

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ КЛАСИЧНИХ ТА ШІ-ОРІЄНТОВАНИХ МЕТОДІВ ПРОГНОЗУВАННЯ НА ПОТОКОВИХ ДАНИХ У РЕАЛЬНОМУ ЧАСІ

¹Іван Хамар, ²Ігор Оленич

STUDY OF THE EFFECTIVENESS OF CLASSICAL AND AI-ORIENTED METHODS FOR REAL-TIME STREAMING DATA PREDICTION

¹Ivan Khamar, ²Ihor Olenych

¹<https://orcid.org/0009-0000-0514-903X>, ²<https://orcid.org/0000-0002-6642-0222>

Львівський національний університет імені Івана Франка,
вул. Драгоманова, 50, Львів, 79005, Україна

Abstract. *This paper presents a comparative analysis of traditional machine learning and AI-based methods for processing large-scale streaming weather data collected via Apache Kafka infrastructure. Five approaches were implemented and evaluated: Random Forest (traditional), Multi-Layer Perceptron, TabNet, LightGBM (AI-based), and Gradient Boosting with MLP (hybrid). The study utilized real-time weather data from 440 geographic locations streamed through Confluent Cloud, with temporal split validation to ensure realistic evaluation. On dataset of 49,494 records, LightGBM achieved the highest accuracy ($R^2 = 0.770$, $MAE = 1.07$), outperforming traditional Random Forest ($R^2 = 0.646$) and deep learning methods. MLP demonstrated competitive performance ($R^2 = 0.547$), while TabNet showed limited effectiveness ($R^2 = 0.306$) despite 1,536 seconds training time. The study confirms that AI-enhanced gradient boosting methods outperform both traditional approaches and pure deep learning on tabular streaming data.*

У сучасних умовах цифрової трансформації спостерігається експоненційне зростання обсягів поточкових даних реального часу з різних джерел. Обробка таких потоків вимагає не лише високої пропускної здатності, але й ефективних алгоритмів прогнозування, здатних адаптуватися до динамічних умов [1]. Традиційні методи машинного навчання (Random Forest, Gradient Boosting) демонструють стабільність та інтерпретованість на табличних даних, проте їх можливості обмежені при виявленні складних нелінійних залежностей [2]. ШІ-орієнтовані підходи — від глибоких нейронних мереж (MLP) до спеціалізованих архітектур для табличних даних (TabNet) та підсилені ШІ методи бустингу як LightGBM — обіцяють вищу точність на великих даних за умови коректних гіперпараметрів [3].

У роботі досліджено ефективність п'яти підходів до регресійного прогнозування метеорологічних параметрів (атмосферний тиск та швидкість вітру) на основі потокових даних, зібраних через розподілену платформу Apache Kafka (Confluent Cloud). Дані отримано з OpenWeather API для 440 географічних локацій (міст Європи, Азії, Америки та Африки) з інтервалом опитування 20-30 секунд, що сформувало датасет обсягом 49,494 записи. Кожен запис містить часову мітку, температуру, вологість, хмарність, координати, а також цільові змінні — атмосферний тиск (hPa) та швидкість вітру (m/s). Для навчання використано часовий розподіл: 70% хронологічно ранніх даних (34,645 records) для тренування, 30% пізніших (14,849 records) — для тестування, що виключає temporal leakage та реалістично оцінює здатності моделей до прогнозування [4]. Feature engineering включав створення поліномних ознак (temperature^2 , humidity^2), взаємодії ($\text{temperature} \times \text{humidity}$), та часових features (hour, day_of_week), що сформувало 9-вимірний простір ознак після StandardScaler нормалізації.

Реалізовано п'ять моделей: Random Forest з параметрами $n_estimators=100$, $max_depth=10$; Multi-Layer Perceptron з архітектурою $128 \rightarrow 64 \rightarrow 32$ нейронів, адаптивною learning rate та early stopping; TabNet [5] з параметрами $n_d=64$, $n_steps=5$ для механізму уваги; LightGBM [6] з $iterations=500$, $depth=8$ для ШІ-покращеного бустингу; та Gradient Boosting [7] з інтеграцією MLP [8] для рекомендацій. Для оцінювання застосовано метрики MAE, RMSE та R^2 , а також показники: latency (ms), throughput (RPS), training time (sec) та memory footprint (MB) [9].

