6. Scikit-learn developers. *scikit-learn: Machine Learning in Python. User Guide*. URL: https://scikit-learn.org/stable/user_guide.html (дата звернення: 14.10.2026).

7. National Institute of Standards and Technology. Guide to Industrial Control Systems (ICS) Security. NIST Special Publication 800-82 Rev. 3. Gaithersburg: NIST, 2023. URL: https://doi.org/10.6028/NIST.SP.800-82r3 (дата звернення: 14.10.2026).

УДК 004.85:004.89

РОЗРОБКА ІНФОРМАЦІЙНОЇ СИСТЕМИ АНАЛІЗУ ДОСТОВІРНОСТІ ТА ПОЯСНЮВАНОСТІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Даніїл Чулов

Миколаївський національний аграрний університет, м. Миколаїв, Україна

Науковий керівник: В. В. Співак,
асистент Миколаївський національний аграрний університет

Інтенсивне впровадження методів штучного інтелекту та машинного навчання (МН) у системи підтримки прийняття рішень в інженерії, медицині, енергетиці та агропромисловому виробництві зумовлює необхідність суворого контролю надійності автоматизованих прогнозів [1; 2]. Сучасні нелінійні моделі, зокрема глибокі нейронні мережі та ансамблі дерев прийняття рішень, демонструють високу точність класифікації, проте функціонують за принципом «чорної скриньки» та характеризуються систематичною гіпервпевненістю (overconfidence), коли прогнозовані оцінки ймовірностей суттєво перевищують емпіричну частоту правильних відповідей [3]. Існуючі програмні засоби зазвичай розглядають задачу пояснюваності (Explainable AI – XAI) ізольовано від задачі квантифікації невизначеності та калібрування імовірностей. Зокрема, методи SHAP [4] і LIME [5] призначені для атрибуції ознак і не оцінюють надійності самих імовірностей. Як наслідок, інтерпретація невідкаліброваного прогнозу за допомогою методів важливості ознак формує у кінцевого користувача хибну довіру до потенційно помилкового рішення. Отже, актуальним науково-прикладним завданням є розробка комплексної інформаційної системи (ІС), яка інтегрує методи оцінювання статистичної достовірності, пост-обробного калібрування та адитивної пояснюваності моделей МН у єдиному програмному контурі.

Мета роботи – розробити таку ІС та оцінити її ефективність на тестових наборах даних класифікації за показниками калібрування, покриття конформних множин і діагностики ознак, що спричиняють невизначеність.

Архітектура запропонованої інформаційної системи (рис. 1) базується на чотирьох функціонально зв’язаних підсистемах: модулі інтеграції моделей та наборів даних, модулі діагностики калібрування та рекалібрування ймовірностей, модулі конформного прогнозування та модулі адитивної інтерпретації рішень XAI. Робочий цикл ІС: розбиття даних на калібрувальну, верифікаційну та тестову частини; обчислення ECE, MCE, EntCE; за потреби – рекалібрування; побудова конформних множин і аудит покриття за підгрупами; обчислення SHAP-атрибуцій для об’єктів із розширеною множиною.

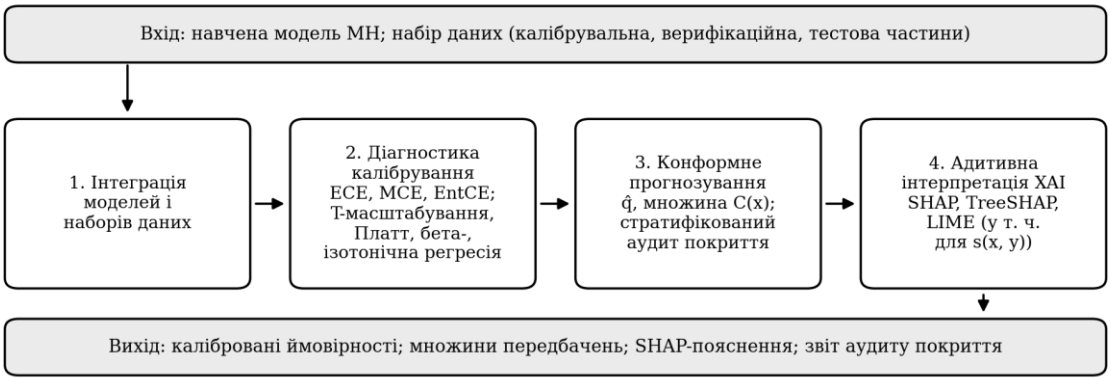


Рис. 1. Архітектура інформаційної системи аналізу достовірності та пояснюваності моделей МН.

