie managementu (Kazachstán)
Veřejná vědecká organizace „Celokrajinské shromáždění lékařů ve veřejné
správě“ (Kyjev, Ukrajina)
Univerzita nových technologií (Kyjev, Ukrajina)
Mezinárodní poradenská společnost "Sidcon" (Kyjev, Ukrajina)
Evropské lyceum "Vědecké perspektivy" (Kyjev, Ukrajina)

v rámci publikační skupiny Publishing Group „ Vědecká perspektiva “

MODERNÍ ASPEKTY VĚDY

Svazek LXX mezinárodní kolektivní monografie

Česká republika
2026

International Economic Institute s.r.o. (Czech Republic)
Central European Education Institute (Bratislava, Slovakia)
National Institute for Economic Research (Batumi, Georgia)
Al-Farabi Kazakh National University (Kazakhstan)
Institute of Philosophy and Sociology of Azerbaijan National Academy of
Sciences (Baku, Azerbaijan)
Institute of Education of the Republic of Azerbaijan (Baku, Azerbaijan)
Regional Academy of Management (Kazakhstan)
Public Scientific Organization "Ukrainian Assembly of Doctors of Sciences in
Public Administration" (Kyiv, Ukraine)
University of New Technologies (Kyiv, Ukraine)
International Consulting company "Sidcon" (Kyiv, Ukraine)
European Lyceum "Scientific Perspectives" (Kyiv, Ukraine)

within the Publishing Group "Scientific Perspectives"

MODERN ASPECTS OF SCIENCE

70-th volume of the international collective monograph

Czech Republic
2026

ISBN 978-617-8614-40-9

<https://doi.org/10.52058/69-2026>

UDC 001.32: 1/3] (477) (02)

C91

Vydavatel:

**Mezinárodní Ekonomický Institut s.r.o.
se sídlem V Lázních 688, Jesenice 252 42
IČO 03562671 Česká republika
Zveřejněno rozhodnutím akademické rady**

Mezinárodní Ekonomický Institut s.r.o. (Zápis č. 231/2026 ze dne 08. červenec 2026)



Monografie jsou indexovány v mezinárodním vyhledávači Google Scholar

Partnerskými akademickými mezinárodními vydavateli monografie jsou: TOJELT, Paradigm Journal a Shanlax International Journal of Education

Recenzenti:

- Karel Nedbálek** - doktor práv, profesor v oboru právo (Česká republika)
Markéta Pavlova - ředitel, Mezinárodní Ekonomický Institut (České republiky)
Iryna Zhukova - kandidátka na vědu ve veřejné správě, docentka (Ukrajina)
Savvas Mavridis - profesor na Mezinárodní řecké univerzitě v oddělení marketingu a cestovního ruchu, Veřejná instituce (Greece)
Yevhen Romanenko - doktor věd ve veřejné správě, profesor, ctěný právník Ukrajiny (Ukrajina)
Humeir Huseyn Akhmedov - doktor pedagogických věd, profesor (Ázerbájdžán)
Jurij Kijkov - doktor informatiky, dr.h.c. v oblasti rozvoje vzdělávání (Česká republika)
Vladimír Bačšín - docent ekonomie (Slovensko)
Ahmet Selcuk Akdemir - doktor věd, profesor, Malatya Turgut Ozal University (Turecko)
Peter Ošváth - docent práva (Slovensko)
Oleksandr Nepomnyashy - doktor věd ve veřejné správě, kandidát ekonomických věd, profesor, řádný člen Vysoké školy stavební Ukrajiny (Ukrajina)
Dina Dashevska - geolog, geochemik Praha, Česká republika (Izrael)
R. Ahmad Zaki El Islami - docent, PhD, profesor katedry vědy a vzdělávání, fakulty pedagogiky a vzdělávání, univerzity Sultan Ageng Tirtayasa (Indonésie)
Dr. Mutmaina - přednášející, univerzita Al Asyariah Mandar Sulawesi Barat, Indonésie, mimořádný profesor, katedra vzdělávání, univerzita GlobalNxt, Malajsie (Malajsie)
Said M. Ismail - PhD, profesor, profesor katedry angličtiny, fakulta přírodních a humanitních věd, univerzita prince Sattama bin Abdulazize, Al-Charj (Saúdská Arábie)
Galandarov, Vagif Kalender - MD, PhD, profesor, profesor katedry chirurgických nemocí, ázerbájdžánská lékařská univerzita (Ázerbájdžán)
Dr. Jane Gakenya Njoroge - profesor, Kenyatta University, (Keňa)
Diana Ibarra Santacruz - profesorka, University of Nariño, (Kolumbie)
Helda Alicia Hidalgo Davila - profesor, University of Nariño, (Kolumbie)
Myujkan Mammadzada - doktor filozofie v oboru pedagogika, Vojenský výzkumný ústav Univerzity národní obrany, Baku (Ázerbájdžán)
Zeynalova Nigar Eldar Kizy - doktorka pedagogických věd, docentka, ředitelka Centra pro inkluzivní vzdělávání, Ázerbájdžánská univerzita jazyků (Ázerbájdžán)
Nani Gvarišvili - doktorka biologických věd, docentka, Státní univerzita Šoty Rustaveliho v Batumi, Batumi (Gruzie)

Tým autorů

C91 Moderní aspekty vědy: LXX. Díl mezinárodní kolektivní monografie / Mezinárodní Ekonomický Institut s.r.o.. Česká republika: Mezinárodní Ekonomický Institut s.r.o., 2026. str. 601

Svazek LXX mezinárodní kolektivní monografie obsahuje publikace o: utváření a rozvoji teorie a historie veřejné správy; formování regionální správy a místní samosprávy; provádění ústavního a mezinárodního práva; finance, bankovníctví a pojišťovnictví; duševní rozvoj osobnosti; rysy lexikálních výrazových prostředků imperativní sémantiky atd.

