МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ВОДНОГО ГОСПОДАРСТВА ТА ПРИРОДОКОРИСТУВАННЯ
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ВОДНОГО ГОСПОДАРСТВА ТА ПРИРОДОКОРИСТУВАННЯ

Кваліфікаційна наукова
праця на правах рукопису

ГАВРИЛЮК МАКСИМ СЕРГІЙОВИЧ

УДК 004.8:631.559

ДИСЕРТАЦІЯ

ІНТЕЛЕКТУАЛЬНІ МОДЕЛІ ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ

122 – Комп’ютерні науки
12 – Інформаційні технології

Подається на здобуття ступеня доктора філософії
Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____ /М.С. Гаврилюк/

Науковий керівник:
Грицюк Петро Михайлович, доктор економічних наук, професор

Рівне – 2026

АНОТАЦІЯ

Гаврилюк М. С. Інтелектуальні моделі прогнозування врожайності на основі методів машинного навчання. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки». – Національний університет водного господарства та природокористування, Рівне, 2026.

Дисертаційна робота присвячена розробленню та дослідженню інтелектуальних моделей прогнозування врожайності на основі методів машинного та глибинного навчання з урахуванням агрокліматичних чинників. Актуальність дослідження зумовлена зростанням кліматичної мінливості, підвищенням частоти екстремальних погодних явищ, просторовою неоднорідністю агрокліматичних умов та необхідністю побудови стійких прогнозних оцінок урожайності для підтримки управлінських рішень у рослинництві. У роботі сформульовано та розв'язано науково-прикладну задачу побудови інтелектуальних моделей прогнозування врожайності, яка охоплює підготовку агрокліматичних даних, формування інформативного ознакового простору, побудову прогнозних і класифікаційних моделей, оцінювання ризику низької врожайності та оптимізаційний аналіз параметрів вегетації. Програмну реалізацію отриманих результатів здійснено у вигляді інформаційно-аналітичної системи «Пшениця України».

Метою дисертаційної роботи є розроблення інтелектуальних моделей прогнозування врожайності на основі методів машинного та глибинного навчання для підвищення точності, стійкості та інформативності прогнозних оцінок в умовах мінливості агрокліматичних чинників. Об'єктом дослідження є процеси інтелектуального аналізу, моделювання та прогнозування врожайності за умов багатофакторного впливу агрокліматичного середовища. Предметом дослідження є методи, моделі та програмні засоби побудови інтелектуальних моделей прогнозування врожайності, що базуються на підготовці агрокліматичних даних, формуванні інформативного ознакового простору, застосуванні методів машинного та глибинного навчання, оцінюванні ризиків і оптимізації параметрів вегетації.

У роботі використано методи системного аналізу, статистичного та кореляційного аналізу, регресійного моделювання, агрокліматичної кластеризації, просторової агрегації кліматичних даних, розвідувального аналізу даних, класифікації, машинного та глибинного навчання, зокрема лінійні моделі, логістичну регресію, Support Vector Machine, Random Forest, рекурентні нейронні мережі GRU та LSTM, а також методи ідентифікації розподілів трендових залишків, квантильного оцінювання ризику та оптимізаційного аналізу. Для формування агрокліматичної бази дослідження використано кліматичні дані ERA5, просторово узгоджені з полігонами областей України рівня ADM1, на основі чого сформовано фінальну матрицю ознак для критичного періоду вегетації.

Наукова новизна одержаних результатів полягає в тому, що удосконалено метод формування агрокліматичного ознакового простору для моделей прогнозування врожайності шляхом просторової агрегації кліматичних даних, побудови трендових залишків і зонального групування областей України; удосконалено методичний підхід до прогнозування врожайності шляхом переходу від числового до класифікаційного прогнозування; набули подальшого розвитку методи ендогенного та екзогенного класифікаційного прогнозування врожайності на основі моделей машинного навчання; удосконалено методику підбору гіперпараметрів прогнозних моделей шляхом поєднання решіткового пошуку з аналізом групи TOP-3 конфігурацій, відібраних за метрикою F1 на валідаційній вибірці; набули подальшого розвитку теоретико-прикладні підходи до застосування рекурентних нейронних мереж GRU та LSTM у задачах прогнозування врожайності; уперше розроблено метод формування нелінійних ознак кумулятивного кліматичного впливу на основі синтезу ознак другого порядку; удосконалено методичні засади оцінювання ризику низької врожайності на основі ідентифікації розподілів трендових залишків і квантильних оцінок ризику; набули подальшого розвитку засоби програмної реалізації інтелектуальних моделей прогнозування врожайності в інформаційно-аналітичній системі.

У результаті дослідження сформовано відтворювану агрокліматичну матрицю ознак для 18 областей України, обґрунтовано доцільність переходу від

абсолютних значень урожайності до трендових залишків як цільової змінної, виконано агрокліматичне групування областей України з виділенням степової, лісостепової та західної зон, що дало змогу збільшити обсяг навчальних вибірок і підвищити стійкість моделей. Побудовано та порівняно ендогенні й екзогенні класифікаційні моделі машинного навчання, а також рекурентні нейромережеві моделі прогнозування. Показано, що для коротких і слабко автокорельованих рядів трендових залишків збільшення складності моделі не гарантує поліпшення якості прогнозу, тоді як використання зовнішніх агрокліматичних і режимних змінних істотно підвищує точність класифікації. Розроблено підхід до оцінювання ризику низької врожайності на основі статистичного та квантильного аналізу трендових залишків, а також підхід до визначення сприятливої температурної траєкторії вегетації. Окремо показано можливість застосування прогнозних моделей для оцінювання втрат урожайності за умов зовнішніх шоків.

Практичне значення одержаних результатів полягає в тому, що розроблені інтелектуальні моделі, методи формування агрокліматичних вибірок, підходи до оцінювання ризику та програмні засоби можуть бути використані для прогнозування врожайності, оцінювання агрокліматичних ризиків і підтримки прийняття управлінських рішень у рослинництві. Інформаційно-аналітична система «Пшениця України» забезпечує автоматизацію підготовки даних, формування ознак, побудови прогнозних моделей, оцінювання ризиків, оптимізації параметрів вегетації, візуалізації результатів, сценарного аналізу та формування аналітичних звітів. Одержані результати можуть бути використані аграрними підприємствами, науково-дослідними установами, органами регіонального управління та в освітньому процесі закладів вищої освіти.

Ключові слова: інтелектуальні моделі, прогнозування врожайності, машинне навчання, глибинне навчання, GRU, LSTM, агрокліматичні дані, класифікаційне прогнозування, трендові залишки, оцінювання ризику, інформаційно-аналітична система.

ABSTRACT

Havryliuk M. S. Intelligent Models for Crop Yield Forecasting Based on Machine Learning Methods. – A qualifying scientific work submitted as a manuscript.

Dissertation for the degree of Doctor of Philosophy in specialty 122 “Computer Science”. – National University of Water and Environmental Engineering, Rivne, 2026.

The dissertation is devoted to the development and investigation of intelligent yield forecasting models based on machine learning and deep learning methods with consideration of agroclimatic factors. The relevance of the research is determined by increasing climate variability, the growing frequency of extreme weather events, the spatial heterogeneity of agroclimatic conditions, and the need to build robust yield forecasts for supporting managerial decision-making in crop production. The dissertation formulates and solves a scientific and applied problem of constructing intelligent yield forecasting models, which includes the preparation of agroclimatic data, the formation of an informative feature space, the development of forecasting and classification models, the assessment of low-yield risk, and the optimisation analysis of vegetation parameters. The software implementation of the obtained results was carried out in the form of the information and analytical system “Wheat of Ukraine”.

The aim of the dissertation is to develop intelligent yield forecasting models based on machine learning and deep learning methods in order to improve the accuracy, robustness, and informativeness of forecasts under conditions of agroclimatic variability. The object of the research is the processes of intelligent analysis, modelling, and forecasting of yield under the multifactorial influence of the agroclimatic environment. The subject of the research is methods, models, and software tools for constructing intelligent yield forecasting models based on the preparation of agroclimatic data, the formation of an informative feature space, the application of machine learning and deep learning methods, risk assessment, and optimisation of vegetation parameters.

The dissertation uses methods of systems analysis, statistical and correlation analysis, regression modelling, agroclimatic clustering, spatial aggregation of climate data, exploratory data analysis, classification, machine learning and deep learning, including linear models, logistic regression, Support Vector Machine, Random Forest,

recurrent neural networks GRU and LSTM, as well as methods for identifying distributions of trend residuals, quantile-based risk assessment, and optimisation analysis. To form the agroclimatic database of the study, ERA5 climate data were used and spatially aligned with ADM1-level polygons of Ukrainian regions, on the basis of which the final feature matrix for the critical vegetation period was formed.

The scientific novelty of the obtained results lies in the following: the method for forming an agroclimatic feature space for yield forecasting models has been improved by means of spatial aggregation of climate data, construction of trend residuals, and zonal grouping of Ukrainian regions; the methodological approach to yield forecasting has been improved by moving from numerical to classification forecasting; methods of endogenous and exogenous classification-based yield forecasting using machine learning models have been further developed; the methodology for selecting hyperparameters of forecasting models has been improved by combining grid search with the analysis of a TOP-3 group of configurations selected by the F1 metric on the validation sample; theoretical and applied approaches to the use of GRU and LSTM recurrent neural networks in yield forecasting tasks have been further developed; a method for forming nonlinear features of cumulative climate impact based on the synthesis of second-order features has been developed for the first time; methodological principles for assessing low-yield risk have been improved on the basis of identifying distributions of trend residuals and quantile-based risk estimates; and software tools for implementing intelligent yield forecasting models in an information and analytical system have been further developed.

As a result of the study, a reproducible agroclimatic feature matrix for 18 regions of Ukraine was formed, the feasibility of moving from absolute yield values to trend residuals as the target variable was substantiated, and agroclimatic grouping of Ukrainian regions was carried out with the identification of the Steppe, Forest-Steppe, and Western zones, which made it possible to increase the size of training samples and improve the robustness of models. Endogenous and exogenous classification models of machine learning, as well as recurrent neural network forecasting models, were constructed and compared. It was shown that for short and weakly autocorrelated series of trend residuals,

increasing model complexity does not guarantee an improvement in forecast quality, whereas the use of external agroclimatic and regime variables significantly improves classification accuracy. An approach to assessing low-yield risk based on statistical and quantile analysis of trend residuals was developed, as well as an approach to determining a favourable temperature trajectory of the vegetation period. The possibility of applying forecasting models to assess yield losses under external shocks was also demonstrated.

The practical significance of the obtained results lies in the fact that the developed intelligent models, methods for forming agroclimatic samples, approaches to risk assessment, and software tools can be used for yield forecasting, assessment of agroclimatic risks, and support of managerial decision-making in crop production. The information and analytical system “Wheat of Ukraine” provides automation of data preparation, feature formation, construction of forecasting models, risk assessment, optimisation of vegetation parameters, visualisation of results, scenario analysis, and generation of analytical reports. The obtained results can be used by agricultural enterprises, research institutions, regional management bodies, and in the educational process of higher education institutions.

Keywords: intelligent models, yield forecasting, machine learning, deep learning, GRU, LSTM, agroclimatic data, classification forecasting, trend residuals, risk assessment, information and analytical system.

Список публікацій здобувача за темою дисертації

Публікації у виданнях, проіндексованих у наукометричній базі Scopus

1. Hrytsiuk P., Babych T., Baranovsky S., Havryliuk M. Assessing of climate impact on wheat yield using machine learning techniques. In: *Information Control Systems & Technologies* : proceedings of the 11th International Conference, ICST-2023, Odesa, Ukraine, September 21–23, 2023. *CEUR Workshop Proceedings*. 2023. Vol. 3513. P. 314–329. URL: https://ceur-ws.org/Vol-3513/paper26.pdf. **Scopus** (1,32/0,33 д.а.; *авторський внесок здобувача — підготовка та попередня обробка даних врожайності й кліматичних показників, формування експериментальної вибірки, реалізація моделей машинного навчання, проведення числових експериментів, аналіз отриманих результатів та підготовка графічних матеріалів*).

2. Hrytsiuk P., Havryliuk M. Modeling of the nonlinear impact of climatic factors on wheat yield using machine learning techniques. In: Ermolayev V. et al. (eds) *Information and Communication Technologies in Education, Research, and Industrial Applications* : 19th International Conference, ICTERI 2024, Lviv, Ukraine, September 23–27, 2024, Proceedings. Communications in Computer and Information Science. Cham : Springer, 2025. Vol. 2359. P. 20–35. https://doi.org/10.1007/978-3-031-81372-6_2. **Scopus** (1,06/0,53 д.а.; *авторський внесок здобувача — формування набору кліматичних факторів для моделювання врожайності пшениці, реалізація нелінійних моделей машинного навчання, проведення обчислювальних експериментів, оцінювання якості моделей, інтерпретація впливу кліматичних змінних на результат прогнозування*).

3. Hrytsiuk P., Havryliuk M., Nehrey M. Constructing the optimal temperature trajectory for the wheat vegetative cycle. In: Ermolayev V. et al. (eds) *Information and Communication Technologies in Education, Research, and Industrial Applications* : 20th International Conference, ICTERI 2025, Nice, France, September 1–4, 2025, Proceedings. Communications in Computer and Information Science. Cham : Springer, 2025. Vol. 2763. P. 101–116. https://doi.org/10.1007/978-3-032-10477-9_8. **Scopus** (0,91/0,3 д.а.; *авторський внесок здобувача — підготовка вихідних даних, участь у*

формалізації задачі побудови оптимальної температурної траєкторії вегетаційного циклу пшениці, виконання числових розрахунків, аналіз модельних результатів та підготовка частини ілюстративного матеріалу).

4. Hrytsiuk P., Babych T., Kardash O., Havryliuk M. Application of machine learning for wheat yield prediction. In: Potapov I. et al. (eds) *Digitalisation and Digital Transformation* : First Research Twinning Conference, RTC-Digital 2023, Liverpool, UK, March 27–30, 2023, Proceedings. Communications in Computer and Information Science. Cham : Springer, 2026. Vol. 2647. P. 3–11. https://doi.org/10.1007/978-3-032-04731-1_1. **Scopus** (0,6/0,15 д.а.; *авторський внесок здобувача — підготовка даних для прогнозування врожайності пшениці, реалізація та тестування моделей машинного навчання, участь у порівнянні результатів прогнозування, підготовка експериментальної частини дослідження*).

Статті у наукових фахових виданнях України категорії «Б»

5. Грицюк П. М., Гаврилюк М. С. Моделювання впливу кліматичних факторів на врожайність пшениці методами машинного навчання. *Штучний інтелект*. 2025. Т. 30, № 1. С. 121–130. https://doi.org/10.15407/jai2025.01.121 (0,92/0,46 д.а.; *авторський внесок здобувача — побудова інформаційної бази дослідження, підготовка кліматичних і врожайних показників, реалізація моделей машинного навчання, проведення обчислювальних експериментів, оцінювання точності моделей та інтерпретація одержаних результатів*).

6. Грицюк П. М., Гаврилюк М. С., Джоші О. І. Ідентифікація закону розподілу трендових залишків врожайності як інструмент моделювання ризиків зерновиробництва. *Штучний інтелект*. 2025. Т. 30, № 2. С. 72–83. https://doi.org/10.15407/jai2025.02.072 (1,02/0,34 д.а.; *авторський внесок здобувача — розрахунок трендових залишків врожайності, підготовка даних для статистичного аналізу, участь в ідентифікації закону розподілу залишків, виконання обчислювальних експериментів, інтерпретація результатів для задач моделювання ризиків зерновиробництва*).

Наукові праці, які засвідчують апробацію матеріалів дисертації

7. Грицюк П. М., Гаврилюк М. С. Класифікаційний підхід до прогнозування врожайності. *Форум молодих економістів-кібернетиків «Моделювання економіки: проблеми, тенденції, досвід»* : тези доповідей X Всеукраїнської науково-практичної конференції, м. Львів, 25 листопада 2022 р. Львів, 2022. С. 47–49. (0,15/0,07 д.а.)

8. Havryliuk M., Hrytsiuk P., Kardash O., Babych T. Forecasting of wheat yield using machine learning techniques. *Digital Theme UK–Ukraine Research Twinning Conference* : UA-DIGITAL 2023, online, March 27–31, 2023. (0,59/0,15 д.а.)

9. Havryliuk M. S. Wheat yield forecasting using machine learning methods. *Проблеми та перспективи розвитку сучасної науки* : збірник тез доповідей Міжнародної науково-практичної конференції молодих науковців, аспірантів і здобувачів вищої освіти, м. Рівне, 11–12 травня 2023 р. Рівне : НУВГП, 2023. С. 139–142. (0,17/0,17 д.а.)

10. Грицюк П. М., Гаврилюк М. С. Оцінка втрат врожаю пшениці внаслідок військової агресії росії. *Інтегровані інтелектуальні робототехнічні комплекси (ІІРТК-2023)* : матеріали XVI Міжнародної науково-практичної конференції, м. Київ, 23–24 травня 2023 р. Київ : НАУ, 2023. С. 374–376. (0,18/0,09 д.а.)

11. Hrytsiuk P., Babych T., Baranovskii S., Havryliuk M. Assessing of climate impact on wheat yield using machine learning techniques. *Information Control Systems and Technologies (ICST-2023)* : materials of the XI International Scientific-Practical Conference, Odesa, Ukraine, September 21–23, 2023. P. 102–105. (0,16/0,04 д.а.)

12. Грицюк П. М., Гаврилюк М. С. Використання машинного навчання для управління процесами аграрної економіки. *Міжнародна конференція з інноваційних рішень у програмній інженерії (ICISSE 2023)*, м. Івано-Франківськ, 29–30 листопада 2023 р. Івано-Франківськ, 2023. С. 150–154. https://doi.org/10.5281/zenodo.10397356 (0,43/0,21 д.а.)

13. Гаврилюк М. С. Використання машинного навчання для управління процесами аграрної економіки. *Науково-практична конференція студентів та аспірантів навчально-наукового інституту кібернетики, інформаційних*

технологій та інженерії, м. Рівне, 23 травня 2024 р. Рівне : НУВГП, 2024. С. 17. (0,08/0,08 д.а.)

14. Гаврилюк М. С., Грицюк П. М. Ідентифікація закону розподілу залишків врожайності сільськогосподарських культур. *Комп'ютерне моделювання та програмне забезпечення інформаційних систем і технологій 2024 (КМПЗ 2024)* : матеріали IV Міжнародної науково-практичної конференції, м. Чернівці, 30 травня – 1 червня 2024 р. Чернівці, 2024. С. 76–80. (0,2/0,1 д.а.)

15. Грицюк П. М., Гаврилюк М. С. Оцінювання нелінійного впливу кліматичних факторів на врожайність пшениці. *Штучний інтелект. Наука. Бізнес* : матеріали Всеукраїнської науково-практичної конференції, м. Одеса, 22 жовтня 2024 р. Одеса, 2024. С. 18–21. (0,19/0,1 д.а.)

16. Гаврилюк М. С., Грицюк П. М. Дослідження впливу кліматичних факторів на коливання врожайності пшениці. *Форум молодих економістів-кібернетиків. Моделювання економіки: проблеми, тенденції, досвід* : матеріали XII Всеукраїнської науково-практичної конференції, м. Львів, 22–23 листопада 2024 р. Львів, 2024. С. 11–14. (0,21/0,1 д.а.)

17. Havryliuk M., Hrytsiuk P. Modeling the nonlinear influence of climatic factors on wheat yield using machine learning methods. *Advances in Science, Technology, and Society* : proceedings of the International Scientific Conference, Oxford, United Kingdom, February 20, 2025. P. 210–214. (0,28/0,14 д.а.)

18. Грицюк П. М., Бабич Т. Ю., Гаврилюк М. С. Побудова оптимальної температурної траєкторії вегетації пшениці. *Математичні методи, моделі та інформаційні технології в економіці* : матеріали IX Міжнародної науково-методичної конференції, м. Чернівці, 24–25 квітня 2025 р. Чернівці, 2025. С. 69–71. (0,1/0,03 д.а.)

19. Гаврилюк М. С. Моделювання нелінійного впливу кліматичних факторів на врожайність пшениці методами машинного навчання. *Науково-практична конференція студентів та аспірантів навчально-наукового інституту кібернетики, інформаційних технологій та інженерії*, м. Рівне, 14 травня 2025 р. Рівне : НУВГП, 2025. С. 49. (0,07/0,07 д.а.)

20. Гаврилюк М. С., Грицюк П. М. Використання закону розподілу трендових залишків врожайності для моделювання ризиків виробництва зерна. *Моделювання, керування та інформаційні технології* : матеріали VIII Міжнародної науково-практичної конференції, м. Рівне, 6–8 листопада 2025 р. Рівне : НУВГП, 2025. С. 286–288. https://doi.org/10.31713/MCIT.2025.089 (0,26/0,13 д.а.)

21. Гаврилюк М. С. Нелінійне моделювання впливу кліматичних факторів на врожайність пшениці в Україні. *Форум молодих економістів-кібернетиків. Моделювання економіки: проблеми, тенденції, досвід* : матеріали XIII Всеукраїнської науково-практичної конференції, м. Львів, 21–22 листопада 2025 р. Львів, 2025. С. 101–102. (0,1/0,1 д.а.)

Праці, які додатково відображають наукові результати дисертації

22. Hrytsiuk P., Babych T., Hladka O., Nehrey M., Havryliuk M. Machine learning-based modeling of climate effects on wheat yield in the Ukrainian Steppe. *Journal of Lifestyle and SDGs Review*. 2026. Vol. 6, No. 1. Art. e08120. https://doi.org/10.47172/2965-730X.SDGsReview.v6.n01.pe08120. (0,88/0,18 д.а.; *авторський внесок здобувача — формування вибірки для степової зони України, підготовка кліматичних ознак, реалізація обчислювальних експериментів із використанням методів машинного навчання, аналіз впливу кліматичних факторів на врожайність пшениці, підготовка таблиць і графіків*).

23. Свідоцтво про реєстрацію авторського права на твір № 127859. Комп’ютерна програма «Математичне моделювання впливу кліматичних факторів на врожайність методами машинного навчання». Автори: Гаврилюк М. С., Грицюк П. М., Бабич Т. Ю., Кардаш О. Л. Дата реєстрації 26 червня 2024 р.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

Accuracy – частка правильних класифікацій; метрика якості класифікаційної моделі.

ADM1 – перший рівень адміністративного поділу; рівень цифрових меж областей, використаний для просторового узгодження даних.

ARIMA – Autoregressive Integrated Moving Average; авторегресійна інтегрована модель ковзного середнього для аналізу часових рядів.

Calinski–Harabasz – індекс оцінювання якості кластеризації, що враховує співвідношення міжкластерної та внутрішньокластерної варіації.

CVaR – Conditional Value-at-Risk; умовна міра ризику, що характеризує середній рівень втрат у нижньому хвості розподілу.

Davies–Bouldin – індекс оцінювання якості кластеризації, менші значення якого відповідають компактнішим і краще відокремленим кластерам.

Differential Evolution – метаевристичний алгоритм диференціальної еволюції, використаний для оптимізаційного аналізу.

Django – вебфреймворк, використаний для серверної частини інформаційно-аналітичної системи.

EDA – Exploratory Data Analysis; розвідувальний аналіз даних.

EOSDA Crop Monitoring – цифрова платформа супутникового аграрного моніторингу.

ERA5 – реаналізний кліматичний набір даних, використаний як джерело температурних показників та опадів.

F1 / F1-score – гармонічне середнє Precision і Recall; метрика якості класифікаційної моделі.

F1_LY – значення метрики F1 для класу низької врожайності (Low Yield).

Gaussian Mixture – ймовірнісна модель суміші гауссових розподілів, використана для кластеризації.

Gradient Boosting – ансамблевий метод машинного навчання, що послідовно уточнює модель за помилками попередніх кроків.

GRU – Gated Recurrent Unit; рекурентна нейронна мережа з механізмами керування прихованим станом.

IAC – інформаційно-аналітична система.

KMeans – метод кластеризації k-середніх.

Logistic Regression / Log Regression – логістична регресія; класифікаційна модель для оцінювання ймовірності належності до класу.

LSTM – Long Short-Term Memory; рекурентна нейронна мережа з коміркою пам’яті та системою воріт.

max_depth – максимальна глибина дерева у моделі Random Forest.

n_estimators – кількість дерев або базових оцінювачів в ансамблевій моделі.

n_steps – ширина ковзного часового вікна, тобто кількість попередніх значень ряду, що подаються на вхід моделі.

OLS – Ordinary Least Squares; метод найменших квадратів.

Precision – точність позитивних передбачень класифікаційної моделі.

Precision_LY – значення Precision для класу низької врожайності (Low Yield).

R1–R4 – показники ранжування областей за різними статистичними мірами ризику.

R10 – сума опадів за квітень.

R20 – сума опадів за травень.

R30 – сума опадів за червень.

R^2 – коефіцієнт детермінації, що характеризує частку варіації цільової змінної, пояснену моделлю.

Random Forest / RF – ансамблевий метод машинного навчання «випадковий ліс».

RBF – Radial Basis Function; радіально-базисна функція, тип ядра в моделі SVM.

Recall – повнота класифікації; здатність моделі виявляти об’єкти цільового класу.

Recall_LY – значення Recall для класу низької врожайності (Low Yield).

RMSE – Root Mean Squared Error; корінь із середньоквадратичної помилки.

RNN – Recurrent Neural Network; рекурентна нейронна мережа.

ROC_AUC – площа під ROC-кривою; метрика здатності моделі розділяти класи за різних порогів прийняття рішення.

Silhouette – коефіцієнт силуету; метрика оцінювання відокремленості кластерів.

Specificity – специфічність класифікації; здатність моделі правильно визначати об’єкти негативного класу.

Specificity_HY – значення Specificity для класу високої врожайності (High Yield).

Spectral Clustering – метод спектральної кластеризації.

Support Vector Machine / SVM – метод опорних векторів.

t1–t9 – декадні температурні предиктори за період квітень–червень.

tmean – середня добова температура, використана для формування декадних температурних предикторів.

units – кількість нейронів у рекурентному шарі нейронної мережі.

VaR – Value-at-Risk; гранична міра ризику втрат за заданого рівня довіри.

y – фактична врожайність; цільова змінна моделі.

ε – трендовий залишок врожайності, тобто відхилення фактичної врожайності від її трендового рівня.

econ – порядкова змінна, що кодує тип аграрної економіки у відповідному історичному періоді.

ВСТУП

Актуальність теми. Прогнозування врожайності сільськогосподарських культур є актуальною задачею сучасних комп'ютерних наук, оскільки потребує розроблення методів інтелектуального аналізу даних та моделей машинного навчання для опрацювання різнорідних просторово-часових даних. У задачах аграрної аналітики використовуються статистичні ряди врожайності, кліматичні, метеорологічні, економічні масиви даних, які мають різну протяжність, часову дискретність та просторовий масштаб. Тому прогнозування врожайності є складною інформаційно-технологічною проблемою, пов'язаною з формуванням ознакового простору, побудовою інтелектуальних моделей, перевіркою їхньої стійкості та інтеграцією результатів у програмні засоби підтримки прийняття рішень.

Складність прогнозування врожайності зумовлена багатофакторною природою її формування. На кінцевий результат одночасно впливають погодно-кліматичні, просторові, ґрунтові, агротехнологічні, економічні та організаційні чинники. Крім того, урожайність має виражену міжрічну мінливість, залежить від фази розвитку культури, просторової неоднорідності територій, локальних агрокліматичних умов, екстремальних температурних явищ, розподілу опадів і тривалості посушливих періодів. Тому побудова прогнозних моделей потребує не лише вибору алгоритму, а й обґрунтованого формування ознакового простору, просторового й часового узгодження даних, виділення трендової та залишкової складових, а також перевірки стійкості моделей в умовах обмежених вибірок.

Сучасні методи машинного та глибинного навчання відкривають нові можливості для прогнозування врожайності, оскільки дають змогу враховувати нелінійні залежності, взаємодію кліматичних чинників, часову структуру даних і складні зв'язки між предикторами та цільовою змінною. У задачах аграрного прогнозування можуть використовуватися регресійні, класифікаційні, ансамблеві, нейромережеві та гібридні моделі. Водночас практичне застосування таких моделей ускладнюється короткими часовими рядами регіональних спостережень, неоднорідністю агрокліматичних умов, можливим перенавчанням складних

алгоритмів, потребою в коректній валідації та необхідністю змістовної інтерпретації результатів. Це зумовлює потребу в розробленні не окремих ізольованих моделей, а цілісного підходу до їх побудови, порівняння, оцінювання й програмної реалізації. Важливим аспектом прогнозування врожайності є забезпечення стійкості отриманих прогнозів. Стійкість прогнозних оцінок забезпечується за рахунок просторової агрегації агрокліматичних даних, зонального формування навчальних вибірок, коректного поділу даних на навчальну, валідаційну та тестову частини, застосування крос-валідації, підбору гіперпараметрів і порівняння альтернативних моделей машинного та глибинного навчання.

Проблематика прогнозування врожайності та оцінювання впливу агрокліматичних чинників на аграрне виробництво є предметом досліджень українських і зарубіжних науковців. Питання кліматичної мінливості, агрометеорологічного аналізу, оцінювання врожайності та аграрних ризиків розглядалися у працях Т. Адаменко, П. Грицюка, Л. Бачишиної та інших дослідників. Розвиток статистичних, економетричних, машинно-навчальних і глибинних підходів до прогнозування врожайності відображено у працях T. van Klompenburg, G. Leng, A. Crane-Droesch, D. Elavarasan, B. S. Sidhu, M. Kalimuthu, S. Veenadhari, M. Kuradusenge, P. Feng, B. Wang, A. Barbosa та інших учених. У цих роботах обґрунтовано доцільність використання кліматичних, супутникових, статистичних, просторових і виробничих даних, а також методів машинного навчання для аналізу та прогнозування врожайності. Водночас активний розвиток цього напряму не усуває низки проблем, пов’язаних із формуванням регіонально агрегованих вибірок, побудовою інформативних ознак, класифікацією ризикових станів, оцінюванням нижньохвостових ризиків і програмною інтеграцією результатів моделювання.

Окремого значення набуває розвиток інформаційних систем і цифрових платформ аграрної аналітики. Сучасні платформи точного землеробства, геоінформаційні та супутникові сервіси, системи підтримки прийняття рішень і дослідницькі програмні рішення забезпечують збирання, збереження, аналіз і

візуалізацію аграрних даних. Вони мають важливе практичне значення для моніторингу стану посівів, управління польовими операціями, аналізу просторової неоднорідності та підтримки агротехнологічних рішень. Однак значна частина таких систем орієнтована переважно на рівень поля або господарства, тоді як регіональне прогнозування врожайності потребує роботи з агрегованими статистичними, кліматичними та просторовими даними. Дослідницькі програмні рішення частіше зосереджуються на окремій моделі або окремому алгоритмічному контурі й не завжди охоплюють повний цикл: підготовку даних, формування ознак, навчання й порівняння моделей, оцінювання ризику, сценарний аналіз, картографічне подання та формування звітних матеріалів.

Отже, попри наявність значної кількості досліджень, інформаційних систем і цифрових платформ, недостатньо опрацьованою залишається задача побудови інтелектуальних моделей прогнозування врожайності, які поєднують просторово узгоджені агрокліматичні дані, зональне формування навчальних вибірок, нелінійне розширення ознакового простору, числове й класифікаційне прогнозування, дослідження меж застосовності рекурентних нейронних мереж, оцінювання ризику низької врожайності та сценарно-оптимізаційний аналіз умов вегетації. Такий підхід є важливим для переходу від точкового прогнозу до комплексної прогнозної аналітики, що дає змогу не лише оцінити очікуваний рівень урожайності, а й виявити ризикові стани, проаналізувати несприятливі відхилення, оцінити вплив кліматичних чинників і сформувати інформаційну основу для прийняття рішень.

У дисертаційній роботі апробацію запропонованих моделей виконано на даних врожайності пшениці в Україні. Вибір цієї культури зумовлений її важливим значенням для аграрного виробництва, чутливістю до агрокліматичних умов, наявністю регіональних статистичних рядів і можливістю змістовної перевірки моделей на реальних даних. При цьому пшениця розглядається не як обмеження теми дисертації, а як прикладна база для апробації розроблених інтелектуальних моделей прогнозування врожайності, які можуть бути адаптовані до інших культур за наявності відповідних статистичних, кліматичних і просторових даних. У

дослідженні використано усереднені статистичні показники врожайності пшениці на рівні адміністративних областей, тому результати моделювання відображають регіональну реакцію врожайності на агрокліматичні умови, а не сортові чи господарські відмінності окремих виробників.

Таким чином, актуальною є науково-прикладна задача розроблення, обґрунтування та експериментальної перевірки інтелектуальних моделей прогнозування врожайності на основі методів машинного навчання. Розв'язання цієї задачі передбачає формування просторово узгодженого агрокліматичного ознакового простору, побудову ендогенних та екзогенних моделей прогнозування, врахування нелінійного й кумулятивного впливу погодних умов, класифікаційне виявлення років низької врожайності, дослідження можливостей моделей глибинного навчання, оцінювання ризику низької врожайності та сценарно-оптимізаційний аналіз сприятливих кліматичних траєкторій. Програмна інтеграція розроблених моделей в інформаційно-аналітичному середовищі забезпечує їх практичне використання для підтримки прогнозних, страхових, інвестиційних та управлінських рішень у рослинництві.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана відповідно до тематичного плану науково-дослідних робіт Національного університету водного господарства та природокористування і пов'язана з виконанням науково-дослідної теми «Інформаційні технології моделювання екологічних, економічних та соціальних процесів» (номер державної реєстрації № 0121U111686), у межах якої здобувач розробив та експериментально апробував інтелектуальні моделі прогнозування врожайності з використанням методів машинного та глибинного навчання, сформував агрокліматичну інформаційну базу дослідження і реалізував програмні засоби аналітичної підтримки прийняття рішень. Експериментальну апробацію розроблених моделей виконано на даних врожайності пшениці в Україні.

Мета і завдання дослідження. Метою дисертаційної роботи є розроблення, обґрунтування та експериментальна перевірка інтелектуальних моделей прогнозування врожайності на основі методів машинного навчання для підвищення

точності, стійкості та інформативності прогнозних оцінок з урахуванням агрокліматичної мінливості, регіональної неоднорідності та ризику низької врожайності.

Для досягнення поставленої мети в роботі необхідно розв’язати такі завдання:

1. Удосконалити метод формування інформаційної бази та агрокліматичного ознакового простору для моделювання врожайності шляхом просторової агрегації кліматичних показників, побудови трендових залишків і зонального групування регіональних спостережень.

2. Удосконалити методичні засади оцінювання ризику низької врожайності на основі ідентифікації розподілів трендових залишків і використання статистичних та квантильних мір ризику.

3. Розробити класифікаційний підхід до прогнозування рівня врожайності за категоріями «низька» / «висока» з використанням трендових залишків як основи для формування цільової змінної.

4. Розробити та експериментально перевірити ендогенні та екзогенні класифікаційні моделі машинного навчання для прогнозування класу врожайності на основі внутрішньої динаміки трендових залишків та екзогенних агрокліматичних ознак у різних агрокліматичних зонах.

5. Розвинути теоретико-прикладні підходи до застосування рекурентних нейронних мереж у задачах прогнозування врожайності шляхом дослідження моделей LSTM і GRU в ендогенній та екзогенній постановках.

6. Удосконалити методику підбору гіперпараметрів прогнозних моделей шляхом поєднання решіткового пошуку з аналізом групи найкращих конфігурацій та незалежним тестовим оцінюванням.

7. Розробити метод синтезу нелінійних кліматичних ознак другого порядку для врахування неадитивного та кумулятивного впливу погодних умов на врожайність.

8. Розробити програмні засоби інтеграції інтелектуальних моделей прогнозування врожайності в інформаційно-аналітичній системі, що забезпечує

підготовку даних, формування ознак, побудову прогнозів, оцінювання ризику, сценарний аналіз і візуалізацію результатів.

Об’єкт дослідження. Процеси інформаційного опрацювання агрокліматичних і врожайнісних даних, інтелектуального аналізу та прогнозного моделювання врожайності сільськогосподарських культур за умов багатофакторного впливу агрокліматичного середовища, просторової неоднорідності та міжрічної мінливості врожайності.

