Міжнародний конкурс студентських робіт зі штучного інтелекту 2026

II тур
Перший етап

Р Е Ц Е Н З І Я

на наукову роботу

назва Benchmarking Loss Functions for Classifier Training on Imbalanced Datasets

шифр Margin Cascade

представлену на Міжнародний конкурс студентських робіт зі штучного інтелекту

№ з/п	Характеристики та критерії оцінки рукопису наукової роботи	Рейтингова оцінка. Максимальна кількість балів (за 120-бальною шкалою)	Бали
1	Актуальність проблеми	10	8,4
2	Новизна та оригінальність ідей	15	10,8
3	Використані методи дослідження	15	11,2
4	Теоретичні наукові результати	10	6,0
5	Практична направленість результатів	20	14,6
6	Рівень використання наукової літератури та інших джерел інформації	5	4,0
7	Ступінь самостійності роботи	10	7,4
8	Якість оформлення	5	4,2
9	Наукові публікації	10	2,0
10	Недоліки роботи (пояснення зниження максимальних балів у пунктах 1-9): Пояснення gradient sparsity суперечить формулі MCM; помилка у формулі Фрідмана та неясність одиниці статистичного аналізу; надмірні висновки про калібрування Focal Loss. Наведено порівняльний аналіз наборів даних, але не описано детально ці набори. Необхідно виправити теоретичну інтерпретацію MCM, уточнити статистичний протокол і перерахувати відповідні висновки.		
10.1	Наявність розробленого програмного забезпечення, у т.ч. веб-сервісу, мобільного застосунку	10	7,2
10.2	Рівень отриманих практичних результатів наукової роботи	10	6,8
Сума балів			82,6

Загальний висновок не рекомендується для захисту на ПНПК
(рекомендується, не рекомендується для захисту на науково-практичній конференції)

Рецензент *члени Конкурсної комісії напряму «Обчислювальний інтелект»*

«06» листопада 2026 року

Порівняння функцій втрат за дисбалансу класів з одночасним аналізом розпізнавання та ймовірнісного калібрування є актуальним. Вибір табличних та зображувальних даних і розгляд ціни обчислень відповідають практичній постановці.

Корисним є єдиний протокол порівняння Log, Focal, Poly та MCM і спроба перенесення MCM на багатокласові нейромережі. Водночас основні функції відомі, джерело походження MCM [6] не ідентифіковане, а заявлена новизна математичного пояснення обмежень MCM не підтверджена через суперечність між формулою та інтерпретацією. Межі нового авторського внеску потрібно уточнити. Сильні сторони: контроль однакових входів CovType, групування Malimg за хешем, офіційний зовнішній тест NSL-KDD, хронологічний поділ Credit Card Fraud, навчання перетворень усередині фолдів, фіксована архітектура та десять парних початкових значень генератора. Однак у статистичному аналізі не узгоджені три CV-фолди й десять фінальних запусків. Формула Фрідмана неправильна для визначених у тексті сум рангів, а omnibus-тест не доводить окрему перевагу чи поразку MCM. Відсутність значущості не доводить еквівалентність методів. Потрібні первинні парні метрики, коректні попарні порівняння та поправки на множинність.

На с. 12 при помилці задано $L=1+\beta\cdot\ln[(p_win+\delta)/(p_t+\delta)]$. Ця формула прямо залежить від p_t незалежно від місця істинного класу в ранжуванні. Отже пояснення на с. 24 і 26, що істинний клас поза top-2 ігнорується, не випливає з визначеного MCM. При $\delta=0$ логарифмічне відношення softmax-імовірностей дорівнює z_win-z_t, тому градієнт за істинним логітом дорівнює $-\beta$ навіть для третього або нижчого класу. За додатного δ змінюється масштаб, але не зникає явна залежність. Також $m\le1$, тож $\exp(-\lambda m)$ при $m\to1$ прямує до $\exp(-\lambda)$, а не до нуля; для обраного $\lambda=2$ це близько 0,135. Потрібне узгодження математики з фактичним кодом, а причини деградації наразі залишаються неперевіреними гіпотезами.