Materiály jsou předkládány v autorském vydání. Autoři odpovídají za obsah a pravopis materiálů.



ODDÍL 10. EKONOMIE A MATEMATICKÉ MODELOVÁNÍ

§10.1 *ECONOMIC AND MATHEMATICAL METHODS IN MODERN ECONOMICS: THEORY, METHODOLOGY AND PRACTICAL APPLICATION UNDER ERA OF DIGITALIZATION* (**Plaskon S.A.**, West Ukrainian National University) 480

§10.2 *ECONOMIC AND MATHEMATICAL MODELING OF BANK ASSET STRUCTURE OPTIMIZATION UNDER RISK AND UNCERTAINTY* (**Ruska R.V.**, Western Ukrainian National University) 508

ODDÍL 11. INFORMAČNÍ TECHNOLOGIE

§11.1 *СТРАТЕГІЇ ВЕРСІОНУВАННЯ У ПРОЦЕСАХ КОЛЕКТИВНОГО ДОКУМЕНТУВАННЯ* (**Ткач Л.М.**, Національний університет «Львівська політехніка», **Стадник О.В.**, Національний університет «Львівська політехніка») 534

ODDÍL 12. POČÍTAČOVÁ VĚDA

§12.1 *ЕНЕРГОЕФЕКТИВНІСТЬ СТРУКТУРНО ОПТИМІЗОВАНИХ НЕЙРОННИХ МЕРЕЖ* (**Симонян В.В.**, Мукачівський державний університет, **Король А.В.**, Мукачівський державний університет) 548

ODDÍL 13. CESTOVNÍ RUCH

§13.1 *КОМЕМОРАТИВНА ПАРАДИГМА: ЯК ВОЄННИЙ ТУРИЗМ ТРАНСФОРМУЄ ДОСВІДИ ВІЙНИ У ГЛОБАЛЬНУ ІСТОРИЧНУ ПАМ'ЯТЬ* (**Гринюк Д.Ю.**, Київський національний університет імені Тараса Шевченка) 564

ODDÍL 14. PRÁVNÍ VĚDY

§14.1 *ЄВРОАТЛАНТИЧНА ІНТЕГРАЦІЯ ТА НАЦІОНАЛЬНА БЕЗПЕКА УКРАЇНИ: ПЕРЕОСМИСЛЕННЯ ПІДХОДІВ В УМОВАХ 2026 РОКУ* (**О. Ромців**, Національний університет «Львівська політехніка») 579



Використання хмарних інструментів вебредагування нівелює бар'єр необхідності встановлення спеціалізованого програмного забезпечення, створюючи тим самим високу мобільність автора та доступність процесу документування з будь-якого пристрою. Це робить запропоновану модель не лише технологічно потужною, а й максимально інклюзивною для фахівців різного профілю.

Висновки. Проведене дослідження дозволяє констатувати, що синергія формату Markdown, розподіленої системи контролю версій Git, колаборативної платформи GitHub та генератора статичних сайтів Hugo формує високоефективний технологічний стек для управління повним життєвим циклом мультимедійної документації.

По міждисциплінарній валідації ми спробували довести, що впровадження VCS-інструментарію є критично необхідним не лише в IT-індустрії, а й у науковій, освітній, правовій та мистецькій сферах. Це забезпечує персистентність даних, верифікацію авторства та прозорість інтелектуальної взаємодії.

Запропонована технологічна модель публікації на основі підходу Documentation-as-Code (DaC) нівелює бар'єри між технічною розробкою та гуманітарним документознавством, забезпечуючи високу швидкість ітерацій та цілісність контенту. Ми визначили, що використання хмарних інструментів вебредагування (github.dev) радикально спрощує вхід у професійну роботу з документацією для фахівців без глибокої технічної підготовки, сприяючи інклюзивності робочих процесів. Запропонована модель публікації документації сприяє гармонізації вітчизняних ділових традицій із вимогами міжнародного стандарту ISO/IEC/IEEE 26514:2022, забезпечуючи єдність технологічних процесів та високу конкурентоспроможність українських фахівців на світовому ринку



На наш погляд, подальший науковий пошук доцільно спрямувати на поглиблене вивчення концепції DocOps як методології безперервної інтеграції документації. Важливим вектором є також розширення застосування хмарних VCS-сервісів в академічному середовищі для створення відкритих наукових архівів та інтерактивних навчальних ресурсів. Трансформація традиційного документообігу у формат «документація як код» залишається стратегічним напрямом для гармонізації українських стандартів із глобальним інформаційним простором.