Предмет дослідження. Методи, моделі та алгоритмічні процедури формування агрокліматичного ознакового простору, побудови ендогенних і екзогенних інтелектуальних моделей прогнозування врожайності, класифікації рівня врожайності, підбору гіперпараметрів, оцінювання ризику низької врожайності, синтезу нелінійних кліматичних ознак та програмної інтеграції результатів моделювання в інформаційно-аналітичній системі.

Методи дослідження. Для досягнення поставленої мети та розв’язання сформульованих завдань у дисертаційній роботі використано сукупність загальнонаукових, статистичних, математичних, інформаційних і спеціальних методів дослідження.

На етапі аналізу предметної області, сучасних підходів до прогнозування врожайності, оцінювання ризиків та застосування методів машинного і глибинного навчання використано методи теоретичного узагальнення, системного аналізу, порівняльного аналізу та формалізації задачі дослідження.

Для формування агрокліматичної бази дослідження використано методи інтеграції даних, просторової агрегації, геоінформаційного аналізу та часової агрегації метеорологічних показників. На їх основі виконано узгодження кліматичних даних ERA5 з полігонами областей України рівня ADM1, формування декадних температурних і опадових ознак для критичного періоду вегетації та побудову фінальної матриці агрокліматичних показників.

Для виділення довгострокової складової врожайності та формування трендових залишків як цільової змінної використано методи регресійного аналізу, зокрема побудову лінійного тренду методом найменших квадратів. Для

агрокліматичної кластеризації областей застосовано методи кореляційного аналізу трендових залишків, кластерного аналізу, аналізу подібності часових рядів і зональної агрегації спостережень.

Для побудови прогнозних моделей урожайності використано методи статистичного моделювання, машинного навчання та глибинного навчання. У межах дослідження застосовано лінійні регресійні моделі, логістичну регресію, Support Vector Machine, Random Forest, класифікаційні моделі прогнозування рівня врожайності, а також рекурентні нейронні мережі типу GRU та LSTM. Підвищення інформативності прогнозних оцінок у роботі забезпечено не лише розширенням агрокліматичного ознакового простору, а й переходом від точкового числового прогнозування до комплексної аналітичної оцінки, що включає класифікацію рівня врожайності, аналіз трендових залишків, виявлення ризику низької врожайності та сценарно-оптимізаційне дослідження кліматичних умов. Для врахування неадитивного та кумулятивного впливу погодних факторів застосовано синтез нелінійних ознак другого порядку, зокрема квадратів базових предикторів і добутків кліматичних показників суміжних часових інтервалів. Це дало змогу підвищити інформативність вхідного простору моделей і формалізувати взаємодію агрокліматичних чинників у послідовних фазах вегетації.

Для підвищення стійкості прогнозних оцінок у роботі використано процедури просторової агрегації агрокліматичних даних, зонального формування навчальних вибірок, поділу даних на навчальну, валідаційну та тестову частини, перехресної крос-валідації, підбору гіперпараметрів моделей, контролю перенавчання та порівняння результатів різних методів машинного і глибинного навчання.

Для оцінювання ризику низької врожайності використано статистичні та квантильні методи аналізу, методи ідентифікації законів розподілу трендових залишків, а також підходи до розрахунку ризикових оцінок VaR і CVaR. Для розв’язання задачі визначення оптимальної температурної траєкторії вегетації застосовано методи оптимізаційного аналізу, зокрема алгоритм Differential Evolution у поєднанні з нелінійними моделями Random Forest та SVM. Для

оцінювання втрат урожайності за умов зовнішніх шоків використано порівняння фактичних, трендових і прогнозних оцінок врожайності, а також методи сценарного та прикладного аналітичного аналізу.

Для програмної інтеграції розроблених інтелектуальних моделей використано методи проєктування інформаційних систем, моделювання структури даних, функціональної декомпозиції програмних модулів, веборієнтованої розробки та інтеграції аналітичних компонентів. Програмну реалізацію інформаційно-аналітичної системи «Пшениця України» виконано з використанням сучасних засобів програмування, візуалізації та керування даними, що забезпечило відтворюваність, модульність і практичну придатність результатів дослідження.

Наукова новизна отриманих результатів. У дисертаційній роботі розв'язано науково-прикладну задачу розроблення інтелектуальних моделей прогнозування врожайності на основі методів машинного навчання, що поєднують просторово узгоджені агрокліматичні дані, зональне формування навчальних вибірок, нелінійне ознакове представлення, класифікаційне та нейромережеве моделювання, оцінювання ризику і програмну інтеграцію результатів.

У межах виконаного дослідження отримано такі нові наукові результати:

вперше:

- розроблено метод формування нелінійних ознак кумулятивного кліматичного впливу на врожайність шляхом синтезу квадратів базових агрокліматичних предикторів і добутків кліматичних показників суміжних часових інтервалів. Це дозволяє формалізувати неадитивні ефекти впливу погодних умов у послідовних фазах вегетації та підвищити інтерпретованість регресійних моделей урожайності;

удосконалено:

- метод формування агрокліматичного ознакового простору для моделей прогнозування врожайності, який, на відміну від використання окремих коротких регіональних рядів, передбачає просторову агрегацію кліматичних даних, формування матриці ознак типу «область–рік–агрокліматичні показники», розрахунок трендових залишків урожайності та групування областей у відносно

однорідні агрокліматичні зони. Це забезпечує підготовку структурованих зональних вибірок для подальшого прогнозування, оцінювання ризиків і сценарного аналізу;

- методичний підхід до прогнозування врожайності шляхом переходу від суто числового прогнозування до класифікаційної постановки задачі, у якій позитивним класом визначено клас низької врожайності. Такий підхід дає змогу спрямувати прогнозну модель на раннє виявлення несприятливих років і використати метрику F1_LY як цільовий критерій якості;

- методику підбору гіперпараметрів прогнозних моделей шляхом поєднання решіткового пошуку з аналізом групи TOP-3 конфігурацій, відібраних за метрикою F1_LY на валідаційній вибірці. На відміну від традиційного підходу, який ґрунтується на виборі однієї найкращої конфігурації, запропонована методика передбачає усереднене оцінювання групи найкращих моделей на тестовій вибірці, що підвищує стійкість інтерпретації результатів і надійність порівняння моделей під час переходу до нових даних;

- методичні засади оцінювання ризику низької врожайності на основі аналізу розподілів трендових залишків урожайності та обчислення нижньохвостових квантильних мір ризику VaR і CVaR. Це дозволяє кількісно оцінювати ризик несприятливих відхилень урожайності від трендового рівня з урахуванням зональної специфіки розподілів;

набули подальшого розвитку:

- методи класифікаційного прогнозування врожайності на основі моделей машинного навчання шляхом побудови та порівняння ендогенної й екзогенної постановок задачі. На відміну від підходів, що розглядають лише внутрішню динаміку ряду або лише зовнішні фактори ізольовано, запропоновано єдину схему оцінювання моделей Logistic Regression, Random Forest і SVM RBF із формуванням цільового класу за трендовими залишками та використанням ендогенних лагових ознак і екзогенних агрокліматичних предикторів. Це дало змогу встановити перевагу екзогенного підходу для виявлення років низької врожайності;

- теоретико-прикладні підходи до застосування рекурентних нейронних мереж LSTM і GRU у задачах прогнозування врожайності шляхом встановлення меж їх застосовності для коротких часових рядів трендових залишків і обґрунтування доцільності переходу від ендогенних до екзогенних RNN-моделей із залученням агрокліматичних та економічних факторів. Це дозволило уточнити роль рекурентних нейронних мереж у системі інтелектуальних моделей прогнозування врожайності та обґрунтувати необхідність використання зовнішніх кліматичних факторів для підвищення стійкості прогнозних результатів;

- засоби програмної реалізації інтелектуальних моделей прогнозування врожайності завдяки їх інтеграції в інформаційно-аналітичній системі, яка поєднує процедури підготовки агрокліматичних даних, формування ознакового простору, класифікаційного прогнозування, оцінювання ризику, сценарного аналізу та візуалізації результатів. Це дозволило забезпечити єдиний програмний контур опрацювання даних, побудови прогнозних моделей, інтерпретації результатів та подання аналітичних оцінок у формі, придатній для практичного використання в задачах підтримки прийняття рішень у рослинництві.

Практичне значення отриманих результатів. Практичне значення одержаних результатів полягає в тому, що розроблені в дисертаційній роботі інтелектуальні моделі, методи формування агрокліматичних вибірок, підходи до оцінювання ризику та програмні засоби можуть бути використані для прогнозування врожайності, аналізу агрокліматичних ризиків і підтримки прийняття управлінських рішень у рослинництві.

Запропонований метод формування агрокліматичних ознак і зональних навчальних вибірок дає змогу отримувати більш репрезентативні набори даних для побудови прогнозних моделей та зменшувати вплив регіональної неоднорідності за умов обмеженої довжини часових рядів урожайності. Побудовані регресійні, класифікаційні, ансамблеві та нейромережеві моделі можуть застосовуватися для прогнозування врожайності з урахуванням агрокліматичних чинників. Класифікаційні моделі прогнозування категорій «низька» / «висока» врожайність

можуть використовуватися як інструмент раннього попередження про ризик несприятливого рівня врожайності.

Практичну цінність має метод синтезу нелінійних ознак кумулятивного кліматичного впливу, який дає змогу враховувати взаємодію погодних факторів у суміжних фазах вегетації, підвищувати інформативність вхідного простору моделей і покращувати якість прогнозних оцінок.

Запропоновані підходи до оцінювання ризику низької врожайності на основі розподілів трендових залишків і квантильних оцінок можуть бути використані для обґрунтування страхових, інвестиційних, фінансових та управлінських рішень у рослинництві. Підхід до визначення сприятливої температурної траєкторії вегетації дає змогу формувати сценарні орієнтири температурних умов, асоційованих із високою врожайністю.

Розроблена інформаційно-аналітична система «Пшениця України» забезпечує автоматизацію підготовки даних, формування зональних вибірок, побудову числових і класифікаційних прогнозів урожайності, оцінювання ризиків низької врожайності, визначення сприятливих параметрів вегетації, сценарний аналіз, візуалізацію результатів і формування аналітичних звітів. Система може бути використана аграрними підприємствами, науково-дослідними установами, органами регіонального управління, страховими й фінансовими організаціями, а також у навчальному процесі закладів вищої освіти.

Результати дисертаційної роботи використані у діяльності ТзОВ «Агротрейд-Рівне» (довідка про практичне використання № 85/26 від 18.05.2026 р.), де практично використані кліматичні дані і прогнози врожайності пшениці для Рівненської та Хмельницької областей, а також у діяльності ТзОВ «Метиз капітал» (довідка про впровадження № 147 від 11.05.2026 р.), де результати використано для аналітичної оцінки впливу кліматичних ризиків та зовнішніх шоків на зниження врожайності пшениці. Результати наукових досліджень також використано при виконанні науково-дослідної роботи Національного університету водного господарства та природокористування (довідка № 16н-26 від 17.04.2026).

Результати роботи впроваджено у навчальний процес Національного університету водного господарства та природокористування (довідка № 011-11 від 10.04.2026).

Особистий внесок здобувача. Дисертаційна робота є самостійно виконаною науковою працею. Основні наукові положення, методичні підходи, експериментальні результати та висновки, що виносяться на захист, отримані здобувачем особисто. У працях, опублікованих у співавторстві, здобувачеві належать постановка і формалізація задач дослідження, формування інформаційної бази, попередня обробка врожайнісних і агрокліматичних даних, побудова матриці ознак для критичного періоду вегетації, реалізація обчислювальних експериментів, аналіз та інтерпретація одержаних результатів.

У працях [1], [2], [4], [5] здобувачем виконано підготовку експериментальних вибірок, формування наборів кліматичних факторів, побудову та порівняння моделей машинного навчання для прогнозування врожайності, оцінювання точності моделей і підготовку графічних матеріалів. У праці [3] здобувач брав участь у формалізації задачі побудови сприятливої температурної траєкторії вегетаційного циклу, виконав частину числових розрахунків та аналіз модельних результатів. У праці [6] здобувачем здійснено розрахунок трендових залишків урожайності, підготовку даних для статистичного аналізу, участь в ідентифікації закону розподілу залишків та інтерпретацію результатів для задач моделювання ризиків зерновиробництва.

У працях [7]–[21] відображено апробацію окремих положень дисертаційного дослідження, зокрема результатів прогнозування врожайності, оцінювання впливу кліматичних факторів, аналізу ризиків і втрат урожаю, а також застосування методів машинного навчання в задачах аграрної економіки. У праці [22] здобувачем виконано формування вибірки для степової зони України, підготовку кліматичних ознак, проведення обчислювальних експериментів і аналіз впливу кліматичних факторів на врожайність пшениці. У праці [23] відображено прикладну програмну реалізацію результатів дослідження у вигляді комп’ютерної програми.

Апробація матеріалів дисертації. Основні положення та результати дисертаційного дослідження були апробовані на всеукраїнських і міжнародних

наукових та науково-практичних конференціях, зокрема: Форумі молодих економістів-кібернетиків «Моделювання економіки: проблеми, тенденції, досвід» (Львів, 2022, 2024, 2025); XVI Міжнародній науково-практичній конференції «Інтегровані інтелектуальні робототехнічні комплекси» (Київ, 2023); XI International Scientific-Practical Conference «Information Control Systems and Technologies» (Odesa, 2023); Міжнародній конференції з інноваційних рішень у програмній інженерії ICISSE 2023 (Івано-Франківськ, 2023); IV Міжнародній науково-практичній конференції «Комп'ютерне моделювання та програмне забезпечення інформаційних систем і технологій» (Чернівці, 2024); Всеукраїнській науково-практичній конференції «Штучний інтелект. Наука. Бізнес» (Одеса, 2024); International Scientific Conference «Advances in Science, Technology, and Society» (Oxford, United Kingdom, 2025); IX Міжнародній науково-методичній конференції «Математичні методи, моделі та інформаційні технології в економіці» (Чернівці, 2025); VIII Міжнародній науково-практичній конференції «Моделювання, керування та інформаційні технології» (Рівне, 2025).

Публікації. За темою дисертації опубліковано 22 наукові праці та отримано 1 свідоцтво про реєстрацію авторського права на комп'ютерну програму. Основні наукові результати дисертації висвітлено у 6 наукових публікаціях, з яких 4 праці опубліковано у виданнях, проіндексованих у наукометричній базі Scopus, та 2 статті — у наукових фахових виданнях України. Апробацію матеріалів дисертації засвідчено у 16 публікаціях у матеріалах всеукраїнських і міжнародних наукових та науково-практичних конференцій. Одна наукова праця та свідоцтво про реєстрацію авторського права на комп'ютерну програму додатково відображають наукові та прикладні результати дисертаційного дослідження.

Структура та обсяг дисертації. Дисертація складається зі вступу, п'яти розділів, висновків, списку використаних джерел та додатків. Загальний обсяг дисертації становить 278 сторінок, з яких 209 сторінок припадає на основний текст. У тексті дисертації наведено 43 рисунки та 38 таблиць. Список використаних джерел містить 180 найменувань. Додатки подано на 28 сторінках.

РОЗДІЛ 1. СУЧАСНІ ПІДХОДИ ДО ІНТЕЛЕКТУАЛЬНОГО ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ

1.1. Інформаційні системи та цифрові платформи в задачах аграрної прогнозної аналітики

Прогнозування врожайності сільськогосподарських культур є складною інформаційно-аналітичною задачею, яка потребує поєднання даних різної природи, методів їх попередньої обробки, алгоритмів моделювання, засобів візуалізації та інструментів підтримки прийняття рішень. У сучасних умовах така задача дедалі частіше реалізується не у вигляді окремого розрахункового модуля, а в межах інформаційних систем і цифрових платформ, які забезпечують збирання, збереження, інтеграцію, аналіз і подання аграрних даних.

Інформаційні системи аграрного призначення можна розглядати як програмно-організаційні середовища, у яких дані про ґрунтові, кліматичні, технологічні, виробничі та просторові характеристики аграрних об’єктів перетворюються на інформаційну основу для моніторингу, прогнозування й обґрунтування управлінських рішень. На відміну від традиційних облікових систем, сучасні аграрні цифрові платформи не обмежуються збереженням даних, а забезпечують їх попередню перевірку, просторову прив’язку, аналітичне опрацювання, побудову індикаторів, формування рекомендацій і, в окремих випадках, застосування методів машинного навчання [25; 26].

У задачах прогнозування врожайності інформаційні системи виконують кілька взаємопов’язаних функцій. По-перше, вони забезпечують накопичення та структурування статистичних, кліматичних, супутникових, ґрунтових і виробничих даних. По-друге, вони створюють умови для попередньої обробки інформації: перевірки повноти даних, виявлення пропусків, пошуку аномальних значень, узгодження просторових і часових масштабів. По-третє, такі системи забезпечують запуск аналітичних процедур, побудову прогнозних оцінок, формування графіків, карт, таблиць і звітних матеріалів. Тому інформаційна система в аграрній

прогнозній аналітиці виступає не лише засобом збереження даних, а й середовищем реалізації повного циклу їх інтелектуального опрацювання.

За функціональним призначенням цифрові рішення аграрного призначення поділяються на кілька груп: платформи точного землеробства, геоінформаційні та супутникові платформи, системи підтримки прийняття рішень, платформи аграрної аналітики, а також дослідницькі програмні середовища для побудови й перевірки прогнозних моделей. Такий поділ є умовним, оскільки окремі системи можуть поєднувати функції кількох груп, однак він дає змогу виокремити їх основне призначення, переваги та обмеження для задач прогнозування врожайності.

Платформи точного землеробства орієнтовані переважно на рівень поля або господарства. Вони забезпечують роботу з картами врожайності, зонами продуктивності, даними сільськогосподарської техніки, історією польових операцій, картами диференційованого внесення добрив і засобів захисту рослин. До таких рішень належать Climate FieldView, John Deere Operations Center, OneSoil Pro, EOSDA Crop Monitoring, xarvio FIELD MANAGER, CropX, Granular Insights та інші цифрові платформи, які використовуються для підтримки агротехнологічних рішень, просторового моніторингу й аграрної аналітики [27–37]. Їх перевагою є практична спрямованість, інтеграція з польовими та виробничими даними, а також можливість оперативного використання результатів у виробничому процесі. Водночас такі платформи переважно орієнтовані на операційне управління господарством, тоді як прогнозування врожайності на регіональному рівні потребує роботи з агрегованими статистичними, кліматичними та просторовими даними.

Окрему групу становлять геоінформаційні й супутникові платформи аграрного моніторингу. Вони використовують просторові шари, межі полів, індекси рослинності, погодні дані та картографічні сервіси для оцінювання стану посівів, аналізу неоднорідності рослинного покриву, виявлення проблемних ділянок і просторового подання аграрних показників. Прикладом такого рішення є EOSDA Crop Monitoring, що поєднує супутниковий моніторинг, погодну аналітику, індекси рослинності та інструменти оцінювання стану посівів [30]. Сильним боком таких платформ є просторове охоплення й наочність результатів. Разом із тим їх

використання для прогнозування врожайності потребує додаткового узгодження супутникових і кліматичних даних із фактичними статистичними рядами, вибору релевантних часових інтервалів, маскування культур, просторової агрегації та перевірки зв’язку між дистанційними індикаторами й цільовим показником.

Системи підтримки прийняття рішень у рослинництві спрямовані на перетворення вхідних даних і результатів моделей у практичні рекомендації. Вони можуть використовувати кліматичні, ґрунтові, географічні, технологічні та виробничі показники для вибору культури, обґрунтування агротехнологічних заходів, формування рекомендацій або попередження про ризикові стани. У таких системах застосовуються алгоритми класифікації, регресії, експертні правила та методи машинного навчання [38–40]. Перевагою цього напряму є прикладна спрямованість і зрозумілість результатів для користувача [41]. Водночас значна частина таких рішень орієнтована на рекомендацію окремої управлінської дії, а не на повний аналітичний цикл прогнозування врожайності, порівняння моделей, оцінювання ризику та документування результатів.

Платформи аграрної аналітики утворюють ширший клас програмних рішень, у яких поєднуються бази даних, аналітичні модулі, карти, графіки, звітність і користувацькі сценарії [25; 42; 43]. На відміну від окремих прикладних інструментів, такі платформи орієнтовані на інтеграцію даних і моделей у межах єдиного середовища. Саме цей підхід є важливим для прогнозування врожайності, оскільки прогнозна оцінка має спиратися не лише на результат роботи моделі, а й на попередні етапи: формування вибірки, перевірку якості даних, побудову ознак, вибір алгоритму, валідацію, інтерпретацію результатів і подання їх у зручній формі.

Дослідницькі програмні рішення для прогнозування врожайності зосереджені на побудові, перевірці та порівнянні моделей. У таких системах можуть використовуватися статистичні моделі, методи машинного навчання, ансамблеві алгоритми, нейронні мережі та гібридні підходи. Їх перевагою є безпосередня орієнтація на прогнозну задачу. Водночас у багатьох випадках вони реалізують окремий алгоритмічний контур і не завжди охоплюють повний процес

роботи з даними, ризиковими оцінками, сценаріями, просторовою інтерпретацією та звітністю.

Окремим обмеженням є те, що прогнозні модулі більшості платформ зосереджені переважно на отриманні числової оцінки врожайності або моніторингу поточного стану посівів. Водночас для аграрної прогнозної аналітики важливими є також класифікація ризикових станів, оцінювання ймовірності низької врожайності, аналіз несприятливих відхилень від тренду, сценарне дослідження зміни кліматичних параметрів і просторове подання результатів. Недостатньо розвиненим залишається також зв’язок між підготовкою даних, формуванням ознак, навчанням моделей, оцінюванням їхньої якості, візуалізацією та формуванням звітних матеріалів.

Зазначені обмеження визначають потребу в розвитку підходів, які поєднують агрокліматичні дані, методи машинного навчання, процедури оцінювання ризику та засоби програмної реалізації в єдиному аналітичному контурі [44]. Такий підхід має забезпечувати не лише побудову прогнозної моделі, а й повний цикл роботи з даними: їх підготовку, перевірку, формування ознакового простору, навчання та порівняння моделей, інтерпретацію результатів, сценарний аналіз і подання отриманих оцінок у зручній для користувача формі.

Після узагальнення основних типів інформаційних систем розглянемо окремі приклади програмних платформ і дослідницьких рішень, які використовуються для моніторингу, аналізу, прогнозування або підтримки прийняття рішень в аграрній сфері. До таких рішень належать комерційні платформи точного землеробства, системи керування польовими операціями, супутникові платформи моніторингу, агрономічні платформи підтримки рішень та дослідницькі системи прогнозування врожайності.

Climate FieldView орієнтована на збирання й аналіз польових даних, роботу з картами врожайності, порівняння шарів даних і підтримку виробничих рішень на рівні поля [27]. John Deere Operations Center забезпечує керування польовими операціями, доступ до даних господарства, моніторинг техніки, планування та аналіз виробничих показників [28]. OneSoil Pro використовується для аналізу зон

продуктивності поля, роботи з історичними даними, побудови карт диференційованого внесення та просторового аналізу посівів [29]. EOSDA Crop Monitoring поєднує супутниковий моніторинг, погодну аналітику, індекси рослинності, оцінювання стану посівів і прогнозні функції [30].

Xarvio FIELD MANAGER використовується для цифрової підтримки рослинництва, зокрема для роботи з картами змінного внесення, оцінювання стану полів і планування агротехнологічних дій [31]. CropX є агрономічною платформою керування господарством, яка агрегує дані від ґрунтових сенсорів, супутників, техніки та інших джерел для формування аналітичних висновків і рекомендацій [32]. Granular Insights орієнтована на планування, аналіз польових рішень, роботу з картографічними шарами, оцінювання продуктивності та формування звітів [33].

Окремо слід виділити дослідницькі рішення, безпосередньо пов’язані з прогнозуванням врожайності. AgriYieldAI подається як AI-орієнтована система підтримки прийняття рішень для прогнозування врожайності на основі мультимодальних даних і методів глибинного навчання [34]. Agricultural Decision System, запропонована R. Aworka та співавторами, використовує кліматичні, виробничі та інші вхідні дані для прогнозування врожайності із застосуванням моделей машинного навчання [35].

Для оглядового аналізу важливо зіставити такі рішення не лише за назвою чи сферою використання, а й за функціональними можливостями, які мають значення для прогнозної аналітики: роботою з польовими, кліматичними та просторовими даними, підтримкою моделей машинного навчання, можливістю аналізу ризикових станів, сценарного дослідження та формування звітних матеріалів.

Порівняльний аналіз окремих інформаційних систем і цифрових платформ аграрної аналітики наведено в табл. 1.1. Позначення в таблиці: «+» — функція реалізована або є однією з основних для системи; «±» — функція реалізована частково або не є основною; «–» — функція не заявлена у відкритому описі системи або не є характерною для відповідного рішення.

Таблиця 1.1

Порівняльний аналіз інформаційних систем і цифрових платформ аграрної аналітики

Функціональні властивості / системи	**Climate FieldView**	**John Deere Operations Center**	**OneSoil Pro**	**EOSDA Crop Monitoring**	**xarvio FIELD MANAGER**	**CropX**	**Granular Insights**	**AgriYieldAI**	**Agricultural Decision System**
Аналіз польових даних і карт урожайності	+	+	±	±	±	±	+	–	–
Використання супутникових або просторових даних	±	±	+	+	+	+	+	+	–
Використання погодних або кліматичних даних	±	±	±	+	+	+	+	+	+
Прогнозування врожайності	±	±	±	+	±	±	±	+	+
Підтримка моделей машинного навчання	–	–	–	+	±	+	±	+	+
Робота з регіонально агрегованими даними	–	–	–	±	–	–	–	+	+
Аналіз ризиків або несприятливих станів	–	±	–	+	+	+	±	±	±
Сценарний аналіз зміни вхідних параметрів	–	–	±	±	±	±	±	±	±
Збереження дослідницьких конфігурацій аналізу	±	±	±	±	±	±	±	±	±
Формування звітів або експорт результатів	+	+	+	+	±	+	+	±	±
Орієнтація на точне землеробство й рівень поля	+	+	+	+	+	+	+	±	–
Орієнтація на дослідницький контур прогнозування	–	–	–	±	–	–	–	+	+

Порівняння наведених рішень свідчить, що сучасні аграрні цифрові платформи забезпечують широкий спектр функцій: від польового моніторингу й управління технікою до супутникової аналітики та прогнозних модулів. Більшість комерційних платформ має високу практичну цінність для виробничого управління, однак їх функціональність здебільшого орієнтована на рівень поля або господарства. Дослідницькі системи, зокрема AgriYieldAI та Agricultural Decision System, більше наближені до задач прогнозування врожайності, проте переважно зосереджуються на окремій моделі або алгоритмічному контурі. Для регіонального прогнозування врожайності залишається актуальним поєднання просторово узгоджених агрокліматичних даних, формування ознак, моделей машинного навчання, оцінювання ризику, сценарного аналізу, картографічного подання та звітності. Далі розглядаються методи прогнозування врожайності та способи формування агрокліматичного ознакового простору, які є основою для побудови інтелектуальних моделей.

1.2. Теоретичні засади прогнозування врожайності та постановки прогнозних задач

Врожайність сільськогосподарських культур є одним із ключових узагальнених показників результативності рослинництва, який відображає кількість одержаної продукції з одиниці площі. У статистичному обліку вона пов’язується з показниками посівних або зібраних площ, валових зборів і середньої продуктивності культури, що забезпечує можливість порівняння результатів у просторі та часі [45]. Для зернових культур врожайність традиційно подається у центнерах з гектара, що дає змогу використовувати її як кількісний цільовий показник у задачах аналізу, моделювання та прогнозування.

Залежно від етапу виробничого циклу та призначення оцінки розрізняють кілька типів врожайності. Потенційна врожайність характеризує максимально можливий рівень продуктивності культури за сприятливих умов і повної реалізації її біологічного потенціалу. Планова врожайність визначається до початку або на ранніх етапах виробничого циклу з урахуванням очікуваних умов вирощування.

Очікувана врожайність уточнюється протягом вегетаційного періоду на основі фактичного стану посівів, погодних умов і поточної агротехнологічної ситуації. Фактична врожайність визначається після збирання врожаю та є підсумковим статистичним показником, який найчастіше використовується для ретроспективного аналізу й побудови прогнозних моделей [45]. У подальшому під урожайністю розуміється фактична врожайність, подана у вигляді регіонального статистичного показника.

У регіональній статистиці врожайність пшениці подається як усереднений показник для відповідної адміністративної території. Такий показник не відображає врожайність окремого сорту, поля або господарства, а є агрегованою характеристикою результативності виробництва культури в межах області. Це означає, що модель працює не з локальними агротехнологічними умовами, а з регіональним статистичним індикатором, який узагальнює вплив кліматичних, ґрунтових, сортових, технологічних і організаційно-економічних чинників. У структурі виробництва пшениці в Україні переважає озима пшениця, що зумовлено її агробіологічними особливостями та поширеністю у виробничій практиці. Водночас культура пшениці представлена значною кількістю сортів, які відрізняються продуктивністю, стійкістю до стресових умов, адаптивністю до регіональних агрокліматичних зон і вимогами до технології вирощування. Оскільки сортовий склад постійно оновлюється, а статистичні дані врожайності на рівні області не деталізують внесок окремих сортів, у дисертації пшениця розглядається як агрегований об’єкт регіонального прогнозування.

Прогнозування врожайності має важливе значення для планування аграрного виробництва, оцінювання продовольчих балансів, формування експортної політики, управління запасами, страхування аграрних ризиків і підтримки управлінських рішень. Для України така задача є особливо актуальною, оскільки зернове виробництво займає важливе місце у структурі аграрного сектору, а пшениця є однією зі стратегічних культур. Водночас урожайність пшениці істотно залежить від погодно-кліматичних умов, зокрема температурного режиму, кількості й розподілу опадів, частоти посушливих періодів і проявів екстремальних погодних

явищ [46–51]. Зміни клімату посилюють невизначеність аграрного виробництва, що підвищує вимоги до якості прогнозних моделей і до способів представлення вхідних даних [52–55].

Урожайність як об’єкт прогнозування має складну природу, оскільки формується під впливом багатьох груп чинників. До природно-кліматичних чинників належать температура повітря, опади, вологозабезпечення, тривалість посушливих періодів, екстремальні температури, радіаційний режим і просторові особливості агрокліматичних умов. До агротехнологічних чинників належать сортовий склад, строки сівби, система удобрення, захист рослин, обробіток ґрунту та рівень технологічної культури виробництва. До економічних і організаційних чинників належать забезпеченість ресурсами, структура господарств, доступність техніки, фінансові умови та зовнішні шоки. Така багатофакторність ускладнює побудову універсальної моделі й потребує чіткого визначення, які саме чинники включаються до прогнозної постановки [56–60].

У загальному вигляді прогнозування врожайності може розглядатися як задача оцінювання майбутнього або умовно майбутнього значення цільової змінної на основі доступних історичних і поточних даних. Залежно від складу вхідної інформації, типу цільової змінної та очікуваного результату можна виокремити кілька основних постановок такої задачі: регресійну, класифікаційну, часово-рядову, екзогенну, ризикову та сценарно-оптимізаційну.

Основні постановки задач прогнозування врожайності можна узагальнити як сукупність напрямів, що відображають різні аспекти прогнозної задачі та відрізняються типом цільової змінної, складом вхідних даних, способом подання результату й аналітичною метою. Їх співвідношення наведено на рис. 1.1.

Прогнозування врожайності не зводиться лише до отримання числової оцінки майбутнього або умовно майбутнього значення. Регресійна, класифікаційна, часово-рядова, екзогенна, ризикова та сценарно-оптимізаційна постановки відображають різні аспекти однієї прогнозної задачі й можуть взаємно доповнювати одна одну. Такий підхід дає змогу поєднати кількісне прогнозування, ідентифікацію несприятливих станів, аналіз часової динаміки, використання

агрокліматичних ознак, оцінювання ризику та дослідження можливих сценаріїв зміни умов вегетації.

Рис. 1.1. Основні постановки задач прогнозування врожайності та їх взаємозв'язок

Регресійна постановка передбачає прогнозування числового значення врожайності або її відхилення від очікуваного рівня. У такому разі модель встановлює залежність між цільовою змінною та набором пояснювальних ознак, наприклад температурних, опадових, супутникових, ґрунтових або виробничих показників. Перевагою регресійної постановки є безпосереднє отримання кількісної прогнозної оцінки, яку можна порівнювати з фактичними значеннями. Обмеженням є те, що точний числовий прогноз може бути нестійким за умов високої міжрічної мінливості, коротких часових рядів і неповноти інформації про всі чинники, що впливають на врожайність [56; 57; 61–63].

Класифікаційна постановка передбачає перехід від прогнозування точного числового значення до визначення категорії врожайності. Наприклад, роки або регіони можуть бути віднесені до класів «низька» / «висока» врожайність, «ризиковий» / «неризиковий» стан або до кількох градацій продуктивності. Такий підхід використовується в ситуаціях, коли для прийняття рішень важливо своєчасно виявити несприятливий стан, навіть якщо точне значення врожайності має певну похибку. Класифікаційна постановка добре узгоджується із задачами раннього

попередження, страхового аналізу та підтримки адаптивних управлінських рішень. Водночас її результат залежить від способу визначення порогів, балансу класів, вибору метрик якості та інформативності вхідних ознак [61; 62; 64–66].

Ендогенна постановка ґрунтується на аналізі динаміки врожайності в часі. У такому підході модель використовує попередні значення ряду для прогнозування наступного значення або класу. Ця логіка відповідає ендогенному підходу, за якого основним джерелом інформації є сам часовий ряд урожайності або його перетворення. До такого класу належать класичні моделі часових рядів, авторегресійні моделі, авторегресійні інтегровані моделі ковзного середнього (Autoregressive Integrated Moving Average, ARIMA), а також рекурентні нейронні мережі, зокрема Gated Recurrent Unit (GRU) та Long Short-Term Memory (LSTM). Перевагою ендогенного підходу є те, що він може бути реалізований навіть за обмеженої кількості зовнішніх змінних. Проте його ефективність істотно залежить від довжини ряду, наявності автокореляції, стабільності структури даних і відсутності різких зовнішніх збурень [67].

Екзогенний підхід, на відміну від ендогенного, передбачає використання зовнішніх пояснювальних змінних, які характеризують умови формування врожайності. У задачах аграрного прогнозування такими змінними можуть бути температурні показники, опади, індекси вегетації, характеристики ґрунтів, агротехнологічні параметри, економічні індикатори або їх поєднання. Саме екзогенна постановка дає змогу враховувати вплив погодно-кліматичних умов на врожайність, виявляти чутливі фази вегетації та формувати інтерпретовані або нелінійні зв’язки між ознаками й цільовою змінною. Разом із тим вона потребує якісного формування ознакового простору, узгодження даних за просторовим і часовим масштабом, а також контролю надлишковості та мультиколінеарності предикторів [67–71].

Ризикова постановка задачі прогнозування врожайності орієнтована не лише на визначення очікуваного значення, а й на оцінювання ймовірності несприятливого результату. У такому підході увага зосереджується на низьковрожайних роках, від’ємних відхиленнях від очікуваного рівня, нижньому

хвості розподілу та можливих втратах. Ризиковий підхід особливо важливий для страхування, фінансового планування, оцінювання продовольчої безпеки та розроблення адаптаційних заходів. Його використання потребує не тільки побудови прогнозної моделі, а й аналізу розподілу помилок або залишків, вибору статистичних мір ризику та інтерпретації отриманих оцінок [72–76].

Сценарно-оптимізаційна постановка пов’язана з аналізом того, як зміна окремих вхідних параметрів або їх траєкторій впливає на прогнозований рівень урожайності. У такому випадку модель використовується не лише для передбачення, а й для дослідження умов, за яких формується сприятливий або несприятливий результат. Для агрокліматичних задач це може означати аналіз температурних і опадових траєкторій у критичні періоди вегетації, оцінювання кумулятивного впливу погодних умов або пошук таких комбінацій ознак, які асоціюються з високими позитивними відхиленнями врожайності. Така постановка розширює прогнозування до рівня аналітичного дослідження можливих станів системи [57; 58; 77].

