

Бажан Тетяна Олександрівна

Старший викладач кафедри Технологій цифрового розвитку

Державний університет інформаційно-комунікаційних технологій, м. Київ

ORCID: 0009-0007-6594-1695

t.bazhan@duikt.edu.ua

Жебка Вікторія Вікторівна

д.т.н. професор, завідувач кафедри Технологій цифрового розвитку

Державний університет інформаційно-комунікаційних технологій, Київ, Україна

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

ORCID: 0000-0003-4051-1190

v.zhebka@duikt.edu.ua

**ПРОГНОЗУВАННЯ ІНВЕСТИЦІЙ ЗА ДОПОМОГОЮ ГІБРИДНОГО АНСАМБЛЮ
LSTM-CATBOOST-RANDOM FOREST З ПОЯСНЮВАНІСТЮ SHAP**

Анотація. У сучасних умовах високої волатильності фінансових ринків завдання прогнозування інвестицій набуває особливого значення. Точність прогнозів безпосередньо впливає на ефективність стратегій управління активами та економічну стабільність у цілому. Традиційні економетричні моделі (ARIMA, VAR, GARCH тощо) мають обмеження у відображенні нелінійних та високорозмірних залежностей, що властиві фінансовим часовим рядам. Сучасні методи машинного навчання, такі як LSTM-мережі, градієнтний бустинг та ансамблеві алгоритми, продемонстрували здатність значно підвищувати точність прогнозів, проте залишається проблема їх «чорного ящика», тобто низької інтерпретованості результатів. У цій роботі запропоновано новий гібридний метод прогнозування інвестицій, який поєднує моделі LSTM, CatBoost і Random Forest у зважений ансамбль із оптимізацією вагових коефіцієнтів та застосуванням підходу SHAP (SHapley Additive exPlanations) для забезпечення пояснюваності результатів. Методологія передбачає комбінування моделей, що враховують як часову структуру даних, так і табличні макроекономічні та ринкові індикатори, з подальшою інтеграцією результатів через вагову схему. SHAP-аналіз забезпечує можливість кількісної оцінки внеску окремих змінних у фінальний прогноз, що робить модель не лише точною, а й придатною для практичного використання у фінансовій аналітиці, де критично важливо розуміти причини прийняття рішень. Експериментальна перевірка здійснена на макроекономічних та фондових даних, при цьому ансамбль продемонстрував вищу точність порівняно з окремими моделями. Отримані результати підтверджують доцільність використання гібридних підходів у прогнозуванні інвестицій і відкривають перспективи подальших досліджень у напрямі Explainable AI для фінансових систем.

Ключові слова: прогнозування інвестицій; машинне навчання; LSTM; CatBoost; Random Forest; ансамбль моделей; SHAP; пояснюваний штучний інтелект; фінансова аналітика, інформаційні технології, інформаційна система.

Tetiana Bazhan

Senior Lecturer, Department of Digital Development Technologies

State University of Information and Communication Technologies, Kyiv

ORCID: 0009-0007-6594-1695

t.bazhan@duikt.edu.ua

Viktoriia Zhebka

Doctor of Technical Sciences, Professor, Chair of the Department of Digital Development Technologies

State University of Information and Communication Technologies, Kyiv, Ukraine

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

ORCID: 0000-0003-4051-1190

v.zhebka@duikt.edu.ua

**INVESTMENT FORECASTING USING A HYBRID LSTM-CATBOOST-RANDOM FOREST
ENSEMBLE WITH SHAP EXPLAINABILITY**

Abstract. In the modern context of high financial market volatility, investment forecasting has become a crucial challenge. Forecast accuracy directly affects asset management strategies and overall economic stability. Traditional econometric models (ARIMA, VAR, GARCH, etc.) have limitations in capturing nonlinear and high-dimensional dependencies inherent to financial time series. Contemporary machine learning methods such as LSTM networks, gradient boosting, and ensemble algorithms have demonstrated significant improvements in forecast accuracy. However, the "black box" problem remains, making it difficult to interpret their outputs. This paper proposes a hybrid investment forecasting method that combines LSTM, CatBoost, and Random Forest within an optimized weighted ensemble and integrates SHAP (SHapley Additive exPlanations) to provide model interpretability. The methodology leverages models capable of capturing both temporal patterns and structured macroeconomic and market indicators, with results integrated via a weighted scheme. SHAP analysis enables quantitative assessment of each feature's contribution to the final prediction, ensuring that the model is not only accurate but also transparent and applicable in financial analytics, where decision explainability is critical. Experimental validation on macroeconomic and stock market data demonstrates

that the ensemble provides higher forecasting accuracy compared with the individual models. These results confirm the effectiveness of hybrid approaches in investment forecasting and highlight the potential of Explainable AI applications in financial systems.

Keywords: investment forecasting; machine learning; LSTM; CatBoost; Random Forest; model ensemble; SHAP; explainable artificial intelligence; financial analytics; information technology; information system..

Вступ

Фінансові ринки характеризуються складною динамікою та високим рівнем невизначеності, що робить завдання прогнозування інвестицій одним із найважливіших як для приватних інвесторів, так і для державних установ. Від якості прогнозів залежить ефективність прийняття рішень у сфері управління портфелями, оцінки ризиків та розробки макроекономічної політики. Традиційні статистичні та економетричні методи, такі як ARIMA чи VAR, дозволяють моделювати базові залежності, проте часто виявляються недостатніми для адекватного відображення складних нелінійних закономірностей та взаємозв'язків у великих даних. Машинне навчання відкриває нові можливості для прогнозування, оскільки сучасні алгоритми здатні виявляти приховані патерни та адаптуватися до мінливих умов ринку. Зокрема, рекурентні нейронні мережі LSTM добре працюють із часовими рядами, Random Forest ефективно моделює нелінійні залежності, а CatBoost оптимізований для роботи з табличними та категоріальними змінними. Поєднання цих моделей в ансамбль дозволяє підвищити точність і стабільність прогнозів завдяки використанню сильних сторін кожного підходу. Водночас застосування Explainable AI, зокрема методу SHAP, забезпечує прозорість прогнозів і дає можливість визначити ключові фактори, які впливають на інвестиційні рішення. Таким чином, розробка гібридного підходу, що поєднує ансамблеві методи машинного навчання з пояснюваністю, є актуальним завданням як з наукової, так і з практичної точки зору.