Список використаних джерел:

1. Bateman S., Gutwin C. Challenges of supporting documentation in collaborative software development. *Proceedings of the ACM on Human-Computer Interaction*. 2020. Vol. 4, No. CSCW2. P. 1–25. DOI: doi.org
2. Morgan K. *Technical Writing Process: The simple, five-step guide that anyone can use to create technical documents such as user guides, manuals, and procedures*. London : Better On Paper Publications, 2015. 248 p.
3. Hadjerrouit S. Wiki as a collaborative writing tool in teacher education: Evaluation and suggestions for effective use. *Computers in Human Behavior*. 2014. Vol. 32. P. 301–312. DOI: doi.org
4. Weiss E. H. *How to write usable user documentation*. Phoenix : Oryx Press, 1991. 265 p.
5. ДСТУ 4163:2020. *Державна уніфікована система документації. Уніфікована система організаційно-розпорядчої документації. Вимоги до оформлення документів*. На заміну ДСТУ 4163-2003 ; чинний від 2021-09-01. Київ : ДП УкрНДНЦ, 2020. 28 с. URL: <https://archives.gov.ua> (дата звернення: 06.07.2026).



6. ISO/IEC/IEEE 26514:2022. *Systems and software engineering — Design and development of information for users*. Geneva : International Organization for Standardization, 2022. 112 p. URL: iso.org (date of access: 06.07.2026).

7. Nagel W. *Subversion Version Control: Using the Subversion Version Control System in Development Projects*. Upper Saddle River : Prentice Hall, 2005. 312 p.

8. Spinellis D. Why and how to use GitHub for your research. *Journal of Software: Evolution and Process*. 2021. Vol. 33, No. 10. Art. e2364. DOI: doi.org

9. Gentle A. *Docs for Developers: An Engineer's Field Guide to Technical Writing*. Berkeley : Apress, 2021. 254 p. DOI: doi.org



ODDÍL 12. POČÍTAČOVÁ VĚDA

§12.1 ЕНЕРГОЕФЕКТИВНІСТЬ СТРУКТУРНО ОПТИМІЗОВАНИХ НЕЙРОННИХ МЕРЕЖ (Симонян В.В., Мукачівський державний університет, Король А.В., Мукачівський державний університет)

Вступ. Стрімке поширення систем штучного інтелекту в промислових, мобільних та IoT-середовищах загострює проблему ресурсної та енергетичної ефективності нейронних мереж. Зростання обчислювальної складності сучасних моделей глибокого навчання призводить до збільшення витрат пам'яті, часу обробки та обчислювальних ресурсів як під час навчання, так і під час інференсу на кінцевих пристроях.

Особливо актуальною ця проблема є для мобільних, вбудованих та edge-систем, де модель повинна працювати за умов обмеженої оперативної пам'яті, продуктивності процесора та енергоспоживання. У таких умовах вибір архітектури нейронної мережі не може ґрунтуватися лише на точності класифікації, оскільки практичне розгортання моделі також залежить від її розміру, затримки інференсу, обчислювальної складності та використання пам'яті.

Методи структурної оптимізації нейронних мереж, зокрема pruning та quantization, дають змогу зменшити окремі ресурсні характеристики моделі. Водночас їхній ефект не є однозначним: зменшення кількості ненульових ваг не завжди приводить до прискорення інференсу, а INT8-квантизація може по-різному впливати на latency залежно від архітектури моделі, процесора та програмного backend. Тому для коректного порівняння оптимізованих моделей потрібна не одна окрема метрика, а узгоджена система оцінювання за кількома критеріями.



Це зумовлює потребу у розробці інтегрального підходу до оцінювання енергоефективності структурно оптимізованих нейронних мереж, який дозволяв би системно порівнювати різні конфігурації моделей у єдиній системі координат. Такий підхід є важливим для задач, у яких необхідно знайти компроміс між точністю моделі та її ресурсною ефективністю.

Мета дослідження — розробити та апробувати багатокритеріальну методику оцінювання енергоефективності структурно оптимізованих нейронних мереж на основі інтегрального показника E_{score} , що враховує точність класифікації, розмір моделі, латентність інференсу, теоретичну обчислювальну складність та використання оперативної пам'яті. Об'єктом дослідження є структурно оптимізовані згорткові нейронні мережі після застосування методів *pruning* та *quantization*. Предметом дослідження є методика порівняльного оцінювання таких мереж за сукупністю проху-показників ресурсної ефективності.

Виклад основного матеріалу. Питання оцінювання енергоефективності нейронних мереж і зниження їх ресурсної складності розглядаються у працях українських науковців. В. Гнатюк і О. Литвинюк [1] досліджують прогнозування завантаження та розподіл ресурсів у телекомунікаційних системах; Г. Завгородня та В. Завгородній [2] аналізують оптимізацію продуктивності комп'ютерних ігор засобами машинного навчання; В. Кучерук, П. Кулаков, Р. Ліщук, С. Концеба та В. Маньковська [3] систематизують критерії ефективності та якості нейронних мереж; О. Мамчич і М. Волк [4] розглядають оцінювання енергетичних витрат мобільних пристроїв у хмарних обчисленнях; О. Супрун, О. Огірко та А. Кримська [5] досліджують гібридні нейромережеві моделі для прогнозування в інтелектуальних системах.