Окремого розгляду у прогнозуванні врожайності потребує вибір цільової змінної. Абсолютне значення врожайності містить як довгострокову тенденцію, так і короткострокові міжрічні коливання. Довгострокова тенденція може бути пов’язана з технологічним розвитком, сортовим оновленням, змінами агротехніки, економічними умовами та загальним рівнем організації виробництва. Натомість короткострокові відхилення значною мірою відображають вплив погодних і кліматичних умов конкретного року. Тому аналіз кліматичного впливу потребує розділення трендової та залишкової складових врожайності [78].

Виділення тренду дає змогу перейти від аналізу абсолютних значень до аналізу відхилень від очікуваного рівня. Такі відхилення можуть інтерпретуватися як трендові залишки або детрендована врожайність. Їх використання має кілька переваг. По-перше, зменшується вплив довгострокових технологічних і організаційно-економічних змін. По-друге, посилюється зв’язок цільової змінної з міжрічною мінливістю агрокліматичних умов. По-третє, з’являється можливість аналізувати не тільки середній рівень урожайності, а й несприятливі відхилення,

що є важливим для оцінювання ризику. Водночас вибір способу детрендування має бути обґрунтованим, оскільки надмірно складна форма тренду може вилучити частину корисної інформації, а надто проста — залишити в даних невраховану систематичну складову.

На якість прогнозної моделі впливає також горизонт прогнозування. Ретроспективне моделювання використовується для пояснення зв’язків між урожайністю та чинниками впливу на основі історичних даних. Поточне або передзбиральне прогнозування орієнтоване на оцінювання врожайності в межах поточного вегетаційного сезону з використанням уже доступних кліматичних або дистанційних показників. Раннє прогнозування передбачає отримання оцінки до завершення вегетаційного періоду, коли частина факторів ще невідома. Чим ранішим є прогноз, тим вищою є його практична цінність, але водночас зростає невизначеність і вимоги до стійкості моделі [56; 61–63; 79–82].

Окрему складність становить просторовий рівень прогнозування. На рівні поля модель може використовувати деталізовані дані про ґрунт, рельєф, агротехнологічні операції, супутникові індекси й локальні погодні умови. На рівні області або іншої адміністративної одиниці прогноз будується за агрегованими статистичними та кліматичними показниками [83]. Такий рівень є важливим для регіонального аналізу, державної статистики, продовольчих балансів і управлінських рішень, однак він супроводжується втратою частини локальної інформації. Тому для регіонального прогнозування особливо важливими є коректна просторова агрегація даних, формування однорідних навчальних вибірок і врахування агрокліматичної неоднорідності територій.

Проблема коротких часових рядів є однією з ключових у регіональному прогнозуванні врожайності. Для окремої області кількість історичних спостережень часто є недостатньою для надійного навчання складних моделей машинного або глибинного навчання. У такій ситуації просте збільшення складності моделі не гарантує підвищення точності, а може призвести до перенавчання. Одним із шляхів подолання цієї проблеми є формування зональних або групових навчальних вибірок на основі подібності регіонів за

агрокліматичними умовами або динамікою трендових залишків. Такий підхід дає змогу збільшити обсяг навчальних даних і водночас зберегти змістовну однорідність вибірки.

У прогнозній постановці потрібно розрізняти пояснювальні й прогностичні задачі. Пояснювальна задача спрямована на встановлення зв'язків між урожайністю та факторами впливу, визначення напряму й сили дії окремих ознак, інтерпретацію важливості предикторів. Прогностична задача орієнтована на отримання максимально стійкої оцінки для нових спостережень. У першому випадку особливо важливими є інтерпретованість і статистична обґрунтованість моделі, у другому — узагальнювальна здатність, контроль перенавчання та коректна схема валідації. Для аграрних задач поєднання цих двох логік підвищує практичну цінність моделі, оскільки результат має бути не лише точним, а й змістовно пояснюваним.

Отже, прогнозування врожайності може бути сформульоване як сукупність взаємопов'язаних задач: числового прогнозування, класифікації рівня врожайності, аналізу часових рядів, екзогенного моделювання за агрокліматичними ознаками, оцінювання ризику та сценарного дослідження умов вегетації. Така багатоваріантність постановок зумовлює необхідність подальшого аналізу даних, ознакового представлення та методів машинного навчання, придатних для роботи з регіонально агрегованими агрокліматичними даними й короткими часовими рядами врожайності.

1.3. Дані та ознакове представлення в задачах прогнозування врожайності

Ефективність моделей прогнозування врожайності значною мірою визначається не лише вибором алгоритму, а й якістю вхідних даних, способом їх попередньої обробки та формою подання ознак. Навіть складні моделі машинного або глибинного навчання не можуть забезпечити стійкий прогноз за умов неповних, неоднорідних, просторово неузгоджених або слабо інформативних даних. Тому формування ознакового простору є одним із ключових етапів інтелектуального прогнозування врожайності.

У задачах аграрної прогнозної аналітики використовуються різні групи даних: статистичні ряди врожайності, кліматичні та метеорологічні показники, супутникові індекси рослинності, ґрунтові характеристики, агротехнологічні параметри, економічні індикатори та просторові дані. Кожна з цих груп відображає окремий аспект формування врожайності [68–70; 84–93]. Статистичні дані характеризують фактичний результат виробництва; кліматичні показники описують умови вегетаційного періоду; супутникові індекси відображають стан рослинного покриву; ґрунтові дані визначають потенціал території; агротехнологічні й економічні змінні відображають рівень організації виробництва та ресурсного забезпечення [68–70; 84–93].

Найпоширенішими джерелами вхідної інформації для прогнозування врожайності є офіційна статистика, метеорологічні спостереження, реаналізні кліматичні дані, дистанційне зондування Землі та польові або виробничі дані. Офіційна статистика забезпечує узгоджені ряди врожайності за адміністративними одиницями, однак зазвичай має річну часову дискретність і не містить детальної інформації про умови формування врожаю. Метеорологічні станції надають фактичні спостереження, проте їх просторове розміщення може бути нерівномірним, а точковий характер вимірювань ускладнює зіставлення з регіональними статистичними показниками. Реаналізні кліматичні дані ERA5, тобто реаналізний кліматичний набір даних п’ятого покоління, забезпечують просторово повне і методично однорідне подання кліматичних показників, але потребують агрегації до потрібного територіального рівня [94]. Для просторового узгодження таких даних із регіональними статистичними показниками важливими є цифрові межі адміністративних одиниць, які дають змогу співвідносити кліматичні або супутникові показники з відповідним територіальним рівнем аналізу [95]. Супутникові дані мають перевагу широкого просторового охоплення, однак їх використання супроводжується потребою в попередній обробці, фільтрації, маскуванні культур і просторово-часовому узгодженні з цільовими показниками [68–70; 96–104].

Важливою особливістю прогнозування врожайності є необхідність узгодження просторового рівня даних. Якщо цільова змінна подана на рівні області, району або іншої адміністративної одиниці, то пояснювальні кліматичні та просторові ознаки також мають бути приведені до відповідного рівня. Використання однієї метеорологічної точки або географічного центру території може спрощувати процедуру розрахунку, але не завжди коректно відображає просторову неоднорідність умов вирощування. Тому в задачах регіонального прогнозування використовують просторову агрегацію, за якої кліматичні або супутникові показники узагальнюються в межах відповідних територіальних полігонів [94; 95].

Проблема просторового узгодження тісно пов’язана з проблемою масштабів. Дані про врожайність можуть бути доступні на рівні країни, області, району, громади, господарства або поля. Кліматичні дані часто мають ґраткову структуру, супутникові індекси — піксельне подання, а статистичні показники — адміністративну прив’язку. Якщо ці рівні не узгоджені між собою, модель може відображати не реальні залежності, а наслідки помилки агрегації. Тому формування ознакового представлення має включати не лише вибір предикторів, а й визначення просторового рівня, на якому поєднуються цільові та пояснювальні змінні [105–113].

Окремого значення набуває часовий рівень узгодження даних. Урожайність зазвичай подається як річний показник, тоді як погодні й кліматичні дані можуть бути годинними, добовими, декадними або місячними. Для прогнозування врожайності недостатньо використати лише середньорічні значення температури чи сумарні річні опади, оскільки вплив погодних умов істотно залежить від фази розвитку культури. Один і той самий температурний або опадовий режим може мати різні наслідки залежно від того, чи спостерігається він у період сходів, кущіння, колосіння, наливу зерна або дозрівання. Тому в агрокліматичному ознаковому просторі важливо враховувати не лише величину показника, а й час його прояву [67; 71; 99].

Формування агрокліматичних ознак передбачає перетворення первинних погодних або кліматичних даних у предиктори, які можуть бути використані моделями машинного навчання. До таких ознак належать середні, мінімальні й максимальні температури за визначені періоди, суми опадів, кількість днів із перевищенням температурних порогів, індекси посухи, гідротермічні коефіцієнти, накопичені температури, а також похідні показники, що характеризують кумулятивний або стресовий вплив погодних умов. Вибір таких ознак має спиратися на біологічну логіку формування врожаю та на здатність показників відображати умови критичних фаз вегетації [59; 67; 71; 114; 115].

У задачах прогнозування врожайності пшениці особливого значення набувають ознаки, пов’язані з весняно-літнім періодом вегетації, коли формується значна частина майбутнього врожаю. Температурний режим і кількість опадів у цей період можуть впливати на інтенсивність ростових процесів, розвиток рослин, забезпеченість вологою, стійкість до стресових умов і формування продуктивних елементів. Тому часове агрегування кліматичних даних за місяцями, декадами або іншими інтервалами дає змогу підвищити інформативність вхідного простору порівняно з використанням надто узагальнених сезонних або річних показників [49; 67; 71].

Ознакове представлення в задачах машинного навчання зазвичай має вигляд матриці, у якій кожне спостереження характеризується набором предикторів і відповідним значенням цільової змінної. Для аграрних задач таким спостереженням може бути комбінація «територія–рік», «поле–сезон», «культура–регіон–рік» або інша одиниця аналізу. Якість такої матриці визначається повнотою даних, узгодженістю просторового й часового рівнів, відсутністю дублювання, коректністю одиниць виміру, інформативністю ознак і збалансованістю цільової змінної для обраної постановки задачі.

У наукових роботах, присвячених прогнозуванню врожайності, формування ознакового простору зазвичай охоплює кілька етапів: визначення джерел даних, очищення та узгодження показників, просторову й часову агрегацію, перетворення первинних змінних у базові предиктори, а також формування похідних ознак,

здатних відображати нелінійні, порогові або кумулятивні ефекти [61–63; 70; 88; 91]. Такий підхід дає змогу перейти від різнорідних аграрних, кліматичних і просторових даних до структурованої навчальної вибірки, придатної для використання в моделях машинного навчання.

Базові ознаки не завжди достатньо повно відображають вплив агрокліматичних умов на врожайність. У реальних агробіологічних процесах дія температури й опадів часто має нелінійний, пороговий або кумулятивний характер. Наприклад, помірне підвищення температури може бути сприятливим у певній фазі розвитку, але надмірне перевищення температурного рівня може спричиняти тепловий стрес. Аналогічно опади можуть мати позитивний вплив за дефіциту вологи, однак надмірне зволоження в окремі періоди може погіршувати умови росту або збирання. Тому підвищення інформативності ознакового простору потребує врахування не лише лінійних предикторів, а й похідних ознак, які відображають квадратичні ефекти, взаємодії між показниками, накопичення впливу та послідовність погодних умов у часі [59; 67; 71; 114; 115].

Синтез нелінійних ознак може здійснюватися різними способами: шляхом додавання квадратів базових предикторів, добутків ознак, різниць між періодами, відношень температурних і опадових показників, кумулятивних індексів або порогових змінних. Такі ознаки можуть підвищувати здатність моделей описувати неадитивний вплив погодних умов, однак водночас збільшують розмірність вхідного простору. Тому їх використання потребує контролю мультиколінеарності, перевірки стійкості моделі, відбору інформативних предикторів і коректної схеми валідації.

Окремою проблемою є формування навчальних вибірок для регіонального прогнозування. Якщо модель будується окремо для кожної області, кількість спостережень часто є обмеженою, особливо за короткого історичного періоду. Це ускладнює використання складних моделей машинного та глибинного навчання, підвищує ризик перенавчання й знижує стійкість оцінок. У такій ситуації об’єднання спостережень за групами територій, подібних за агрокліматичними

умовами або за характером динаміки врожайності, дає змогу збільшити обсяг навчальної вибірки без механічного змішування різнорідних регіонів.

Формування навчальної вибірки має враховувати також тип прогнозної задачі. Для регресійного прогнозування важливо зберегти кількісну шкалу цільової змінної та забезпечити достатню варіативність значень. Для класифікаційної постановки необхідно визначити правила віднесення спостережень до класів, наприклад «низька» або «висока» врожайність, а також перевірити баланс класів. Для ризикового аналізу важливими є спостереження з несприятливими відхиленнями, оскільки саме вони формують нижній хвіст розподілу. Для сценарно-оптимізаційної постановки важливо, щоб ознаки мали змістовну інтерпретацію та могли бути змінені в межах допустимих сценаріїв.

Таким чином, ознакове представлення в задачах прогнозування врожайності є проміжною ланкою між предметною аграрною інформацією та алгоритмами машинного навчання. Воно визначає, які аспекти кліматичного, просторового, часово-динамічного та виробничого впливу будуть доступні моделі. Недостатньо інформативний або просторово неузгоджений набір ознак обмежує якість прогнозу незалежно від складності алгоритму. Натомість правильно сформований ознаковий простір створює основу для побудови регресійних, класифікаційних, нейромережевих, ризикових і сценарно-оптимізаційних моделей, що розглядаються в подальших підрозділах.

1.4. Методи машинного та глибинного навчання у прогнозуванні врожайності

Вибір методу прогнозування врожайності залежить від типу вхідних даних, розміру навчальної вибірки, просторового рівня аналізу, характеру цільової змінної та очікуваної форми результату. Для одних задач достатнім може бути інтерпретований регресійний підхід, який дає змогу оцінити напрям впливу окремих факторів. Для інших задач використовують алгоритми машинного навчання, здатні враховувати нелінійні залежності та взаємодії між ознаками. Якщо дані мають виражену часову структуру, можуть застосовуватися моделі часових

рядів і рекурентні нейронні мережі. Тому в задачах аграрної прогнозної аналітики важливим є не лише вибір однієї моделі, а й порівняння груп методів за точністю, стійкістю, інтерпретованістю та придатністю до конкретної структури даних [56; 61–63; 116–119].

Класичні статистичні та економетричні моделі становлять базовий рівень прогнозування. До них належать лінійна регресія, множинна регресія, трендові моделі, авторегресійні моделі та інші підходи, у яких цільова змінна подається як функція часу або набору пояснювальних ознак. Їх перевагою є прозорість, простота реалізації та можливість змістовної інтерпретації коефіцієнтів. У задачах прогнозування врожайності такі моделі дають змогу оцінити вплив температури, опадів або інших факторів, а також сформувати базову модель для подальшого порівняння зі складнішими алгоритмами.

Разом із тим лінійні моделі мають обмеження, пов’язані з припущенням про відносно просту форму зв’язку між предикторами та цільовою змінною. У реальних агробіологічних процесах вплив кліматичних чинників часто є нелінійним, пороговим або кумулятивним. Наприклад, підвищення температури може бути сприятливим у межах певного діапазону, але негативним після перевищення критичного рівня. Опади також можуть мати різний ефект залежно від фази вегетації та попереднього вологозабезпечення. Тому лінійні моделі є важливими для інтерпретації, але не завжди достатніми для отримання стійкого прогнозу за складної структури даних [120; 121].

Методи машинного навчання розширюють можливості прогнозування врожайності завдяки здатності працювати з багатовимірними наборами ознак, виявляти нелінійні залежності та враховувати взаємодії між предикторами. У сучасних дослідженнях аграрного прогнозування широко застосовуються дерева рішень, Random Forest, Gradient Boosting, XGBoost, Support Vector Machine, Logistic Regression, нейронні мережі та гібридні моделі [57; 61–63; 122–133]. Такі методи можуть використовувати як кліматичні й статистичні показники, так і супутникові, ґрунтові, виробничі та мультиджерельні дані.

Метод опорних векторів (Support Vector Machine, SVM) належить до методів, які можуть застосовуватися як для регресії, так і для класифікації. Його перевагою є здатність будувати розділювальну поверхню у просторі ознак і використовувати ядрові функції для опису нелінійних залежностей. У задачах прогнозування врожайності SVM може бути корисним тоді, коли кількість спостережень є відносно невеликою, але ознаки мають складну структуру. Водночас ефективність цього методу залежить від вибору ядра, параметрів регуляризації та масштабу ознак. За невдалого налаштування SVM може бути нестійким або складним для інтерпретації.

Метод випадкового лісу (Random Forest, RF) є ансамблевим методом, який поєднує результати багатьох дерев рішень [134; 135]. Його перевагами є стійкість до шуму, здатність описувати нелінійні залежності, робота з різними типами ознак і можливість оцінювання важливості предикторів. У задачах прогнозування врожайності Random Forest часто використовується для роботи з кліматичними, супутниковими та мультиджерельними даними. Обмеженням методу є те, що він не завжди забезпечує просту змістовну інтерпретацію, а за невеликої вибірки може відображати випадкові особливості навчальних даних.

Gradient Boosting і його модифікації також належать до ансамблевих методів. На відміну від Random Forest, де дерева будуються відносно незалежно, бустингові алгоритми послідовно уточнюють модель, зменшуючи помилки попередніх кроків. Такі методи можуть демонструвати високу точність у задачах прогнозування, але потребують ретельного налаштування гіперпараметрів, контролю перенавчання та коректної валідації. Для аграрних задач це особливо важливо, оскільки часові ряди врожайності часто є короткими, а кількість зовнішніх ознак може бути значною [136; 137].

Logistic Regression займає особливе місце серед методів машинного навчання, оскільки поєднує простоту, інтерпретованість і придатність до класифікаційних задач. У прогнозуванні врожайності вона може використовуватися для визначення ймовірності належності року або регіону до класу низької чи високої врожайності. Перевагою логістичної регресії є зрозуміла інтерпретація напряму впливу ознак і

можливість отримати ймовірнісну оцінку. Обмеженням є переважно лінійний характер межі класифікації, що може бути недостатнім за складних нелінійних залежностей між кліматичними предикторами та врожайністю.

Класифікаційний підхід до прогнозування врожайності є важливим доповненням до традиційного числового прогнозування. У практичних задачах не завжди необхідно визначити точне значення врожайності; часто важливіше своєчасно ідентифікувати ризик низького рівня врожайності. Перехід до категорій «низька» / «висока» дає змогу формувати моделі раннього попередження, порівнювати різні алгоритми за здатністю виявляти несприятливі стани та використовувати результати для підтримки страхових, фінансових або управлінських рішень. Водночас класифікаційна постановка потребує обґрунтованого вибору порогу, контролю балансу класів і використання відповідних метрик якості [61; 62; 64–66].

Для оцінювання класифікаційних моделей використовують матрицю помилок, $Accuracy$, $Precision$, $Recall$, $F1-score$, площу під ROC-кривою (ROC_AUC) та інші показники. $Accuracy$ характеризує загальну частку правильних класифікацій, але може бути недостатньо інформативною за дисбалансу класів. $Precision$ показує, наскільки надійними є позитивні передбачення моделі. $Recall$ відображає здатність моделі виявляти об’єкти цільового класу, що особливо важливо для низьковрожайних або ризикових років. $F1-score$ поєднує $Precision$ і $Recall$, а ROC_AUC характеризує здатність моделі розділяти класи за різних порогів прийняття рішення. Тому в аграрних ризикових задачах використовують не одну, а кілька метрик, оскільки різні показники відображають різні аспекти якості класифікації [64; 65].

Окрему групу становлять моделі глибинного навчання. Вони можуть бути ефективними для великих наборів даних, складних нелінійних залежностей, зображень, просторово-часових структур і мультиджерельних ознак [62; 68; 69; 138–150]. У задачах прогнозування врожайності глибинне навчання застосовується для аналізу супутникових знімків, часових рядів кліматичних показників, індексів рослинності та комбінованих наборів даних. До таких моделей належать

багатошарові нейронні мережі, згорткові нейронні мережі, рекурентні нейронні мережі, LSTM, GRU, attention-архітектури та гібридні моделі.

Рекурентні нейронні мережі призначені для роботи з послідовними даними, де значення поточного стану може залежати від попередніх спостережень. У задачах прогнозування врожайності така властивість є корисною для аналізу часових рядів, сезонної динаміки, послідовностей агрокліматичних показників або індексів рослинності. Проте класичні рекурентні мережі мають труднощі з довготривалими залежностями, тому в практиці частіше використовуються архітектури LSTM і GRU, які мають спеціальні механізми керування передаванням інформації.

LSTM використовує комірку пам’яті та систему воріт, що регулюють збереження, оновлення й передавання інформації між часовими кроками. Завдяки цьому модель може враховувати залежності між віддаленими елементами послідовності. GRU є компактнішою архітектурою, у якій використовується менша кількість механізмів керування, що може зменшувати кількість параметрів і спрощувати навчання. Обидві архітектури є перспективними для часових рядів, однак їх ефективність істотно залежить від довжини ряду, кількості спостережень, структури автокореляції, складу вхідних ознак і схеми валідації [62; 64; 151–156].

У задачах прогнозування врожайності використання GRU та LSTM потребує обережної інтерпретації. Якщо часовий ряд є коротким, має слабку автокореляційну структуру або містить значну частку випадкової міжрічної мінливості, рекурентна модель може не мати достатньої інформаційної основи для навчання [64; 78; 157–161]. У таких умовах збільшення складності архітектури не гарантує підвищення точності, а може призвести до перенавчання. Тому рекурентні моделі потрібно порівнювати з простішими статистичними та машинними підходами, а також перевіряти перехід від уніваріативної постановки до мультиваріативної.

Підбір гіперпараметрів є важливою складовою побудови моделей машинного та глибинного навчання. До таких параметрів належать тип ядра й коефіцієнти регуляризації для SVM, кількість дерев і глибина дерев для Random Forest, швидкість навчання та кількість ітерацій для бустингових моделей, кількість шарів,

нейронів, епох, розмір вікна та параметри регуляризації для нейронних мереж. Некоректний вибір гіперпараметрів може істотно погіршити якість прогнозу або створити ілюзію високої точності на навчальних даних. Тому налаштування моделей має виконуватися разом із процедурою валідації.

Валідація моделей у задачах прогнозування врожайності повинна враховувати структуру даних. Для незалежних табличних спостережень можуть застосовуватися навчальна, валідаційна та тестова вибірки, а також k-fold cross-validation. Для часових рядів важливо уникати витоку інформації з майбутніх періодів у минулі, тому випадкове перемішування спостережень не завжди є коректним. Якщо модель має використовуватися для прогнозування майбутніх років, схема перевірки повинна зберігати часовий порядок. Для регіональних або зональних вибірок додатково важливо контролювати, щоб результати не відображали лише особливості окремої області або групи спостережень [162–164].

Для оцінювання ефективності регресійних моделей найчастіше використовують такі метрики: MAE, MSE, $RMSE$, $MAPE$ та коефіцієнт детермінації R^2. MAE характеризує середню абсолютну помилку, $RMSE$ сильніше штрафує великі відхилення, $MAPE$ подає помилку у відсотковій формі, а R^2 відображає частку варіації цільової змінної, пояснену моделлю. Водночас жодна з цих метрик не є універсальною. Наприклад, високе R^2 не завжди означає придатність моделі для прогнозування низьковрожайних років, а низька середня помилка може приховувати великі помилки в критичних ризикових ситуаціях. Тому оцінювання якості моделей має відповідати постановці задачі.

З огляду на це порівняння методів прогнозування врожайності не може обмежуватися лише формальним типом алгоритму або окремим показником точності. Одна й та сама модель може демонструвати різну ефективність залежно від обсягу вибірки, кількості предикторів, наявності часової структури, просторової неоднорідності даних, балансу класів і характеру цільової змінної. Для регіональних аграрних задач особливо важливими є здатність моделі працювати з короткими часовими рядами, стійкість до перенавчання, можливість використання агрокліматичних ознак, придатність до класифікації несприятливих станів і

потенціал для подальшого сценарного або ризикового аналізу. Саме тому вибір методу має спиратися не лише на очікувану точність, а й на відповідність моделі структурі даних і меті прогнозування.

Узагальнення відмінностей між основними групами методів охоплює не лише типові алгоритми, а й функціональні властивості, важливі для прогнозування врожайності: здатність працювати з малими вибірками, інтерпретованість, урахування нелінійних залежностей, придатність для регресійної, класифікаційної, ендогенної та екзогенної постановок, а також можливість використання в ризиковому, сценарному або оптимізаційному аналізі. Відповідне порівняння наведено в табл. 1.2. Позначення в таблиці: «+» — властивість характерна для відповідної групи методів; «±» — властивість реалізується частково або залежить від конкретної моделі, даних і налаштувань; «–» — властивість не є типовою для відповідної групи методів.

Наведене порівняння показує, що жодна з груп методів не є універсальною для всіх постановок задачі прогнозування врожайності. Статистичні та економетричні моделі є придатними як базовий інтерпретований рівень аналізу, однак мають обмеження щодо врахування нелінійних і кумулятивних ефектів. Методи машинного навчання, зокрема SVM, Random Forest і Gradient Boosting, краще пристосовані до роботи з багатовимірними агрокліматичними ознаками та нелінійними залежностями, але потребують налаштування параметрів і контролю перенавчання. Рекурентні нейронні мережі є перспективними для часових рядів і послідовностей, проте їх застосування за коротких регіональних рядів потребує окремої перевірки. Гібридні та мультиджерельні моделі мають найбільший потенціал для інтеграції різних джерел даних, але водночас потребують складнішої підготовки ознакового простору та програмної реалізації.

Аналіз методів машинного та глибинного навчання свідчить, що вибір моделі має визначатися не лише її потенційною точністю, а й відповідністю структурі даних. Для коротких регіональних часових рядів важливо враховувати ризик перенавчання, обмеженість вибірки та можливу слабкість автокореляційної структури.

Таблиця 1.2

Порівняльний аналіз груп методів прогнозування врожайності за функціональними властивостями

Функціональні властивості / групи методів	**Статистичні та економетричні моделі**	**SVM / SVR**	**Logistic Regression**	**Random Forest**	**Gradient Boosting / XGBoost**	**RNN / LSTM / GRU**	**Гібридні та мультиджерельні моделі**
Робота з малими вибірками	+	+	+	±	±	–	±
Інтерпретованість результатів	+	±	+	±	±	–	±
Урахування нелінійних залежностей	–	+	–	+	+	+	+
Урахування взаємодії між ознаками	±	+	–	+	+	+	+
Придатність для регресійного прогнозування	+	+	–	+	+	+	+
Придатність для класифікаційної постановки	±	+	+	+	+	+	+
Придатність для ендогенного прогнозування часових рядів	±	±	–	±	±	+	+
Придатність для екзогенного прогнозування за агрокліматичними ознаками	+	+	±	+	+	+	+
Робота з мультиджерельними даними	±	±	±	+	+	+	+
Оцінювання важливості ознак	±	–	±	+	+	–	±
Чутливість до масштабування ознак	±	+	+	±	±	+	+
Потреба в підборі гіперпараметрів	±	+	±	+	+	+	+
Ризик перенавчання за коротких рядів	±	±	±	±	+	+	+
Придатність для виявлення низьковрожайних / ризикових років	±	+	+	+	+	+	+
Придатність для сценарного або оптимізаційного аналізу	±	±	–	±	+	±	+

Для екзогенних агрокліматичних моделей визначальним є якість ознакового простору, коректність просторово-часової агрегації та здатність моделі враховувати нелінійний вплив кліматичних чинників. Для класифікаційних задач особливе значення мають баланс класів і здатність моделі виявляти саме низьковрожайні роки.

Порівняльне використання моделей різної складності – від інтерпретованих статистичних і логістичних моделей до ансамблевих алгоритмів та рекурентних нейронних мереж — забезпечує ширше оцінювання їх придатності для задач прогнозування врожайності. Такий підхід дає змогу не тільки вибрати модель із кращими метриками, а й встановити межі застосовності окремих класів алгоритмів для коротких часових рядів, регіонально агрегованих агрокліматичних даних і класифікаційного прогнозування низького рівня врожайності.

1.5. Оцінювання ризиків, сценарний аналіз та оптимізаційні підходи в аграрному прогнозуванні

Прогнозування врожайності не обмежується визначенням очікуваного числового значення. Для аграрного виробництва важливими є також імовірність несприятливого результату, величина можливого відхилення від очікуваного рівня, чутливість урожайності до зміни кліматичних параметрів і можливість аналізу альтернативних сценаріїв. За таких умов прогнозна модель має виконувати не лише функцію передбачення, а й функцію аналітичної підтримки рішень, пов’язаних з управлінням ризиками, адаптацією до кліматичної мінливості та оцінюванням можливих наслідків зміни умов вегетації.

У межах такого аналізу поєднуються квантильні міри ризику, сценарне моделювання й оптимізаційні алгоритми: для кількісного опису несприятливих відхилень використовуються VaR і CVaR [72; 73], для пошуку сприятливих комбінацій вхідних ознак можуть застосовуватися евристичні методи глобальної оптимізації, зокрема Differential Evolution [165; 166], а загальне трактування ризику пов’язується з імовірністю, невизначеністю та величиною можливих втрат [167].

Ризик низької врожайності можна розглядати як імовірність або очікувану величину несприятливого відхилення фактичного результату від певного базового, трендового або нормативного рівня. Такий підхід відрізняється від звичайного прогнозування тим, що увага зосереджується не на середньому значенні, а на нижній частині розподілу, де розміщуються роки з найгіршими результатами. Для аграрного сектору це має важливе значення, оскільки саме низьковрожайні роки створюють найбільші проблеми для виробників, регіонального планування, страхування, продовольчої безпеки та фінансової стабільності.

Урожайність як часовий ряд містить довгострокову та короткострокову складові. Довгострокова складова може бути пов’язана з технологічним розвитком, сортовим оновленням, зміною агротехніки, економічними умовами та організацією виробництва. Короткострокова міжрічна мінливість значною мірою відображає вплив погодних і кліматичних умов конкретного року. Тому ризиковий аналіз має спиратися не лише на абсолютні значення врожайності, а й на трендові залишки, які показують відхилення фактичної врожайності від очікуваного трендового рівня.

Трендові залишки є зручним об’єктом для оцінювання ризику, оскільки дають змогу відокремити несприятливу міжрічну мінливість від загальної тенденції зміни врожайності. Від’ємні залишки можуть інтерпретуватися як роки, у яких фактичний результат був нижчим за очікуваний рівень. Саме такі спостереження формують основу для аналізу ризику низької врожайності. Позитивні залишки, навпаки, характеризують роки зі сприятливими відхиленнями. Такий поділ є важливим для побудови як ризикових моделей, так і сценарно-оптимізаційного аналізу умов, пов’язаних із високою або низькою продуктивністю.

Оцінювання ризику низької врожайності може здійснюватися різними способами. Найпростішим підходом є аналіз частоти низьковрожайних років або частки спостережень, у яких урожайність була нижчою за певний поріг. Такий підхід є зрозумілим, але залежить від вибору порогу й не завжди враховує величину можливих втрат. Іншим підходом є використання статистичних характеристик розподілу залишків: середнього значення, стандартного відхилення, асиметрії, ексцесу та параметрів апроксимуючих розподілів. Проте для ризикових задач

особливо важливо аналізувати не центральну частину розподілу, а його нижній хвіст.

Для кількісного оцінювання нижньохвостових ризиків можуть використовуватися квантильні міри ризику, зокрема Value-at-Risk та Conditional Value-at-Risk. Value-at-Risk характеризує такий рівень несприятливого відхилення, який не буде перевищено з заданою ймовірністю за припущенням про відповідний розподіл або емпіричну структуру даних. Conditional Value-at-Risk, на відміну від VaR, оцінює середню величину втрат у найгіршій частині розподілу, тобто за межами відповідного квантиля [72; 73]. Тому CVaR є більш інформативним для аналізу екстремально несприятливих років, коли важливим є не лише факт потрапляння в ризикову зону, а й очікувана глибина відхилення.

У задачах прогнозування врожайності використання VaR і CVaR потребує попереднього аналізу розподілу трендових залишків. Якщо залишки мають приблизно нормальний розподіл, ризикові оцінки можуть бути отримані на основі параметричної апроксимації. Якщо розподіл має асиметрію, важкі хвости або істотні відхилення від нормальності, альтернативні розподіли або емпіричні квантилі забезпечують гнучкішу оцінку нижньохвостових ризиків [72; 73]. Вибір розподілу має важливе значення, оскільки саме нижній хвіст визначає оцінку ризику низької врожайності. Недооцінювання важкості хвоста може призвести до заниження ризику, тоді як надмірно консервативна апроксимація може завищити очікувані втрати.

Ризиковий аналіз тісно пов’язаний із класифікаційною постановкою прогнозування. Якщо класифікаційна модель дає змогу віднести спостереження до класу «низька» або «висока» врожайність, то ризикові міри дають кількісну характеристику несприятливого відхилення. У такому разі класифікація відповідає на питання про наявність ризикового стану, а VaR і CVaR — про можливу глибину цього ризику. Поєднання цих підходів підсилює аграрну аналітику, оскільки дозволяє одночасно виявляти низьковрожайні роки та оцінювати можливу величину негативного відхилення.

Окрім кліматичної мінливості, аграрне виробництво може зазнавати впливу зовнішніх шоків, зокрема воєнних, логістичних або інституційних порушень, що посилює потребу в поєднанні прогнозних моделей із ризиковими оцінками та сценарним аналізом [57; 58].

Крім ризикового аналізу, важливою складовою прогнозної аналітики є сценарне дослідження. На відміну від звичайного прогнозування, яке спрямоване на оцінювання очікуваного результату за наявними вхідними даними, сценарний аналіз дає змогу дослідити, як зміна окремих параметрів або їх комбінацій може вплинути на прогнозований рівень урожайності. У задачах агрокліматичного моделювання сценаріями можуть бути зміни температури в окремі періоди вегетації, зміни кількості опадів, посилення посушливості, зміщення сприятливих температурних інтервалів або поєднання кількох погодних факторів [57; 58].

Сценарний аналіз має особливе значення у випадках, коли модель використовується не лише для передбачення, а й для дослідження чутливості врожайності до зміни умов. Наприклад, якщо модель показує, що окремі температурні інтервали мають суттєвий вплив на відхилення врожайності від тренду, можна дослідити, які зміни температурної траєкторії асоціюються зі сприятливими або несприятливими результатами. Такий підхід не слід трактувати як пряме керування погодою; його призначення полягає в модельному виявленні умов, пов’язаних із підвищеним або зниженим рівнем урожайності.

У сценарному аналізі ключову роль відіграє поняття кліматичної траєкторії вегетаційного періоду. Під такою траєкторією можна розуміти впорядкований у часі набір температурних, опадових або інших агрокліматичних показників, що характеризують умови формування врожайності протягом визначених фаз розвитку культури. На відміну від одиничного середнього показника, траєкторія відображає послідовність умов у часі. Це важливо, оскільки врожайність може залежати не лише від середньої температури або сумарних опадів за сезон, а й від того, у які саме періоди спостерігалися сприятливі чи стресові умови.

Оптимізаційний підхід у прогнозуванні врожайності полягає у використанні побудованої моделі як цільової функції або аналітичної залежності, за допомогою

якої здійснюється пошук таких значень вхідних ознак, що відповідають певному критерію [165]. У найпростішому випадку таким критерієм може бути максимізація прогнозованої врожайності або позитивного трендового залишку. У складніших постановках можуть враховуватися обмеження на допустимі значення ознак, реалістичність сценаріїв, взаємозв'язок між показниками та стабільність отриманого результату.

Для пошуку сприятливих кліматичних траєкторій можуть використовуватися як класичні оптимізаційні методи, так і евристичні алгоритми. Одним із таких підходів є Differential Evolution, який застосовується для глобальної оптимізації в неперервних просторах і може бути корисним тоді, коли цільова функція є нелінійною, недиференційовною або складною для аналітичного дослідження [165]. У поєднанні з моделями машинного навчання Differential Evolution може використовуватися для пошуку таких комбінацій агрокліматичних ознак, які асоціюються з високими прогнозними значеннями цільової змінної.