Робота містить практично корисний benchmark із оцінкою Balanced Accuracy, Macro F1, NLL, ECE і часу навчання. Проте рекомендації уникати Focal Loss у безпекових задачах надмірно загальні: оцінено конкретні архітектури, сітки параметрів і відсутність післянавчального калібрування. Не перевірено temperature scaling, класові ваги та альтернативні пороги рішень. Для рідкісного шахрайства бракує PR-AUC, precision/recall позитивного класу та оцінки ціни помилок; загальний ECE може приховувати проблеми меншості.

Деталізований протокол і лістинг оркестратора підтримують позитивну оцінку самостійної експериментальної роботи. Автор відкрито декларує використання генеративного ШІ (с. 28); сама декларація не є підставою для зниження.

Рукопис має чітку структуру, опис архітектур і налаштувань, таблиці та графіки. Недоліки: у табл. 5 розкид наведено лише для Balanced Accuracy, хоча текст обіцяє середні та стандартні відхилення для первинних і вторинних метрик; статус BatchNorm у таблицях позначено N/A без пояснення.

У додатку А наведено багатопроцесний GPU-оркестратор із фазами smoke/CV/final, журналюванням, вибором параметрів і викликом аналізу. Це змістовне підтвердження програмної частини. Однак без src.paper_experiment, реалізацій втрат, analyze_paper_results.py, конфігурацій середовища та даних цей лістинг не є самодостатнім. Особливо потрібна перевірка фактичного MCM щодо формули розділу 2.

Подано повний набір порівнянь чотирьох втрат на чотирьох датасетах. Balanced Accuracy MCM на Malimg 0,7491±0,0793 проти Log Loss 0,9200±0,0052 є помітним емпіричним результатом; негативні результати теж мають наукову цінність. Водночас доказ статистичної значущості потребує перерахунку, а причинне пояснення MCM – перегляду. Твердження, що Focal Loss завжди погіршує калібрування, суперечить навіть табл. 5: на NSL-KDD його ECE 0,1850 краще за 0,2084 для Log Loss, а NLL 1,2076 краще за 1,9156. ECE не визначає напрям помилки — надмірну чи недостатню впевненість.

Для сум рангів R_j стандартна статистика без поправки на зв'язки має вигляд $Q = 12\cdot\Sigma R_j^2 / [N\cdot k\cdot(k+1)] - 3N(k+1)$. У рукописі наведено множник $12N / [k(k+1)]$, який відповідає іншому визначенню рангових величин, але R_j прямо названі сумами. Треба виправити або формулу, або визначення. Якщо блоками є саме три CV-фолди й порівнюються чотири втрати, максимальна стандартна статистика Q дорівнює $N(k-1)=9$. Для χ^2-наближення з трьома ступенями вільності це дає мінімальне $p\approx0,0293$. Значення 0,0002 і 0,0003 з табл. 6 тому не узгоджуються з описаною трифолдовою χ^2-процедурою. Вони можуть стосуватися іншої кількості блоків, зокрема десяти фінальних запусків, - це слід установити за кодом і первинними даними. За трьох блоків саме асимптотичне наближення також потребує обережності.

Питання та рекомендації

1. Показати код MCM і тест градієнта на прикладі трьох класів, де істинний клас має найменшу ймовірність. Якщо код використовує лише top-2, визнати його відмінність від формули; якщо формулу, переглянути пояснення gradient sparsity.
2. Надати матрицю парних результатів, визначення статистичних блоків, Q, ступені вільності та поправку на зв'язки. Розмежувати мінливість від ініціалізації на фіксованому тесті й невизначеність узагальнення на нові дані.
3. Уточнити вибір гіперпараметрів і використання валідації всередині CV. Переконатися, що одна й та сама валідація не використана як незалежний фінальний тест.
4. Додати reliability diagrams, середню впевненість проти точності, класове калібрування та temperature scaling. ECE з 15 бінів саме по собі не доводить overconfidence.
5. Обмежити висновки фактичними даними: на NSL-KDD Focal Loss поліпшує ECE/NLL; на Credit Card відсутність значущої різниці не доводить еквівалентності втрат. Порівнювати практичну величину ефекту разом зі статистичною невизначеністю.

Робота є змістовним експериментальним дослідженням із сильними рішеннями щодо ізоляції даних і порівнюваності запусків. Основні числові результати становлять інтерес, проте теоретичне пояснення MCM, статистична значущість і загальні рекомендації щодо калібрування потребують суттєвого уточнення.