Серед зарубіжних досліджень К. М. Т. А. Anick [8] аналізує AI-рішення для оптимізації енергетичних процесів у промисловості; М. S. Arimbi,



R. Amini, A. K. Asri та N. Martin [10] розглядають AI у відновлюваній енергетиці; P. Biswas та співавтори [11] узагальнюють підходи до оптимізації енергоспоживання; K. T. Chui, M. D. Lytras і A. Visvizi [9] досліджують smart monitoring та оптимізацію енергоспоживання; Y. B. Melnyk [13] розглядає розвиток AI у цифровому суспільстві; S. Naveen і M. R. Kounte [12] досліджують pruning і quantization CNN на IoT edge; E. C. Onukwulu, M. O. Agho та N. L. Eyo-Udo [7] пропонують AI-framework для енергетичного сектору; A. Stecyk та I. Miciuła [6] аналізують AI-платформи енергетичної оптимізації.

У роботі використовуються такі метрики: accuracy, model size, CPU latency, dense FLOPs, memory footprint, кількість параметрів, кількість ненульових параметрів і weight sparsity. Кількість параметрів та sparsity подаються як звітні метрики, а E_{score} розраховується за п'ятьма базовими компонентами: accuracy, model size, latency, FLOPs і memory footprint.

Quantization у роботі означає зменшення розрядності представлення ваг і частини операцій моделі: замість 32-бітного формату FP32 використовується компактніший 8-бітний формат INT8. Це зменшує розмір моделі на диску, але не завжди гарантує прискорення inference. Вплив quantization на latency залежить від процесора та реалізації INT8-операцій у PyTorch. У проведених експериментах для квантизації використовувався CPU quantization backend x86. Latency і CPU latency показують час одного запуску моделі на процесорі; перед вимірюванням виконуються warm-up ітерації. FLOPs, dense FLOPs і dense theoretical FLOPs трактуються як теоретична кількість операцій для повної щільної структури моделі; MACs використовуються для їх розрахунку. Memory footprint означає приблизне використання RAM під час inference. Усі ці метрики є проху-показниками ресурсної поведінки моделі, а E_{score} використовується для інтегрального ранжування моделей за accuracy, model size, latency, FLOPs і memory footprint.



Запропонована методика виходить з того, що структурно оптимізована модель не може оцінюватися лише за accuracy [10]. Наприклад, квантизована модель може мати майже ту саму точність, але значно менший розмір; pruning може підвищити sparsity, але не дати реального прискорення без sparse-kernels. Тому потрібне узгоджене порівняння за кількома критеріями.

Для забезпечення порівнянності різнорідних метрик усі показники приводяться до діапазону (0; 1). Для accuracy використовується пряма нормалізація, оскільки збільшення цього показника відповідає вищій якості класифікації:

$$n_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Для model size, latency, FLOPs і memory footprint використовується обернена нормалізація, оскільки зменшення цих показників відповідає нижчій ресурсній складності моделі:

$$n_i = \frac{x_{\max} - x_i}{x_{\max} - x_{\min}} \quad (2)$$

Якщо $x_{\max} = x_{\min}$, відповідна метрика не розрізняє конфігурації між собою і не впливає на їх ранжування. У такому випадку всім конфігураціям задається однакове нормоване значення.

Інтегральний показник E_{score} використовується для узагальненого порівняння структурно оптимізованих нейронних мереж за сукупністю нормованих метрик. Він не є абсолютною фізичною мірою енергоспоживання, а відображає співвідношення між точністю класифікації та ресурсними характеристиками моделі в межах заданих умов експерименту. Показник визначається за формулою:

$$E_{\text{score}} = w_1 \times \text{Acc}_{\text{norm}} + w_2 \times \text{Size}_{\text{norm}} + w_3 \times \text{Latency}_{\text{norm}} + w_4 \times \text{FLOPS}_{\text{norm}} + w_5 \times \text{Mem}_{\text{norm}}$$



де Acc_{norm} — нормоване значення точності класифікації, де більше значення є кращим; $Size_{norm}$ — нормований розмір моделі на диску, де менше значення є кращим; $Latency_{norm}$ — нормована затримка інференсу, де менше значення є кращим; $FLOPS_{norm}$ — нормована теоретична обчислювальна складність моделі, де менше значення є кращим; Mem_{norm} — нормований memory footprint під час інференсу, де менше значення є кращим; w_1, w_2, w_3, w_4, w_5 — вагові коефіцієнти відповідних метрик, що визначають їхню відносну важливість у підсумковій оцінці, причому $w_1 + w_2 + w_3 + w_4 + w_5 = 1$, а всі ваги належать діапазону $[0; 1]$. У базовому експерименті використано рівні ваги 0,20 для кожної складової.

Інтегральний показник E_{score} є безрозмірною порівняльною оцінкою, що набуває значень у діапазоні від 0 до 1. Він не вимірюється у фізичних одиницях, таких як ват, джоуль, мілісекунда або мегабайт, оскільки розраховується на основі нормалізованих метрик: accuracy, model size, latency, FLOPs і memory footprint. Чим більше значення E_{score} , тим кращим є узагальнений баланс між якістю моделі та її ресурсними характеристиками в межах заданого експерименту.

Методи структурної оптимізації/ У межах експериментальної частини основну увагу приділено pruning, quantization та їх комбінації [12].