На наборі даних 49,494 записів найвищу точність продемонстрував LightGBM ($R^2 = 0.770$, MAE = 1.07, RMSE = 1.44), показавши позитивний generalization gap ($R^2 \text{ train} = 0.762$), що свідчить про відсутність перенавчання. Random Forest як представник традиційного підходу досяг $R^2 = 0.646$ (MAE = 1.29), поступаючись LightGBM на 0.124 R^2 , але забезпечуючи швидке тренування (2.39 сек проти 8.05 сек). MLP продемонстрував конкурентний результат $R^2 = 0.547$ (MAE = 1.66) з хорошою генералізацією ($R^2 \text{ train} = 0.538$), підтверджуючи спроможність нейронних мереж на достатніх обсягах даних, проте ціною тривалого тренування (91.51 сек). Gradient Boosting з MLP показав $R^2 = 0.409$, що нижче за чисті tree-based методи. TabNet продемонстрував найгірший результат $R^2 = 0.306$ (MAE = 2.07) попри екстремально довге тренування (1,536.84 сек), що вказує на невідповідність attention-based архітектури специфіці метеорологічних даних. Аналіз latency показав, що LightGBM (0.45 ms, 2,069 RPS) та Random Forest (0.54 ms, 2,307 RPS) забезпечують оптимальний баланс точності та швидкості для production систем.

Дослідження підтвердило перевагу LightGBM над традиційними та deep learning методами на табличних streaming даних, де він забезпечує найвищу

accuracy при прийнятних latency та training time. MLP показав конкурентні результати на великому датасеті, підтверджуючи важливість обсягу даних для нейронних мереж. TabNet виявився неефективним для метеорологічного прогнозування через надмірну складність архітектури та тривале тренування без відповідного приросту точності. Для production streaming систем на табличних даних оптимальним вибором є LightGBM як посилений III метод бустингу.

Бібліографія

1. Bifet A., Gavalda R., Holmes G., Pfahringer B. Machine Learning for Data Streams. *MIT Press*, 2018. <https://doi.org/10.7551/mitpress/10654.001.0001>
2. Breiman L. Random Forests. *Machine Learning*, 2001 Vol. 45, pp. 5–32. <https://doi.org/10.1023/A:1010933404324>
3. Arik S., Pfister T. TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. Vol. 35, No. 8, pp. 6679–6687. <https://doi.org/10.1609/aaai.v35i8.16826>
4. Bergmeir C., Benítez J. On the use of cross-validation for time series prediction. *Information Sciences*, 2012. Vol. 191, pp. 192–213. <https://doi.org/10.1016/j.ins.2011.12.028>
5. Ali Aalianvari, Shirin Jahanmiri. A Comparative Study of TabNet and Classical Machine Learning Models for Landslide Prediction, 17 October 2025, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-7694229/v1>]
6. Wang, Q., Lu, H., Li, F. *et al.* Advancing LightGBM with data augmentation for predicting the residual strength of corroded pipelines. *npj Mater Degrad* 9, 128 (2025). <https://doi.org/10.1038/s41529-025-00673-9>
7. Elshaarawy, M.K., Hamed, A.K. & Alsaadawi, M.M. Hybrid gradient boosting models for concrete compressive strength classification and prediction. *Int. J. Mach. Learn. & Cyber.* 16, 8857–8891 (2025). <https://doi.org/10.1007/s13042-025-02776-w>
8. Wei, S., Wu, Z., Zhang, Z., Zhou, J. (2023). Boosting Adaptive Graph Augmented MLPs via Customized Knowledge Distillation. In: Koutra, D., Plant, C., Gomez Rodriguez, M., Baralis, E., Bonchi, F. (eds) *Machine Learning and Knowledge Discovery in Databases: Research Track. ECML PKDD 2023. Lecture Notes in Computer Science()*, vol 14171. Springer, Cham. https://doi.org/10.1007/978-3-031-43418-1_6
9. Hancock, J.T., Khoshgoftaar, T.M. CatBoost for big data: an interdisciplinary review. *J Big Data* 7, 94 (2020). <https://doi.org/10.1186/s40537-020-00369-8>