1.Актуальність дослідження

У сучасних умовах цифрової економіки ефективне прогнозування інвестицій набуває особливої ваги для формування державної політики, стратегій розвитку підприємств та прийняття управлінських рішень у фінансовій сфері. Зростаюча волатильність глобальних ринків, вплив макроекономічних показників, технологічних змін та геополітичних факторів зумовлюють потребу у застосуванні більш гнучких та адаптивних методів аналізу. Традиційні статистичні моделі часто виявляються недостатніми для відображення нелінійних закономірностей та складних взаємодій між економічними показниками.

Методи машинного навчання, зокрема LSTM-мережі, демонструють високу ефективність у виявленні довготривалих часових залежностей. Разом із цим, Random Forest та CatBoost забезпечують стійкість до шумів, здатність працювати з гетерогенними та категоріальними даними, а також добре узагальнюють результати. Поєднання цих методів у гібридному ансамблі дозволяє комплексно враховувати як часові, так і структурні характеристики даних, підвищуючи точність прогнозування інвестиційних показників.

Не менш важливою є проблема пояснюваності моделей штучного інтелекту. Використання методології SHAP дає змогу не лише отримати прогноз, а й зрозуміти внесок кожної ознаки у прийняте рішення. Це критично важливо для сфери інвестицій, де прозорість та обґрунтованість прогнозів впливають на довіру користувачів, інвесторів та регуляторів.

Таким чином, розробка та дослідження гібридного ансамблю LSTM-CatBoost-Random Forest з інтегрованою SHAP-пояснюваністю є актуальним напрямом наукових пошуків, що поєднує переваги різних класів моделей машинного навчання та відповідає вимогам практики щодо точності, стійкості й інтерпретованості прогнозів в умовах цифрової трансформації економіки.

2.Огляд літератури

Прогнозування інвестиційних та фінансово-економічних показників належить до складних задач аналізу часових рядів, для яких характерні нелінійність, нестационарність, структурні зміни та залежність від значної кількості взаємопов'язаних факторів. У зв'язку з цим поряд із традиційними статистичними методами дедалі ширше застосовуються методи машинного та глибокого навчання, здатні моделювати складні залежності у багатовимірних даних. Сучасні огляди методів прогнозування часових рядів підтверджують активний розвиток нейромережевих, ансамблевих та гібридних підходів, а також тенденцію до комбінування моделей різної природи з метою підвищення стійкості прогнозування [8], [12].

Одним із базових підходів до моделювання послідовних даних є рекурентні нейронні мережі типу Long Short-Term Memory (LSTM), запропоновані S. Hochreiter та J. Schmidhuber [1]. Використання спеціальних механізмів керування станом пам'яті дозволяє LSTM моделювати довготривалі часові залежності та частково вирішувати проблему зникнення градієнта, характерну для класичних рекурентних нейронних мереж. Завдяки цим властивостям LSTM широко застосовуються для прогнозування економічних і фінансових часових рядів.

Водночас використання складніших нейромережевих моделей не гарантує безумовного підвищення точності для будь-якого набору даних. Зокрема, T. Rygh та співавтори [9] під час дослідження прогнозування інфляції порівняли LSTM із традиційними моделями та показали, що ефективність нейронної мережі істотно залежить від структури даних, набору вхідних змінних і умов прогнозування. В окремих експериментальних сценаріях традиційні методи забезпечували конкурентні або кращі результати. Це підтверджує доцільність не орієнтуватися на одну модель, а досліджувати можливості комбінування алгоритмів, які використовують різні принципи формування прогнозу.

Для аналізу структурованих даних значного поширення набули ансамблеві методи на основі дерев рішень. Random Forest, запропонований L. Breiman [2], формує множину дерев на різних підмножинах спостережень та ознак із подальшим агрегуванням отриманих результатів. Такий механізм дає змогу

моделювати нелінійні залежності, знижувати дисперсію прогнозу та підвищувати стійкість моделі до шуму у вхідних даних.

Іншим ефективним представником деревних ансамблів є CatBoost, запропонований L. Prokhorenkova та співавторами [4]. Алгоритм реалізує градієнтний бустинг дерев рішень та використовує механізм ordered boosting, спрямований на зменшення зміщення під час навчання. Особливістю CatBoost також є ефективна робота з категоріальними ознаками без необхідності їх попереднього перетворення за допомогою традиційних методів кодування. Завдяки цим властивостям алгоритм використовується для задач прогнозування на основі структурованих багатовимірних даних.

Перспективність використання CatBoost та інших ансамблевих методів підтверджується сучасними прикладними дослідженнями. А. М. Abed [7] досліджує гібридизацію CatBoost з оптимізаційним алгоритмом у задачі прогнозування попиту, демонструючи можливість підвищення ефективності прогнозування за рахунок поєднання машинного навчання та оптимізаційних процедур. E. Supriyadi, N. Danila та Z. Ali [10] розглядають ансамблеві моделі у задачах прогнозування фінансових продуктів та використовують комплекс показників RMSE, MAE, MAPE, MSE і $\sqrt{R^2}$ для оцінювання їх точності. Отримані результати свідчать про доцільність застосування ансамблевих алгоритмів для аналізу складних фінансових залежностей [10].