Pruning 30% і 50% реалізовано як global unstructured pruning. Це означає, що окремі ваги з меншими абсолютними значеннями занулюються, але фільтри, канали або нейрони фізично не видаляються. Після pruning виконується fine-tuning із активними pruning masks, щоб занулені ваги не «оживали» під час донавчання. Лише після fine-tuning pruning робиться постійним.

INT8 quantization реалізовано як post-training static quantization з калібрувальними запусками. Такий підхід дає змогу зменшити розмір



представлення ваг і активацій, але фактичний вплив на latency залежить від CPU, backend, інструкцій процесора та розміру моделі. У роботі використано CPU-бенчмарк, оскільки edge/IoT-сценарії часто пов'язані саме з обмеженими CPU-ресурсами. У таблиці 1 наведено узагальнену характеристику методів структурної оптимізації, які розглядаються в роботі. Для кожного методу подано очікуваний ресурсний ефект і основні застереження, які необхідно враховувати під час інтерпретації експериментальних результатів.

Таблица 1

Характеристика методів оптимізації

Метод	Очікуваний ефект	Застереження
Pruning	Менше ненульових ваг, вища sparsity.	Unstructured pruning не гарантує прискорення.
Quantization INT8	Менший model size, можливий speedup.	Latency залежить від backend та CPU.
Pruning + INT8	Поєднання sparsity і компактного зберігання.	Можливе накопичення похибок accuracy.
Knowledge distillation	Менша student-модель.	Потрібна teacher-модель.
Low-rank approximation	Скорочення параметрів у матричних шарах.	Ефект залежить від архітектури.

Важливе методологічне уточнення: у всіх експериментах FLOPs подаються як dense theoretical FLOPs. Для unstructured pruning вони не зменшуються, оскільки форма тензорів не змінюється. Для quantization кількість операцій також може залишатися тією самою, хоча змінюється тип арифметики.

Експериментальний протокол, моделі та датасети

Експериментальна частина має демонстраційний характер і спрямована на перевірку методики, а не на досягнення максимальної accuracy. Було використано два датасети: Fashion-MNIST як простий відтворюваний benchmark і CIFAR-10 як складніший приклад із кольоровими зображеннями.



Усі latency-вимірювання виконувалися на CPU при batch size = 1, фіксованій кількості потоків, у режимі model.eval() та torch.inference_mode(). Перед вимірюванням виконувалися 20 warm-up ітерацій, після чого проводилося 100 вимірюваних запусків. Подано mean, std, median і p95, але в основних таблицях наведено mean $\pm$ std.

Таблиця 2

Датасети та моделі

Датасет	Вхід	Модель	Параметри	Dense FLOPs
Fashion-MNIST	1×28×28, 10 класів	Simple CNN: 2 conv + FC	206 922	2.436 M
CIFAR-10	3×32×32, 10 класів	SmallCIFARCNN: 3 conv + FC	620 362	21.698 M

Як видно з таблиці 2, для перевірки методики використано два набори даних різної складності: Fashion-MNIST із одноканальними зображеннями та CIFAR-10 із кольоровими зображеннями. Це дає змогу оцінити, чи зберігається логіка порівняння моделей не лише на простому benchmark, а й на складнішому датасеті.

Для кожного датасету було сформовано однаковий набір експериментальних конфігурацій, наведений у таблиці 3.

Таблиця 3

Конфігурації експерименту

№	Конфігурація	Опис
1	Baseline FP32	Модель без структурної оптимізації.
2	Pruning 30%	Global unstructured pruning 30% + fine-tuning.
3	Pruning 50%	Global unstructured pruning 50% + fine-tuning.
4	Quantization INT8	Post-training static quantization FP32 $\rightarrow$ INT8.
5	Pruning 50% + INT8	Fine-tuned pruning 50%, потім INT8 quantization.

Memory footprint вимірювався як приблизна зміна RSS процесу під час інференсу. Ця метрика менш стабільна, ніж model size, оскільки



Python/PyTorch та операційна система можуть кешувати пам'ять. Тому model size розглядається як основний стабільний показник компактності.

Для практичної перевірки методики було сформовано єдиний експериментальний конвеєр, однаковий для Fashion-MNIST і CIFAR-10. Спочатку навчалася baseline-модель у форматі FP32 без структурної оптимізації. Після цього на її основі створювалися окремі оптимізовані конфігурації: модель із pruning 30%, модель із pruning 50%, модель після INT8-квантизації та комбінована модель Pruning 50% + INT8. Такий порядок дає змогу порівнювати всі варіанти з однаковою початковою архітектурою і окремо аналізувати вплив кожного методу оптимізації.

Для pruning використовувалося глобальне неструктуроване відсікання ваг. Це означає, що рівень pruning задавався для всієї моделі загалом, а не окремо для кожного шару. Ваги з меншими абсолютними значеннями занулювалися, але розміри шарів залишалися незмінними. Після цього виконувалося коротке fine-tuning-навчання з активними pruning masks. Такий етап потрібний для того, щоб модель адаптувалася до втрати частини ваг, а занулені значення не відновлювалися під час оновлення параметрів. Після завершення fine-tuning маски робилися постійними, і модель оцінювалася як окрема оптимізована конфігурація.