Використання оптимізаційних алгоритмів разом із моделями машинного навчання потребує обережної інтерпретації. Отримана траєкторія не є прогнозом фактичної погоди і не означає, що відповідні умови можуть бути безпосередньо реалізовані в агровиробництві. Вона має розглядатися як модельний сценарій, який показує, за яких комбінацій вхідних ознак побудована модель очікує сприятливий результат. Тому такі траєкторії використовуються як аналітичні орієнтири для інтерпретації моделі, порівняння регіональних умов, оцінювання чутливості врожайності та формування припущень для подальших досліджень.

Оптимізаційний аналіз найбільш змістовно працює у поєднанні з нелінійними моделями, оскільки саме вони можуть відображати складніші залежності між агрокліматичними показниками та врожайністю. Лінійна модель дає змогу оцінити напрям впливу ознак, але не завжди відображає порогові, квадратичні або кумулятивні ефекти. Нелінійна регресія, ансамблеві методи, SVM або інші алгоритми машинного навчання можуть бути використані для опису складніших поверхонь відгуку. На основі таких моделей можна досліджувати, які

поєднання температурних або опадових характеристик відповідають сприятливим сценаріям.

Таким чином, для задач інтелектуального прогнозування врожайності важливо поєднувати кілька аналітичних рівнів. Перший рівень пов’язаний із побудовою числового або класифікаційного прогнозу. Другий рівень передбачає оцінювання ризику низької врожайності на основі трендових залишків і нижньохвостових характеристик розподілу. Третій рівень спрямований на сценарне дослідження зміни агрокліматичних умов. Четвертий рівень пов’язаний з оптимізаційним пошуком сприятливих траєкторій вегетаційного періоду. Така багаторівнева постановка створює основу для формулювання завдань дослідження та подальшої побудови моделей у наступних розділах.

1.6. Узагальнення проблем сучасних підходів і постановка завдань дисертаційного дослідження

Розглянуті інформаційні системи, цифрові платформи, джерела даних, ознакові представлення та методи прогнозування врожайності свідчать про активний розвиток аграрної прогнозної аналітики. Сучасні підходи охоплюють широкий спектр рішень: від платформ точного землеробства й геоінформаційного моніторингу до моделей машинного та глибинного навчання, систем підтримки прийняття рішень, ризикового аналізу й сценарного моделювання. Водночас наявні рішення не повною мірою вирішують задачу побудови цілісного підходу до прогнозування врожайності на основі регіонально агрегованих агрокліматичних даних, коротких часових рядів, нелінійних кліматичних ознак, класифікаційних моделей, рекурентних нейронних мереж, процедур підбору гіперпараметрів, ризикових оцінок і програмної реалізації результатів.

Однією з ключових проблем є обмеженість коротких регіональних рядів урожайності. Якщо моделі будуються окремо для кожної області, кількість спостережень часто є недостатньою для стійкого навчання, особливо у випадку використання складних алгоритмів машинного або глибинного навчання. Механічне об’єднання всіх регіонів також не є достатньо обґрунтованим, оскільки області можуть істотно відрізнятися за агрокліматичними умовами, структурою

міжрічної мінливості та характером реакції врожайності на погодні чинники. Тому важливим є формування таких навчальних вибірок, які одночасно збільшують кількість спостережень і зберігають змістовну однорідність даних.

Другою проблемою є обмеженість сучасних підходів до прогнозування врожайності, що не враховують нижньохвостовий ризик. Низьковрожайні роки формують окрему групу спостережень, що має особливе значення для управління ризиками. Для їх аналізу потрібні не тільки прогнозні або класифікаційні моделі, а й кількісні оцінки нижньохвостових характеристик розподілу, які відображають імовірність і глибину несприятливих відхилень. У цьому разі трендові залишки можуть бути використані як основа для ідентифікації розподілів, оцінювання квантилів і розрахунку мір ризику.

Третьою проблемою є те, що традиційне числове прогнозування врожайності не завжди відповідає потребам практичної аналітики. У багатьох випадках важливим є не лише точне передбачення значення, а своєчасне визначення класу врожайності, зокрема виявлення років із низьким рівнем урожайності. Така постановка має значення для раннього попередження, страхового аналізу, фінансового планування та підтримки управлінських рішень. Це обґрунтовує розвиток класифікаційного підходу, у якому врожайність розглядається за категоріями, а моделі оцінюються за здатністю виявляти саме несприятливі стани.

Четверта проблема пов’язана з недостатнім зіставленням ендогенної та екзогенної постановок прогнозування. Ендогенний підхід спирається на попередню динаміку ряду врожайності або трендових залишків, тоді як екзогенний підхід використовує зовнішні агрокліматичні ознаки. Для коротких і неоднорідних регіональних рядів важливо не лише побудувати окремі моделі, а й експериментально порівняти їхню ефективність у різних агрокліматичних зонах.

П’ята проблема пов’язана із застосуванням моделей глибинного навчання до коротких часових рядів урожайності. Рекурентні нейронні мережі, зокрема GRU та LSTM, є перспективними для аналізу послідовностей, однак їх ефективність істотно залежить від довжини ряду, наявності автокореляційної структури, обсягу навчальної вибірки й залучення зовнішніх пояснювальних змінних. За коротких

рядів трендових залишків збільшення складності архітектури може не підвищувати точність, а спричиняти перенавчання. Тому актуальним є не лише застосування рекурентних моделей, а й встановлення меж їх придатності та обґрунтування переходу до мультиваріативної постановки.

Шостою проблемою є нестійкість результатів прогнозування за умов вибору однієї конфігурації гіперпараметрів. У задачах із короткими вибірками окрема найкраща конфігурація на валідаційній вибірці може відображати випадкові особливості поділу даних, а не справжню перевагу моделі. Тому потрібна методика підбору гіперпараметрів, яка поєднує систематичний пошук, аналіз групи найкращих конфігурацій і незалежне тестове оцінювання.

Сьомою проблемою є недостатня інформативність базових кліматичних показників. Температура та опади впливають на врожайність не лише прямо й незалежно, а й через взаємодії, накопичення ефектів, порогові стани та послідовність погодних умов у часі. Використання лише початкових предикторів може не відображати неадитивного та кумулятивного впливу погодних факторів у критичні фази вегетації. Це зумовлює потребу в розширенні ознакового простору шляхом синтезу нелінійних ознак другого порядку.

Восьма проблема пов’язана з практичним використанням результатів моделювання. Навіть обґрунтовані моделі мають обмежену прикладну цінність, якщо вони не інтегровані в програмне середовище, яке забезпечує підготовку даних, формування ознак, побудову прогнозів, оцінювання ризику, сценарний аналіз, візуалізацію та відтворення аналітичних процедур. Тому результати дослідження потребують реалізації у вигляді інформаційно-аналітичної системи, що поєднує дані, моделі, ризикові оцінки, сценарії, карти, графіки й звітні матеріали.

Виявлені обмеження сучасних підходів формують логічну основу для постановки завдань дисертаційного дослідження. Їх зв’язок із відповідними напрямами розв’язання та завданнями роботи наведено на рис. 1.2.

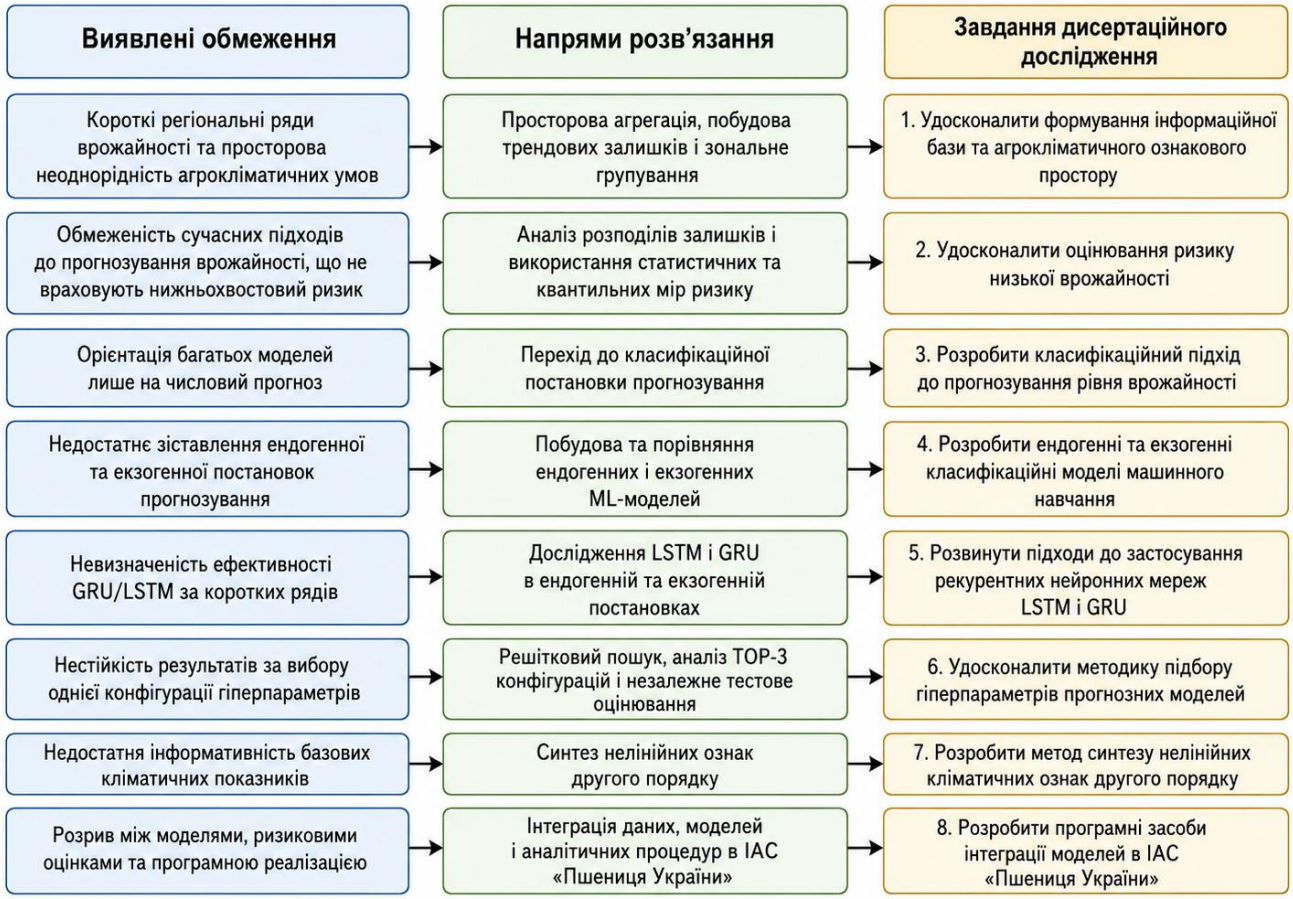

Рис. 1.2. Зв'язок виявлених обмежень сучасних підходів із завданнями дисертаційного дослідження

Схема конкретизує перехід від виявлених у розділі обмежень до восьми завдань дисертаційного дослідження. На основі проведеного аналізу в дисертаційній роботі поставлено такі завдання:

1. Удосконалити метод формування інформаційної бази та агрокліматичного ознакового простору для моделювання врожайності шляхом просторової агрегації кліматичних показників, побудови трендових залишків і зонального групування регіональних спостережень.

2. Удосконалити методичні засади оцінювання ризику низької врожайності на основі ідентифікації розподілів трендових залишків і використання статистичних та квантильних мір ризику.

3. Розробити класифікаційний підхід до прогнозування рівня врожайності за категоріями «низька» / «висока» з використанням трендових залишків як основи для формування цільової змінної.

4. Розробити та експериментально перевірити ендогенні та екзогенні класифікаційні моделі машинного навчання для прогнозування класу врожайності на основі внутрішньої динаміки трендових залишків та екзогенних агрокліматичних ознак у різних агрокліматичних зонах.

5. Розвинути теоретико-прикладні підходи до застосування рекурентних нейронних мереж у задачах прогнозування врожайності шляхом дослідження моделей LSTM і GRU в ендогенній та екзогенній постановках.

6. Удосконалити методику підбору гіперпараметрів прогнозних моделей шляхом поєднання решіткового пошуку з аналізом групи найкращих конфігурацій та незалежним тестовим оцінюванням.

7. Розробити метод синтезу нелінійних кліматичних ознак другого порядку для врахування неадитивного та кумулятивного впливу погодних умов на врожайність.

8. Розробити програмні засоби інтеграції інтелектуальних моделей прогнозування врожайності в інформаційно-аналітичній системі, що забезпечує підготовку даних, формування ознак, побудову прогнозів, оцінювання ризику, сценарний аналіз і візуалізацію результатів.

Виконання зазначених завдань визначає подальшу логіку дисертаційного дослідження. У другому розділі розглядається формування агрокліматичного ознакового простору, трендових залишків, зональних вибірок і ризикових характеристик. У третьому розділі досліджується ендогенний підхід до прогнозування врожайності з використанням моделей часових рядів, класифікаційних моделей, процедур підбору гіперпараметрів і рекурентних нейронних мереж. У четвертому розділі розглядається екзогенний підхід, що ґрунтується на агрокліматичних ознаках, нелінійному розширенні ознакового простору, моделях машинного та глибинного навчання, а також сценарно-оптимізаційному аналізі кліматичних траєкторій. У п’ятому розділі подано програмну реалізацію запропонованих методів і моделей в інформаційно-аналітичній системі.

Узагальнений зв’язок між сформульованими завданнями, основними напрямами моделювання, ризиковим та сценарно-оптимізаційним аналізом і програмною реалізацією результатів наведено на рис. 1.3.

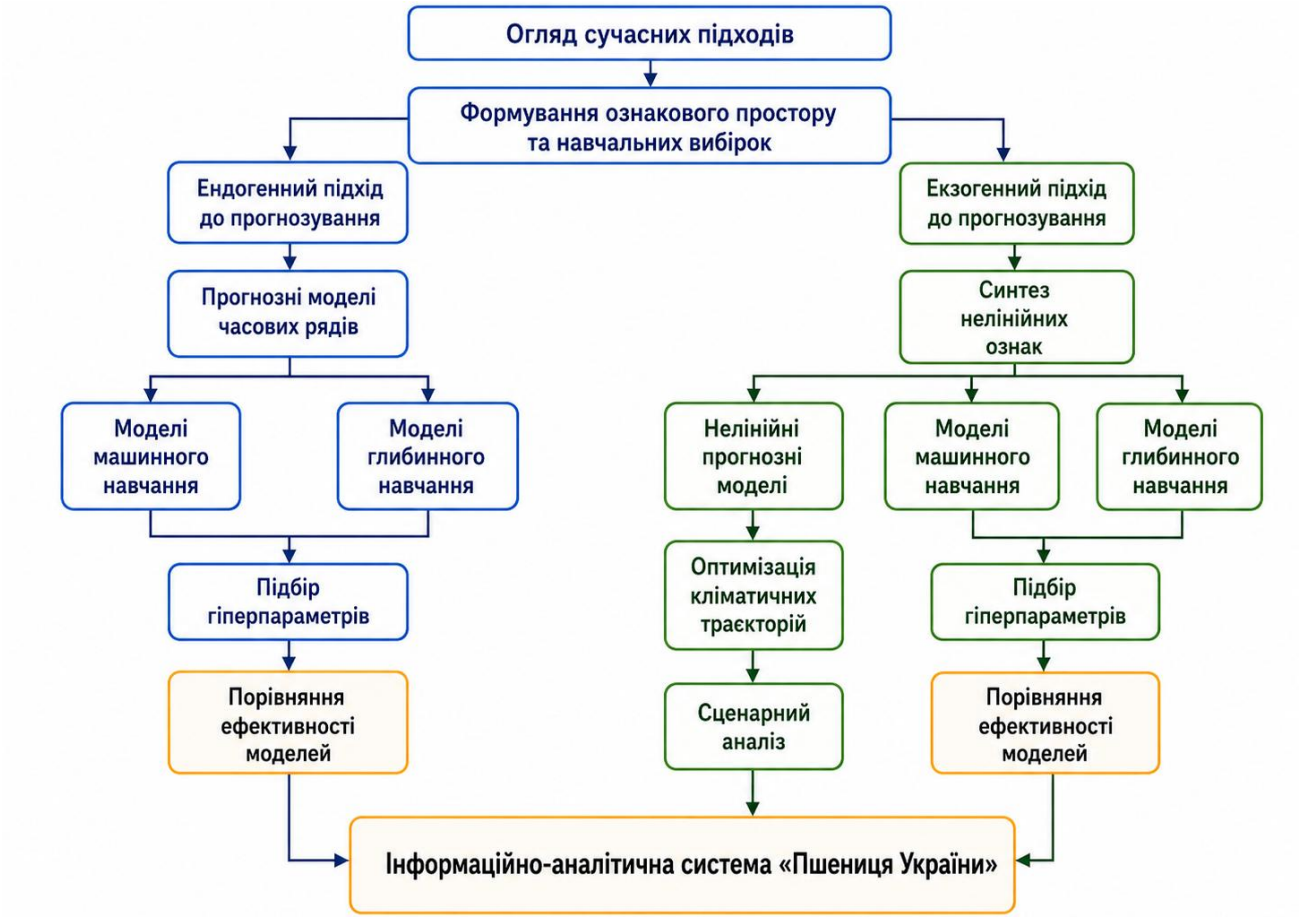

Рис. 1.3. Узагальнена структурно-логічна схема дисертаційного дослідження

Структура дослідження охоплює два взаємопов’язані напрями прогнозування: ендогенний, орієнтований на аналіз часової динаміки врожайності, та екзогенний, що базується на використанні агрокліматичних ознак. Окремим складником дослідження є синтез нелінійних ознак і визначення сприятливих кліматичних траєкторій. Отримані результати формують основу для подальшої програмної реалізації в інформаційно-аналітичній системі «Пшениця України».

Висновки до розділу 1

1. Проаналізовано сучасні інформаційні системи, цифрові платформи та програмні рішення, що використовуються в аграрній аналітиці, зокрема платформи точного землеробства, геоінформаційні й супутникові сервіси, системи підтримки

прийняття рішень та дослідницькі системи прогнозування врожайності. Встановлено, що значна частина таких рішень орієнтована переважно на рівень поля або господарства, тоді як задача регіонального прогнозування врожайності потребує спеціалізованого підходу до агрегованих статистичних, кліматичних і просторових даних.

2. Розглянуто основні постановки задачі прогнозування врожайності: регресійну, класифікаційну, ендогенну, екзогенну, ризикову та сценарно-оптимізаційну. Показано, що для задач аграрної прогнозної аналітики недостатньо обмежуватися лише числовим прогнозом, оскільки важливими є також класифікація років із низькою врожайністю, оцінювання ризику, аналіз можливих сценаріїв зміни агрокліматичних умов і пошук сприятливих траєкторій вегетаційного періоду.

3. Узагальнено основні групи даних, що використовуються для прогнозування врожайності: статистичні, кліматичні, метеорологічні, супутникові, ґрунтові, агротехнологічні, економічні та просторові. Обґрунтовано, що якість прогнозування значною мірою залежить від просторово-часового узгодження даних, коректного формування навчальних вибірок і побудови інформативного ознакового простору.

4. Проаналізовано методи машинного та глибинного навчання, які застосовуються у прогнозуванні врожайності, зокрема лінійні й нелінійні регресійні моделі, SVM, Random Forest, Gradient Boosting, Logistic Regression, класифікаційні моделі, GRU та LSTM. Встановлено, що складні моделі не є універсально кращими, оскільки їх ефективність залежить від довжини часових рядів, структури ознак, схеми валідації та ризику перенавчання.

5. Обґрунтовано доцільність поєднання прогнозного, ризикового та сценарно-оптимізаційного аналізу в задачах аграрної прогнозної аналітики. Показано, що трендові залишки можуть бути використані як основа для оцінювання несприятливих відхилень урожайності, а квантильні міри ризику VaR і CVaR — для кількісної характеристики нижньохвостового ризику низької врожайності.

6. Виявлені обмеження сучасних підходів дали змогу сформулювати завдання дисертаційного дослідження, спрямовані на формування інформаційної бази та агрокліматичного ознакового простору, оцінювання ризику низької врожайності, розроблення класифікаційного підходу, побудову ендогенних і екзогенних моделей машинного навчання, дослідження рекурентних нейронних мереж, удосконалення методики підбору гіперпараметрів, синтез нелінійних кліматичних ознак і програмну інтеграцію результатів в інформаційно-аналітичну систему.

РОЗДІЛ 2. МЕТОДИКА ФОРМУВАННЯ ПРОСТОРУ АГРОКЛІМАТИЧНИХ ОЗНАК ДЛЯ ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ

2.1. Формування простору кліматичних ознак

Метою формування простору агрокліматичних ознак у цьому розділі є створення інформаційного базису для подальшої побудови моделей прогнозування врожайності, оцінювання ризику низької врожайності та програмної реалізації результатів дослідження в інформаційно-аналітичній системі. Розглянуті в попередньому розділі методи прогнозування та інформаційні системи аграрної аналітики показують, що якість прогнозних результатів залежить не лише від вибору математичної або машинної моделі, а й від способу формування вхідної інформаційної бази. Тому в дисертації особливу увагу приділено побудові узгодженого набору агрокліматичних і статистичних даних, який забезпечує формування ознакового простору, побудову ендогенних та екзогенних прогнозних моделей врожайності та подальшу інтеграцію результатів у програмні засоби підтримки прийняття рішень.

У цьому контексті формування простору агрокліматичних ознак розглядається як початковий етап майнінгу даних для майбутньої інформаційно-аналітичної системи. Його зміст полягає у збиранні, просторовому узгодженні, очищенні, структуризації та попередньому аналізі агрокліматичних і статистичних даних з метою подальшого виявлення закономірностей, оцінювання ризиків і формування інформаційної основи для підтримки прийняття рішень у сфері аграрного виробництва.

У цьому дослідженні під простором агрокліматичних ознак розуміється впорядкована структура даних, у якій кожне спостереження характеризується трьома основними компонентами: територіальною належністю, роком спостереження та набором агрокліматичних показників. Тобто вихідні дані мають структуру типу «область–рік–ознака», де ознаками виступають декадні температурні показники, місячні суми опадів та інші змінні, що можуть бути використані для моделювання врожайності. У термінах машинного навчання така

структура надалі перетворюється у матрицю «спостереження–ознаки», у якій кожному рядку відповідає певна область у певному році, а кожному стовпцю — окрема вхідна змінна прогнозної моделі.

Для створення інтелектуальної інформаційної системи необхідно сформувати матрицю агрокліматичних ознак, у якій регіональні показники врожайності поєднуються з температурними та опадовими характеристиками критичного періоду вегетації. На початковому етапі дослідження були проаналізовані дані за областями України, після чого для подальшого моделювання були виокремлені 18 областей України, які можуть бути згруповані за агрокліматичною подібністю і схожим характером динаміки врожайності пшениці. Для цих областей було сформовано просторово узгоджені ряди агрокліматичних і врожайнісних показників за період 2000–2021 роки, що становлять основу для регресійного, класифікаційного та оптимізаційного аналізу. Обмеження основної просторово-кліматичної матриці 2021 роком зумовлене тим, що починаючи з 2022 р. на аграрне виробництво України істотно вплинув зовнішній воєнний шок, наслідки якого потребують окремого модельного оцінювання. Водночас для побудови ендогенних моделей, що спираються лише на внутрішню динаміку врожайності, у роботі використано довші часові ряди врожайності за період 1955–2021 рр.

У межах роботи ознаковий простір формується на основі поєднання трьох груп даних: статистичних показників урожайності пшениці, реаналізних кліматичних даних і цифрових меж адміністративних областей України. Статистичні дані врожайності використовуються як регіональний цільовий показник, кліматичні дані — як джерело температурних та опадових предикторів, а полігони адміністративних меж — як просторова основа для узгодження кліматичних показників із рівнем подання врожайності.

Фактичні статистичні показники врожайності пшениці взято з даних Державної служби статистики України [168]. У роботі врожайність розглядається як регіональний річний показник, поданий на рівні області [45]. Саме тому кліматичні предиктори також мають бути приведені до обласного рівня, щоб

уникнути невідповідності між просторовим масштабом цільової змінної та просторовим масштабом пояснювальних факторів.

Джерелом кліматичних показників обрано набір даних ERA5, доступний через Copernicus Climate Data Store [94]. Реаналізні дані ERA5 являють собою просторово й часово узгоджені кліматичні оцінки, сформовані на основі поєднання даних спостережень і чисельного моделювання атмосфери. Для задач дослідження важливими є кілька властивостей ERA5: просторове покриття всієї території України, однорідність методики формування даних, наявність довгих часових рядів і можливість відтворюваного отримання добових статистичних показників.

У задачах регіонального моделювання використання лише даних наземних метеорологічних станцій може бути обмежене нерівномірністю їх просторового розміщення, можливими пропусками, відмінностями в режимах спостережень і складністю узгодження точкових вимірювань з адміністративними межами областей. Реаналізні дані ERA5 забезпечують єдину просторово-часову основу для розрахунку температурних показників і показників опадів у межах усіх досліджуваних областей України.

Альтернативними джерелами для подібних задач можуть бути ERA5-Land, NASA POWER, CHIRPS, E-OBS, супутникові індекси рослинності або дані наземних метеорологічних станцій. Проте ці джерела мають різну просторову роздільність, часову доступність, набір змінних, методику формування та ступінь придатності до територіального усереднення на рівні адміністративних областей. Тому ERA5 використано як базове джерело кліматичних даних. Узагальнену характеристику можливих джерел агрокліматичних даних наведено в табл. 2.1.

Для просторового узгодження кліматичних показників із регіональними статистичними даними використано полігони адміністративних меж областей України першого рівня адміністративного поділу (ADM1) з бази geoBoundaries [95]. Перед просторовим опрацюванням геометрії було приведено до єдиної географічної системи координат EPSG:4326, що забезпечило узгодження геопросторових даних із координатною структурою кліматичного набору.

Таблиця 2.1

Порівняльна характеристика можливих джерел агрокліматичних даних для задач прогнозування врожайності

Джерело даних	**Основні переваги**	**Обмеження**	**Доцільність використання в роботі**
Дані метеорологічних станцій	Фактичні наземні спостереження, висока довіра до локальних вимірювань	Нерівномірне просторове розміщення, можливі пропуски, складність узгодження з полігонами областей	Можуть використовуватися для локальної перевірки, але не є основою регіональної матриці ознак
ERA5	Довгий часовий ряд, просторово повне покриття, єдина методика формування, наявність температурних і опадових показників	Ґраткові дані потребують просторової агрегації до рівня областей	Використано як базове джерело для формування просторово узгодженої агрокліматичної матриці ознак
ERA5-Land	Вища деталізація для суходолу, орієнтація на land-surface змінні	Інший набір даних і часові межі; потребує окремої процедури обробки та порівняння	Може бути використано в подальших дослідженнях для уточнення просторової деталізації
NASA POWER	Зручний доступ до агрометеорологічних показників, орієнтація на прикладні задачі	Інша методика формування даних і просторового подання; потребує окремої перевірки для регіонального рівня	Розглядається як альтернативне джерело, але не використовується як базове
CHIRPS / супутникові опадові продукти	Деталізація опадів, придатність для аналізу посушливих умов	Переважно фокус на опадах; не покриває повний набір температурних ознак	Може доповнювати аналіз опадів, але не замінює повну агрокліматичну матрицю
Супутникові індекси рослинності	Відображають фактичний стан рослинного покриву, придатні для моніторингу посівів	Потребують окремої обробки, маскування культур, узгодження з площами посівів і періодами спостереження	Доцільні для подальшого розширення моделі, але не є основою цієї роботи

Значення температури та опадів у роботі визначалися не для координат географічного центру області, а шляхом усереднення всіх пікселів кліматичного поля, що потрапляють у межі відповідного полігона ADM1. Такий підхід зменшує залежність результату від випадкового вибору однієї точки й краще відповідає

регіональному характеру статистичних даних про врожайність. Оскільки цільовий показник подається на рівні області, то і кліматичні предиктори мають бути агреговані до того самого просторового рівня.

Далі для кожної області виконувалося відсікання кліматичного поля за її полігоном і розрахунок середніх значень температури та опадів за всіма пікселями, розташованими в межах відповідної адміністративної одиниці. У результаті було отримано регіональні добові кліматичні ряди, узгоджені з обласним рівнем подання врожайності.

У дослідженні використано дані за квітень–червень 2000–2021 рр., оскільки цей період охоплює важливі етапи весняно-літньої вегетації пшениці та є інформативним для аналізу впливу температурного режиму й водозабезпечення на врожайність [1; 5; 22; 67; 71]. До фінальної вибірки включено 18 областей України, для яких сформовано узгоджені за просторовим і часовим рівнем ряди врожайнісних та агрокліматичних показників. Такий склад вибірки забезпечує можливість подальшого зонального групування областей і побудови навчальних вибірок більшого обсягу, що розглядається у підрозділі 2.3.

Для температурного блоку залучено добові максимальні та мінімальні значення температури повітря на висоті 2 м. Оскільки температурні значення в наборі ERA5 подаються в абсолютній шкалі Кельвіна, їх було переведено в градуси Цельсія за формулою

$$T_{^\circ\mathrm{C}} = T_K - 273{,}15 \tag{2.1}$$

де T_K – температура у Кельвінах, $T_{^\circ\mathrm{C}}$ – температура в градусах Цельсія.

Після цього для кожної області й кожної доби обчислювалася середньодобова температура:

$$tmean_{\,d} = (Tmax_d + Tmin_d)/2, \tag{2.2}$$

де $Tmax_d$ і $Tmin_d$ — відповідно максимальна та мінімальна температура за добу d.

Для переходу від добових до декадних значень використано правило поділу кожного місяця на три інтервали: перша декада охоплює 1–10 число, друга — 11–20 число, третя — 21 число і до кінця місяця. Таким чином, для квітня, травня та

червня було сформовано дев’ять температурних ознак: t1–t3 для квітня, t4–t6 для травня та t7–t9 для червня.

Декадна температура визначалася як середнє значення середньодобової температури за відповідні дні декади:

$$t_{m,k} = \frac{1}{n_{m,k}} \sum_{d \in D_{m,k}} tmean_d, \tag{2.3}$$

де m — місяць, k — номер декади, $D_{m,k}$ — множина днів відповідної декади, $n_{m,k}$ — кількість днів із наявними спостереженнями.

Для опадового блоку використано добові значення total_precipitation. Оскільки ця змінна в ERA5 подається в метрах водного шару, значення було переведено в міліметри за формулою

$$R_{mm} = 1000 \cdot R_m, \tag{2.4}$$

де R_m — сума опадів у метрах, R_{mm} — сума опадів у міліметрах.

На відміну від температурних показників, які агрегувалися за декадами, опади було подано у вигляді місячних сум. Такий підхід обрано з огляду на накопичувальний характер впливу опадів на водозабезпечення рослин і більшу стійкість місячних сум порівняно з коротшими інтервалами. Для кожної області було сформовано три опадові ознаки:

$$R_m = \sum R_{d,mm}, \tag{2.5}$$

де $R_{d,mm}$ — добова сума опадів у міліметрах; m — відповідний місяць. У результаті отримано R10 — суму опадів за квітень, R20 — суму опадів за травень, R30 — суму опадів за червень.

На рис. 2.1 показано загальну логіку формування агрокліматичної матриці ознак для прогнозування врожайності пшениці. Структурна схема відображає поєднання трьох груп вихідних даних: статистичних показників врожайності; реаналізних кліматичних даних ERA5 та цифрових меж областей України рівня ADM1, а також подальший перехід від просторового узгодження та агрегації добових показників до формування агрокліматичних предикторів t1–t9, R10, R20, R30. Результатом цього етапу є фінальний набір даних виду «область–рік–у–t1–t9–R10, R20, R30», який

використовується для розвідувального аналізу, формування трендових залишків, агрокліматичної кластеризації та побудови прогнозних моделей.

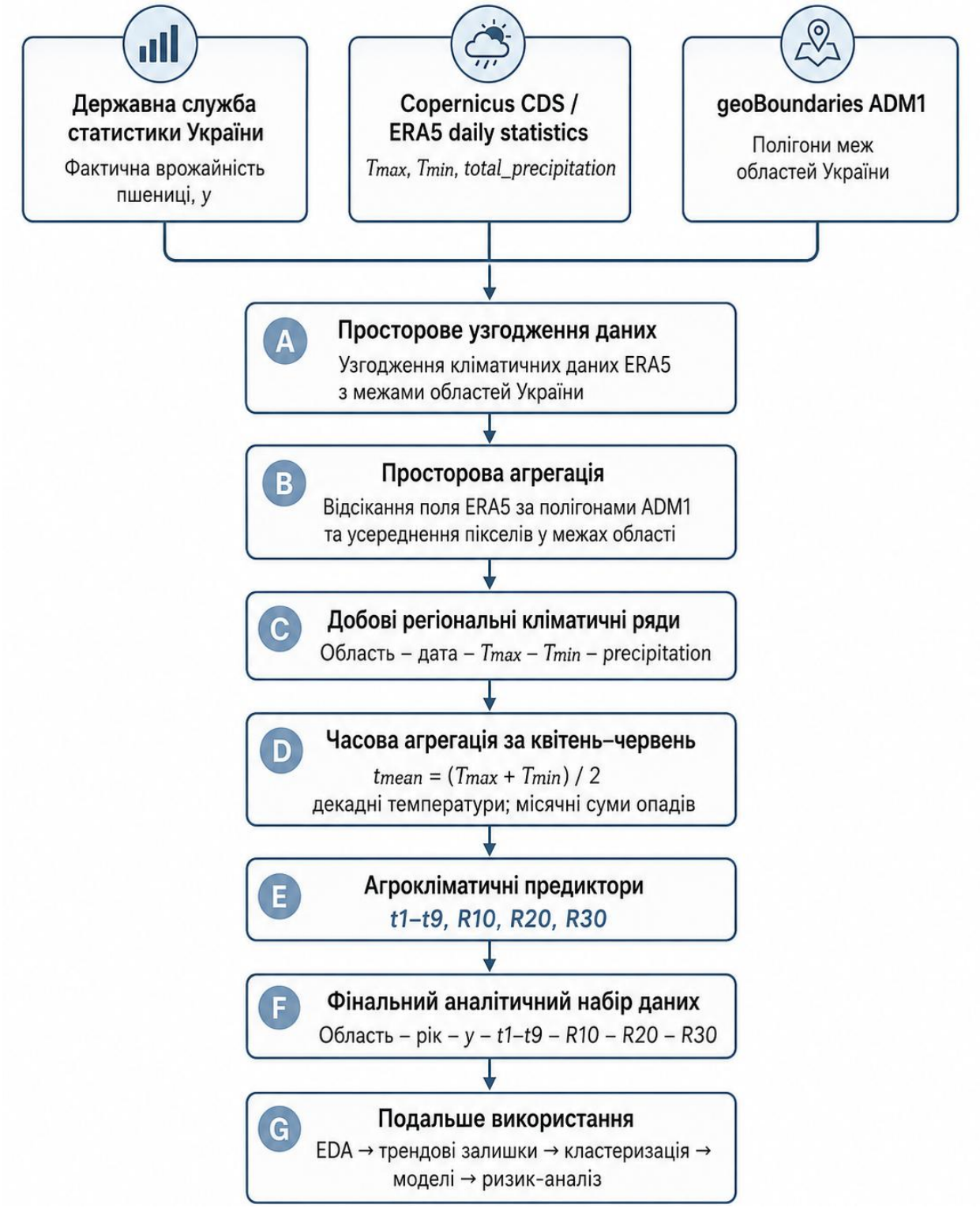

Рис. 2.1. Структурна схема формування агрокліматичної матриці ознак для прогнозного моделювання

Структуру сформованого фінального аналітичного набору даних та зміст його основних змінних наведено в табл. 2.2.