Таким чином, LSTM, Random Forest і CatBoost характеризуються різними механізмами формування прогнозу. LSTM орієнтована на моделювання часових залежностей у послідовностях [1], Random Forest забезпечує стійке агрегування множини дерев рішень [2], а CatBoost реалізує градієнтний бустинг та ефективно моделює нелінійні залежності у структурованих даних [4]. Поєднання моделей різної природи в межах єдиного ансамблю створює передумови для комплексного врахування часової та факторної структури інвестиційних даних.

Окремою проблемою використання складних моделей машинного навчання у фінансовому прогнозуванні є недостатня інтерпретованість отриманих результатів. Висока прогнозна точність не дає можливості безпосередньо встановити, які саме фактори та якою мірою вплинули на конкретне прогнозоване значення. Для фінансових та інвестиційних систем це має особливе значення, оскільки практичне використання результатів моделювання потребує не лише формування прогнозу, а й пояснення причин його отримання.

Одним із найбільш поширених підходів Explainable Artificial Intelligence є SHAP (SHapley Additive exPlanations), запропонований S. M. Lundberg та S.-I. Lee [3]. Метод базується на значеннях Шеплі та дозволяє кількісно оцінювати внесок кожної ознаки у формування окремого результату моделі. Адитивна структура SHAP забезпечує можливість отримання як локальних пояснень для конкретних прогнозів, так і глобального оцінювання важливості факторів на множині спостережень [3].

Подальший розвиток SHAP пов'язаний з адаптацією методу до різних класів моделей та специфічних структур даних. A. Zern, K. Broelemann та G. Kasneci [6] досліджують interventional SHAP values та interaction values для дерев регресії, розширюючи можливості аналізу взаємодії факторів у моделях на основі дерев. Для часових рядів A. Nayeib та співавтори запропонували WindowSHAP – підхід, у якому враховується часова структура вхідних даних і забезпечується ефективніше пояснення прогнозів моделей, що працюють із послідовностями [5].

Особливо актуальним є використання XAI у фінансових системах. P.-D. Arsenault, S. Wang та J.-M. Patenaude [11] у дослідженні методів пояснюваного штучного інтелекту для прогнозування фінансових часових рядів систематизують сучасні підходи до інтерпретації складних моделей і підкреслюють необхідність поєднання прогновної ефективності з прозорістю результатів. Це свідчить про поступовий перехід від задачі виключно підвищення точності прогнозування до побудови моделей, результати яких можуть бути інтерпретовані та обґрунтовані.

Проведений аналіз наукових праць показує, що наявні дослідження окремо підтверджують перспективність LSTM для моделювання часових залежностей [1], [8], [12], Random Forest і CatBoost – для аналізу нелінійних залежностей у структурованих даних [2], [4], [7], [10], а SHAP та його модифікацій – для підвищення пояснюваності результатів складних моделей [3], [5], [6], [11]. Водночас питання спільного використання цих підходів у межах єдиного методу прогнозування інвестицій потребує подальшого дослідження.

Отже, актуальним є розроблення підходу, який поєднує здатність LSTM моделювати часову структуру даних, можливості CatBoost і Random Forest щодо виявлення нелінійних залежностей між факторними ознаками та механізм SHAP для пояснення сформованого прогнозу. У цій роботі для вирішення зазначеного завдання пропонується зважений гібридний ансамбль LSTM-CatBoost-Random Forest, у якому внесок базових моделей визначається на основі їх прогновної точності з подальшою оптимізацією вагових коефіцієнтів, а інтерпретація результуючого прогнозу здійснюється за допомогою SHAP.

3. Мета і завдання дослідження.

Метою дослідження є розробка гібридного ансамблю моделей LSTM-CatBoost-Random Forest із застосуванням методології SHAP для підвищення точності та пояснюваності прогнозування інвестиційних показників.

Завдання дослідження:

- 1) Розробити архітектуру гібридного ансамблю та інтегрувати в нього метод пояснюваності SHAP.
- 2) Провести експериментальне дослідження на основі макроекономічних та фінансових даних.
- 3) Оцінити точність та інтерпретованість запропонованої моделі порівняно з окремими методами.

4. Результати дослідження.

Запропонована архітектура гібридного ансамблю LSTM-CatBoost-Random Forest на рис. 1 побудована за принципом комбінування моделей різної природи з подальшою інтеграцією їх результатів у єдиний прогноз. Загальна структура моделі включає три основні рівні: рівень підготовки даних, рівень базових моделей та рівень ансамблювання.

Першим етапом реалізації запропонованого методу є формування та попередня обробка вхідного набору даних. Як результативний показник розглядається обсяг інвестицій, а до множини факторних ознак включаються макроекономічні, фінансові та цифрові показники, які характеризують стан економічного середовища та можуть впливати на динаміку інвестиційної активності. До таких показників належать ВВП, рівень інфляції, облікова ставка, валютний курс, фондові індекси та показники цифрового розвитку.

Для математичної формалізації вхідних даних вектор ознак у момент часу t представимо у вигляді:

$$x_t = [I_{t-1}, I_{t-2}, \dots, I_{t-p}, GDP_t, INF_t, IR_t, ER_t, SI_t, DI_t], \quad (1)$$

де $I_{t-1}, I_{t-2}, \dots, I_{t-p}$ – попередні значення інвестиційного показника з глибиною лагу p ; GDP_t – значення валового внутрішнього продукту; INF_t – рівень інфляції; IR_t – облікова ставка; ER_t – валютний курс; SI_t – показник стану фондового ринку; DI_t – показник цифрового розвитку у момент часу t .

Таке представлення вхідного вектора дає змогу одночасно врахувати власну часову динаміку інвестиційного показника та вплив зовнішніх економічних і цифрових факторів. Включення лагових значень дозволяє відобразити залежність майбутнього рівня інвестицій від їх попередньої динаміки, тоді як макроекономічні, фінансові та цифрові характеристики забезпечують урахування зовнішнього середовища, у якому відбуваються інвестиційні процеси.