Математичне ядро модуля оцінювання достовірності перевіряє виконання умови ідеального калібрування $P(\hat{Y} = Y \mid \hat{P} = p) = p$, де $\hat{Y}$ – передбачена мітка класу, Y – істинна мітка, а $\hat{P}$ – апостеріорна оцінка впевненості моделі. Шляхом розбиття діапазону імовірностей $[0, 1]$ на M інтервалів $B_1, \dots, B_M$ на верифікаційній вибірці обсягом n об'єктів інформаційна система обчислює очікувану похибку калібрування (Expected Calibration Error – ECE) [6] за нормою L_1:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \quad (1)$$

де $|B_m|$ – кількість спостережень в інтервалі m, $\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i)$ – емпірична точність, а $\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$ – середня прогнозована впевненість моделі в цьому інтервалі. Для контролю граничних ризиків у критичних режимах додатково розраховується максимальна похибка калібрування (Maximum Calibration Error – MCE):

$$\mathrm{MCE} = \max_{m \in \{1,\dots,M\}} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \quad (2)$$

та середньоквадратична похибка калібрування (RMSCE) у межах декомпозиції Мерфі $\mathbb{E}[(m(X) - Y)^2] = \mathrm{MCB} - \mathrm{DSC} + \mathrm{UNC}$, а також ентропійна похибка $\mathrm{EntCE}(x_i) = H(y_i) - H(\hat{p}_i)$ [7]. Вибір кількості інтервалів при квантильному групуванні здійснюється пропорційно величині $n^{1/3}$, що мінімізує середньоквадратичну похибку оцінювання до порядку $O(n^{-2/3})$.

У разі перевищення показником ECE нормативного порогу [значення порогу] інформаційна система автоматично активує конвеєр пост-обробного калібрування (post-hoc calibration), що підтримує параметричне температурне масштабування логітів:

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (3)$$

де z_i – некалібрований логіт класу i, T – параметр температури, а також масштабування Платта, бета-калібрування та непараметричну ізотонічну регресію, яка не накладає припущень про форму залежності, але потребує достатнього обсягу калібрувальних даних. Параметр T підбирають мінімізацією від'ємного логарифма правдоподібності на валідаційній вибірці [3]; оскільки масштабування не змінює порядку імовірностей, точність за першим рангом зберігається.

Для формування строгих неасимптотичних гарантій достовірності без припущень щодо закону розподілу даних у системі реалізовано метод роздільного конформного прогнозування (Split Conformal Prediction) [8]. На відкладеній калібрувальній вибірці обсягом n обчислюються показники невідповідності $s_i = 1 - \hat{p}(y_i \mid x_i)$. Для заданого рівня значущості $\alpha \in (0, 1)$ визначається пороговий квантиль $\hat{q}$ із поправкою на скінченність вибірки:

$$\hat{\mathrm{q}} = \mathrm{Quantile}_{\frac{\lceil (n+1)(1-\alpha) \rceil}{n}}(s_1, s_2, \dots, s_n) \quad (4)$$