Таблиця 2.2

Структура фінального аналітичного набору даних для прогнозного моделювання

Позначення	**Зміст показника**	**Роль у наборі даних**	**Період / спосіб агрегації**	**Одиниця виміру**	**Джерело / спосіб формування**
Область	Адміністративна область України	Просторовий ідентифікатор	Річний запис	–	Просторова прив’язка до полігону області ADM1 [95]
Рік	Календарний рік спостереження	Часовий ідентифікатор	2000–2021 рр.	рік	Часова прив’язка статистичних і кліматичних даних
y	Фактична врожайність пшениці	Цільова змінна	Річний показник на рівні області	ц/га	Дані Державної служби статистики України [168]
t1	Середня температура першої декади квітня	Кліматичний предиктор	1–10 квітня	°C	Середнє з добових значень tmean по області на основі ERA5 [94]
t2	Середня температура другої декади квітня	Кліматичний предиктор	11–20 квітня	°C	Середнє з добових значень tmean по області на основі ERA5 [94]
t3	Середня температура третьої декади квітня	Кліматичний предиктор	21–30 квітня	°C	Середнє з добових значень tmean по області на основі ERA5 [94]
t4	Середня температура першої декади травня	Кліматичний предиктор	1–10 травня	°C	Середнє з добових значень tmean по області на основі ERA5 [94]
t5	Середня температура другої декади травня	Кліматичний предиктор	11–20 травня	°C	Середнє з добових значень tmean по області на основі ERA5 [94]
t6	Середня температура третьої декади травня	Кліматичний предиктор	21–31 травня	°C	Середнє з добових значень tmean по області на основі ERA5 [94]
t7	Середня температура першої декади червня	Кліматичний предиктор	1–10 червня	°C	Середнє з добових значень tmean по області на основі ERA5 [94]
t8	Середня температура другої декади червня	Кліматичний предиктор	11–20 червня	°C	Середнє з добових значень tmean по області на основі ERA5 [94]
t9	Середня температура третьої декади червня	Кліматичний предиктор	21–30 червня	°C	Середнє з добових значень tmean по області на основі ERA5 [94]
R10	Сума опадів за квітень	Кліматичний предиктор	Місячна агрегація	мм	Сума добових опадів по області на основі ERA5 [94]
R20	Сума опадів за травень	Кліматичний предиктор	Місячна агрегація	мм	Сума добових опадів по області на основі ERA5 [94]
R30	Сума опадів за червень	Кліматичний предиктор	Місячна агрегація	мм	Сума добових опадів по області на основі ERA5 [94]

У табл. 2.2 наведено структуру фінального аналітичного набору даних, сформованого для подальшого прогнозування врожайності пшениці. Змінні «область» і «рік» забезпечують просторово-часову ідентифікацію кожного запису, показник у виступає цільовою змінною, а ознаки t1–t9 та R10–R30 утворюють базовий агрокліматичний простір предикторів. Така структура є узгодженою з подальшими етапами дослідження, зокрема з розвідувальним аналізом даних, формуванням трендових залишків врожайності, агрокліматичною кластеризацією областей та побудовою прогнозних моделей.

Наступним етапом дослідження став розвідувальний аналіз даних (Exploratory data analysis, EDA), спрямований на перевірку повноти, коректності, варіативності та кореляційної структури змінних. У цьому дослідженні EDA розглядається як один із базових етапів інтелектуального аналізу даних (data mining), спрямований на виявлення структури, якості, варіативності та взаємозв’язків у сформованому агрокліматичному наборі. На цьому етапі перевіряється придатність даних до подальшого автоматизованого опрацювання в інформаційно-аналітичній системі.

На першому етапі EDA перевірялася однозначність ключа «область–рік» і наявність дублювань. Для вибірки, використаної в аналізі, дублікати за природним ключем відсутні, повні дублікати рядків також не виявлені, а критичних виходів за допустимі фізичні межі не зафіксовано. Це підтверджує коректність структури сформованого набору даних.

На другому етапі аналізувалася повнота даних. Для сформованої матриці ознак частка пропусків у ключових змінних дорівнює нулю, що свідчить про повноту даних і дає можливість використовувати їх без додаткових процедур імпутації.

На третьому етапі виконувалося описове статистичне узагальнення ознак. Для температурних і опадових показників визначалися середні значення, медіани, квартилі, стандартні відхилення, мінімальні та максимальні значення. Одержані результати показують, що окремі ознаки характеризуються різною варіативністю, а це є важливим для подальшого аналізу інформативності факторів і вибору методів моделювання. Як показують аналогічні таблиці описової статистики у пов’язаних

роботах, температурні декадні показники та місячні опади мають різні масштаби мінливості, що необхідно враховувати при інтерпретації моделей.

Четвертим етапом було виявлення потенційних викидів. Аналіз показав наявність окремих спостережень з екстремальними значеннями метеопоказників, характерними для специфічних років або регіонів, що в подальшому потребує робастної перевірки при побудові моделей.

П’ятим етапом EDA був кореляційний аналіз між ознаками з метою оцінювання мультиколінеарності та виявлення стійких взаємозв’язків між температурами різних декад і опадами в різні місяці. Результати цього блоку вказують на наявність корельованості між частиною температурних змінних, що є очікуваним наслідком часової близькості декад одного вегетаційного періоду. Для наочного представлення структури взаємозв’язків між агрокліматичними ознаками кореляційну матрицю для факторів t1–t9 і R10, R20, R30 наведено на рис. 2.2. Це дає підстави в подальших розділах контролювати мультиколінеарність факторів, зокрема при побудові лінійних моделей та інтерпретації ваг ознак.

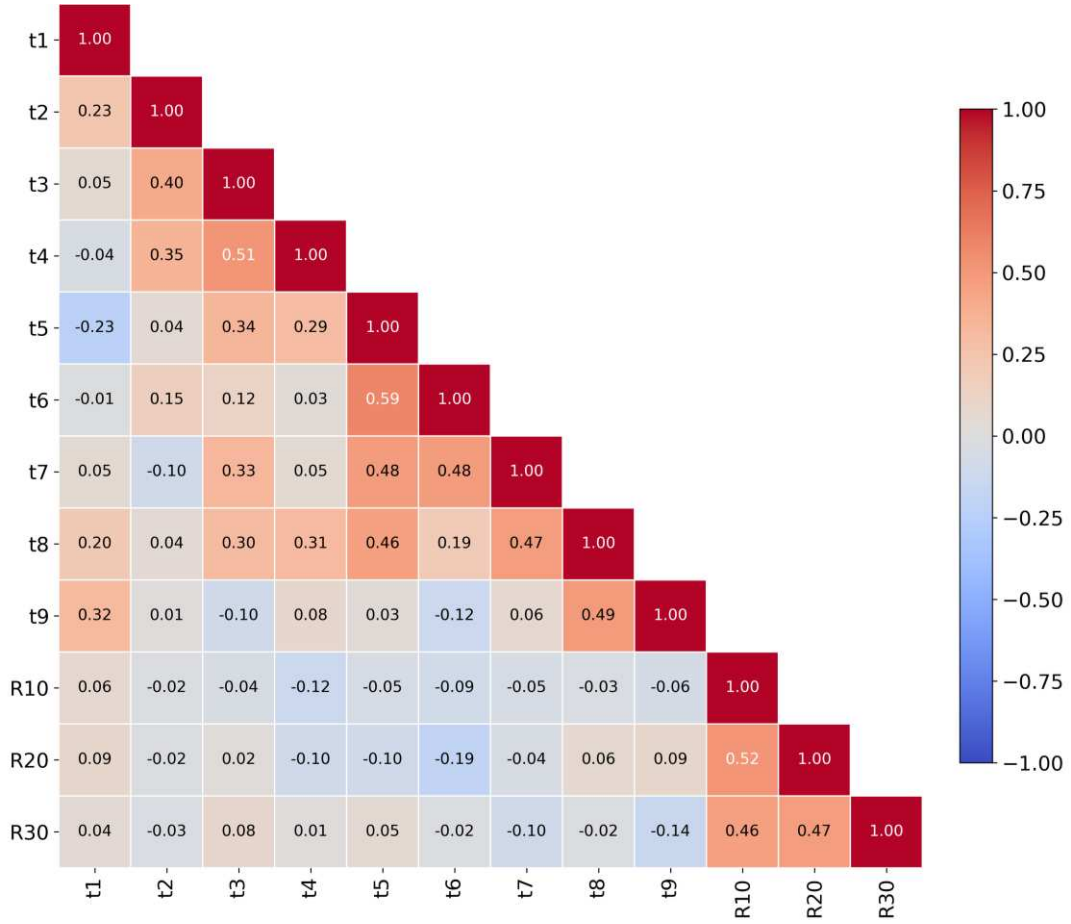

Рис. 2.2. Кореляційна матриця агрокліматичних ознак t1–t9 та R10, R20, R30

Особливу увагу в аналізі кліматичних факторів приділено травню, оскільки температурні та опадові умови цього місяця можуть істотно впливати на формування врожайності. Тому цікавими для аналізу є температурні та опадові показники цього місяця.

На рис. 2.3 зображено динаміку річних змін температурної ознаки t5 для Херсонської області України. Аналіз рис. 2.3 засвідчує значні міжрічні коливання температурного показника t5 з тенденцією до зростання середньої температури другої декади травня. Така динаміка узгоджується із загальними проявами потепління клімату та може впливати на ріст рослин і майбутню врожайність.

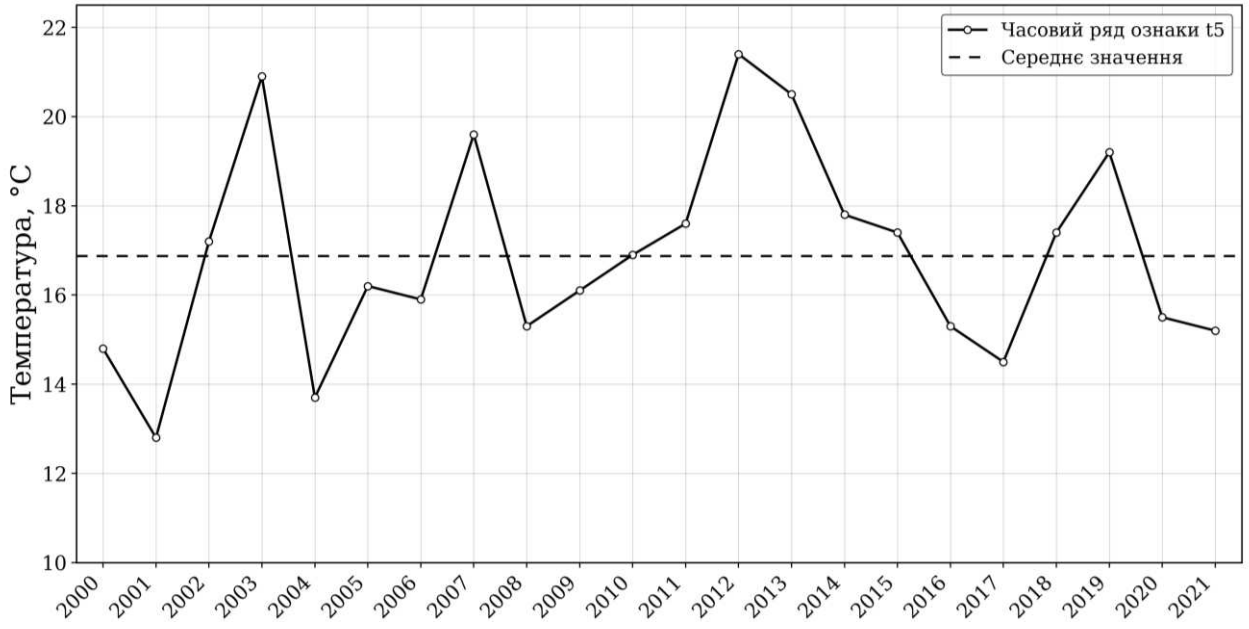

Рис. 2.3. Міжрічні зміни температурного показника t5 для Херсонської області України. Штрихова лінія ілюструє середнє значення за весь період

На рис. 2.4 подано динаміку міжрічних змін опадової ознаки R20 для Херсонської області України. Графік показує значну міжрічну мінливість травневих опадів. Для 2002, 2007 та 2013 років місячна сума опадів є нижчою від позначки 10 мм. Такі значення фактора R20 можуть бути пов’язані з підвищеним ризиком зниження врожайності пшениці у відповідні роки спостережень.

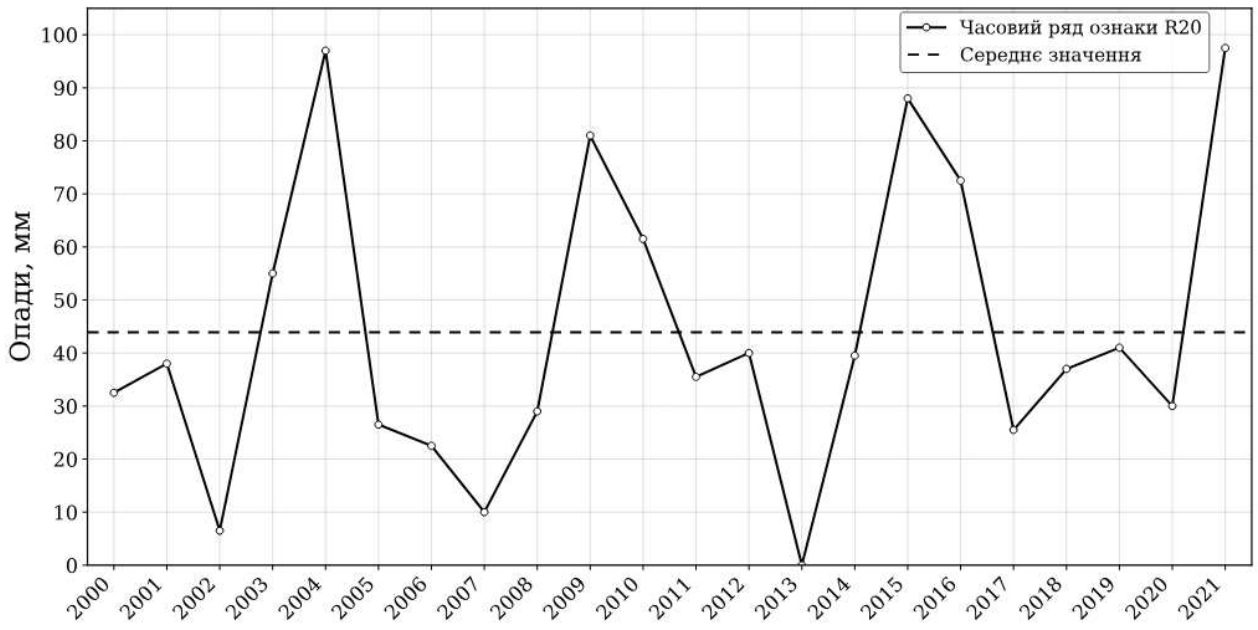

Рис. 2.4. Міжрічні зміни опадового показника R20 для Херсонської області України. Штрихова лінія ілюструє середнє значення за весь період

Узагальнення результатів EDA дає підстави для кількох важливих висновків. По-перше, сформований набір даних є повним, однозначним і придатним до подальшого аналізу без додаткових процедур очищення. По-друге, температурні та опадові ознаки характеризуються різною дисперсією та по-різному впливають на варіативність регіональної врожайності, що вказує на доцільність подальшого відбору інформативних змінних [16]. По-третє, виявлена кореляційна залежність окремих температурних факторів свідчить про необхідність контролю мультиколінеарності в регресійних моделях. По-четверте, часові профілі окремих ознак підтверджують наявність міжрічної варіативності та трендів, що може бути використано при побудові пояснювальних і прогностичних моделей.

Таким чином, нами сформовано просторово узгоджену агрокліматичну матрицю ознак для 18 областей України за період 2000–2021 рр. Вона поєднує декадні температурні показники та місячні суми опадів за критичний період вегетації, агреговані до рівня адміністративних областей. Виконаний розвідувальний аналіз підтвердив повноту, однозначність і придатність сформованого набору даних для подальшого аналізу трендових залишків,

агрокліматичного групування областей та побудови моделей машинного навчання [1; 4; 5; 11].

2.2. Формування часових рядів трендових залишків врожайності як цільової змінної

Для аналізу впливу кліматичних факторів на врожайність необхідно відокремити довгострокову тенденцію її зміни від міжрічних коливань. Абсолютні значення врожайності відображають не лише погодні умови, а й технологічний розвиток, зміну агротехніки, організаційно-економічні трансформації та інші чинники. Тому в подальшому аналізі як цільову змінну доцільно використовувати трендові залишки врожайності, отримані після вилучення довгострокової трендової складової. Такий підхід дає змогу перейти від аналізу рівнів урожайності до аналізу її відхилень від очікуваного трендового рівня, які більш безпосередньо відображають міжрічну мінливість, пов’язану з агрокліматичними умовами.

Отже, у роботі трендові залишки розглядаються як детрендована цільова змінна, придатна для подальшого прогнозування, агрокліматичного групування областей та оцінювання ризику низької врожайності [6; 78; 157–159].

У межах дисертаційного дослідження особливий інтерес становить степова зона України, яка характеризується підвищеною вразливістю до посух, високих температур і дефіциту вологи [22; 46–49]. Саме для цього регіону кліматичні умови квітня–червня значною мірою визначають підсумковий рівень урожайності [1; 5; 11; 16; 22; 67; 71]. Тому аналіз динаміки врожайності для степової зони є важливим етапом переходу від загального опису часових рядів до побудови моделей, орієнтованих на виявлення кліматичних закономірностей.

Як ілюстрацію динаміки врожайності доцільно розглянути дані врожайності пшениці для Херсонської області, яка є типовим представником південного посушливого регіону [22; 46–49]. Для реалізації ендогенних прогнозних моделей врожайності було використано часові ряди врожайності пшениці за період 1955–2021 рр. [168]. Динаміка врожайності пшениці у Херсонській області за цей період,

показана на рис. 2.5, має складний характер і загалом відображає основні етапи змін урожайності, характерні для більшості областей України.

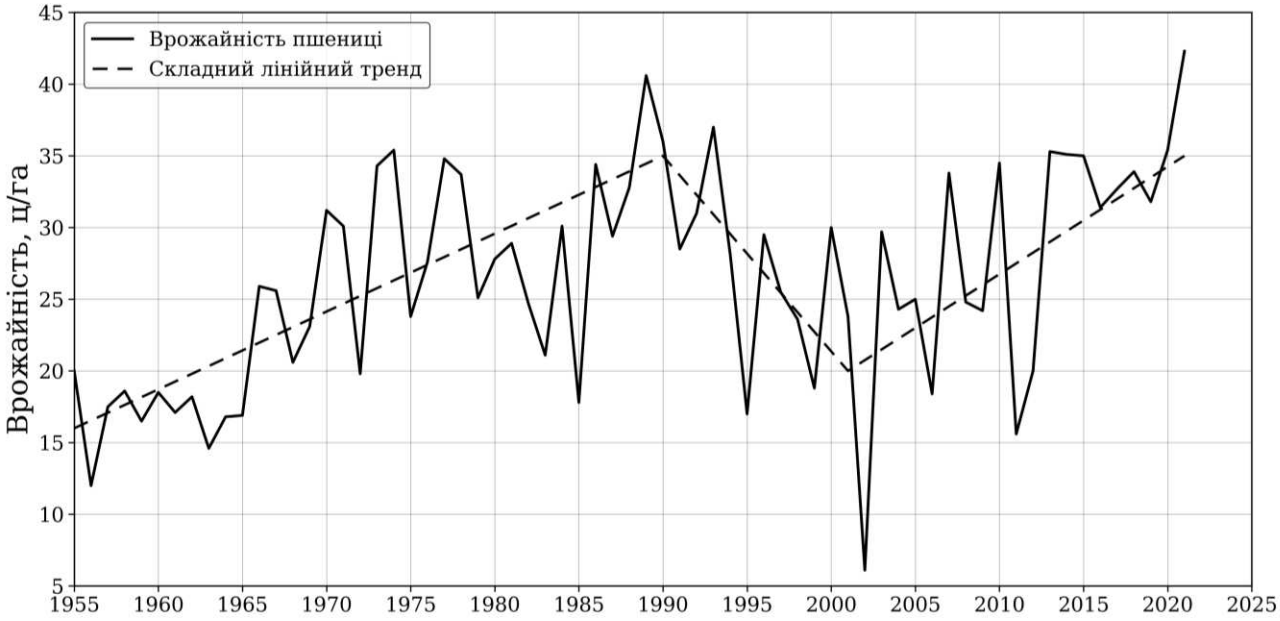

Рис. 2.5. Динаміка врожайності пшениці у Херсонській області та кусково-лінійний тренд

З 1955 по 1990 рік врожайність зростала завдяки впровадженню нових технологій, зокрема мінеральних добрив, засобів захисту рослин та більш продуктивної сільськогосподарської техніки. Період 1990–1999 років відзначився складними змінами в сільськогосподарській економіці України, спричиненими розпадом колективної системи господарювання. Цей період супроводжувався значним зниженням врожайності. З кінця 1990-х років основними виробниками зерна в Україні стали великі агрохолдинги. Врожайність знову почала зростати, що можна пояснити зростанням інвестиційної привабливості зернового сектору та значними інвестиціями, які він залучав. Внаслідок цього покращилася насіннєва база та агротехнології, логістична мережа розвинулася, включаючи залізничні вагони, зерновози та порти.

Інституційна структура сільськогосподарської економіки суттєво впливала на ефективність сільськогосподарського сектору України. Для врахування інституційно-економічних відмінностей між виділеними періодами в екзогенних моделях надалі використовується порядкова змінна $econ$, яка кодує тип аграрної

економіки [78; 157–159]. Її зміст і роль у прогнозних моделях розглянуто в розділі 4.

Тенденція до зростання врожайності зерна за період 2000–2021 роки супроводжується значними коливаннями, спричиненими переважно кліматичними факторами. Посухи, що спостерігалися у 2003 та 2012 роках, були одними з найважливіших чинників зниження врожаю. Це свідчить про те, що в степовій зоні зростаючий тренд врожайності поєднується з високою чутливістю до екстремальних погодних коливань [22; 49; 169].

З огляду на наявність структурних змін у розвитку аграрного виробництва часовий інтервал 1955–2021 рр. поділено на три періоди: 1955–1989 рр., 1990–1999 рр. та 2000–2021 рр. Така сегментація відповідає різним інституційно-технологічним умовам функціонування аграрного сектору: періоду планово-колективної системи, періоду трансформаційного спаду та періоду ринково-орієнтованого розвитку аграрного виробництва. Для кожного сегмента тренд урожайності апроксимовано окремою лінійною моделлю

$$tr_\tau = a_i + b_i \cdot \tau, \quad \tau \in T_i. \tag{2.6}$$

Тут T_i позначає розглянутий часовий інтервал, визначений як один із наступних періодів: 1955–1989, 1990–1999 або 2000–2021. Коефіцієнти тренду a_i, b_i оцінювалися шляхом підгонки моделі до статистичних даних за допомогою методу найменших квадратів [170]. Кліматичний вплив на врожайність доцільно досліджувати не через абсолютні значення врожайності, а через її відхилення від тренду. Для цього вводиться показник детрендованої врожайності ε_τ:

$$\varepsilon_\tau = y_\tau - tr_t \tag{2.7}$$

Саме ця величина надалі використовується як базова характеристика впливу погодних і кліматичних умов на коливання врожайності. Такий підхід дозволяє відокремити ефекти, пов’язані з технологічним прогресом і структурними змінами в аграрному виробництві, від безпосереднього впливу природних умов.

Графік детрендованої врожайності для Херсонської області наведено на рис. 2.6. Часовий ряд трендових залишків врожайності коливається навколо нульового рівня що дає підстави розглядати його як ряд міжрічних відхилень врожайності від

очікуваного трендового значення. Середнє значення трендових залишків врожайності становить –0,04 ц/га. Для вибору методу прогнозування ряду трендових залишків врожайності необхідно оцінити ступінь випадковості цього ряду за допомогою кореляційного аналізу [160]. На рис. 2.7 представлена автокореляційна функція та часткова автокореляційна функція ряду трендових залишків врожайності.

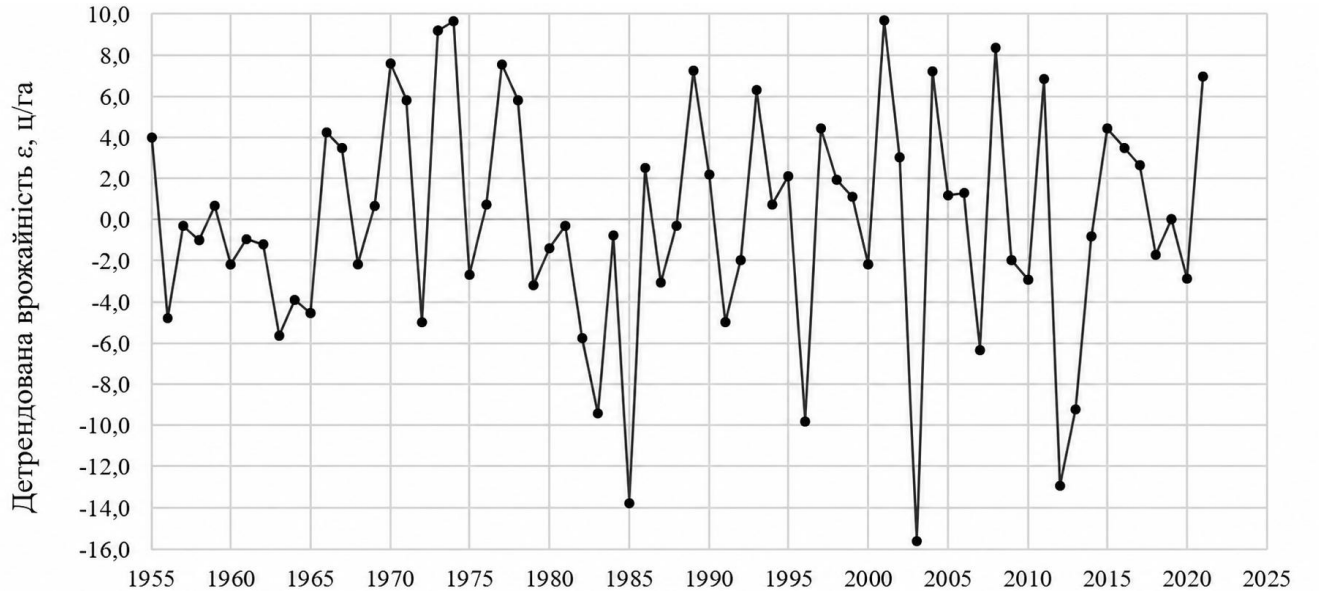

Рис. 2.6. Детрендований ряд врожайності пшениці для Херсонської області

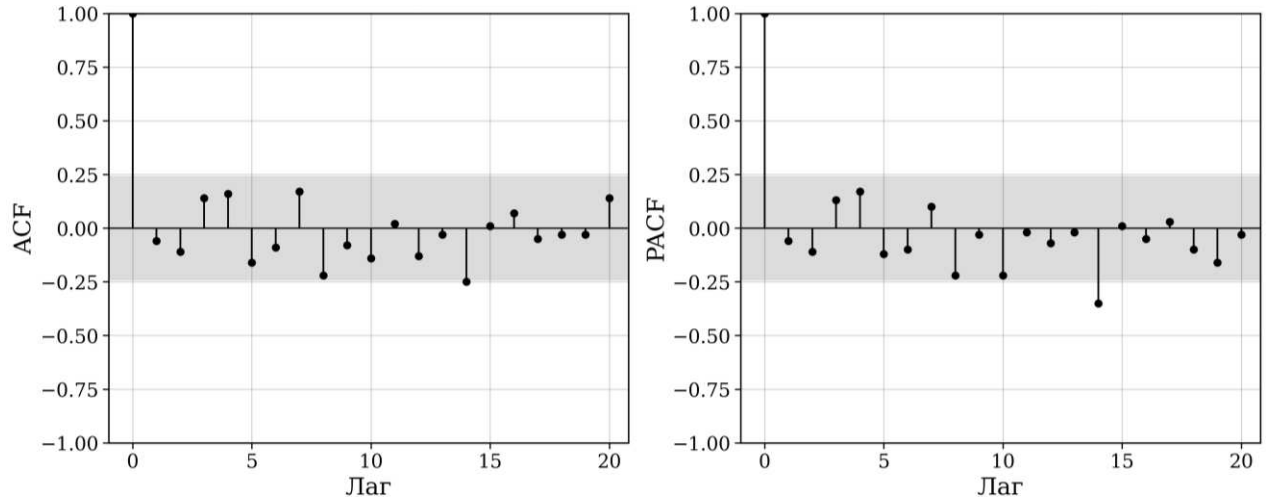

Рис. 2.7. Автокореляційна функція (зліва) і часткова автокореляційна функція (справа) для ряду трендових залишків врожайності пшениці Херсонської області

Аналіз ACF і PACF вказує на практичну відсутність внутрішніх кореляційних зв'язків у часових рядах трендових залишків врожайності. Слабкі кореляції між рівнями ряду присутні лише у чотирьох найближчих лагах. Це вказує на те, що цей

часовий ряд є близьким до випадкового шуму. З одного боку, це свідчить про високу якість побудованої моделі тренду. З іншого боку, це може викликати певні труднощі при побудові прогнозних моделей врожайності авторегресійного типу.

Перевірка трендових залишків ε для Херсонської області на відповідність нормальному закону розподілу була виконана за критеріями Shapiro–Wilk та Jarque–Bera (табл. 2.3).

Таблиця 2.3

Перевірка відповідності трендових залишків до нормального закону розподілу

Критерій	**Статистика**	**p-value**	**Висновок**
Shapiro–Wilk	0,9704	0,1111	Нормальність не відхиляється
Jarque–Bera	2,8530	0,2402	Нормальність не відхиляється

За обома використаними критеріями нульова гіпотеза про нормальний розподіл не була відхилена на рівні значущості 0,05. Отже, емпіричний розподіл трендових залишків не має статистично значущих відхилень від нормального розподілу. Для наочної перевірки відповідності трендових залишків ε нормальному закону розподілу було побудовано гістограму щільності емпіричного розподілу з накладеною кривою нормальної щільності. Графічний аналіз підтверджує, що форма емпіричного розподілу є близькою до нормальної (рис. 2.8).

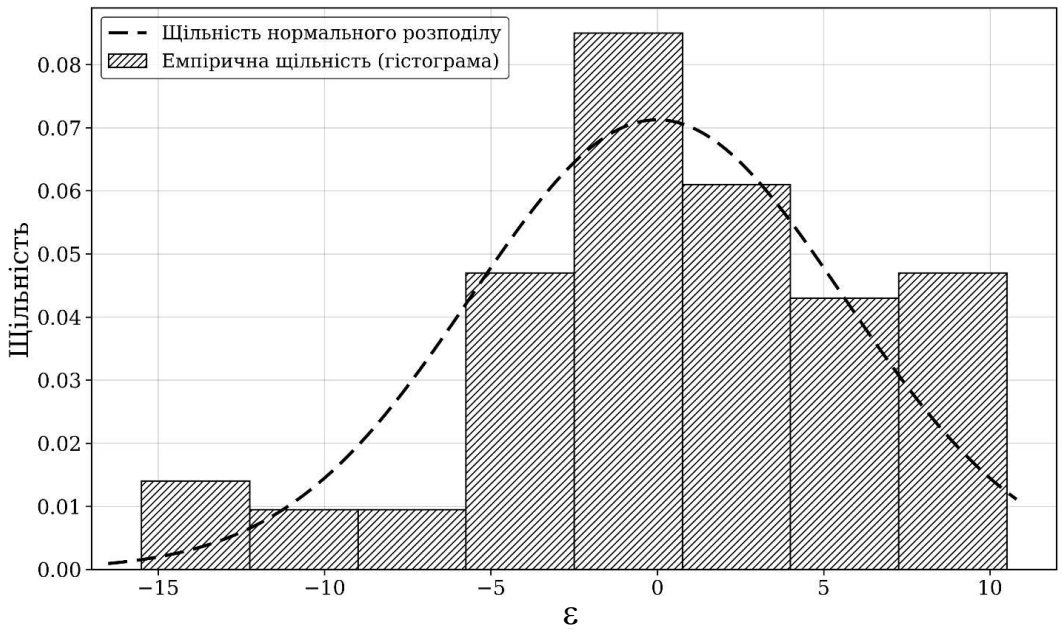

Рис. 2.8. Гістограма щільності розподілу для трендових залишків врожайності

Зауважимо, що отриманий висновок стосується ряду трендових залишків для Херсонської області. Для зональних вибірок розподільчі властивості залишків можуть відрізнятися, що враховується в підрозділі 2.4 під час оцінювання ризику низької врожайності.

Таким чином, формування трендових залишків урожайності створює цільову змінну, більш придатну для аналізу міжрічної агрокліматично зумовленої мінливості, ніж абсолютні рівні врожайності. Використання ε дає змогу відокремити довгострокові технологічні та організаційно-економічні зміни від короткострокових відхилень, пов’язаних із погодними умовами. Надалі ця змінна використовується як основа для агрокліматичної кластеризації областей, оцінювання ризику низької врожайності та побудови ендогенних і екзогенних моделей машинного навчання.

2.3. Агрокліматична кластеризація областей України та регіональна агрегація даних

Важливою передумовою побудови стійких моделей прогнозування врожайності є зменшення впливу регіональної неоднорідності агрокліматичних умов [1; 5; 22; 171; 172]. У зв’язку з цим у роботі поставлено завдання виділити групи областей України, які характеризуються подібним характером динаміки врожайності пшениці. Для цього ми виконали агрокліматичну кластеризацію областей України на основі подібності часових рядів трендових залишків урожайності [6; 7; 22]. Такий підхід дає змогу перейти від аналізу окремих коротких регіональних рядів до формування зональних вибірок більшого обсягу, придатних для подальшого регресійного, класифікаційного та машинно-навчального моделювання [1; 22; 171; 172].

Таким чином, кластеризація областей проводиться не за абсолютними значеннями врожайності, а за часовими рядами її трендових залишків [6; 78; 157–159]. Саме трендові залишки, як було показано в підрозділі 2.2, найповніше відображають вплив погодно-кліматичних умов на міжрічні коливання врожайності. Використання цього показника дозволяє усунути довгострокову

компоненту зростання врожайності, пов’язану з технологічними, організаційними та економічними змінами, і зосередитися на просторових відмінностях реакції врожайності на кліматичні фактори. Саме тому для групування областей у роботі використано кореляційний аналіз часових рядів детрендованої врожайності [6; 78].

Сутність підходу полягає в тому, що для кожної області формується часовий ряд трендових залишків урожайності, після чого між областями обчислюються парні коефіцієнти кореляції Пірсона. Якщо часові ряди двох областей виявляються достатньо схожими, тобто мають високі кореляційні зв’язки, це свідчить про подібність механізму реакції врожайності на зміни погодних умов. Сукупність таких кореляційних зв’язків дає змогу виявити групи областей зі схожими агрокліматичними характеристиками та об’єднати їх у спільні зони. Такий спосіб кластеризації є змістовно обґрунтованим, оскільки він спирається не на формальну географічну близькість, а на емпірично встановлену подібність міжрічної поведінки врожайності. Остаточне групування виконувалося з урахуванням структури кореляційної матриці та природно-кліматичної інтерпретації отриманих блоків.

У результаті проведеного аналізу для відібраних 18 областей було сформовано три агрокліматичні зони: степову, лісостепову та західну. Степова зона охоплює області південної та південно-східної частини досліджуваної вибірки, для яких характерна більша роль температурного режиму, дефіциту вологи та посушливих умов. Лісостепова зона об’єднує області центральної та північно-східної частини вибірки, у яких умови формування врожайності мають більш перехідний характер. Західна зона включає області з відносно вищим рівнем зволоження та відмінною структурою міжрічної мінливості врожайності.

Кореляційний аналіз дав змогу виділити в межах досліджуваної вибірки три групи областей зі схожими закономірностями міжрічних коливань детрендованої врожайності. Таке групування створює основу для переходу від аналізу окремих областей до моделювання врожайності на рівні узагальнених просторових регіонів, для яких можна сформувати статистично достатні вибірки спостережень. Оскільки кожна із зон містить по шість областей, об’єднання даних у межах зон забезпечує

методичну основу для порівняння ефективності прогнозних моделей, побудованих за зональними вибірками. Склад виділених агрокліматичних зон узагальнено в табл. 2.4.