У загальному вигляді задачу прогнозування інвестицій сформульовано як визначення функціональної залежності:

$$\hat{I}_{t+h} = F(x_{t-p+1}, x_{t-p+2}, \dots, x_t; \Theta), \quad (2)$$

де $\hat{I}_{t+h}$ – прогнозоване значення інвестиційного показника на горизонті h ; h – горизонт прогнозування; $F(\cdot)$ – функція прогнозування, що реалізується запропонованою гібридною моделлю; Θ – сукупність параметрів моделі.

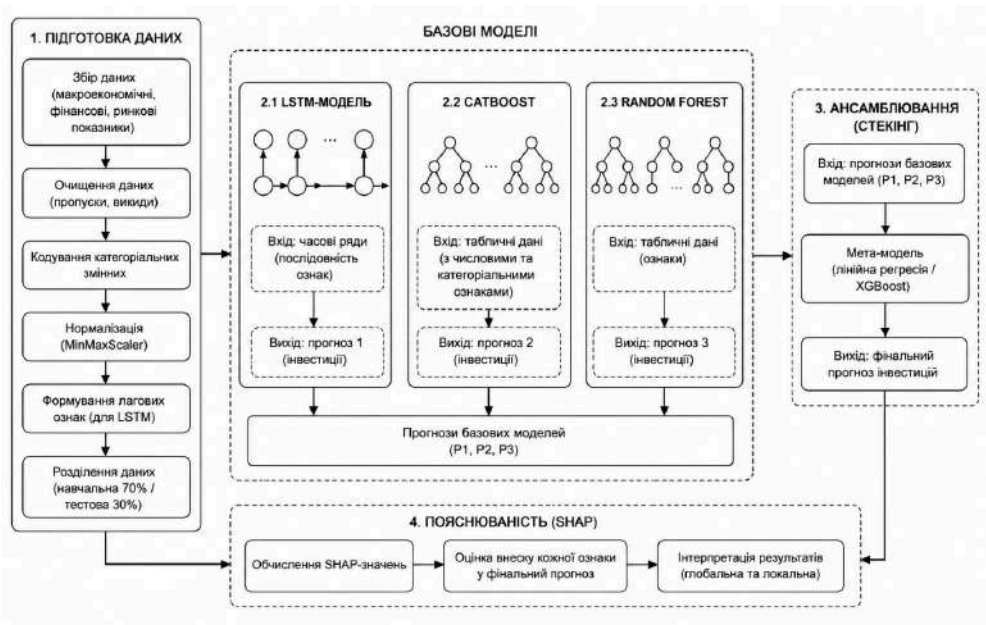


Рис. 1 Архітектура гібридного ансамблю LSTM-CatBoost-Random Forest

Перед передаванням даних до базових моделей виконується їх попередня обробка. Вона передбачає перевірку структури набору даних, обробку пропущених значень, виявлення аномальних спостережень, нормалізацію числових показників, формування лагових ознак і часових послідовностей, а також поділ даних на навчальну та тестову вибірки зі збереженням хронологічної послідовності спостережень. Збереження часової структури даних є принципово важливим під час прогнозування часових рядів, оскільки використання інформації майбутніх періодів на етапі навчання може призвести до витоку даних і завищення оцінок точності моделі [8], [12].

Для приведення числових показників, що характеризуються різними одиницями вимірювання та масштабами, до єдиного діапазону використовується нормалізація:

$$x_{t,j}^{norm} = \frac{x_{t,j} - x_j^{min}}{x_j^{max} - x_j^{min}}, \quad (3)$$

де $x_{t,j}$ – початкове значення j -ї ознаки у момент часу t ; x_j^{min} та x_j^{max} – відповідно мінімальне та максимальне значення цієї ознаки, визначені на навчальній вибірці.

Визначення параметрів нормалізації лише за навчальною вибіркою дає змогу уникнути використання інформації тестового періоду під час підготовки моделі та забезпечує коректність подальшого експериментального оцінювання.

Однією з основних складових запропонованого гібридного методу є модель LSTM, запропонована S. Hochreiter та J. Schmidhuber [1], яка використовується для моделювання довготривалих залежностей у послідовних даних. На відміну від стандартних рекурентних нейронних мереж, архітектура LSTM використовує механізм керування інформацією за допомогою спеціальних вентилів, що дає змогу зберігати значущу інформацію протягом тривалих часових інтервалів і зменшувати проблему зникнення градієнта [1].

На кожному часовому кроці t вентиль забування визначає, яку частину інформації попереднього стану необхідно зберегти:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \quad (4)$$

де f_t – вектор значень вентиля забування; $\sigma(\cdot)$ – сигмоїдна функція активації; W_f – матриця ваг; h_{t-1} – прихований стан LSTM на попередньому часовому кроці; x_t – вектор вхідних ознак; b_f – вектор зміщення.

Вхідний вентиль визначає частину нової інформації, яка буде записана до стану комірки:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i). \quad (5)$$

Одночасно формується вектор кандидатних значень нового стану комірки:

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c). \quad (6)$$

На основі отриманих значень здійснюється оновлення внутрішнього стану LSTM [1]:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \quad (7)$$

де C_t та C_{t-1} – відповідно поточний і попередній стани комірки; $\odot$ – операція поелементного множення.

Вихідний вентиль визначається як:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \quad (8)$$

після чого формується прихований стан:

$$h_t = o_t \odot \tanh(C_t). \quad (9)$$

Сукупність співвідношень (4)-(9) описує механізм функціонування LSTM-комірки, який забезпечує контрольоване збереження, оновлення та передавання інформації між часовими кроками [1]. У межах запропонованої моделі на основі отриманого прихованого стану формується прогноз інвестиційного показника:

$$I_{t+h}^{LSTM} = W_y h_t + b_y, \quad (10)$$

де I_{t+h}^{LSTM} – прогноз, отриманий моделлю LSTM на горизонті h ; W_y – матриця ваг вихідного шару; b_y – параметр зміщення.