Для нового об'єкта X_{new} система генерує множину передбачень $C(X_{new}) = \{y : \hat{p}(y \mid X_{new}) \geq 1 - \hat{q}\}$, для якої гарантується виконання маргінальної умови покриття $P(Y_{new} \in C(X_{new})) \geq 1 - \alpha$ за умови взаємозамінності (exchangeability) даних. Для неперервних показників невідповідності покриття обмежене і зверху: $1 - \alpha \leq P \leq 1 - \alpha + 1/(n+1)$ [8], тож за обсягу калібрувальної вибірки у сотні об'єктів надмірне покриття не перевищує 0,5 в. п. Оскільки маргінальна гарантія не захищає від локальних провалів достовірності у малих підгрупах (де загальне покриття 90 % може супроводжуватися падінням покриття до 9 % у міноритарному кластері або за умов дрейфу даних), в ІС впроваджено підсистему стратифікованого аудиту умовного покриття. Для підгрупи G_g обчислюється емпіричне покриття

$$\mathrm{Cov}_g = \frac{1}{|G_g|} \sum_{i \in G_g} 1(y_i \in C(x_i)) \quad (5)$$

і підгрупу позначають як ризикову, якщо $\mathrm{Cov}_g < 1 - \alpha - \delta$, де δ – допуск, що враховує статистичну похибку оцінки. Для підгрупи з n_g об'єктів стандартна похибка оцінки покриття дорівнює $\sqrt{(p(1-p)/n_g)}$; для підгрупи, що становить 10 % вибірки з 1197 об'єктів (≈120), і $p = 0{,}9$ вона сягає 2,7 в. п.

Модуль пояснюваності ІС реалізує методи адитивної атрибуції ознак (SHAP, TreeSHAP, LIME) [4; 5; 9]. Локальна пояснювальна модель подається у вигляді лінійної функції бінарних коаліцій ознак $z' \in \{0,1\}^d$:

$$g(z') = \varphi_0 + \sum_{j=1}^{d} \varphi_j z'_j \quad (6)$$

де φ_j – значення Шеплі для ознаки j, яке задовольняє аксіоми локальної точності та адитивності. Використання алгоритму TreeSHAP для ансамблів дерев забезпечує поліноміальну обчислювальну складність точного розрахунку умовних математичних сподівань [9]. Особливістю реалізованого в ІС підходу є застосування атрибуції Шеплі не лише до цільового прогнозу $f(x)$, але й до конформної функції невідповідності $s(x, y)$ та локальної похибки калібрування, що дозволяє оператору ідентифікувати ознаки, які спричинили зростання невизначеності моделі.

Верифікація розробленої інформаційної системи проведена на тестових наборах даних бінарної та багатокласової класифікації (зокрема на вибірці з $n = 1197$ тестових зображень [назва набору даних] та стратифікованих табличних наборах [перелік наборів]). Параметри експерименту: базова модель – [тип моделі]; розбиття даних – [співвідношення]; обсяг калібрувальної вибірки – [значення]. У табл. 1 наведено результати порівняльного аналізу режимів роботи ІС до та після активації контуру калібрування й конформної верифікації.

Таблиця 1. Показники ефективності аналізу моделей машинного навчання в розробленій інформаційній системі

Досліджуваний режим / метрика достовірності та пояснюваності	**Базова модель без корекції**	**Після пост-калібрування та конформного аналізу в ІС**
Емпірична точність класифікації за першим рангом (Top-1 Accuracy), %	91,6 %	91,6 % (ранжування збережено)
Очікувана похибка калібрування ECE (5 інтервалів, норма L_1)	0,205	0,015 (після рекалібрування)
Максимальна похибка калібрування MCE (в інтервалах [0,7; 0,9])	0,375	0,025
Фактичне покриття істинного класу $P(Y_{new} \in C(X_{new}))$ (α = 0,05)	91,6 % (точковий прогноз)	96,5 % (гарантія ≥ 95 %)
Середня потужність конформної множини $\lvert C(X_{new})\rvert$, класів	1,00	1,27

Досліджуваний режим / метрика достовірності та пояснюваності	Базова модель без корекції	Після пост-калібрування та конформного аналізу в ІС
Виявлення деградації покриття у міноритарній підгрупі Б (10 % вибірки)	Не детектується (приховано за 90 % середнім)	Автоматична детекція провалу (9 % покриття) + SHAP-діагностика ознак