Таблиця 2.4

Агрокліматичні зони України, виділені за результатами кореляційного аналізу трендових залишків врожайності

Агрокліматична зона	**Склад зони**
Степова	Одеська, Херсонська, Миколаївська, Дніпропетровська, Запорізька, Кіровоградська
Лісостепова	Сумська, Харківська, Полтавська, Київська, Черкаська, Вінницька
Західна	Волинська, Рівненська, Житомирська, Львівська, Тернопільська, Хмельницька

Важливо зазначити, що не всі області України були включені до наведених зон через якісно відмінний характер динаміки врожайності, який не дозволив віднести їх до вищеназваних груп. Донецька, Луганська, Чернігівська, Чернівецька, Закарпатська, Івано-Франківська області не віднесені до жодної із зон через відмінну динаміку врожайності, що пов’язано з природно-кліматичними умовами цих регіонів. Цей результат показує, що виділення агрокліматичних зон в роботі не є штучним розподілом усіх областей України, а відображає лише ті об’єднання, для яких подібність динаміки врожайності була емпірично підтверджена кореляційним аналізом.

Для наочного підтвердження виділених агрокліматичних зон побудовано кореляційну матрицю трендових залишків врожайності для 18 областей досліджуваної вибірки. Вона відображає парні коефіцієнти кореляції між часовими рядами детрендованої врожайності окремих областей за період 2000–2021 рр. Такий спосіб подання дає змогу оцінити не лише індивідуальні зв’язки між окремими областями, а й загальну блочну структуру подібності міжрічних коливань урожайності.

На рис. 2.9 видно, що області, віднесені до степової, лісостепової та західної зон, утворюють групи з підвищеним рівнем внутрішньої корельованості. Це підтверджує змістовність попереднього групування, наведеного в табл. 2.4, і показує, що виділені зони ґрунтуються не лише на географічній або природно-

кліматичній інтерпретації, а й на емпірично встановленій подібності часової динаміки трендових залишків. Отже, кореляційна матриця виконує роль візуальної перевірки сформованого зонального поділу та обґрунтовує подальше використання об’єднаних зональних вибірок у прогнозному моделюванні.

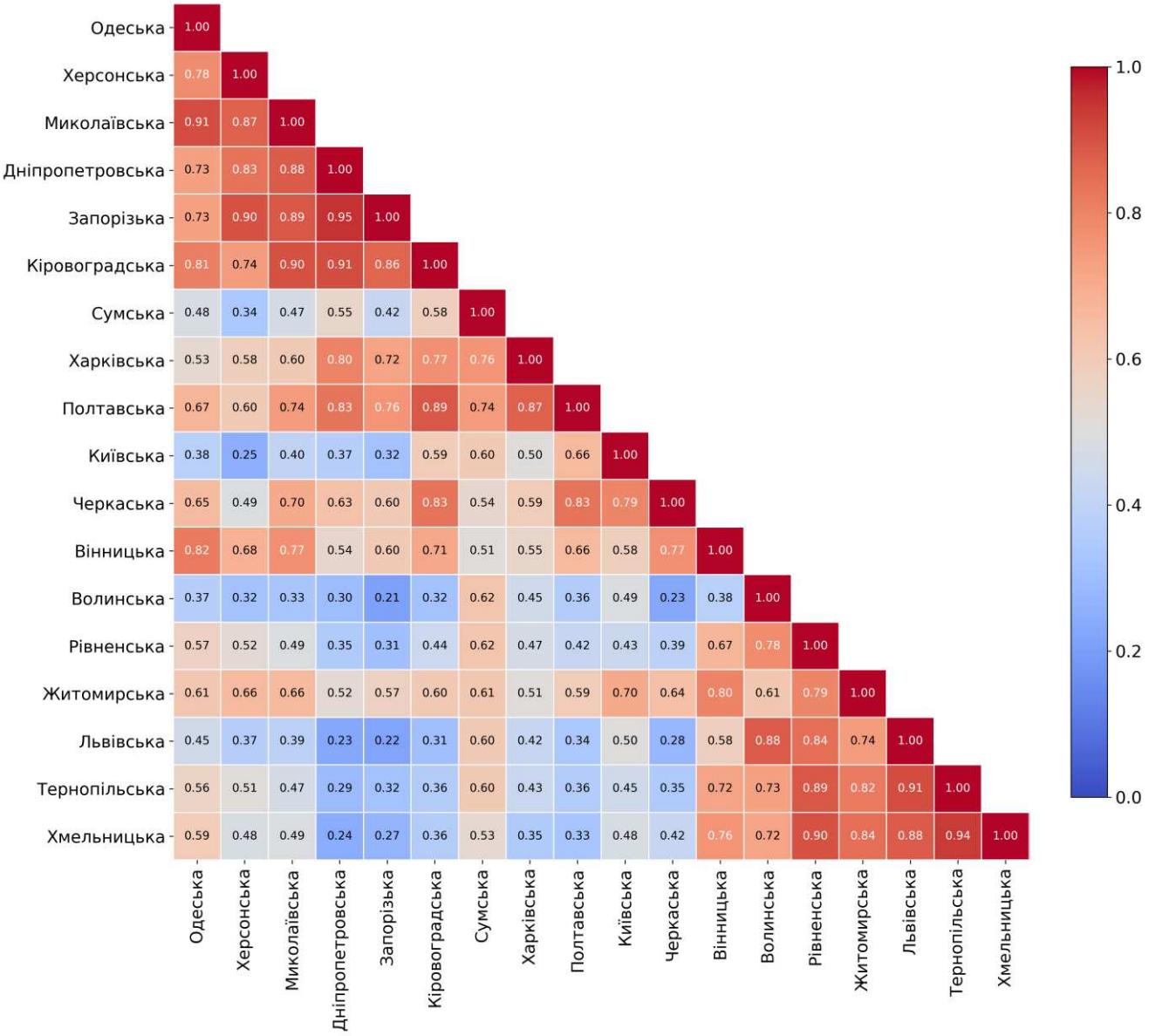

Рис. 2.9. Кореляційна матриця трендових залишків врожайності пшениці для областей України

Виділення агрокліматичних зон використовується нами для формування укрупнених зональних вибірок на подальших етапах моделювання. По-перше, такий підхід дає змогу суттєво збільшити обсяг вибірки. Якщо на рівні однієї області за період 2000–2021 рр. доступно лише 22 річні спостереження, то об’єднання шести областей у межах однієї зони дає змогу сформувати вибірку

обсягом до 132 спостережень. Це підвищує репрезентативність навчальних даних, зменшує вплив випадкових локальних коливань і створює передумови для використання регресійних, класифікаційних та інших моделей машинного навчання. По-друге, зональний підхід забезпечує більш коректну інтерпретацію впливу кліматичних чинників, оскільки дає змогу аналізувати не випадкові регіональні особливості, а стійкі закономірності в межах однорідніших агрокліматичних груп.

Кластеризація областей у цьому дослідженні виконує роль одного з етапів майнінгу агрокліматичних даних, оскільки дозволяє виявити приховану просторову структуру у сформованому наборі ознак без попереднього задання класів. Отримані групи областей надалі можуть розглядатися як структурні одиниці майбутньої інформаційно-аналітичної системи, для яких формуються окремі аналітичні, прогнозні та ризикові оцінки.

Кореляційний аналіз трендових залишків використано як основний інструмент змістовного групування областей за подібністю міжрічної реакції врожайності. Для додаткової перевірки обґрунтованості виділення трьох зон виконано кластерний аналіз за узагальненими агрокліматичними характеристиками областей і параметрами динаміки врожайності. Таким чином, кореляційний аналіз визначає змістову основу зонування, а кластерний аналіз виконує роль додаткової незалежної перевірки групування областей України.

На відміну від кореляційного аналізу, який ґрунтується на подібності часових рядів трендових залишків врожайності, кластерний аналіз дає змогу оцінити структуру групування областей за узагальненими характеристиками агрокліматичного середовища та параметрами динаміки врожайності [173]. У межах цього етапу кожна область розглядалася як окремий об’єкт кластеризації. Для формування матриці ознак кластерного аналізу використовувалися середні значення агрокліматичних факторів за квітень–червень 2000–2021 років, а також коефіцієнт нахилу тренду врожайності, отриманий під час виділення трендової складової. Структура набору даних охоплювала назву області, середні значення факторів t1–t9, R10, R20, R30 та параметр довгострокової динаміки врожайності.

Набір даних містив 18 об’єктів, що відповідали областям України, відібраним раніше за результатами кореляційного аналізу трендових залишків врожайності. Таким чином, кластеризація виконувалася не за окремими річними спостереженнями, а за узагальненим профілем області, що характеризує її температурно-опадовий режим і довгострокову динаміку врожайності.

Перед кластеризацією ознаки було нормовано, оскільки температурні показники, опади та параметри тренду мають різні одиниці вимірювання й масштаби варіації. Нормування виконувалося за стандартною схемою:

$$z_{ij} = (x_{ij} - \bar{x}_j)/s_j, \tag{2.8}$$

де x_{ij} — значення j-ї ознаки для i-ї області; $\bar{x}_j$ — середнє значення j-ї ознаки; s_j — стандартне відхилення j-ї ознаки; z_{ij} — нормоване значення ознаки. Така процедура забезпечує порівнюваність змінних і зменшує ризик домінування ознак із більшими числовими масштабами.

Оптимальну кількість кластерів оцінювали за кількома критеріями: ліктьовим методом, коефіцієнтом силуету, індексом Davies–Bouldin та індексом Calinski–Harabasz. Використання кількох критеріїв дає змогу зменшити залежність результату від одного показника якості кластеризації та отримати більш стійке рішення щодо кількості груп. Результати графічного оцінювання оптимальної кількості кластерів наведено на рис. 2.10.

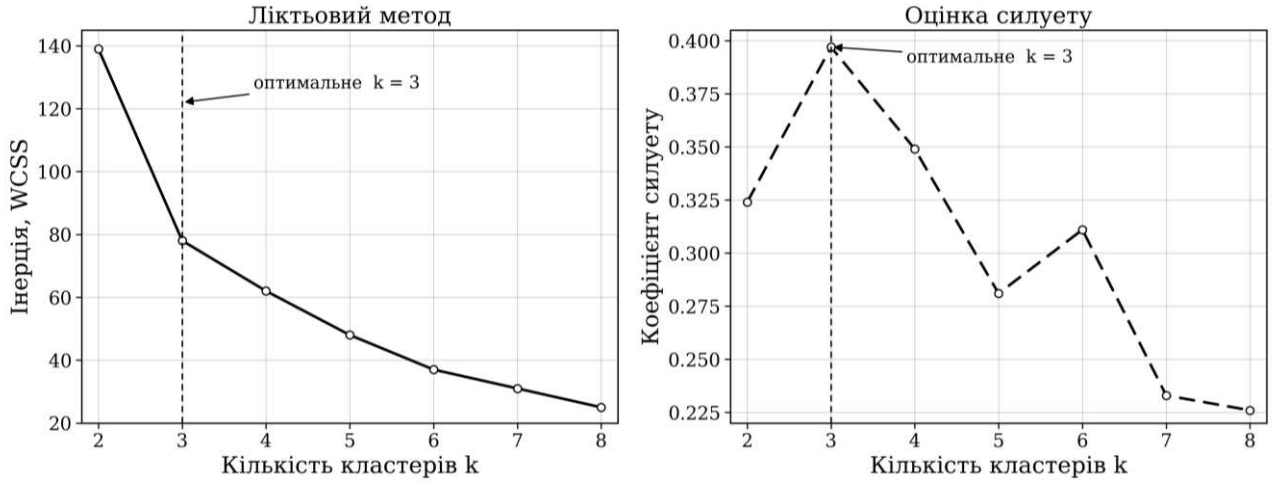

Рис. 2.10. Оцінювання оптимальної кількості кластерів: ліктьовий метод (зліва) та коефіцієнт силуету (справа)

Як видно з рис. 2.10, за ліктьовим методом після трьох кластерів спостерігається помітне сповільнення зменшення інерції WCSS, а коефіцієнт силуету досягає найвищого значення при k=3. Це підтверджує доцільність розбиття областей досліджуваної вибірки на три основні групи та узгоджується з результатами кореляційного аналізу трендових залишків.

Для узагальнення результатів оцінювання кількості кластерів використано також індекси Davies–Bouldin та Calinski–Harabasz. Результати оцінювання оптимальної кількості кластерів за сукупністю критеріїв наведено в табл. 2.5.

Таблиця 2.5

Оцінювання оптимальної кількості кластерів для групування областей України

Метод оцінювання	**Оптимальна кількість кластерів**
Ліктьовий метод	3
Silhouette	3
Davies–Bouldin	8
Calinski–Harabasz	3
Середній ранг за 4 критеріями	3

Як видно з табл. 2.5, більшість використаних критеріїв вказує на доцільність виділення трьох кластерів. Ліктьовий метод, коефіцієнт силуету та індекс Calinski–Harabasz узгоджено підтверджують структуру з трьох груп. Індекс Davies–Bouldin дає інше значення, що може бути пов’язано з наявністю перехідних областей і неоднорідністю окремих регіонів за частиною ознак. Однак за узагальненим ранговим підходом найбільш обґрунтованим є виділення трьох кластерів, що добре узгоджується з попереднім кореляційним аналізом трендових залишків врожайності.

Для додаткової перевірки обґрунтованості агрокліматичного групування областей України використано методи кластерного аналізу. Кластеризація належить до методів навчання без учителя і застосовується тоді, коли необхідно виявити внутрішню структуру даних без попередньо заданих класів або груп. У контексті цього дослідження кластеризація використовується для виявлення груп областей, подібних за сукупністю агрокліматичних характеристик і параметрів динаміки врожайності. На відміну від кореляційного аналізу трендових залишків, який

відображає подібність міжрічної реакції врожайності, кластерний аналіз дозволяє оцінити подібність областей у багатовимірному просторі узагальнених ознак.

Вхідними даними для кластеризації є набір числових характеристик, що описують кожну область України. До такого набору можуть входити середні значення температурних показників, суми опадів за критичний період вегетації, характеристики тренду врожайності та інші узагальнені агрокліматичні показники. Оскільки ці змінні мають різні одиниці вимірювання і різні масштаби варіації, перед застосуванням алгоритмів кластеризації доцільно виконувати стандартизацію даних [173]. Це дозволяє уникнути ситуації, коли змінні з більшими числовими значеннями штучно домінують під час обчислення відстаней або подібності між областями.

Для перевірки стійкості зонального поділу використовувалися метод k-середніх (KMeans), модель гауссової суміші (Gaussian Mixture) та метод спектральної кластеризації (Spectral Clustering) [173]. Використання кількох методів зумовлене тим, що кожен із них ґрунтується на різних припущеннях щодо структури даних. Це дозволяє не лише отримати конкретне розбиття областей на групи, а й перевірити стійкість виявленої кластерної структури до зміни алгоритму кластеризації.

Метод KMeans є одним із найпоширеніших алгоритмів кластеризації. Він ґрунтується на припущенні, що об’єкти можна поділити на компактні групи, кожна з яких характеризується певним центром. Алгоритм ітераційно виконує два основні кроки: спочатку кожний об’єкт відноситься до найближчого центра кластера, а потім центри кластерів перераховуються як середні значення об’єктів, що до них належать. Процедура повторюється до стабілізації кластерів. Формально метод KMeans мінімізує суму квадратів відстаней від об’єктів до центрів відповідних кластерів:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2, \tag{2.9}$$

де K — кількість кластерів, C_k — множина об’єктів, що належать до k-го кластера, x_i — вектор ознак i-го об’єкта, μ_k — центр k-го кластера. Перевагами методу KMeans є простота реалізації, обчислювальна ефективність та добра

інтерпретованість результатів. Водночас метод найкраще працює тоді, коли кластери є відносно компактними та близькими до сферичної форми у просторі ознак.

Метод Gaussian Mixture ґрунтується на припущенні, що емпіричні дані можна подати як суміш кількох багатовимірних нормальних розподілів. Кожен компонент такої суміші відповідає окремому кластеру, а кожний об’єкт має певні ймовірності належності до всіх кластерів. На відміну від KMeans, який виконує жорстке віднесення об’єкта лише до одного кластера, Gaussian Mixture забезпечує імовірнісне групування. Це є важливою перевагою у випадку агрокліматичних даних, оскільки окремі області можуть мати проміжні характеристики і розташовуватися поблизу меж між кластерами. У такому разі імовірнісна інтерпретація дозволяє більш гнучко оцінити їх належність до певної агрокліматичної групи.

У загальному вигляді модель Gaussian Mixture описує густину розподілу даних як суму кількох нормальних компонент:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k), \tag{2.10}$$

де π_k — вага k-го компонента суміші, $N(x|\mu_k, \Sigma_k)$ — багатовимірний нормальний розподіл із вектором середніх μ_k та коваріаційною матрицею Σ_k. Параметри моделі зазвичай оцінюються за допомогою *EM*-алгоритму. У задачі агрокліматичного групування цей метод є корисним для перевірки того, наскільки чітко області відокремлюються між собою та чи існують області з проміжним типом агрокліматичних характеристик.

Метод Spectral Clustering використовує інший підхід до кластеризації. Він базується не безпосередньо на центрах кластерів або параметрах розподілу, а на матриці подібності між об’єктами. Спочатку для всіх пар областей обчислюється міра подібності, після чого формується граф, у якому вершинами є області, а ребра відображають ступінь їх подібності. Далі на основі спектральних властивостей матриці Лапласа цього графа виконується перетворення даних у новий простір, де подальше розбиття на кластери стає більш наочним. Перевагою цього методу є

здатність виявляти складніші кластерні структури, які не обов’язково мають компактну або сферичну форму.

У цьому дослідженні Spectral Clustering використано переважно як додатковий інструмент перевірки стійкості результатів. Якщо різні за своєю логікою методи кластеризації дають подібну кількість кластерів або близьке групування областей, це підвищує довіру до отриманої агрокліматичної структури. Водночас з огляду на невелику кількість об’єктів кластеризації, якими є області України, результати Spectral Clustering потребують обережної інтерпретації та мають розглядатися не ізольовано, а разом з іншими критеріями якості.

Таким чином, KMeans, Gaussian Mixture та Spectral Clustering дозволяють оцінити структуру агрокліматичної подібності областей з різних методичних позицій. KMeans перевіряє наявність компактних груп, Gaussian Mixture дає імовірнісну інтерпретацію кластерної належності, а Spectral Clustering аналізує структуру подібності між областями через графові перетворення. Сукупне використання цих методів дає змогу підвищити обґрунтованість висновку про доцільність виділення трьох укрупнених агрокліматичних зон.

Для оцінки якості кластеризації використовують метрики: Silhouette, Davies–Bouldin та Calinski–Harabasz [173]. Метрика Silhouette оцінює, наскільки об’єкт ближчий до свого кластера, ніж до інших кластерів. Метрика Silhouette набуває значень від –1 до 1, причому більші значення відповідають кращій відокремленості кластерів. Для реальних агрокліматичних даних із малою кількістю об’єктів значення близько 0,35–0,40 можна інтерпретувати як помірно виражену кластерну структуру. Метрика Davies–Bouldin є меншою для компактніших і краще відокремлених кластерів, а Calinski–Harabasz зростає зі збільшенням міжкластерної відокремленості відносно внутрішньокластерного розсіювання. Сукупність використаних критеріїв підтверджує доцільність вибору трьох кластерів.

Для розбиття областей України на агрокліматичні зони нами були використані методи KMeans, Gaussian Mixture та Spectral Clustering. Оцінка кожного з методів наведена у табл. 2.6.

Таблиця 2.6

Оцінка якості методів кластеризації областей України

Метод	**Silhouette**	**Davies–Bouldin**	**Calinski–Harabasz**	**Оцінка**
KMeans	≈ 0,3958	≈ 0,8705	≈ 15,2161	найкращий або один із найкращих
Gaussian Mixture	≈ 0,3930	≈ 0,8660	≈ 14,8518	дуже близький до KMeans
Spectral Clustering	≈ 0,3681	≈ 0,8626	≈ 13,8300	дещо слабший за сукупністю метрик

Дані табл. 2.6 показують, що методи KMeans і Gaussian Mixture дають близькі результати за більшістю показників якості. KMeans має дещо вищі значення Silhouette та Calinski–Harabasz, тоді як Gaussian Mixture демонструє дуже близьку якість кластеризації й водночас є зручним для графічної інтерпретації областей, розташованих поблизу меж кластерів. Spectral Clustering поступається за сукупністю метрик, тому використовувався переважно як додаткова перевірка стійкості кластерної структури.

Для наочного подання результатів кластеризації багатовимірний простір ознак було спроєктовано на площину перших двох головних компонент. Така візуалізація не використовується для побудови самих кластерів, але дозволяє графічно оцінити взаємне розташування областей і відокремленість отриманих груп. Двовимірна ілюстрація кластеризації областей України, виконаної методом Gaussian Mixture представлена на рис. 2.11.

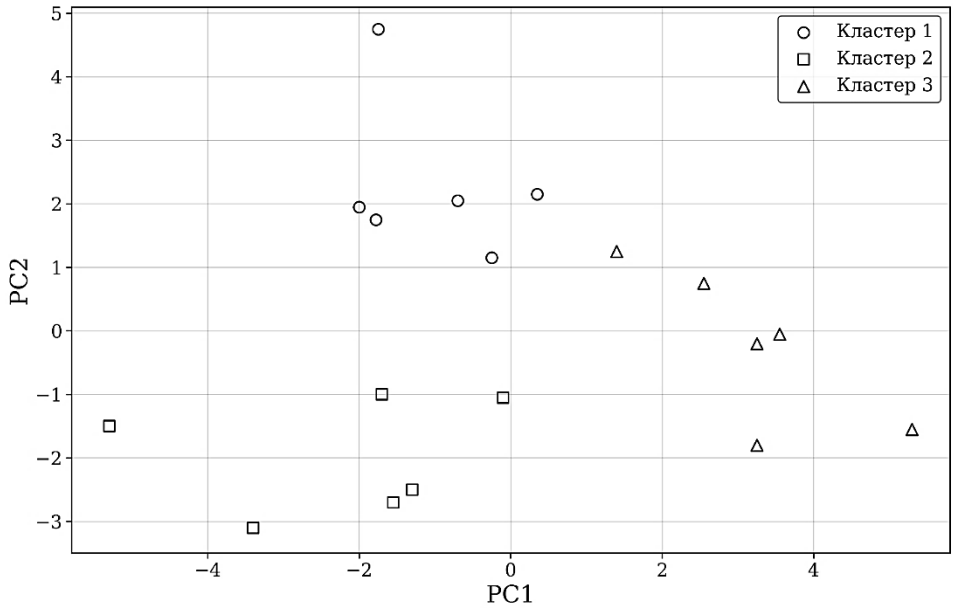

Рис. 2.11. Двовимірна графічна ілюстрація кластеризації областей України, виконаної методом Gaussian Mixture

Отже, кластерний аналіз підтвердив висновки кореляційного аналізу, виконаного раніше, щодо кількості кластерів (агрокліматичних зон України) у кількості 3. Що стосується складу кластерів, визначених методами кластерного аналізу, то він, переважно, збігається з результатами кореляційного аналізу, виконаного раніше. Деякі розходження можуть стосуватися областей, які знаходяться на межі двох зон. У подальших розділах роботи як основа для формування зональних вибірок використовується групування, отримане за результатами кореляційного аналізу трендових залишків врожайності і узгоджене з природно-кліматичною інтерпретацією областей. Зауважимо, що у випадках, коли результати кластеризації можна було трактувати двояко (відстані між об'єктами майже рівні) ми віддавали перевагу географічному принципу об'єднання областей у зони.

Таким чином, агрокліматична кластеризація областей України та подальша регіональна агрегація даних забезпечили формування зональних вибірок для степової, лісостепової та західної зон. Таке групування дозволило збільшити обсяг навчальних даних, зменшити вплив випадкових локальних коливань і створити основу для побудови узагальнених зональних моделей прогнозування врожайності. Сформовані зональні вибірки надалі використовуються для регресійного та класифікаційного моделювання, а також для оцінювання ризику низької врожайності в підрозділі 2.4.

2.4. Оцінювання ризиків низької врожайності та наслідків зовнішніх шоків

Сільськогосподарське виробництво є високоризиковим видом економічної діяльності. Причинами ризиків можуть бути несприятливі кліматичні умови, природні катаклізми (зливи, град, посухи, екстремальні температури), невдала агрономічна діяльність, зміни попиту та цін на аграрну продукцію [46–49; 167]. У підрозділі 2.2 було показано, що абсолютні значення врожайності містять довгострокову трендову складову, пов'язану з технологічними, організаційними та економічними змінами. Тому для оцінювання ризику низької врожайності в роботі

використано трендові залишки, які характеризують відхилення фактичної врожайності від очікуваного трендового рівня [6; 78; 157–159]. Саме від'ємні значення таких залишків інтерпретуються як несприятливі відхилення, що можуть бути використані для кількісного оцінювання аграрного ризику.

У сучасній економічній теорії ризик зазвичай оцінюється на основі статистичного аналізу дохідності активу [167; 176]. Для зерновиробництва як аналог дохідності часто використовується рентабельність виробництва, значення якої наводяться у статистичних джерелах [168]. Проте в умовах української економіки офіційні дані про рентабельність зерновиробництва не завжди дозволяють об'єктивно оцінити дохідність цього виду діяльності та ризики, пов'язані з ним. З огляду на це у даному дослідженні для оцінювання ризику зерновиробництва використовується статистика врожайності, надійність якої є значно вищою [6; 157]. За умови близького рівня технологій у різних регіонах врожайність може розглядатися як адекватна характеристика результативності її вирощування.

Аналіз часових рядів врожайності в областях України показує, що впродовж останніх десятиліть врожайність має стійку тенденцію до зростання [78; 157]. Це пояснюється не лише природно-кліматичними факторами, а й розвитком технологій зерновиробництва, покращенням насіннєвої бази, удосконаленням агротехніки, логістики та інвестиційним пожвавленням галузі. За таких умов безпосереднє використання абсолютних значень врожайності для оцінювання ризику є некоректним, оскільки сама врожайність виступає нестаціонарною змінною [78; 157–159]. З огляду на це для аналізу ризику доцільно перейти від абсолютних значень врожайності до її відхилень від довгострокового тренду [6; 78; 157–159]. Тренд врожайності моделюється лінійною регресійною залежністю (2.6). Коливання врожайності відносно тренду описуються трендовими залишками (2.7). Для оцінки ризиків виробництва у роботі використано статистичні дані про врожайність пшениці в областях України за період 2000–2021 роки [168].

Такий підхід має кілька важливих переваг. По-перше, він дає змогу усунути довгострокову компоненту зростання врожайності та зосередитися на міжрічній

варіативності, яка і є основним носієм ризику. По-друге, трендові залишки дозволяють більш коректно порівнювати регіони між собою, оскільки різниця між ними відображає не стільки рівень технологічного розвитку, скільки реакцію врожайності на змінні зовнішні умови. По-третє, саме трендові залишки найбільш повно відображають вплив погодно-кліматичних умов на динаміку врожайності, що зумовлює їх використання і при агрокліматичному групуванні областей України [6; 22].

Найбільш поширеним статистичним показником ризику є стандартне відхилення [167; 176]. Для нестаціонарних часових рядів врожайності у нашому дослідженні використано модифіковані статистичні міри ризику, що базуються на аналізі трендових залишків [6; 78; 167]. Міра ризику у вигляді стандартного відхилення трендових залишків визначається як

$$R_1 = \sigma_\varepsilon = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^2}, \tag{2.11}$$

де ε_t — трендовий залишок, T — кількість років спостережень. Співвідношення (2.11) дещо перебільшує роль спостережень, які сильно відхиляються від тренду, тому для зменшення впливу екстремальних значень часто використовують середнє абсолютне відхилення

$$R_2 = d_\varepsilon = \frac{1}{T}\sum_{t=1}^{T} |\varepsilon_t|. \tag{2.12}$$

На відміну від стандартного відхилення, цей показник є менш чутливим до поодиноких великих відхилень і дає більш стійку оцінку ризику за наявності аномальних років.

Використаний вище підхід однаково трактує як додатні, так і від'ємні відхилення врожайності від тренду. Це не повністю відповідає економічній природі ризику, який пов'язаний саме з можливістю настання несприятливої події. Оскільки позитивні відхилення врожайності не несуть економічних втрат, для оцінювання аграрного ризику доцільно використовувати семіваріаційні показники, що враховують лише негативні значення трендових залишків [6; 167; 176]. Семіваріація може бути записана у вигляді

$$R_3 = SV = \frac{1}{T}\sum_{t=1}^{T} min(\varepsilon_t, 0)^2. \tag{2.13}$$

На основі семіваріації можна визначити ще дві похідні міри ризику — напівстандартне відхилення та середнє абсолютне негативне відхилення:

$$R_4 = SSV = \sqrt{SV}. \quad (2.14)$$

У співвідношеннях (2.12)–(2.13) для оцінювання ризику використовуються лише негативні відхилення врожайності від тренду, що є особливо важливим у випадку асиметрії закону розподілу трендових залишків. У подальшому використовуються такі позначення: R_1 — стандартне відхилення трендових залишків; R_2 — середнє абсолютне відхилення; R_3 — напівстандартне відхилення; R_4 — середнє абсолютне негативне відхилення.

Використання кількох статистичних мір ризику дає змогу не лише кількісно охарактеризувати мінливість трендових залишків урожайності, а й виконати міжрегіональне порівняння ступеня ризику зерновиробництва. Для цього доцільно застосовувати два взаємодоповнювальні підходи. Перший підхід полягає у безпосередньому порівнянні значень статистичних показників ризику для кожної області. Другий підхід ґрунтується на ранжуванні областей за ступенем спадання ризику, коли ранг 1 присвоюється регіону з найвищим ступенем ризику. Саме така логіка використана в наявних матеріалах дослідження, де ризик оцінювався за кількома показниками R_1–R_4, а підсумкові висновки формувалися на основі їх зіставлення.

Для узагальнення результатів порівняльного аналізу виконано ранжування областей України за ступенем ризику вирощування пшениці. Показники R_1–R_4 були визначені за співвідношеннями (2.11)–(2.14) і відображають різні аспекти мінливості трендових залишків урожайності. Зокрема, частина показників характеризує загальну варіативність ряду, тоді як інші акцентують увагу на несприятливих відхиленнях, тобто на випадках, коли фактична врожайність є нижчою за очікуваний трендовий рівень.

Оскільки окремі статистичні міри можуть по-різному реагувати на амплітуду коливань, асиметрію розподілу та глибину від’ємних залишків, для інтегрального порівняння використано ранговий підхід. Для кожного показника області впорядковувалися за спаданням рівня ризику, причому ранг 1 присвоювався області

з найвищим значенням відповідної міри ризику. Після цього для кожної області обчислювалася сума рангів за всіма використаними показниками. Менша сума рангів відповідає вищому узагальненому рівню ризику, оскільки така область посідає високі позиції одразу за кількома критеріями оцінювання.

На відміну від задач зонального прогнозування, для загального ранжування ризику використано ширшу сукупність областей України. Це пояснюється тим, що статистичні міри ризику можуть бути розраховані для ряду врожайності окремої області незалежно від її належності до однієї з агрокліматичних зон. Такий підхід дає змогу не лише порівняти області всередині сформованих зон, а й отримати загальну картину просторової диференціації ризику низької врожайності. Результати ранжування наведено в табл. 2.7.

Таблиця 2.7

Ранжування областей України за ступенем ризику вирощування пшениці

Область	**R_1**	**R_2**	**R_3**	**R_4**	**Ранг R_1**	**Ранг R_2**	**Ранг R_3**	**Ранг R_4**	**Сума рангів**
Харківська	8,18	6,00	6,78	3,39	2	4	1	1	8
Дніпропетровська	8,29	6,31	5,79	2,90	1	1	4	4	10
Кіровоградська	8,12	6,07	6,31	3,15	3	3	2	2	10
Одеська	7,73	6,21	5,91	2,95	4	2	3	3	12
Черкаська	6,97	5,60	5,66	2,83	6	5	5	5	21
Миколаївська	7,03	5,59	5,24	2,62	5	6	8	8	27
Полтавська	6,67	4,96	5,30	2,65	7	10	6	6	29
Київська	6,28	5,04	5,29	2,65	11	8	7	7	33
Херсонська	6,63	5,17	5,02	2,51	8	7	12	12	39
Донецька	6,61	4,98	5,12	2,56	9	9	11	11	40
Хмельницька	6,32	4,75	5,20	2,60	10	12	10	10	42
Луганська	6,23	4,11	5,23	2,62	12	16	9	9	46
Запорізька	6,11	4,78	4,60	2,30	13	11	14	14	52
Сумська	5,94	4,40	4,78	2,39	14	14	13	13	54
Вінницька	5,78	4,64	4,29	2,14	15	13	16	16	60
Чернівецька	5,46	4,17	4,29	2,15	16	15	15	15	61
Тернопільська	4,82	3,62	4,00	2,00	17	17	17	17	68
Житомирська	4,28	3,21	3,79	1,89	19	19	18	18	74
Чернігівська	4,41	3,38	3,22	1,61	18	18	20	20	76
Івано-Франківська	4,05	2,88	3,29	1,65	20	21	19	19	79
Рівненська	3,82	2,71	3,09	1,54	22	22	21	21	86
Закарпатська	3,89	2,98	2,27	1,14	21	20	24	24	89
Львівська	3,58	2,60	2,83	1,41	24	23	22	22	91
Волинська	3,60	2,56	2,69	1,35	23	24	23	23	93

Дані табл. 2.7 показують, що немає істотної різниці між методами оцінки ризику R_3 та R_4, а відмінність між методами R_1 та R_2 також є невеликою. Близькість отриманих ранжувань свідчить, що показник R_1 може бути використаний як базова узагальнена міра мінливості трендових залишків, тоді як семіваріаційні показники доцільно застосовувати для уточнення ризику саме несприятливих відхилень. Найвищі оцінки ризику за цим підходом характерні для Харківської, Дніпропетровської, Кіровоградської та Одеської областей, тоді як найменші ризики зафіксовано для Рівненської, Волинської, Львівської та Закарпатської областей.

Статистичні методи, розглянуті вище, дають усереднену оцінку ризику для кожної області України, однак не відповідають на питання про ймовірність настання екстремально несприятливих відхилень урожайності. Для переходу до квантильних оцінок ризику необхідно ідентифікувати закон розподілу трендових залишків ε_t. Саме ця задача є центральною для побудови кількісних показників ризику, подібних до тих, що використовуються у фінансовій ризикології. Головна проблема полягає у тому, що для трендових залишків врожайності не можна автоматично приймати гіпотезу про нормальний розподіл; вона повинна бути перевірена емпірично для кожної агрокліматичної зони.

Ми пропонуємо досліджувати закон розподілу не для окремої області, а для об’єднаної вибірки областей, що входять до однієї агрокліматичної зони. Такий підхід обґрунтовується тим, що трендові залишки врожайності найбільш адекватно відображають вплив погодно-кліматичних умов на врожайність, а об’єднання областей за результатами кореляційного аналізу дає змогу отримати статистично більш надійну сукупність спостережень. У попередньому параграфі було виділено три агрокліматичні зони: степову, лісостепову та західну. Для кожної з них було сформовано об’єднану вибірку трендових залишків, що використовується для перевірки гіпотез щодо типу розподілу. Кожна з об’єднаних вибірок містила 132 спостереження, що є достатнім для статистично значущої оцінки закону розподілу.

Для перевірки гіпотези про закон розподілу використовуємо критерій Пірсона:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - m_i)^2}{m_i}, \quad (2.15)$$

де n_i — емпіричні частоти, m_i — теоретичні частоти у i-му інтервалі, k — кількість інтервалів групування. Значення статистики χ^2 порівнювалося з критичним значенням відповідного розподілу з урахуванням кількості інтервалів групування та числа оцінених параметрів. За результатами перевірки визначалося, чи може відповідний теоретичний розподіл бути прийнятий як статистично обґрунтована апроксимація емпіричних даних.

Результати ідентифікації закону розподілу показали, що для західної зони трендові залишки достатньо добре описуються нормальним розподілом. Для лісостепової зони нормальний розподіл також може бути використаний як прийнятна апроксимація, хоча якість узгодження є дещо нижчою. Для степової зони гіпотеза про нормальний розподіл була відхилена, а сам розподіл трендових залишків виявився розподілом із «важкими хвостами», який найкраще апроксимується розподілом Лапласа [6; 14; 20; 174; 175]. Використання асиметричного розподілу Лапласа дає лише незначні поправки, що свідчить про малу асиметрію емпіричного розподілу.