Таким чином, LSTM-компонента запропонованого ансамблю забезпечує формування прогнозу з урахуванням послідовної структури даних та залежності поточного стану інвестиційного процесу від попередніх часових періодів. Доцільність використання моделей глибокого навчання для прогнозування часових рядів підтверджується сучасними дослідженнями у цій галузі [8], [12], а застосування LSTM до макроекономічного прогнозування демонструє перспективність такого підходу порівняно з традиційними моделями [9].

Водночас використання лише LSTM не забезпечує повного врахування складних нелінійних взаємозв'язків між структурованими макроекономічними, фінансовими та цифровими показниками. Тому другим і третім компонентами запропонованого ансамблю є моделі Random Forest та CatBoost.

Random Forest, запропонований L. Breiman [2], являє собою ансамблевий метод, заснований на побудові множини дерев рішень із використанням випадкових підмножин даних та ознак. Для задачі регресії результуючий прогноз Random Forest визначається шляхом агрегування прогнозів окремих дерев [2]:

$$\hat{I}_{t+h}^{RF} = \frac{1}{B} \sum_{b=1}^B T_b(x_t), \quad (11)$$

де B – кількість дерев у моделі Random Forest; $T_b(x_t)$ – прогноз, сформований b -м деревом для вектора ознак x_t ; $\hat{I}_{t+h}^{RF}$ – результуючий прогноз Random Forest.

Усереднення результатів окремих дерев сприяє зменшенню дисперсії прогнозу та підвищенню стійкості моделі до випадкових коливань вхідних даних [2]. У запропонованому методі Random Forest використовується для виявлення складних нелінійних взаємозв'язків між інвестиційним показником та сукупністю макроекономічних, фінансових і цифрових факторів.

На відміну від Random Forest, CatBoost належить до методів градієнтного бустингу дерев рішень. Особливістю алгоритму є використання механізмів ordered boosting та спеціальних підходів до роботи з категоріальними ознаками, спрямованих на зменшення зміщення під час навчання [4]. Загальну прогнозу функцію бустингової моделі після M ітерацій у межах запропонованої формалізації можна представити як:

$$\hat{I}_{t+h}^{CB} = F_M(x_t) = F_0(x_t) + \sum_{m=1}^M \eta \gamma_m T_m(x_t), \quad (12)$$

де $F_0(x_t)$ – початкове наближення прогнозу; M – кількість ітерацій бустингу; $T_m(x_t)$ – дерево рішень, сформоване на m -й ітерації; γ_m – коефіцієнт внеску відповідного дерева; η – швидкість навчання.

Параметри моделі визначаються шляхом мінімізації функції втрат:

$$L_{CB} = \frac{1}{N} \sum_{t=1}^N L(I_t, F_M(x_t)), \quad (13)$$

де N – кількість спостережень у навчальній вибірці; I_t – фактичне значення інвестиційного показника; $F_M(x_t)$ – прогнозоване значення; $L(\cdot)$ – функція втрат. Основні принципи побудови та навчання CatBoost детально представлені в роботі Prokhorenkova та співавторів [4].

Використання CatBoost у запропонованій архітектурі дозволяє доповнити часову складову LSTM моделлю, орієнтовану на виявлення нелінійних залежностей у структурованих даних. Random Forest, своєю чергою, забезпечує додаткову стійкість ансамблю завдяки формуванню множини дерев та агрегуванню отриманих прогнозів [2]. Поєднання моделей різної природи в межах ансамблевого підходу дає змогу одночасно використовувати здатність нейромережових моделей до виявлення часових закономірностей та переваги деревних алгоритмів під час моделювання нелінійних залежностей у структурованих даних [8], [10], [12].

Таким чином, після завершення навчання базових моделей для кожного моменту часу формуються три незалежні прогнозні оцінки:

$$\hat{I}_{t+h} = \begin{bmatrix} \hat{I}_{t+h}^{RF} \\ \hat{I}_{t+h}^{CB} \\ \hat{I}_{t+h}^{LSTM} \end{bmatrix}. \quad (14)$$

Отриманий вектор прогнозів є вхідною інформацією для наступного рівня запропонованого методу – ансамблювання. На цьому рівні визначається внесок кожної базової моделі та формується результуючий прогноз інвестиційного показника.

Для формування результуючого прогнозу отримані значення базових моделей об'єднуються за допомогою зваженого ансамблю. На відміну від простого усереднення, за якого внесок усіх моделей вважається однаковим, у запропонованому підході враховується різна прогнозна здатність LSTM, CatBoost та Random Forest.

Результуючий прогноз інвестиційного показника визначається як

$$\hat{I}_{t+h}^{ens} = w_L \hat{I}_{t+h}^{LSTM} + w_C \hat{I}_{t+h}^{CB} + w_R \hat{I}_{t+h}^{RF}, \quad (15)$$

де $\hat{I}_{t+h}^{ens}$ – результуючий прогноз гібридного ансамблю; $\hat{I}_{t+h}^{LSTM}$, $\hat{I}_{t+h}^{CB}$, $\hat{I}_{t+h}^{RF}$ – прогнози відповідно LSTM, CatBoost та Random Forest; w_L , w_C , w_R – вагові коефіцієнти базових моделей.

Для забезпечення інтерпретованості ваг та стабільності результуючого прогнозу вводяться обмеження

$$w_L + w_C + w_R = 1, \quad (16)$$

$$w_L \geq 0, w_C \geq 0, w_R \geq 0. \quad (17)$$

Умови (16)-(17) забезпечують формування результуючого прогнозу як опуклої комбінації прогнозів базових моделей. При цьому більша вага відповідає моделі, внесок якої у формування кінцевого прогнозу є більшим.