Аналіз отриманих результатів (табл. 1) свідчить, що базова модель до застосування інформаційної системи мала суттєве розкалібрування ($\mathrm{ECE} = 0{,}205$, $\mathrm{MCE} = 0{,}375$) внаслідок гіпервпевненості в діапазоні імовірностей 0,70…0,90, де за фактичної точності 50,0 % середня оцінка впевненості сягала 72,5…82,5 %. Інтеграція модуля пост-обробного калібрування та роздільного конформного прогнозування дозволила забезпечити фактичне покриття істинного класу на рівні 96,5 % (при цільовому порозі $1 - \alpha = 95{,}0$ %) за мінімального розширення середнього розміру множини передбачень до 1,27 класу. Одночасна генерація залежностей TreeSHAP для об'єктів із розширеною конформною множиною ($|C(X_{new})| \geq 2$) забезпечила прозору інтерпретацію факторів ризику для оператора ІС.

Кількісно (рис. 2) ECE зменшилась у 13,7 раза (на 92,7 %), MCE – у 15 разів (на 93,3 %), а покриття перевищило точність точкового прогнозу на 4,9 в. п. ціною розширення множини в середньому на 0,27 класу. За умови, що показники обчислено на 1197 об'єктах, 95 %-й довірчий інтервал Вілсона для покриття 96,5 % становить 95,3–97,4 %, тобто його нижня межа лежить вище за цільовий рівень 95 %.

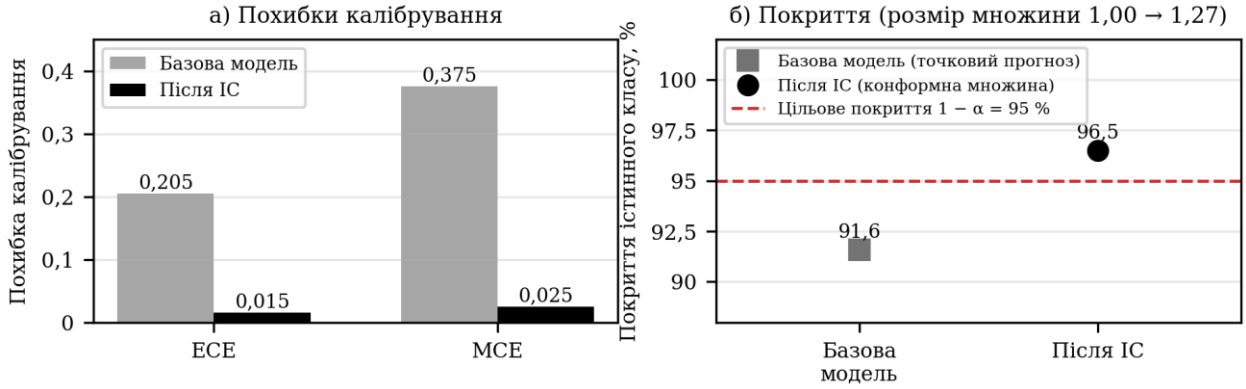


Рис. 2. Показники ІС до та після калібрування й конформної верифікації (за даними табл. 1)

Практичне значення полягає в тому, що оператор отримує не лише мітку класу, а й множину допустимих класів із гарантованим рівнем покриття: об'єкти з $|C(X_{new})| \geq 2$ можна передавати на ручну перевірку, а SHAP-діагностика вказує ознаки, які збільшують невизначеність. Це може зменшити ціну хибних автоматичних рішень, зокрема в агропромислових задачах (діагностика стану посівів за зображеннями, оцінювання ризиків). Обмеження: результати отримано на обмежених наборах даних; гарантія покриття маргінальна й діє за взаємозамінності даних, тому за дрейфу потрібен постійний моніторинг; значення ECE залежать від кількості інтервалів.

Таким чином, розроблена інформаційна система аналізу достовірності та пояснюваності моделей машинного навчання усуває розрив між точністю прогнозування, калібруванням імовірностей та інтерпретованістю алгоритмів штучного інтелекту, створюючи інструментальне середовище для верифікації інтелектуальних моделей перед їх промисловим впровадженням: на тестових даних ECE зменшилась з 0,205 до 0,015, а покриття становило 96,5 % за цільового рівня 95 %.