INT8-квантизація виконувалася як post-training static quantization. Для цього вже навчена FP32-модель переводилася в режим підготовки до квантизації: перед калібруванням виконувалося злиття суміжних операцій Conv/ReLU та Linear/ReLU, після чого задавався quantization backend і конфігурація квантизації. Далі через підготовлену модель пропускала частина навчальних даних для калібрування діапазонів активацій. Калібрування потрібне для визначення числових меж, у яких FP32-значення можуть бути наближено представлені у форматі INT8. Після калібрування



модель конвертувалася у квантизований варіант. У комбінованій конфігурації INT8-квантизація застосовувалася не до baseline-моделі, а до моделі після Pruning 50% і fine-tuning.

Оцінювання кожної конфігурації виконувалося за однаковою схемою. Спочатку визначалася accuracy на тестовій вибірці. Далі обчислювалися загальна кількість параметрів, кількість ненульових параметрів, overall sparsity та weight sparsity. Для квантизованих моделей ці структурні показники визначалися за відповідною FP32-моделлю-джерелом, оскільки після INT8-квантизації частина параметрів у PyTorch зберігається у packed-форматі й не завжди представлена як звичайні параметри моделі. Окремо зберігався файл кожної конфігурації, після чого визначався його розмір на диску. Для latency використовувався CPU-сценарій з batch size = 1, оскільки такий режим ближчий до інференсу окремих запитів у мобільних, вбудованих та edge-системах. Перед основним вимірюванням виконувалися warm-up-запуски, а потім фіксувалася серія вимірюваних запусків із розрахунком mean, standard deviation, median і p95.

Підрахунок FLOPs у дослідженні виконувався як dense theoretical FLOPs для фіксованої архітектури. Для Fashion-MNIST Simple CNN цей показник становив 2,436 М FLOPs, а для CIFAR-10 SmallCIFARCNN — 21,698 М FLOPs. Оскільки застосований pruning є unstructured, форма згорткових і повнозв'язних шарів не змінювалася. Тому dense FLOPs залишалися однаковими для baseline, pruning, quantization і комбінованих конфігурацій. Це відрізняє unstructured pruning від structured pruning, де фізичне видалення каналів або фільтрів може змінити кількість операцій.

Описаний порядок експерименту узгоджується з підходами, у яких pruning та 8-bit quantization розглядаються як засоби зменшення розміру моделі, memory footprint і latency для edge-сценаріїв [12]. Використання



кількох метрик відповідає ідеї порівняльного оцінювання нейронних мереж за точністю, швидкодією та ресурсними витратами [1]. Оскільки пряме енергоспоживання в роботі не вимірювалося, latency, FLOPs, model size і memory footprint трактуються як проху-показники ресурсної поведінки моделі; загальна доцільність урахування енергетичних витрат та AI-driven оптимізації підтверджується працями [13; 1; 6].

Результати Fashion-MNIST

Для Fashion-MNIST базова CNN досягла accuracy 90,85% при model size 0,7931 MB і середній CPU latency 0,3125 ms. Pruning 30% після fine-tuning підвищив accuracy до 91,67% і дав невелике зменшення latency. Pruning 50% забезпечив accuracy 91,79%, але latency зросла до 0,3332 ms. INT8-квантизація суттєво зменшила розмір моделі, проте latency у цьому CPU-сценарії збільшилася.

Таблица 4

Результати Fashion-MNIST

Конфігурація	Acc. %	Spars. %	Size MB	Latency ms	FLOPs M	E _{score}
Baseline FP32	90.85	0.00	0.7931	0.3125 ± 0.0711	2.4361	0.5396
Pruning 30% FP32	91.67	30.00	0.7933	0.3007 ± 0.0293	2.4361	0.5758
Pruning 50% FP32	91.79	50.00	0.7933	0.3332 ± 0.0608	2.4361	0.7559
Quantization INT8	90.80	0.00	0.2093	0.9075 ± 0.1918	2.4361	0.6000
Pruning 50% + INT8	91.68	50.00	0.2101	0.5790 ± 0.0688	2.4361	0.8524

Найвище значення E_{score} отримала конфігурація Pruning 50% + Quantization INT8: 0,8524. Вона поєднала accuracy 91,68%, weight sparsity 50% і зменшення model size приблизно на 73,5%. Хоча latency цієї конфігурації була вищою за baseline, інтегральна оцінка показала найкращий баланс між якістю та ресурсними характеристиками.



Цей результат демонструє, що аналіз лише accuracy або лише model size є неповним. Якщо орієнтуватися тільки на accuracy, перевагу мала б Pruning 50% FP32; якщо лише на розмір, найкращою виглядала б INT8-конфігурація. Багатокритеріальний підхід дозволяє врахувати обидва аспекти одночасно.

Результати CIFAR-10

На CIFAR-10 було перевірено, чи зберігається логіка методики на складнішому датасеті. Базова SmallCIFARCNN досягла accuracy 76,75%, model size 2,3705 MB і CPU latency 0,8732 ms. Після pruning 30% і 50% accuracy зросла до 80,23% та 80,41% відповідно, що пояснюється fine-tuning після pruning.