Принциповим елементом запропонованого методу є визначення вагових коефіцієнтів не експертним шляхом, а на основі прогнозової похибки базових моделей. Для початкового визначення ваг використовується обернена залежність від RMSE відповідної моделі:

$$w_j^{(0)} = \frac{RMSE_j^{-1}}{\sum_{k=1}^K RMSE_k^{-1}}, j = 1, \dots, K, \quad (18)$$

де $K = 3$ – кількість базових моделей; $RMSE_j$ – значення кореня середньоквадратичної похибки j -ї моделі; $w_j^{(0)}$ – її початкова вага.

Такий підхід забезпечує автоматичне збільшення внеску моделей із меншою похибкою прогнозування та зменшення внеску моделей, які демонструють нижчу точність. При цьому сума отриманих ваг дорівнює одиниці.

Разом із тим використання лише обернено пропорційного зважування не гарантує мінімальної похибки саме результуючого ансамблю. Тому в запропонованому методі вагові коефіцієнти уточнюються шляхом розв'язання оптимізаційної задачі:

$$w^* = \underset{w}{\operatorname{argmin}} \frac{1}{N_v} \sum_{t=1}^{N_v} \left[\left(w_R^{\wedge RF} I_t \right) \right]^2, \quad (19)$$

за обмежень

$$w_L + w_C + w_R = 1, w_L, w_C, w_R \geq 0, \quad (20)$$

де N_v – кількість спостережень валідаційної вибірки; I_t – фактичне значення інвестиційного показника; $w^* = (w_L^*, w_C^*, w_R^*)$ – вектор оптимальних ваг.

Таким чином, оптимальні вагові коефіцієнти визначаються за критерієм мінімізації середньоквадратичного відхилення ансамблевого прогнозу від фактичних значень на валідаційних даних. Важливо, що тестова вибірка при визначенні ваг не використовується. Вона залишається незалежною та застосовується виключно для кінцевого оцінювання прогнозової здатності сформованої моделі.

Такий спосіб формування ансамблю дозволяє враховувати не лише різну природу базових алгоритмів, але й їх фактичну ефективність на досліджуваному наборі даних. Якщо одна з моделей демонструє більшу прогнозову похибку, оптимізаційний алгоритм може зменшити її вагу. Водночас модель, яка забезпечує точніше відтворення динаміки інвестиційного показника, отримує більший внесок у результуючий прогноз.

Отже, запропонований підхід відрізняється від простого арифметичного усереднення прогнозів. Його особливістю є формування оптимізованої зваженої композиції моделей різної природи, що забезпечує одночасне врахування часових закономірностей, нелінійних взаємозв'язків між ознаками та стійкості деревних ансамблів.

Для оцінювання ефективності запропонованого методу результати ансамблю порівнюються з прогнозами окремих моделей LSTM, CatBoost та Random Forest. Якість прогнозування визначається за сукупністю метрик, що дозволяють оцінити абсолютну та квадратичну похибки, а також ступінь відповідності прогнозованої динаміки фактичним значенням.

Середньоквадратична похибка визначається як

$$MSE = \frac{1}{N} \sum_{t=1}^N \left(\hat{I}_t \right)^2, \quad (21)$$

де N – кількість спостережень тестової вибірки; I_t – фактичне значення інвестиційного показника; $\hat{I}_t$ – прогнозоване значення.

Корінь середньоквадратичної похибки визначається за формулою

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N \left(\hat{I}_t \right)^2}. \quad (22)$$

Метрика RMSE є особливо чутливою до значних відхилень прогнозу від фактичних значень, оскільки помилки підносяться до квадрату. Менше значення RMSE відповідає вищій точності прогнозування.

Для додаткової оцінки використовується середня абсолютна похибка:

$$MAE = \frac{1}{N} \sum_{t=1}^N |I_t - \hat{I}_t|. \quad (23)$$

MAE характеризує середню абсолютну величину відхилення прогнозованих значень від фактичних та, на відміну від RMSE, меншою мірою залежить від окремих великих помилок.

Для оцінювання відносної величини прогнозної похибки використовується середня абсолютна відсоткова похибка:

$$MAPE = \frac{100\%}{N} \sum_{t=1}^N \left| \frac{I_t - \hat{I}_t}{I_t} \right|. \quad (24)$$

Використання MAPE дає можливість представити похибку у відсотковій формі, що спрощує порівняння точності прогнозування показників різного масштабу. Водночас застосування цієї метрики потребує обережності за наявності нульових або близьких до нуля фактичних значень I_t .

Коефіцієнт детермінації визначається як

$$R^2 = 1 - \frac{\sum_{t=1}^N (I_t - \hat{I}_t)^2}{\sum_{t=1}^N (I_t - \bar{I})^2}, \quad (25)$$

де $\bar{I}$ – середнє фактичне значення інвестиційного показника на відповідній вибірці.

На відміну від показників похибки, для яких менші значення свідчать про кращу якість прогнозу, більше значення R^2 означає кращу відповідність прогнозованих значень фактичним даним.

Комплексне використання RMSE, MAE, MAPE та R^2 дозволяє уникнути оцінювання моделі лише за одним критерієм. Зокрема, RMSE сильніше реагує на значні помилки, MAE характеризує середню абсолютну похибку, MAPE відображає її відносну величину, а R^2 характеризує відповідність прогнозованої та фактичної варіативності даних.

Для кількісного визначення переваги гібридного ансамблю над окремою базовою моделлю доцільно також використовувати відносне зменшення RMSE:

$$\Delta RMSE_j = \frac{RMSE_j - RMSE_{ens}}{RMSE_j} \cdot 100\%, \quad (26)$$

де $RMSE_j$ – значення RMSE для відповідної базової моделі; $RMSE_{ens}$ – значення RMSE запропонованого ансамблю.