Кумулятивна функція нормального розподілу задається співвідношенням

$$F(z) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^z e^{-t^2} dt\,. \quad (2.16)$$

Тут $z = \frac{x-\mu}{\sigma\sqrt{2}}$; μ — математичне сподівання випадкової величини x; σ — середньоквадратичне відхилення x.

Кумулятивна функція розподілу Лапласа має вигляд

$$F_L(x) = \begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{\gamma}\right), & x < \mu, \\ 1 - \frac{1}{2}\exp\left(\frac{\mu-x}{\gamma}\right), & x \geq \mu, \end{cases} \quad (2.17)$$

де μ — параметр положення, γ — параметр масштабу. Для степової зони як параметр положення доцільно використовувати медіану розподілу трендових залишків, яка за наявними розрахунками становить 0,81 ц/га. Використання оптимізаційного підходу дозволило нам визначити оптимальне значення параметра γ розподілу Лапласа.

Аналіз табл. 2.8 показує, що закон розподілу трендових залишків не є однаковим для всіх агрокліматичних зон України. Це означає, що квантильні оцінки ризику мають будуватися з урахуванням регіональної специфіки розподілу, а автоматичне використання нормального закону для всіх зон є методично некоректним.

Таблиця 2.8

Ідентифікація закону розподілу трендових залишків врожайності за агрокліматичними зонами

Агрокліматична зона	**Склад зони**	**Найкраща апроксимація**
Степова	Одеська, Херсонська, Миколаївська, Запорізька, Дніпропетровська, Кіровоградська	Розподіл Лапласа
Лісостепова	Сумська, Харківська, Полтавська, Київська, Черкаська, Вінницька	Нормальний розподіл, дещо гірша узгодженість
Західна	Волинська, Рівненська, Житомирська, Львівська, Тернопільська, Хмельницька	Нормальний розподіл

Для кількісного оцінювання нижньохвостового ризику низької врожайності у роботі використано квантильні міри VaR та CVaR. Оскільки ризик низької врожайності пов’язаний саме з від’ємними трендовими залишками, для подальших розрахунків введено величину втрат урожайності відносно тренду:

$$L_t = \max(0, -\varepsilon_t), \tag{2.18}$$

де ε_t — трендовий залишок урожайності у році t, а L_t — величина несприятливого відхилення врожайності від трендового рівня. Такий запис дає змогу перейти від аналізу залишків, які можуть бути як додатними, так і від’ємними, до додатної величини втрат, що має безпосередню економічну інтерпретацію.

Для рівня довіри α показник VaR_α визначається як квантиль розподілу втрат:

$$VaR_\alpha = q_\alpha(L). \tag{2.19}$$

У цьому дослідженні VaR_α інтерпретується як гранична величина втрат урожайності відносно тренду, яка з імовірністю α не буде перевищена. На відміну від стандартного відхилення, цей показник характеризує не загальну мінливість

урожайності, а саме нижній хвіст розподілу, тобто область несприятливих відхилень.

Для глибшої характеристики хвостового ризику використано також показник $CVaR_\alpha$, який визначає середню величину втрат за умови, що вони перевищують поріг VaR_α:

$$CVaR_\alpha = \mathrm{E}(L|L \geq VaR_\alpha). \quad (2.20)$$

Отже, $CVaR_\alpha$ показує очікуваний рівень втрат у найбільш несприятливих випадках і є більш інформативною мірою ризику для аналізу екстремально низької врожайності. У роботі розрахунки VaR та CVaR виконано для рівня довіри 90 %. При цьому $VaR_{90\%}$ — це порогове значення втрат врожайності, яке дорівнює втратам врожайності, які настають з імовірністю 10 %; $CVaR_{90\%}$ — середнє значення втрат серед найгірших 10 % випадків. Для лісостепової та західної зон використано нормальну апроксимацію розподілу трендових залишків, а для степової зони — розподіл Лапласа, що краще враховує важчі хвости емпіричного розподілу.

Таблиця 2.9

Міри ризику VaR і CVaR для різних агрокліматичних зон (рівень довіри 90 %)

	Степова зона	**Лісостепова зона**	**Західна зона**
Розподіл залишків	Лапласа	нормальний	нормальний
VaR, (ц/га)	9,43	8,40	5,67
CVaR, (ц/га)	15,84	11,51	7,76

Наведене у табл. 2.9 значення CVaR дає оцінку можливих втрат врожайності (ц/га) у порівнянні з очікуваним трендовим значенням, які можуть настати з імовірністю 10 % і менше. Як показує аналіз табл. 2.9, міра ризику CVaR дає оцінку ризику на 37% вищу, ніж міра ризику VaR у випадку нормального розподілу; міра ризику CVaR дає оцінку ризику на 68% вищу, ніж міра ризику VaR у випадку розподілу Лапласа. З таблиці видно, що степова зона є найбільш ризикованим регіоном з точки зору вирощування пшениці. Найнижчі значення VaR і CVaR отримано для західної зони, що узгоджується з меншою вираженістю посушливих умов порівняно зі степовою зоною.

Квантильна логіка є важливою також для подальшої класифікаційної постановки задачі прогнозування, оскільки дозволяє формально визначати область несприятливих значень трендових залишків і використовувати її для побудови цільового класу низької врожайності. У пов'язаних матеріалах низькі значення врожайності визначалися на основі положення спостереження на інтегральній кривій розподілу трендових залишків. Це демонструє, що квантильна логіка є корисною не лише для безпосереднього оцінювання ризику, а й для побудови класифікаційних моделей урожайності.

Практичне значення квантильного підходу полягає в тому, що він дозволяє перейти від загальної характеристики мінливості до оцінювання імовірності конкретних втрат. На відміну від стандартного відхилення чи семіваріації, які дають усереднену оцінку нестабільності врожайності, показник VaR має чітку інтерпретацію як поріг можливого несприятливого відхилення. Це робить його зручним інструментом для аграрного менеджменту, регіонального планування та побудови сценаріїв підтримки зерновиробництва в умовах кліматичної невизначеності.

Узагальнену послідовність оцінювання ризику низької врожайності на основі трендових залишків наведено на рис. 2.12.

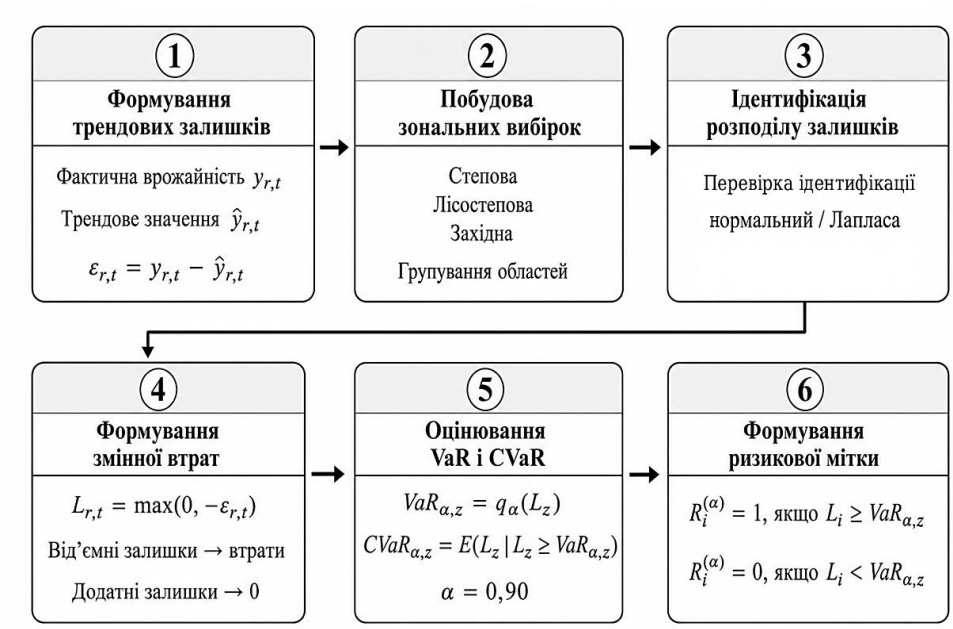

Рис. 2.12. Послідовність оцінювання ризику низької врожайності на основі трендових залишків і змінної втрат

На рис. 2.12 узагальнено послідовність переходу від фактичної врожайності до кількісного оцінювання ризику низької врожайності. Спочатку для кожної області формуються трендові залишки як різниця між фактичним і трендовим значеннями врожайності. Далі залишки об’єднуються в межах агрокліматичних зон, для яких ідентифікується відповідний закон розподілу. Для безпосередньої економічної інтерпретації ризику від’ємні залишки перетворюються на величину втрат відносно тренду. На основі розподілу цієї величини визначаються показники VaR і CVaR, а також може формуватися ризикова мітка для подальших класифікаційних моделей.

Таким чином, статистичні та квантильні методи оцінювання ризику низької врожайності, побудовані на основі розподілів трендових залишків, дозволили формалізувати поняття ризику в задачах зерновиробництва [61; 73; 176; 177]. Це створює підґрунтя для кількісного порівняння несприятливих сценаріїв і може бути використано для підтримки страхових, інвестиційних і управлінських рішень у рослинництві [6; 14; 20].

Отримані ризикові оцінки характеризують типові несприятливі відхилення врожайності, однак для прикладного аналізу важливо також розглянути випадки, коли такі відхилення посилюються дією позакліматичних чинників.

Окремим напрямом прикладного використання прогнозних моделей є оцінювання втрат врожайності за умов зовнішніх шоків. Оцінювання наслідків зовнішніх шоків у роботі розглядається як приклад прикладного використання трендових і прогнозних моделей. Повномасштабна війна, яка розпочалася у 2022 році, суттєво трансформувала умови функціонування зерновиробництва в Україні, зумовивши не лише руйнування виробничої та логістичної інфраструктури, а й скорочення площ обробітку, зростання витрат на ресурси, дефіцит робочої сили та порушення технологічної дисципліни. Для оцінювання втрат врожайності, пов’язаних із військовими діями, ми порівняли фактичну врожайність для 2022 року з очікуваним рівнем, сформованим на основі трендової складової та кліматичної поправки. Різниця між фактичним значенням врожайності і модельним очікуванням інтерпретується як орієнтовна додаткова втрата, що не пояснюється

звичайною кліматичною мінливістю [10; 178]. Така оцінка не претендує на повне причинне вимірювання воєнного впливу, але дозволяє кількісно окреслити масштаб можливих втрат і показати придатність розробленого інструментарію для сценарного аналізу кризових подій.

Опишемо методику оцінювання втрат врожайності внаслідок військових дій. На першому етапі для кожної області за даними 2000–2021 рр. було побудовано лінійну трендову модель врожайності. На другому етапі визначалося відхилення фактичної врожайності від тренду. На третьому етапі для кожної агрокліматичної зони нами побудовано регресійні моделі кліматичної складової ε з використанням множинної лінійної регресії. Використовуючи кліматичні дані 2022 року, ми обчислили теоретичне значення кліматичної поправки $\hat{\varepsilon}_{r,2022}$ для кожної області. Після цього було оцінено військово зумовлену втрату врожайності:

$$dy_r = y_{r,2022} - \hat{y}^{tr}_{r,2022} - \hat{\varepsilon}_{r,2022}. \quad (2.21)$$

Якщо фактична врожайність 2022 р. виявлялася нижчою за суму трендового прогнозу та кліматичної поправки, різниця інтерпретувалася як додаткова втрата врожайності, пов’язана з військовими діями. Оцінювання втрат валового збору здійснювалося за формулою:

$$\Delta Q_r = dy_r \cdot S_r/10, \quad (2.22)$$

де S_r — посівна площа пшениці у 2022 р., тис. га; ΔQ_r — втрати валового збору, тис. т. Ділення на 10 використовується для перерахунку з центнерів у тонни, оскільки dy_r вимірюється в ц/га (1 ц = 0,1 т).

Найбільшу оцінку втрат врожайності dy^* отримано для Херсонської області — 36,90 ц/га. Водночас у розрахунку втрат валового збору ΔQ_r для цієї області результат не враховано через відсутність повних даних щодо посівної площі пшениці у 2022 році. Серед областей, для яких розрахунок ΔQ_r є повним, найбільші втрати врожайності зафіксовано у Вінницькій (10,67 ц/га), Житомирській (9,03 ц/га), Харківській (9,00 ц/га), Київській (8,70 ц/га), Тернопільській (6,51 ц/га) та Полтавській (6,01 ц/га) областях. Для Сумської області значення dy було додатним, що означає відсутність додаткових втрат урожайності.

У перерахунку на втрати валового збору найбільші оцінки втрат нами отримано для Одеської області — 374,23 тис. т, Харківської — 369,49 тис. т, Вінницької — 341,90 тис. т, Дніпропетровської — 213,23 тис. т та Київської — 181,47 тис. т. Значні втрати також характерні для Полтавської (161,68 тис. т), Тернопільської (144,24 тис. т), Житомирської (141,19 тис. т), Миколаївської (136,51 тис. т) та Хмельницької (134,58 тис. т) областей (рис. 2.13). Сумарні втрати валового збору пшениці за розрахованою вибіркою становили 2580,17 тис. т, або 2,58 млн т. Отримані нами результати узгоджуються з макрорівневими висновками, поданими у статті [178].

Отже, використання трендових залишків як основи ризикового аналізу дало змогу перейти від оцінювання загальної мінливості врожайності до кількісного описання несприятливих нижньохвостових відхилень.

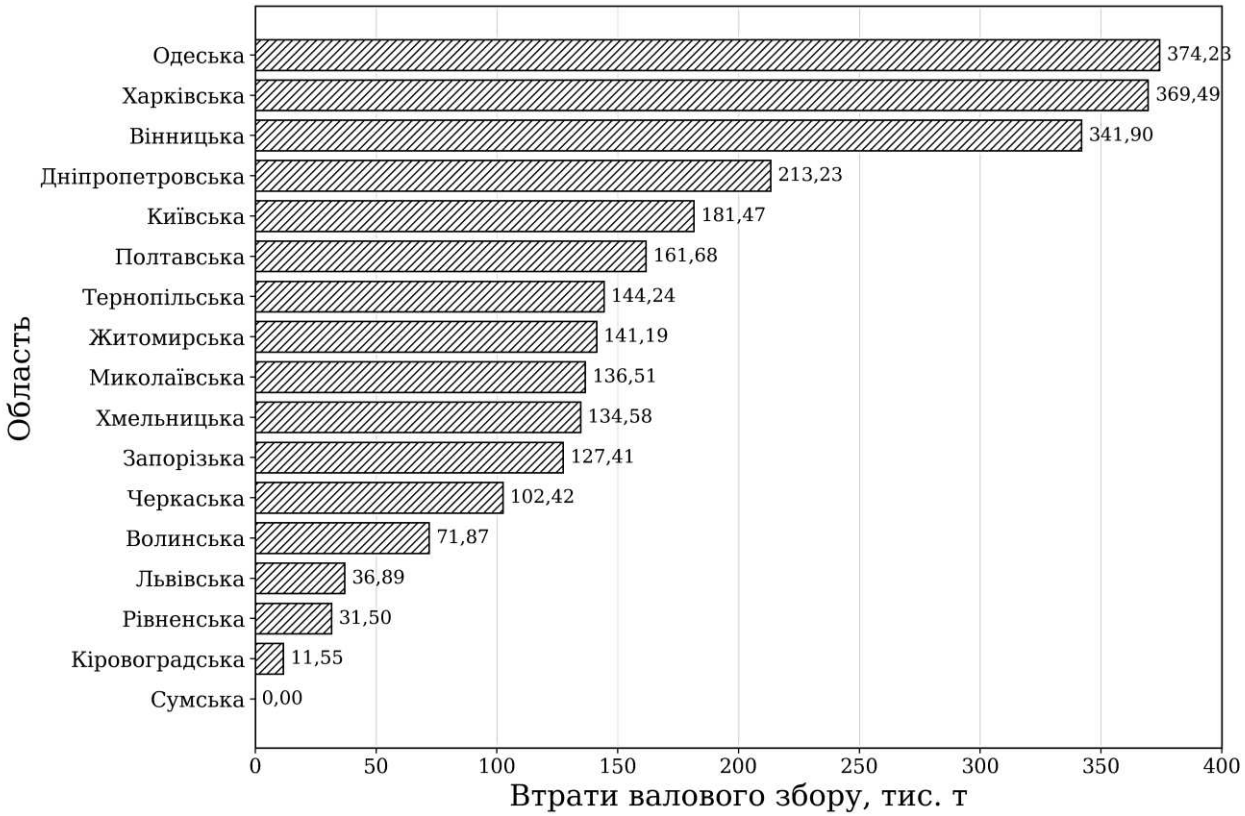

Рис. 2.13. Втрати валового збору пшениці за областями України у 2022 році, тис. т

Ідентифікація розподілів залишків за агрокліматичними зонами дозволила обґрунтувати застосування нормального розподілу для лісостепової та західної зон і розподілу Лапласа для степової зони, що створило основу для розрахунку VaR і CVaR. Приклад оцінювання наслідків зовнішнього шоку 2022 року показав, що

запропонований інструментарій може бути використаний також для сценарного аналізу кризових відхилень, однак такі оцінки мають розглядатися як модельні та орієнтовні.

Висновки до розділу 2

У розділі 2 сформовано інформаційну основу для подальшої розробки інформаційно-аналітичної системи прогнозування та оцінювання ризиків урожайності. Для цього побудовано агрокліматичну матрицю ознак виду «область–рік–t1–t9–R10, R20, R30» для 18 областей України за період 2000–2021 рр. на основі просторово агрегованих даних ERA5 і полігонів ADM1. Сформована матриця є структурованим інформаційним ресурсом, придатним для використання у програмних модулях зберігання, попереднього опрацювання, аналітичного оцінювання, прогнозування та візуалізації результатів в інформаційно-аналітичній системі.

Розвідувальний аналіз даних підтвердив повноту, однозначність і коректність сформованої вибірки, а також дозволив виявити характер варіативності та кореляційних зв’язків між агрокліматичними ознаками. Це забезпечило підготовку вхідного масиву даних до подальшого автоматизованого опрацювання та створило передумови для його використання як єдиного інформаційного шару в системі підтримки прийняття рішень у сфері аграрного виробництва.

Обґрунтовано використання трендових залишків урожайності як інформаційного показника, що характеризує відхилення фактичної врожайності від її довгострокового трендового рівня. Такий підхід дозволяє відокремити довгострокову тенденцію зростання врожайності, зумовлену технологічними, організаційними та економічними чинниками, від міжрічної мінливості, пов’язаної переважно з агрокліматичними умовами. Сформовані ряди трендових залишків можуть використовуватися в інформаційно-аналітичній системі як основа для ідентифікації сприятливих і несприятливих років, оцінювання ризику низької врожайності та формування аналітичних повідомлень для користувача.

Виконано агрокліматичне групування областей України та показано доцільність виділення трьох укрупнених агрокліматичних зон: степової, лісостепової та західної. Таке групування забезпечує регіональну структуризацію інформаційної бази, дає змогу врахувати просторову неоднорідність агрокліматичних умов і створює основу для побудови зональних аналітичних модулів інформаційної системи. Регіональна агрегація даних також зменшує вплив випадкових локальних коливань і підвищує надійність подальшого аналізу врожайності.

Додаткова перевірка методами кластерного аналізу підтвердила обґрунтованість виділення трьох укрупнених агрокліматичних зон. Результати оцінювання оптимальної кількості кластерів за критеріями Elbow, Silhouette, Davies–Bouldin і Calinski–Harabasz загалом узгоджуються з результатами кореляційного аналізу трендових залишків. Це дозволяє розглядати виділені зони не лише як географічні або природно-кліматичні групи, а і як структурні одиниці інформаційної системи, для яких можуть формуватися окремі аналітичні, прогнозні та ризикові оцінки. Таким чином, нами виконано поставлене раніше завдання щодо удосконалення формування агрокліматичного ознакового простору для моделювання врожайності шляхом просторової агрегації кліматичних показників, побудови трендових залишків і зонального групування регіональних спостережень.

У рамках завдання щодо удосконалення методичних засад оцінювання ризику низької врожайності запропоновано статистичний і квантильний підходи до оцінювання ризику низької врожайності на основі трендових залишків. Показано, що закон розподілу залишків має зональну специфіку: для степової зони характерні важчі хвости розподілу, що зумовлює доцільність використання розподілу Лапласа, тоді як для лісостепової та західної зон прийнятною апроксимацією є нормальний розподіл. За показниками VaR і CVaR встановлено, що степова зона характеризується найвищим рівнем ризику несприятливих відхилень урожайності від трендового рівня, а західна зона — найнижчим. Отримані ризикові показники можуть бути використані як кількісні індикатори в інформаційно-аналітичній системі для виявлення зон підвищеного ризику та підтримки управлінських рішень.

Показано можливість використання сформованої інформаційної бази для орієнтовного оцінювання наслідків зовнішніх шоків. На прикладі впливу військових дій на виробництво пшениці продемонстровано, що відхилення фактичної врожайності від очікуваного рівня може бути використане для модельного оцінювання додаткових втрат, які не пояснюються довгостроковим трендом і кліматичною складовою. Такий підхід має прикладне значення для розробки сценарного модуля інформаційно-аналітичної системи, призначеного для аналізу кризових відхилень у зерновиробництві.

Отже, у другому розділі сформовано ознаковий, часовий, просторовий і ризиковий базис інформаційно-аналітичної системи прогнозування врожайності. Отримані результати забезпечують підготовку структурованого набору даних, виділення регіональних агрокліматичних зон, формування показників трендових відхилень і кількісних індикаторів ризику. У сукупності ці процедури утворюють етап майнінгу агрокліматичних даних для майбутньої інформаційної системи, що створює інформаційно-методичну основу для подальшої реалізації програмних модулів прогнозування, оцінювання ризиків, сценарного аналізу та підтримки прийняття рішень у системі аграрної аналітики.

РОЗДІЛ 3. ЕНДОГЕННІ МОДЕЛІ МАШИННОГО ТА ГЛИБИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ

3.1. Методика побудови прогнозної моделі врожайності

Використовуючи сформований у розділі 2 ознаковий простір та відповідну базу даних, перейдемо до побудови моделей машинного навчання з метою прогнозування врожайності пшениці. Спочатку виконаємо постановку задач прогнозування, опишемо структуру вхідних і вихідних даних, схему експериментальних досліджень, а також критерії оцінювання якості моделей.

3.1.1. Ендогенний та екзогенний підходи до прогнозування врожайності

У нашому дослідженні прогнозування врожайності розглянуто два підходи до формування вхідних даних для моделей машинного та глибинного навчання: ендогенний та екзогенний.

Ендогенний підхід передбачає використання лише внутрішньої інформації самого часового ряду врожайності або його похідних характеристик. У межах цього підходу модель навчається на попередніх значеннях показника, зокрема трендових залишках урожайності, і намагається виявити внутрішні закономірності їх зміни в часі. Перевагою ендогенного підходу є його простота та відсутність потреби у залученні додаткових факторних змінних. Водночас його ефективність може бути обмеженою у випадку коротких часових рядів або слабкої автокореляції досліджуваного показника.

Екзогенний підхід ґрунтується на включенні до моделі зовнішніх факторів, які потенційно впливають на формування врожайності. До таких факторів можуть належати температурні показники, кількість опадів, агрометеорологічні умови в окремі періоди вегетації, а також економічні або технологічні змінні. У цьому випадку модель прогнозує клас або рівень урожайності не лише на основі попередньої динаміки самого показника, а й з урахуванням зовнішніх умов, що визначають розвиток рослин. Перевагою екзогенного підходу є можливість змістовної інтерпретації впливу факторів та підвищення прогнозної точності,

особливо тоді, коли врожайність істотно залежить від кліматичних і економічних умов.

У цьому розділі основну увагу зосереджено на ендогенній класифікаційній постановці, у якій прогнозування класу врожайності виконується на основі попередніх значень трендових залишків. Екзогенний підхід із використанням агрокліматичних та економічних факторів розглядається у розділі 4.

Поділ прогнозних моделей на підходи, що використовують лише внутрішню динаміку досліджуваного часового ряду, та підходи, які додатково включають зовнішні предиктори, узгоджується із загальною логікою сучасного прогнозування часових рядів, зокрема з концепцією динамічної регресії та моделей із зовнішніми змінними [160; 161].

3.1.2. Класифікаційний підхід до прогнозування врожайності

Методичною основою подальшого моделювання є припущення, що для аналізу впливу агрокліматичних факторів доцільно використовувати не лише абсолютні значення врожайності, а й її детрендовану форму. Як було обґрунтовано в підрозділі 2.2, трендові залишки дозволяють відокремити довгострокову тенденцію зростання врожайності, зумовлену розвитком технологій та організаційних умов виробництва, від міжрічної мінливості, що найбільшою мірою відображає дію погодно-кліматичних чинників. Саме тому в дисертації використовуються дві основні постановки задач: прогнозування числового значення трендових залишків врожайності та прогнозування її належності до одного з двох класів — низької або високої врожайності.

У класифікаційній постановці задача формулюється як визначення належності спостереження до класу «низька врожайність» або «висока врожайність». Такий підхід має особливе прикладне значення, оскільки для управлінських рішень, пов’язаних з оцінюванням ризиків, логістикою, фінансовим плануванням та страхуванням, часто більш важливим є саме раннє виявлення несприятливого сценарію, ніж точне числове передбачення врожайності. У цьому разі на основі розподілу трендових залишків або іншого обраного критерію

формується бінарна цільова змінна, яка й використовується для навчання класифікаційних моделей.

Методичною основою класифікаційного підходу є використання трендових залишків врожайності ε, визначених як відхилення фактичної врожайності від її лінійного тренду [6; 24; 78]. Саме такий показник, як було обґрунтовано в підрозділі 1.3, найбільш повно відображає дію погодно-кліматичних факторів. Для побудови моделей класифікації знову використаємо об’єднану вибірку для шести областей лісостепової зони України: Сумської, Харківської, Полтавської, Київської, Черкаської та Вінницької. Сукупний обсяг цієї вибірки становить 132 спостереження, що забезпечує достатню статистичну основу для застосування методів машинного навчання.

Першим етапом класифікаційного підходу є перевірка гіпотези про закон розподілу трендових залишків врожайності ε. У праці [1] наведено результати, згідно з якими для об’єднаної вибірки лісостепової зони найбільш близьким до фактичного розподілу трендових залишків є нормальний закон, хоча рівень надійності цього висновку є дещо нижчим за 95%. Для подальшої класифікації в дослідженні прийнято припущення про нормальний розподіл трендових залишків ε, що дозволяє використати інтегральну криву розподілу для визначення порога між низькою та високою врожайністю. Інтегральну криву розподілу трендових залишків врожайності для лісостепової зони наведено на рис. 3.1.

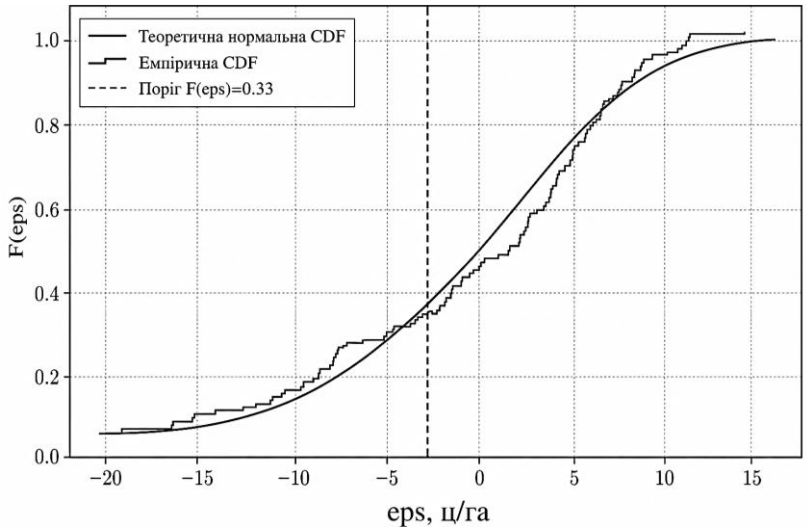

Рис. 3.1. Інтегральна крива розподілу трендових залишків врожайності для лісостепової зони України

На рис. 3.1 показано емпіричну та теоретичну інтегральні криві розподілу трендових залишків врожайності ε для сукупності областей лісостепової зони. Вертикальна лінія відповідає порогу $F(\varepsilon)$=0,33, який використовується для поділу спостережень на класи «low yield» та «high yield». Такий підхід дозволяє перейти від задачі числового прогнозування до задачі категоріального розпізнавання рівня майбутньої врожайності.

У нашому дослідженні основна увага приділяється прогнозуванню саме низьких урожаїв, оскільки такі випадки створюють найбільші ризики для виробників зерна. У роботі використано два способи формування класів врожайності. У більш ранніх експериментах [1] клас «low yield» визначався за квантильним правилом

$$F(\varepsilon) < 0{,}33, \tag{3.1}$$

де $F(\varepsilon)$ — значення інтегральної функції розподілу для відповідного значення ε. Якщо умова (3.1) не виконується, спостереження відноситься до групи «high yield». Для реалізації класифікаційного підходу вводиться бінарна змінна ε_1, яка визначається за правилом

$$\varepsilon_1 = \begin{cases} 1, \text{якщо } F(\varepsilon) < 0.33, \\ 0, \text{в іншому випадку.} \end{cases} \tag{3.2}$$

У роботі [1] наведено оцінку часток класів: кількість випадків «low yield» становить близько 28,8 % від усіх спостережень, а «high yield» — 71,2%. Це означає, що задача класифікації є незбалансованою, а тому при оцінюванні якості моделей особливу увагу слід приділяти не лише загальній точності, а й чутливості до розпізнавання низьких урожаїв.

У більш пізніх наших дослідженнях [24] був використаний більш простий підхід до класифікації трендових залишків врожайності. При цьому правило класифікації визначається наступним чином:

$$class = \begin{cases} 1, \text{якщо } \varepsilon < 0 \ (low\ yield), \\ 0, \text{якщо } \varepsilon \geq 0 \ (high\ yield). \end{cases} \tag{3.3}$$

Якщо $\varepsilon < 0$ спостереження відносять до групи «низька врожайність»; інакше воно класифікується як «висока врожайність». Саме це правило буде застосовано нами нижче при дослідженні класифікаційних моделей врожайності.

3.1.3. Методи прогнозування часових рядів

Бінаризація трендових залишків врожайності відкриває можливість застосування різних прогнозних моделей класифікації. При цьому часовий ряд трендових залишків розглядається як послідовність спостережень, що накладає певні обмеження на формування навчальної та контрольної вибірки. Зазвичай у моделях машинного навчання відбір даних до навчальної вибірки виконується випадковим чином, але для моделей часових рядів потрібно застосовувати хронологічне розбиття, що унеможливлює потрапляння майбутніх спостережень до навчальної вибірки. Контрольна вибірка у часовому розрізі повинна бути розташована після навчальної вибірки.

Для розв'язання задачі бінарної класифікації врожайності використовують методи машинного навчання: логістичну регресію, лінійний дискримінантний аналіз, Support Vector Machine з лінійним ядром, Support Vector Machine з радіальною функцією ядра та Random Forest. Для підвищення статистичної надійності отриманих прогнозних оцінок використовують процедуру перехресної перевірки з подальшим усередненням результатів. Такий підхід дає змогу зменшити залежність отриманої моделі від випадкового розбиття вибірки на навчальну та тестову частини. Ефективність моделей машинного навчання залежить від наявності часових залежностей у динаміці врожайності. Одним із завдань нашого дослідження є спроба вияснити, чи може глибинне навчання підвищити якість прогнозування порівняно з класичними методами машинного навчання. З цією метою у роботі проведено дослідження моделей LSTM та GRU в ендогенній постановці. Водночас у дисертації враховується обмеження, пов'язане з короткою довжиною часових рядів, що потребує обережної інтерпретації результатів нейромережевого моделювання.

Крім моделей машинного та глибинного навчання до прогнозування часових рядів можна застосувати інші методи. Модель ARIMA є класичною статистичною моделлю прогнозування часових рядів, яка поєднує авторегресійну складову, операцію інтегрування та складову ковзного середнього. Назва ARIMA походить від *Autoregressive Integrated Moving Average*. Модель позначається як

ARIMA(p, d, q), де p — порядок авторегресійної частини, d — порядок чисельного диференціювання часового ряду для досягнення стаціонарності, а q — порядок моделі ковзного середнього [160; 161].

Авторегресійна складова враховує залежність поточного значення ряду від його попередніх значень, інтегрована складова використовується для усунення тренду та приведення ряду до стаціонарного вигляду, а складова ковзного середнього описує вплив попередніх випадкових похибок на поточне значення. У загальному вигляді модель ARIMA застосовується тоді, коли прогнозування базується лише на внутрішній динаміці самого часового ряду, без залучення зовнішніх факторів [160; 161].

У задачах прогнозування врожайності модель ARIMA може розглядатися як базовий ендогенний підхід, оскільки вона використовує лише попередні значення врожайності або трендових залишків. Водночас її ефективність істотно залежить від наявності автокореляційної структури в часовому ряді. Якщо ряд трендових залишків є коротким і слабкокорельованим, можливості ARIMA-моделей щодо побудови точного прогнозу суттєво обмежуються [160; 161].

3.1.4. Метрики оцінювання якості класифікаційних моделей

Для оцінювання якості прогнозних моделей класифікаційного типу у роботі використовують наступну матрицю помилок [64] (табл. 3.1)

Таблиця 3.1

Матриця помилок у загальному вигляді

		Predicted / Classified		
		– (Negative)	**+ (Positive)**	
True	**(Negative)**	True Negative (TN)	False Positive (FP)	Overall True Negative
	(Positive)	False Negative (FN)	True Positive (TP)	Overall True Positive
		Overall Predicted Negative	Overall Predicted Positive	

Матриця помилок містить інформацію про такі результати класифікації: істинно позитивний (True Positive) — це результат, коли модель правильно

прогнозує позитивний клас. Зауважимо, що позитивним результатом при класифікації вважається результат, який є негативним у побутовому розумінні (низька врожайність). Істинно негативним (True Negative) вважається результат, при якому модель правильно прогнозує негативний клас (висока врожайність). Хибнопозитивний результат (False Positive) — це результат, коли модель неправильно прогнозує позитивний клас (прогноз — низька врожайність; фактично — висока врожайність). Хибнонегативним (False Negative) вважається результат при якому модель неправильно прогнозує негативний клас.

Для класифікаційних моделей основними критеріями якості є загальна точність класифікації $Accuracy$, чутливість $Recall$, специфічність $Specificity$, точність позитивна $Precision$ та площа під ROC-кривою AUC [64; 65]. Загальна точність визначається як частка правильно класифікованих спостережень:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{3.4}$$

де TP — істинно позитивні, TN — істинно негативні, FP — хибнопозитивні, FN — хибнонегативні результати.

Якщо позитивним результатом вважати низьку врожайність, то при цьому: чутливість характеризує частку правильно виявлених випадків низької врожайності:

$$Recall = \frac{TP}{TP+FN}, \tag{3.5}$$

специфічність — частку правильно виявлених випадків високої врожайності:

$$\text{Specificity} = \frac{TN}{TN+FP}, \tag{3.6}$$

точність позитивна — частку випадків низької врожайності серед усіх випадків, віднесених класифікатором до цього класу:

$$Precision = \frac{TP}{TP+FP}. \tag{3.7}$$

Крім наведених вище, використовується також оцінка $F1$, яка поєднує показники чутливості та точності в один показник ефективності [64]. $F1$ — це середньозважене значення точності та чутливості. Тому ця оцінка враховує як

хибнопозитивні, так і хибнонегативні результати. Метрика $F1$ зазвичай є кориснішою за точність, особливо якщо маємо нерівномірний розподіл за класами.

$$F1 = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}. \quad (3.8)$$

Використовується також показник AUC (площа під ROC-кривою [65]) як інтегральний критерій якості бінарного класифікатора, який дозволяє порівнювати класифікаційні моделі.