Таблиця 5

Результати CIFAR-10

Конфігурація	Acc. %	Spars. %	Size MB	Latency ms	FLOPs M	E _{score}
Baseline FP32	76.75	0.00	2.3705	0.8732 ± 0.1075	21.6975	0.4033
Pruning 30% FP32	80.23	30.00	2.3709	0.8765 ± 0.1263	21.6975	0.5903
Pruning 50% FP32	80.41	50.00	2.3709	0.8767 ± 0.0653	21.6975	0.6000
Quantization INT8	76.87	0.00	0.6108	0.6645 ± 0.0613	21.6975	0.8066
Pruning 50% + INT8	80.18	50.00	0.6117	0.6873 ± 0.1235	21.6975	0.9659

На відміну від Fashion-MNIST, у CIFAR-10 INT8-квантизація не лише зменшила model size приблизно на 74,2%, а й знизилу CPU latency з 0,8732 ms до 0,6645 ms. Комбінована конфігурація Pruning 50% + INT8 отримала accuracy 80,18%, weight sparsity 50%, model size 0,6117 MB і найвище E_{score} = 0,9659.

Цей експеримент підтверджує апаратну та модельну залежність quantization. На одному датасеті INT8 погіршив latency, на іншому — покращив. Саме тому методика не робить універсальних тверджень щодо окремого методу, а порівнює конфігурації за сукупністю критеріїв.



Обговорення результатів

Обидва експерименти показали, що найкращою за E_{score} стала комбінована конфігурація Pruning 50% + Quantization INT8. Проте це не слід трактувати як універсальне доведення переваги комбінації для всіх моделей. Правильніший висновок полягає в тому, що запропонована методика здатна виявляти збалансовані рішення, які не очевидні при аналізі однієї метрики. У таблиці 6 узагальнено результати ранжування моделей за інтегральним показником E_{score} . Для обох датасетів найвищу оцінку отримала конфігурація Pruning 50% + INT8, що свідчить про її кращий баланс між точністю класифікації та ресурсними характеристиками в межах проведених експериментів.

Таблица 6

Порівняння конфігурацій із найвищим значенням E_{score}

Датасет	Конфігурація	Accuracy %	Size MB	Latency ms	E_{score}
Fashion-MNIST	Pruning 50% + INT8	91.68	0.2101	0.5790	0.8524
CIFAR-10	Pruning 50% + INT8	80.18	0.6117	0.6873	0.9659

Результати підтверджують, що збільшення розрідженості моделі після unstructured pruning не є тотожним реальному прискоренню інференсу. У результатах weight sparsity досягала 30% і 50%, однак dense FLOPs не змінювалися. Розмір FP32-файлу після pruning також майже не зменшувався, оскільки стандартне dense-збереження фіксує нулі як звичайні значення.

Quantization мала різний вплив на latency. У Fashion-MNIST вона зменшила розмір моделі, але збільшила затримку; у CIFAR-10 — зменшила і розмір, і latency. Це свідчить, що ефект INT8-квантизації залежить не лише від самого методу, а й від архітектури моделі, розміру операцій, CPU та реалізації INT8-операцій у програмному backend. У цьому контексті



релевантною є робота S. Naveen і M. R. Kounte, де pruning та 8-bit quantization розглядаються як засоби зменшення розміру моделі, memory footprint і latency для edge-сценаріїв [7].

Методика також передбачає правило прийнятності за accuracy: модель не повинна обиратися лише через високий E_{score} , якщо втрата якості є неприйнятною для задачі. У наведених експериментах усі моделі залишилися в межах прийнятного accuracy drop, а комбінована конфігурація мала найкраще ранжування.

Висновки. Було запропоновано багатокритеріальну методику оцінювання структурно оптимізованих нейронних мереж на основі інтегрального показника E_{score} . Її відмінність від суто алгоритмічних досліджень полягає в тому, що вона не пропонує новий спосіб інференсу чи нову архітектуру, а систематизує правила порівняння моделей після застосування pruning, quantization та їх комбінації [12] за єдиною нормованою шкалою п'яти критеріїв: accuracy, model size, CPU latency, FLOPs і memory footprint [2].

Експерименти на Fashion-MNIST і CIFAR-10 підтвердили працездатність запропонованого підходу. По-перше, аналіз лише accuracy є недостатнім для вибору оптимальної конфігурації: у Fashion-MNIST найвищу точність мала Pruning 50% FP32, тоді як найкращий баланс за E_{score} — конфігурація Pruning 50% + INT8 (0,8524). По-друге, вплив quantization на латентність є апаратно залежним: у Fashion-MNIST INT8 збільшив CPU latency, тоді як у CIFAR-10 — зменшив. Це підтверджує необхідність емпіричного тестування на цільовому обладнанні, а не екстраполяції теоретичних очікувань. По-третє, unstructured pruning, незважаючи на зростання weight sparsity до 30–50%, не зменшує dense FLOPs і не гарантує реального прискорення без підтримки sparse-kernels. Зменшення кількості ненульових параметрів є структурним артефактом і не тотожне зниженню



обчислювального навантаження в стандартних умовах виконання. По-четверте, комбінована конфігурація Pruning 50% + INT8 на обох датасетах отримала найвище значення E_score (0,8524 для Fashion-MNIST і 0,9659 для CIFAR-10). Це свідчить, що в межах проведених експериментів поєднання pruning та INT8-квантизації забезпечило найкращий інтегральний баланс між accuracy та ресурсними характеристиками за умови дотримання прийнятного рівня accuracy drop.