Аналогічно відносне зменшення MAE визначається як

$$\Delta MAE_j = \frac{MAE_j - MAE_{ens}}{MAE_j} \cdot 100\%. \quad (27)$$

Введення показників (26)-(27) дозволяє кількісно оцінити, наскільки використання запропонованого ансамблю покращує прогнозування порівняно з кожною окремою моделлю.

Однак висока точність прогнозування сама по собі не забезпечує достатньої прозорості моделі. Це особливо важливо для гібридного ансамблю, у якому кінцевий прогноз є результатом взаємодії декількох складних алгоритмів. Для вирішення цієї проблеми до запропонованого методу інтегровано SHAP (SHapley Additive exPlanations), запропонований Lundberg та Lee [3]. Метод базується на значеннях Шеплі та дозволяє кількісно визначити внесок окремих ознак у конкретний прогноз моделі [3].

Для множини ознак FSHAP-значення i -ї ознаки визначається як

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)], \quad (28)$$

де ϕ_i – SHAP-значення i -ї ознаки; F – повна множина факторних ознак; S – підмножина ознак, що не містить i -ту ознаку; $f(S)$ – очікуваний результат моделі для відповідної множини ознак. Така конструкція оцінює граничний внесок ознаки для можливих коаліцій інших ознак із відповідним ваговим усередненням.

Для окремого прогнозу адитивне SHAP-пояснення можна представити як

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i, \quad (29)$$

де ϕ_0 – базове значення прогнозу, а ϕ_i – внесок i -ї ознаки у відхилення конкретного прогнозу від базового значення. SHAP належить до класу адитивних методів пояснення та призначає ознакам значення важливості для конкретного прогнозу [3].

У контексті запропонованого методу SHAP-аналіз використовується не лише для визначення загальної важливості макроекономічних, фінансових та цифрових показників, але й для пояснення окремих прогнозованих значень інвестицій. Це дозволяє встановити, які фактори сприяли підвищенню прогнозованого обсягу інвестицій, а які, навпаки, зумовили його зменшення.

Оскільки результуюча функція запропонованого ансамблю має вигляд

$$F_{ens}(x) = w_L f_L(x) + w_C f_C(x) + w_R f_R(x), \quad (30)$$

де $f_L(x)$, $f_C(x)$ та $f_R(x)$ – прогнознi функції LSTM, CatBoost та Random Forest відповідно, для лінійно зваженого ансамблю внесок i -ї ознаки можна агрегувати відповідно до ваг базових моделей:

$$\phi_i^{ens} = w_L \phi_i^{LSTM} + w_C \phi_i^{CB} + w_R \phi_i^{RF}. \quad (31)$$

Формула (31) є особливо важливою для математичного опису запропонованого підходу, оскільки забезпечує зв'язок між механізмом ансамблювання та механізмом пояснюваності. За умови використання узгодженого визначення SHAP для компонентів лінійної комбінації вона дозволяє інтерпретувати кінцевий прогноз не лише через ваги базових моделей, а й через внески конкретних факторних ознак. Адитивна природа SHAP є фундаментальною властивістю цього підходу [3].

Для визначення глобальної значущості i -ї ознаки на множині спостережень використовується середнє абсолютне SHAP-значення:

$$Imp_i = \frac{1}{N} \sum_{t=1}^N |\phi_{i,t}^{ens}|, \quad (32)$$

де Imp_i – глобальна важливість i -ї ознаки; $\phi_{i,t}^{ens}$ – її SHAP-внесок для t -го спостереження.

Чим більше значення Imp_i , тим більшим є середній вплив відповідної ознаки на результати прогнозування ансамблю. На основі цього показника можна сформулювати впорядкований перелік факторів, що найбільше впливають на прогноз інвестицій.

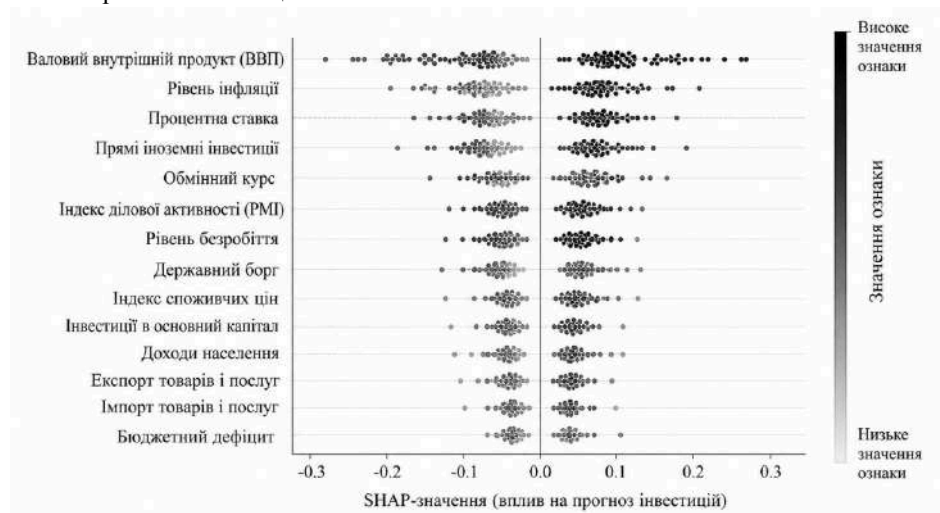


Рис. 2 SHAP summary plot впливу факторів на прогноз інвестицій

Аналіз отриманих SHAP-значень показує, що найбільший вплив на результати прогнозування мають ВВП, рівень інфляції, процентна ставка, прямі іноземні інвестиції та обмінний курс. Значущими також є індекс ділової активності (PMI), рівень безробіття, державний борг та інші макроекономічні показники.