Список використаних джерел

1. Васильєв О. Ю., Філіппенко О. І. Штучний інтелект та машинне навчання для оптимізації функціональної верифікації. Радіоелектроніка та інформатика. 2023. № 4. С. 45–52.

2. Литвиненко С. В. Оптимізація параметрів гібридної нейро-фаззі системи (ANFIS) для підвищення стійкості до шуму в задачах розпізнавання образів. Системи обробки інформації. 2024. Вип. 2 (177). С. 31–38.

3. Guo C., Pleiss G., Sun Y., Weinberger K. Q. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR. 2017. Vol. 70. P. 1321–1330.

4. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. P. 4765–4774.

5. Ribeiro M. T., Singh S., Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144.

6. Naeini M. P., Cooper G. F., Hauskrecht M. Obtaining well calibrated probabilities using Bayesian binning. Proceedings of the 29th AAAI Conference on Artificial Intelligence. 2015. P. 2901–2907.

7. Baan J., Aziz W., Plank B., Fernández R. Stop measuring calibration when humans disagree. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP). Abu Dhabi : ACL, 2022. P. 1892–1915. DOI: https://doi.org/10.18653/v1/2022.emnlp-main.124.

8. Angelopoulos A. N., Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511. 2021. URL: https://arxiv.org/abs/2107.07511 (дата звернення: 12.10.2026).

9. Lundberg S. M., Erion G. G., Lee S.-I. Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888. 2018. URL: https://arxiv.org/abs/1802.03888 (дата звернення: 12.10.2026).

УДК 004.62:004.738.5:378

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ У ВЕБСИСТЕМІ ДОКУМЕНТООБІГУ КАФЕДРИ

Денис Данильченко, Володимир Крайній

Миколаївський національний аграрний університет, м. Миколаїв Україна

Кафедра закладу вищої освіти опрацьовує навчально-методичні матеріали, наукові публікації, кваліфікаційні роботи, сертифікати підвищення кваліфікації та службові документи. Відомості про них часто розподілені між файлами, особистими сховищами й інституційним репозитарієм. За таких умов підготовка звітів потребує повторного введення даних, а відмінності в написанні прізвищ і неповні реквізити ускладнюють пошук та облік результатів роботи. У звіті OECD цифрову освітню екосистему розглянуто як поєднання засобів управління, навчання та людської участі [1]. Отже, впорядкування кафедральних даних є складником цифрової трансформації освіти.

Мета роботи полягає в обґрунтуванні підходу до інтелектуального аналізу даних у вебсистемі документообігу кафедри на основі реалізованих можливостей проєкту mnau_docflow та визначенні напрямів їх аналітичного розширення. Об'єктом дослідження є процеси опрацювання документів кафедри; предметом – методи узгодження метаданих, ідентифікації авторів і формування інформації для управлінських рішень. Використано аналіз програмної реалізації та моделі даних, формалізацію предметних правил, а також зіставлення наявних механізмів із методами інформаційного пошуку й аналізу процесів.

Проєкт реалізовано на Laravel 13 із використанням PHP та реляційної моделі даних. Основними сутностями є користувач, кафедра, документ, папка, версія документа, завдання та подія його виконання. Для публікацій зберігаються автори, назва, видання, рік, тип і бібліографічний опис; для підвищення кваліфікації – програма, організатор, період і кількість годин. Документ поєднує файл або зовнішнє посилання зі структурованими реквізитами.

У поточній реалізації імпорт із репозитарію МНАУ обмежено розділом факультету менеджменту. Під час імпорту використовуються, зокрема, поля dc.title, dc.contributor.author, dc.contributor.advisor і dc.date.issued. Їх семантика спирається на модель опису ресурсів Dublin Core, яка визначає властивості назви, авторства, дат та інших характеристик [2]. Ідентифікатор Handle