Разом із тим необхідно чітко окреслити обмеження отриманих результатів. По-перше, у дослідженні не застосовувалися засоби прямого вимірювання енергоспоживання — ватметри, лічильники RAPL/NVML або спеціалізовані профілери споживання. Тому E_score є порівняльним прохупоказником, а не абсолютною фізичною мірою енергії у ватах або джоулях. По-друге, застосований unstructured pruning не змінює фізичну структуру тензорів, що унеможливорює реальне скорочення FLOPs без спеціалізованих sparse-операторів. По-третє, memory footprint, виміряний як приріст RSS у Python/PyTorch, є нестабільним показником через кешування пам'яті операційною системою; надійнішим залишається model size на диску. По-четверте, всі експерименти проведено на компактних CNN із використанням Fashion-MNIST та CIFAR-10, що не репрезентують сучасні промислові архітектури: результати не можна автоматично переносити на ResNet, ViT, Transformer або виробничі edge-обладнання без додаткових досліджень. По-п'яте, вагові коефіцієнти у формулі E_score задані рівними ($w = 0,20$ для кожного компонента), тоді як у реальних задачах їх необхідно визначати відповідно до конкретних вимог застосунку.



Список використаних джерел:

1. Гнатюк В., Литвинюк О. Метод прогнозування завантаження та розподілу ресурсів з використанням нейронних мереж у телекомунікаційних системах. *Вісник Хмельницького національного університету. Технічні науки*. 2026, № 363(2), 224–229.

2. Завгородня Г. А., Завгородній В. В. . Оптимізація продуктивності комп'ютерних ігор на основі методів машинного навчання. *Системи та технології*. 2026. № 1(71), 45–51.

3. Кучерук В., Кулаков П., Ліщук Р., Концеба С., Маньковська В. . Критерії ефективності та якості нейронних мереж. *Вимірювальна та обчислювальна техніка в технологічних процесах*. 2025, № 3, 30–39.

4. Мамчич О., Волк М. . Оцінювання енергетичних витрат у процесі використання мобільних пристроїв для хмарних обчислень. *Innovative Technologies and Scientific Solutions for Industries*, 2023, 1(23), 72–82.

5. Супрун О. М., Огірко О. І., Кримська А. О. . Розроблення гібридних нейромережевих моделей для прогнозування в інтелектуальних системах. *Вісник Херсонського національного технічного університету*. 2025. № 1(92), Ч. 2, 222 – 229.

6. A. Stecyk, I. Miciuła. Harnessing the Power of Artificial Intelligence for Collaborative Energy Optimization Platforms. *Energies*. 2023. Vol. 16(13). Article 5210.

7. E. C. Onukwulu, M. O. Agho, N. L. Eyo-Udo. Developing a framework for AI-driven optimization of supply chains in energy sector. *Global Journal of Advanced Research and Reviews*. 2023. Vol. 1(2). P. 82–101.

8. K. M. T. A. Anick. AI-Enabled Decision Support Systems for Industrial Energy Optimization in U.S. Manufacturing. *ASRC Procedia: Global Perspectives in Science and Scholarship*. 2025. Vol. 1(1). P. 2160–2201.



9. K. T. Chui, M. D. Lytras, A. Visvizi. Energy Sustainability in Smart Cities: Artificial Intelligence, Smart Monitoring, and Optimization of Energy Consumption. *Energies*. 2018. Vol. 11(11). Article 2869.

10. M. S. Arimbi, R. Amini, A. K. Asri, N. Martin. Artificial Intelligence in Renewable Energy: A Review of Predictive Maintenance and Energy Optimization. *Journal of Clean Technology*. 2025. Vol. 2(1). P. 29–44.

11. P. Biswas, A. Rashid, A. Biswas, M. A. Al Nasim, S. Chakraborty, K. D. Gupta, R. George. AI-driven approaches for optimizing power consumption: a comprehensive survey. *Discover Artificial Intelligence*. 2024. Vol. 4. Article 116.

12. S. Naveen, M. R. Kounte. Optimized Convolutional Neural Network at the IoT edge for image detection using pruning and quantization. *Multimedia Tools and Applications*. 2025. Vol. 84. P. 5435–5455.

13. Y. B. Melnyk. Introduction to the Monograph on Artificial Intelligence in Digital Society. *Artificial Intelligence in Digital Society*. 2026. Vol. 1. P. 5–9.

Vydavatel:

Mezinárodní Ekonomický Institut s.r.o.
se sídlem V Lázních 688, Jesenice 252 42
IČO 03562671 Česká republika

MODERNÍ ASPEKTY VĚDY

Svazek LXX mezinárodní kolektivní monografie

Podepsáno k tisku 11. červenec 2026
Formát 60x90/8. Ofsetový papír a tisk
Headset Times New Roman.
Mysl. tisk. oblouk. 8.2. Náklad 100 kopií.



МУКАЧІВСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

89600, м. Мукачево, вул. Ужгородська, 26

тел./факс +380-3131-21109

Веб-сайт університету: www.msu.edu.ua

E-mail: info@msu.edu.ua, pr@mail.msu.edu.ua

Веб-сайт Інституційного репозитарію Наукової бібліотеки МДУ: <http://dspace.msu.edu.ua:8080>

Веб-сайт Наукової бібліотеки МДУ: <http://msu.edu.ua/library/>