Графік порівняння фактичних та прогнозованих значень на рис. 3 демонструє високу точність моделі на тестовій вибірці та здатність ансамблю адекватно відтворювати реальну динаміку інвестиційних показників.

Висновки

У результаті проведеного дослідження розроблено гібридний ансамбль моделей LSTM-CatBoost-Random Forest для прогнозування інвестиційних показників в умовах цифрового розвитку економіки. Запропонований підхід поєднує переваги глибоких нейронних мереж, здатних ефективно працювати з часовими рядами, та ансамблевих методів машинного навчання, орієнтованих на обробку табличних і нелінійних даних.

Запропоновано механізм формування результуючого прогнозу як оптимізованої зваженої комбінації прогнозів LSTM, CatBoost та Random Forest, у якій вагові коефіцієнти визначаються на валідаційній вибірці за критерієм мінімізації середньоквадратичної похибки. Це свідчить про доцільність інтеграції різних типів моделей для врахування як часової динаміки, так і структурних характеристик економічних даних.

Важливим науковим результатом є інтеграція методології SHAP у процес прогнозування, що забезпечує інтерпретованість отриманих результатів. Проведений SHAP-аналіз дозволив визначити ключові фактори, які впливають на інвестиційні показники, зокрема макроекономічні індикатори, фінансові змінні та показники цифрового розвитку. Це підвищує практичну цінність моделі, оскільки дає змогу не лише отримувати прогноз, а й обґрунтовувати його.

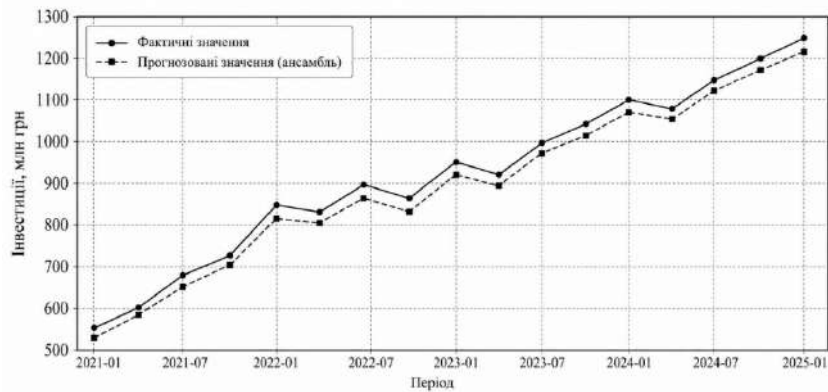


Рис. 3 Порівняння фактичних та прогнозованих значень інвестицій

Запропонований підхід може бути використаний у фінансовій аналітиці, системах підтримки прийняття рішень, державному управлінні та інвестиційному плануванні. Його застосування є особливо актуальним в умовах високої волатильності ринків та зростання ролі цифрових технологій в економіці.

До обмежень дослідження можна віднести залежність результатів від якості та повноти вхідних даних, а також необхідність налаштування параметрів моделей для різних економічних середовищ. Крім того, використання ансамблевих моделей потребує більших обчислювальних ресурсів порівняно з окремими алгоритмами.

Перспективи подальших досліджень полягають у розширенні набору вхідних даних за рахунок альтернативних джерел (зокрема, Big Data та відкритих фінансових платформ), удосконаленні методів пояснюваності для часових моделей (наприклад, WindowSHAP), а також дослідженні можливостей використання трансформерних архітектур у задачах прогнозування інвестицій.

Список використаних джерел:

1. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
2. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
3. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
4. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.
5. Nayeibi, A., Tipirneni, S., Reddy, C. K., Foreman, B., & Subbian, V. (2023). WindowSHAP: An efficient framework for explaining time-series classifiers based on Shapley values. *Journal of Biomedical Informatics*, 144, 104438. <https://doi.org/10.1016/j.jbi.2023.104438>
6. Zern, A., Broelemann, K., & Kasneci, G. (2023). Interventional SHAP values and interaction values for piecewise linear regression trees. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9), 11164–11173. <https://doi.org/10.1609/aaai.v37i9.26322>
7. Abed, A. M. (2024). Accelerate demand forecasting by hybridizing CatBoost with the dingo optimization algorithm to support supply chain conceptual framework precisely. *Frontiers in Sustainability*, 5, 1388771. <https://doi.org/10.3389/frsus.2024.1388771>
8. Kong, X., Chen, Z., Liu, W., Ning, K., Zhang, L., Marier, S. M., Liu, Y., Chen, Y., & Xia, F. (2025). Deep learning for time series forecasting: A survey. *International Journal of Machine Learning and Cybernetics*, 16, 5079–5112. <https://doi.org/10.1007/s13042-025-02560-w>
9. Rygh, T., Vaage, C., Westgaard, S., & de Lange, P. E. (2025). Inflation forecasting: LSTM networks vs. traditional models for accurate predictions. *Journal of Risk and Financial Management*, 18(7), 365. <https://doi.org/10.3390/jrfm18070365>
10. Supriyadi, E., Danila, N., & Ali, Z. (2025). Enhancing financial product forecasting accuracy using EMD and feature selection with ensemble models. *Journal of Open Innovation: Technology, Market, and Complexity*, 11(2), 100531. <https://doi.org/10.1016/j.joitmc.2025.100531>
11. Arsenault, P.-D., Wang, S., & Patenaude, J.-M. (2025). A survey of explainable artificial intelligence (XAI) in financial time series forecasting. *ACM Computing Surveys*, 57(10), 1–37. <https://doi.org/10.1145/3729531>
12. Liao, K., Xuan, X., & Ma, K.-L. (2026). Deep learning for time series forecasting: A survey of recent advances. *Frontiers of Computer Science*, 20, 2011359.

Надійшла до редакції: 05.05.2026

Прийнята до друку: 15.09.2026

Опубліковано: 30.09.2026