Слід пам'ятати, що у класифікаційних моделях прогнозування врожайності результатом роботи моделі є не безпосередньо клас, а прогнозна ймовірність належності спостереження до класу низької врожайності. Для перетворення цієї ймовірності у бінарний клас використовується поріг класифікації [64; 66]. Базовим варіантом є фіксований поріг τ=0,5; якщо прогнозна ймовірність низької врожайності не менша за 0,5, спостереження належить до класу низької врожайності, інакше — до класу високої врожайності.

Крім того, часто використовують оптимізаційний поріг класифікації, який визначається на валідаційній вибірці. У цьому випадку перевіряється кілька можливих значень порогу, а оптимальним вважається те, за якого досягається найбільше значення метрики $F1$ для класу низької врожайності. Такий підхід дозволяє краще збалансувати точність і повноту виявлення несприятливих років. Недоліком цього підходу є не зовсім чітка інтерпретованість, оскільки для різних вибірок значення цього порогу може бути різним.

Для моделей типу ARIMA, які у стандартній постановці прогнозують числове значення часового ряду, поріг τ=0,5 безпосередньо не використовується. У цьому випадку спочатку прогнозується значення трендового залишку ε_t, а класифікаційне рішення може бути прийняте за знаком прогнозу: *якщо* $\varepsilon_t \leq 0$, прогнозується клас низької врожайності, *якщо* $\varepsilon_t > 0$ — клас високої врожайності.

3.1.5. Підбір гіперпараметрів прогнозних моделей

Якість прогнозної моделі значною мірою залежить не лише від типу використаного алгоритму, а й від вибору його гіперпараметрів [58]. На відміну від

параметрів моделі, які оцінюються безпосередньо під час навчання, гіперпараметри задаються до початку навчання і визначають структуру моделі, складність апроксимації, ширину часового вікна та здатність моделі узагальнювати закономірності на нових даних. У роботі для підбору гіперпараметрів використано решітковий пошук, який передбачає перебір наперед заданих комбінацій параметрів і вибір тієї комбінації, що забезпечує найкраще значення цільової метрики.

У межах ендогенної класифікаційної постановки для різних груп моделей використано такі гіперпараметри. Для моделі ARIMA основними гіперпараметрами є $p,\ d,\ q$, де p — порядок авторегресійної складової, d — порядок чисельного диференціювання ряду, а q — порядок складової ковзного середнього. Для моделей машинного навчання часовий ряд трендових залишків попередньо перетворюється у лаговий простір ознак, тому спільним гіперпараметром є n_steps, тобто ширина ковзного часового вікна вхідних даних. Для *Random Forest* додатково варіюється кількість дерев $n\_estimators$. Для *Support Vector Machine* з *RBF*-ядром використовуються параметри C і $gamma$, де C визначає силу регуляризації, а $gamma$ — вплив окремих спостережень у просторі ознак. Для моделі *Logistic Regression* основним гіперпараметром є C, який також визначає інтенсивність регуляризації. Для рекурентних нейронних мереж LSTM і GRU ключовими гіперпараметрами є n_steps та $units$, де $units$ задає кількість нейронів у рекурентному шарі.

У комп’ютерних експериментах використано типові решітки значень гіперпараметрів, наведені у табл. 3.2.

Ширина ковзного вікна n_steps визначає кількість попередніх значень трендових залишків, які використовуються для прогнозування наступного стану. Збільшення n_steps дає змогу врахувати довшу передісторію ряду, однак для коротких і слабкокорельованих часових рядів надмірне збільшення цього параметра може призвести до зростання розмірності вхідного простору та погіршення узагальнювальної здатності моделі [78; 157–161]. Тому в роботі

розглядалися невеликі значення n_steps, що відповідає обмеженій довжині ряду трендових залишків.

Таблиця 3.2

Підбір гіперпараметрів прогнозної класифікаційної моделі

Модель	**Гіперпараметр**	**Зміст гіперпараметра**	**Значення, що перебиралися**
ARIMA	p	порядок авторегресійної складової	0, 1, 2, 3, 4, 5
ARIMA	d	порядок різницювання часового ряду	0, 1
ARIMA	q	порядок складової ковзного середнього	0, 1, 2, 3, 4, 5
Random Forest	n_steps	ширина ковзного часового вікна	2, 3, 4, 5, 6
Random Forest	n_estimators	кількість дерев у ансамблі	40, 80, 120, 160, 200
SVM, RBF kernel	n_steps	ширина ковзного часового вікна	2, 3, 4, 5, 6
SVM, RBF kernel	C	параметр регуляризації	0,01, 0,1, 1, 10, 100
SVM, RBF kernel	Gamma	параметр ширини RBF-ядра	scale, 0,01, 0,1, 1
Logistic Regression	n_steps	ширина ковзного часового вікна	2, 3, 4, 5, 6
Logistic Regression	C	параметр регуляризації	0,01, 0.1, 1, 10, 100
LSTM	n_steps	довжина вхідної послідовності	2, 3, 4, 5, 6
LSTM	Units	кількість нейронів у рекурентному шарі	16, 32, 64, 128
GRU	n_steps	довжина вхідної послідовності	2, 3, 4, 5, 6
GRU	units	кількість нейронів у рекурентному шарі	16, 32, 64, 128

Цільовою функцією під час підбору гіперпараметрів у класифікаційній постановці обрано метрику $F1$ для класу низької врожайності, тобто $F1_{LY}$. Такий вибір зумовлений тим, що для задачі прогнозування врожайності особливо важливим є виявлення несприятливих років. Метрика $F1$ поєднує точність позитивного прогнозу та повноту виявлення позитивного класу.

У цій роботі позитивним класом вважається клас низької врожайності. Тому максимізація $F1_{LY}$ дозволяє збалансувати дві вимоги: не пропускати роки з низькою врожайністю та водночас не відносити надмірну кількість сприятливих років до ризикового класу.

Алгоритм решіткового пошуку має такий вигляд. Спочатку для кожної моделі задається скінченна множина допустимих значень гіперпараметрів. Далі формується декартовий добуток цих множин, тобто набір усіх можливих

комбінацій параметрів. Для кожної комбінації виконується навчання моделі на навчальній вибірці, після чого якість моделі оцінюється на валідаційній вибірці або за процедурою крос-валідації. Для класифікаційних моделей розраховується значення $F1_{LY}$, а оптимальною вважається та комбінація гіперпараметрів, для якої це значення є максимальним. Після вибору оптимальних гіперпараметрів модель перевіряється на тестовій вибірці, яка не використовувалася під час навчання і підбору параметрів [64; 65].

У випадку моделей, які формують прогнозну ймовірність належності спостереження до класу низької врожайності, додатково може виконуватися підбір порогу класифікації. Базовим є поріг τ=0,5. Водночас для підвищення якості виявлення класу низької врожайності може використовуватися оптимізаційний поріг, який визначається на валідаційній вибірці за максимумом $F1_{LY}$. Такий підхід дає змогу адаптувати модель до дисбалансу класів і підвищити її прикладну цінність у задачах раннього виявлення ризику низької врожайності [64–66].

Отже, підбір гіперпараметрів у роботі розглядається як обов'язковий етап побудови прогнозної моделі. Він забезпечує порівнянність результатів для різних класів моделей, дозволяє контролювати складність моделей і створює основу для об'єктивного вибору найефективнішої моделі за цільовою метрикою $F1_{LY}$.

3.1.6. Верифікація прогнозних моделей

Верифікація прогнозних моделей є необхідним етапом побудови моделей машинного та глибинного навчання, оскільки дозволяє оцінити не лише якість апроксимації навчальних даних, а й здатність моделі узагальнювати виявлені закономірності на нові спостереження. Основними ризиками під час навчання прогнозних моделей є перенавчання та недонавчання. Перенавчання виникає тоді, коли модель надмірно пристосовується до навчальної вибірки і втрачає здатність коректно працювати на нових даних. Недонавчання, навпаки, означає, що модель є надто простою і не здатна відобразити суттєві закономірності у даних [64].

Одним із поширених інструментів перевірки якості моделей є крос-валідація, або перехресна перевірка. Її основна ідея полягає в тому, що вихідний набір даних

поділяється на кілька частин, одна з яких використовується для перевірки моделі, а решта — для її навчання. Процедура повторюється кілька разів таким чином, щоб кожна частина даних по черзі використовувалася як перевірочна вибірка. Після цього значення метрик якості усереднюються, що дозволяє отримати більш стійку оцінку ефективності моделі порівняно з одноразовим поділом даних.

Найпростішим способом перевірки є метод відкладеної вибірки, або holdout-перевірка. У цьому випадку весь набір даних поділяється на дві неперетинні частини: навчальну та тестову. Найчастіше використовуються пропорції 70 % / 30 %, 75 % / 25 % або 80 % / 20 %. Перевагою цього підходу є простота реалізації, однак його недоліком є залежність результату від конкретного способу поділу даних. Якщо вибірка є малою, випадковий склад навчальної або тестової частини може суттєво впливати на значення метрик якості.

Більш надійним підходом є *K-fold cross-validation*. У цьому випадку набір даних поділяється на K приблизно рівних частин, які називаються складками. На кожній ітерації модель навчається на K-1 складках, а перевіряється на тій складці, яка залишилася. Процедура повторюється K разів, після чого результати усереднюються. Для задач бінарної класифікації доцільно використовувати стратифіковану крос-валідацію, за якої у кожній складці зберігається приблизно однакове співвідношення класів низької та високої врожайності.

У задачах прогнозування часових рядів процедура верифікації має певні особливості. Випадкове перемішування спостережень може призвести до порушення причинно-часової структури даних, коли інформація з майбутніх років опосередковано потрапляє до навчальної вибірки. Тому для моделей часових рядів використовується хронологічний поділ даних: модель навчається на попередніх спостереженнях, а перевіряється на наступних у часі. Такий підхід є особливо важливим для ендогенних моделей, у яких прогноз базується на попередніх значеннях трендових залишків.

Якщо потрібно провести оптимізацію гіперпараметрів моделі, використовують поділ даних на навчальну, валідаційну та контрольну вибірки. Навчальна вибірка призначена для оцінювання параметрів моделі, тобто

безпосереднього навчання алгоритму. Валідаційна вибірка використовується для підбору гіперпараметрів, вибору структури моделі та, за необхідності, оптимізації порогу класифікації. Контрольна, або тестова, вибірка застосовується лише на завершальному етапі для незалежного оцінювання якості вже обраної моделі. Такий поділ дозволяє уникнути завищення прогнозної якості, оскільки контрольні дані не беруть участі ні в навчанні, ні в підборі гіперпараметрів.

Для класичних моделей машинного навчання, побудованих на лагових ознаках, можливе використання як holdout-перевірки, так і крос-валідації. Водночас для нейромережевих моделей LSTM і GRU, які працюють із послідовностями спостережень, основна увага приділяється збереженню хронологічного порядку даних. У цьому випадку навчальна частина ряду використовується для оцінювання ваг нейронної мережі, валідаційна — для контролю якості навчання, підбору гіперпараметрів і запобігання перенавчанню, а контрольна — для остаточного порівняння результатів з іншими моделями.

Особливу роль у верифікації прогнозних моделей відіграє оцінювання стійкості прогнозних результатів. Під стійкістю прогнозної моделі у роботі розуміється її здатність зберігати прийнятний рівень точності та узгодженості результатів на даних, які не використовувалися під час навчання. Стійка модель не повинна демонструвати істотного погіршення якості під час переходу від навчальної вибірки до валідаційної або контрольної, а також має забезпечувати близькі за змістом результати за різних процедур перевірки, зокрема за зміни складу навчальної вибірки, часових інтервалів або значень гіперпараметрів у допустимих межах [64].

У задачах прогнозування врожайності стійкість моделі має особливе значення, оскільки аграрні дані характеризуються обмеженою довжиною часових рядів, міжрічною мінливістю, регіональною неоднорідністю та впливом екстремальних агрокліматичних умов. Модель, яка демонструє високі показники лише на навчальних даних, але істотно втрачає якість на валідаційній або контрольній вибірці, не може вважатися достатньо стійкою. Натомість стійкою є така модель, яка забезпечує збалансовані значення основних метрик класифікації,

зокрема $Accuracy$, $Precision$, $Recall$, $F1_{LY}$, $Specificity$ та ROC_AUC, на незалежних вибірках і не виявляє ознак надмірного пристосування до окремого набору спостережень.

Отже, верифікація прогнозних моделей у роботі базується на поєднанні крос-валідації, хронологічного поділу часових рядів, використання навчальної, валідаційної та контрольної вибірок, а також аналізу стійкості прогнозних оцінок. Такий підхід забезпечує більш об’єктивне порівняння моделей машинного та глибинного навчання і дозволяє оцінити їхню придатність для прогнозування ризику низької врожайності.

Загальна схема експериментальних досліджень у межах розділу 3 є такою. Спочатку, з використанням методів машинного навчання та часових рядів врожайності пшениці за період 1955–2021 рр. будуються прогнозні класифікаційні моделі. Після цього аналізуються можливості рекурентних нейронних мереж для прогнозування часової динаміки врожайності. Оптимізація класифікаційних моделей врожайності здійснюється шляхом решіткового пошуку гіперпараметрів моделі. Завершальним етапом є порівняльне оцінювання ефективності моделей машинного та глибинного навчання за узгодженою системою критеріїв.

Таким чином, у підрозділі 3.1 сформовано єдину методичну основу для подальшого порівняння ендогенних моделей машинного та глибинного навчання. У наступних підрозділах ця методика застосовується до часового ряду трендових залишків урожайності Херсонської області.

3.2. Класифікаційні моделі прогнозування врожайності на основі методів машинного навчання

Найбільш поширеним підходом до прогнозування врожайності у даному дослідженні є класифікаційна постановка задачі, у межах якої прогнозується не саме значення врожайності, а його належність до певного класу [4; 7–9; 24; 61; 62; 64–66]. Для аграрної практики та управлінських рішень така постановка є особливо корисною, оскільки дозволяє заздалегідь ідентифікувати ризик формування низької врожайності та використовувати результати як інструмент попередження про

несприятливий сценарій. Методична основа класифікаційного підходу полягає у використанні трендових залишків урожайності ε, що характеризують відхилення фактичних значень урожайності від відповідного лінійного тренду.

Ендогенний підхід до прогнозування врожайності передбачає використання лише часового ряду врожайності і виключає врахування впливу зовнішніх факторів. Такий підхід дещо збіднює прогнозні можливості моделі. Нашим завданням було вивчити можливості та межі застосовності ендогенного підходу до прогнозування врожайності на прикладі часових рядів врожайності пшениці. Алгоритм класифікаційного прогнозування наведено на рис. 3.2.

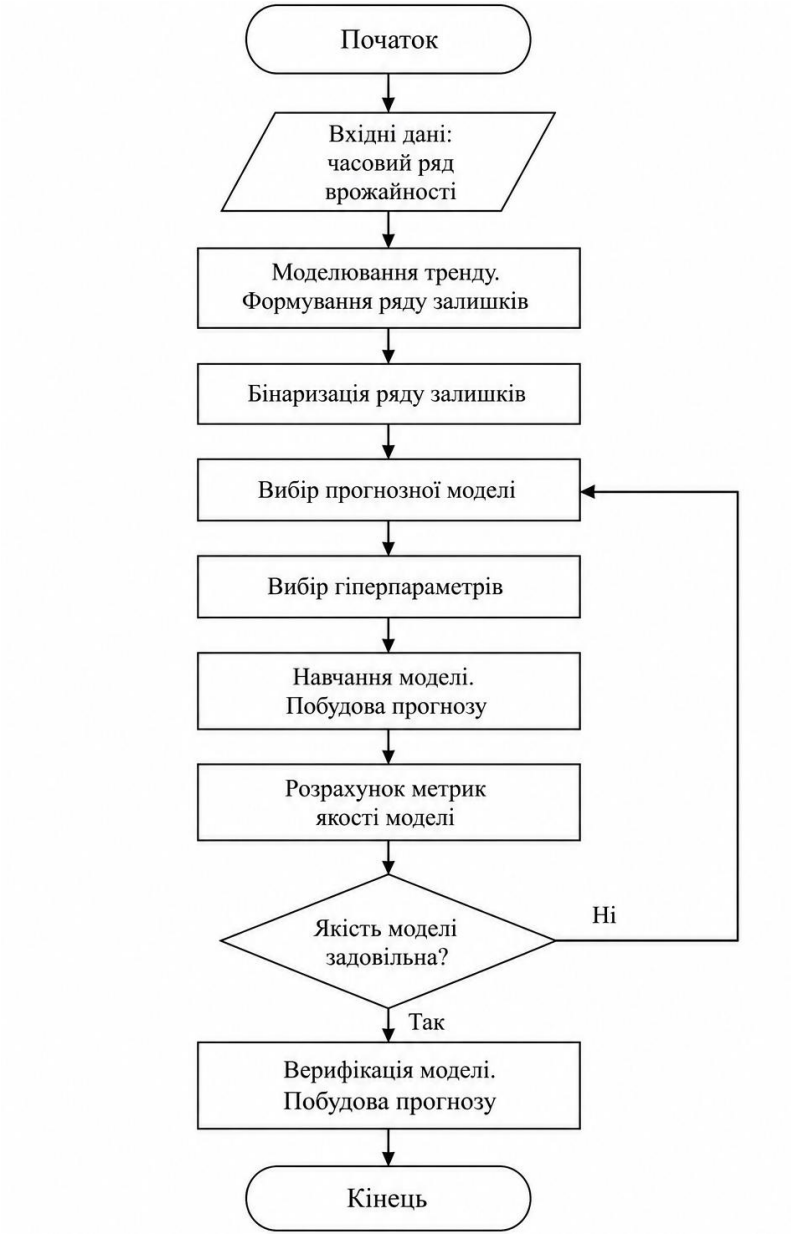

Рис. 3.2. Алгоритм класифікаційного прогнозування врожайності

Як вхідні дані для прогнозної моделі використано часовий ряд урожайності пшениці за період 1955–2021 рр. Об’єктом дослідження обрано Херсонську область України. Такий вибір зумовлений тим, що вплив кліматичних факторів на врожайність тут особливо відчутний і це спричиняє значні міжрічні коливання врожайності [22; 46–51]. Графічна ілюстрація ряду трендових залишків врожайності представлена на рис. 3.3.

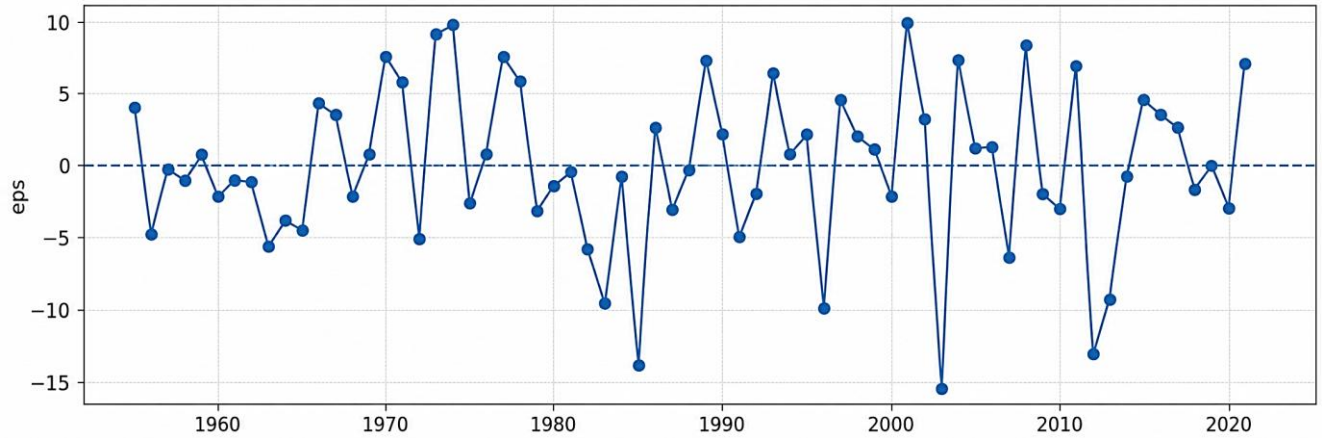

Рис. 3.3. Трендові залишки врожайності пшениці для Херсонської області

В ендогенному підході до прогнозування часових рядів використовується метод ковзного вікна. Кілька початкових значень часового ряду трендових залишків врожайності, позначених як n_steps, розглядаються як пояснювальні змінні, тоді як наступний елемент ряду з номером $n\_steps + 1$ слугує результуючою змінною. Переміщуючи ковзне вікно вздовж часового ряду трендових залишків, ми отримуємо набір з $N - n\_steps$ вибірок, які можна використовувати для навчання моделі. Кожна вибірка містить n_steps предикторів і одну результуючу змінну.

Для побудови ендогенних класифікаційних моделей прогнозування врожайності у роботі використано кілька класичних методів машинного навчання: Random Forest, Support Vector Machine та Logistic Regression. На відміну від статистичних моделей часових рядів, ці методи не задають наперед конкретної авторегресійної структури, а навчаються виявляти залежності між лаговими ознаками, сформованими на основі попередніх значень трендових залишків урожайності, та цільовим класом. У такій постановці кожне спостереження

описується вектором попередніх значень ряду, а результатом класифікації є прогноз належності наступного року до класу низької або високої врожайності.

Використання цих трьох моделей дає змогу порівняти різні підходи до ендогенної класифікації врожайності: ансамблевий нелінійний підхід Random Forest, ядровий нелінійний підхід SVM та лінійний імовірнісний підхід Logistic Regression. Таке порівняння є важливим для оцінювання того, чи містить короткий часовий ряд трендових залишків достатньо інформації для побудови стійкого прогнозу класу врожайності лише на основі його внутрішньої динаміки.

Для вибору найкращого варіанта кожної моделі ми застосовуємо метод решіткового пошуку. Можливі значення гіперпараметрів моделей Random Forest, Support Vector Machine з RBF kernel, Logistic Regression наведені у табл. 3.3.

Таблиця 3.3

Гіперпараметри прогнозних класифікаційних моделей

Random Forest	n_steps	ширина ковзного часового вікна	2, 3, 4, 5, 6
Random Forest	n_estimators	кількість дерев у ансамблі	40, 80, 120, 160, 200
SVM, RBF kernel	n_steps	ширина ковзного часового вікна	2, 3, 4, 5, 6
SVM, RBF kernel	C	параметр регуляризації	0,01, 0.1, 1, 10, 100
SVM, RBF kernel	gamma	параметр ширини RBF-ядра	scale, 0,01, 0.1, 1
Logistic Regression	n_steps	ширина ковзного часового вікна	2, 3, 4, 5, 6
Logistic Regression	C	параметр регуляризації	0,01, 0,1, 1, 10, 100

Класифікація трендових залишків здійснюється за правилом (3.2). Вхідний ряд трендових залишків врожайності ділиться на навчальну вибірку, валідаційну вибірку та контрольну вибірку у співвідношенні 70 % / 15 % / 15 %. Навчальна вибірка використовується для навчання моделі, валідаційна вибірка використовується для підбору гіперпараметрів, тестова вибірка використовується для оцінки якості (розрахунку метрик) моделі.

Метод Random Forest є ансамблевим алгоритмом машинного навчання, який ґрунтується на побудові множини дерев рішень. Кожне дерево навчається на певній підвибірці даних і формує власний класифікаційний прогноз, а підсумкове рішення визначається шляхом агрегування результатів окремих дерев. Перевагою Random Forest є здатність враховувати нелінійні залежності між ознаками, зменшувати ризик перенавчання порівняно з окремим деревом рішень і забезпечувати достатньо

стійкі результати на невеликих вибірках. У цьому дослідженні основними гіперпараметрами Random Forest є ширина лагового вікна (n_steps) та кількість дерев ($n\_estimators$).

Ми ввели удосконалення у методику вибору конфігурації найкращої моделі та наступної її перевірки на тестовій вибірці. Орієнтація на лише одну найкращу конфігурацію моделі може призвести до нестійкого оцінювання якості моделі на тестовій вибірці. Тому, ми спочатку згідно з описаною вище методикою здійснюємо решітковий пошук на валідаційній вибірці для пошуку найкращих гіперпараметрів прогнозних моделей. Для кожного класу моделей машинного навчання відбираються три моделі з різною конфігурацією та максимальним значенням цільового показника. Такий підхід на відміну від традиційного підходу, який ґрунтується на виборі однієї найкращої конфігурації, передбачає усереднене оцінювання групи найкращих моделей на тестовій вибірці, що підвищує стійкість інтерпретації результатів і надійність порівняння моделей під час переходу до нових даних. Метрика $F1\_LY$ була нами обрана як найбільш об'єктивний показник якості прогнозної моделі, який узгоджує точність позитивного прогнозу з повнотою виявлення об'єктів позитивного класу.

Алгоритмічну послідовність решіткового пошуку гіперпараметрів моделі Random Forest наведено на рис. 3.4. Зовнішній цикл відповідає перебору ширини лагового вікна n_steps, а внутрішній цикл — перебору кількості дерев $n\_estimators$. Для кожної конфігурації модель навчається на навчальній вибірці, після чого обчислюється значення метрики $F1\_LY$ на валідаційній вибірці. Використання саме цієї метрики зумовлене тим, що в межах класифікаційної постановки основна увага приділяється виявленню класу низької врожайності, який має найбільше прикладне значення для задач раннього попередження. Якщо отримане значення $F1\_LY$ перевищує поточний найкращий результат, відповідна комбінація гіперпараметрів зберігається як нова найкраща конфігурація. Така процедура забезпечує послідовне порівняння всіх допустимих варіантів моделі та дає змогу обґрунтовано вибрати параметри, які забезпечують найкращу якість класифікації на валідаційних даних.

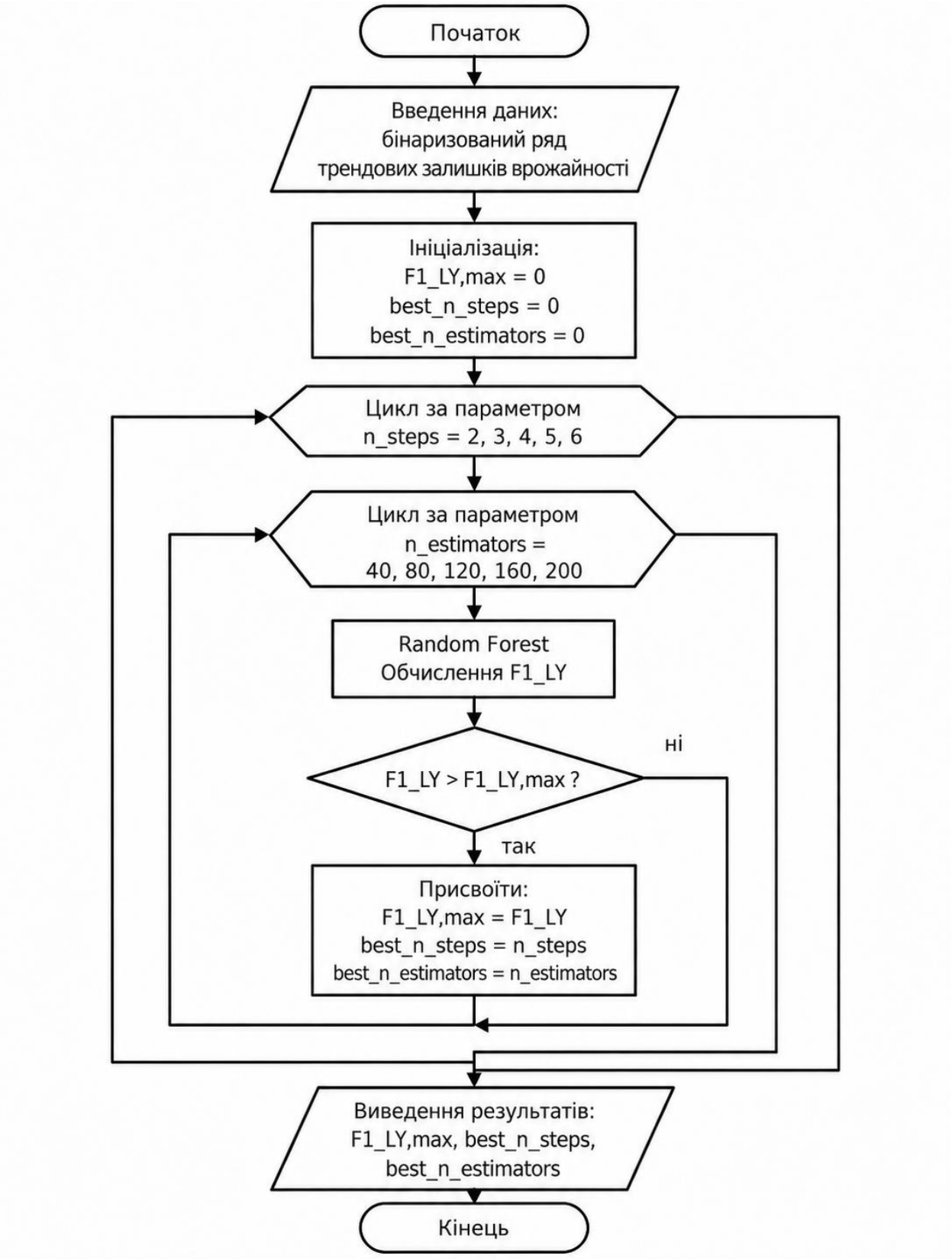

Рис. 3.4. Блок-схема алгоритму решіткового пошуку гіперпараметрів моделі Random Forest

Результати решіткового пошуку гіперпараметрів моделі Random Forest наведені у табл. 3.4.

Таблиця 3.4

Значення метрики $F1_{LY}$ для моделі Random Forest, розраховані на решітці гіперпараметрів

n_steps \ n_estimators	40	80	120	160	200
2	0,4444	0,4444	0,4444	0,4444	0,4444
3	0,6667	0,6000	0,6000	0,6667	0,6667
4	0,7500	0,8571	0,8571	0,8571	0,8571
5	0,7500	0,7500	0,7500	0,7500	0,7500
6	0,6667	0,6667	0,6667	0,6667	0,7500

Як видно з табл. 3.4, найвище значення метрики $F1_{LY}$ = 0,8571 отримано для набору гіперпараметрів n_{steps} = 4; $n_{estimators}$ = 80 ÷ 200. Метрики оптимальної моделі мають такі значення: $Accuracy$ = 0,8889; $Precision$ = 1,0000; $Recall$ = 0,7500; $Specificity$ = 1,0000; $F1$ = 0,8571; ROC_AUC = 0,8500. Підкреслимо, що всі метрики розраховані на валідаційній вибірці. Графічну ілюстрацію решіткового пошуку наведено на рис. 3.5.

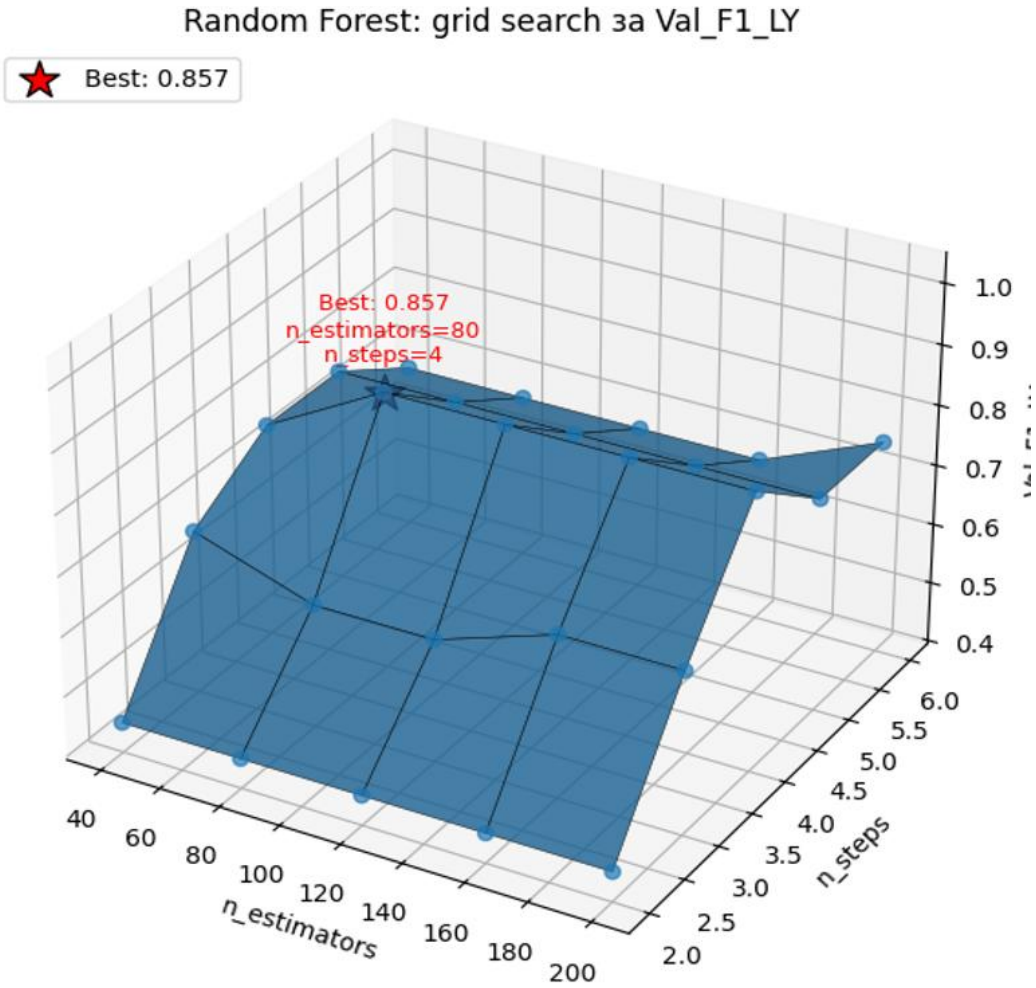

Рис. 3.5. Решітковий пошук гіперпараметрів моделі Random Forest

Logistic Regression є класичною лінійною моделлю бінарної класифікації, яка оцінює ймовірність належності спостереження до одного з двох класів. У цій роботі вона використовується як базова інтерпретована модель, що дозволяє оцінити, наскільки якісним може бути класифікаційний прогноз за умови лінійного зв'язку між лаговими ознаками та ймовірністю низької врожайності. Перевагою Logistic Regression є простота, стійкість, невелика кількість гіперпараметрів і можливість порівняння з більш складними нелінійними моделями. Основними параметрами, що аналізуються для цієї моделі, є ширина лагового вікна n_steps та параметр регуляризації C.

Результати решіткового пошуку гіперпараметрів моделі Logistic Regression наведені у табл. 3.5.

Таблиця 3.5

Значення метрики $F1_{LY}$ для моделі Logistic Regression, розраховані на решітці гіперпараметрів

n_steps \ C	**0,01**	**0,10**	**1,00**	**10,00**	**100,00**
2	0,4000	0,4000	0,4000	0,4000	0,4000
3	0,4000	0,4444	0,4444	0,4444	0,4444
4	0,6000	0,6000	0,6000	0,6000	0,6000
5	0,6000	0,6000	0,6000	0,6000	0,6000
6	0,6000	0,6000	0,6000	0,6000	0,6000

Згідно з даними табл. 3.5, найвище значення метрики $F1_{LY}$ = 0,6000 отримано для широкої групи гіперпараметрів n_{steps} =4 ÷ 6. При цьому значення гіперпараметра C не впливає на якість класифікації. Метрики оптимальної моделі мають такі значення $Accuracy$ = 0,5555; $Precision$ = 0,5000; $Recall$ = 0,7500; $Specificity$ = 0,4000; $F1$ = 0,6000; ROC_AUC = 0,6500. Графічну ілюстрацію решіткового пошуку наведено на рис. 3.6.

Метод Support Vector Machine належить до ефективних методів класифікації, які будують роздільну поверхню між класами з максимальним зазором. У лінійному випадку така поверхня є гіперплощиною, а у випадку використання ядрових функцій модель може формувати нелінійну межу між класами. У роботі

використовується SVM із радіально-базисним ядром (Radial Basis Function, RBF), що дозволяє враховувати можливу нелінійну залежність між попередніми значеннями трендових залишків і класом майбутнього стану врожайності. Основними гіперпараметрами цієї моделі є n_steps, параметр регуляризації C та параметр ядра $gamma$, який визначає локальність впливу окремих навчальних спостережень.

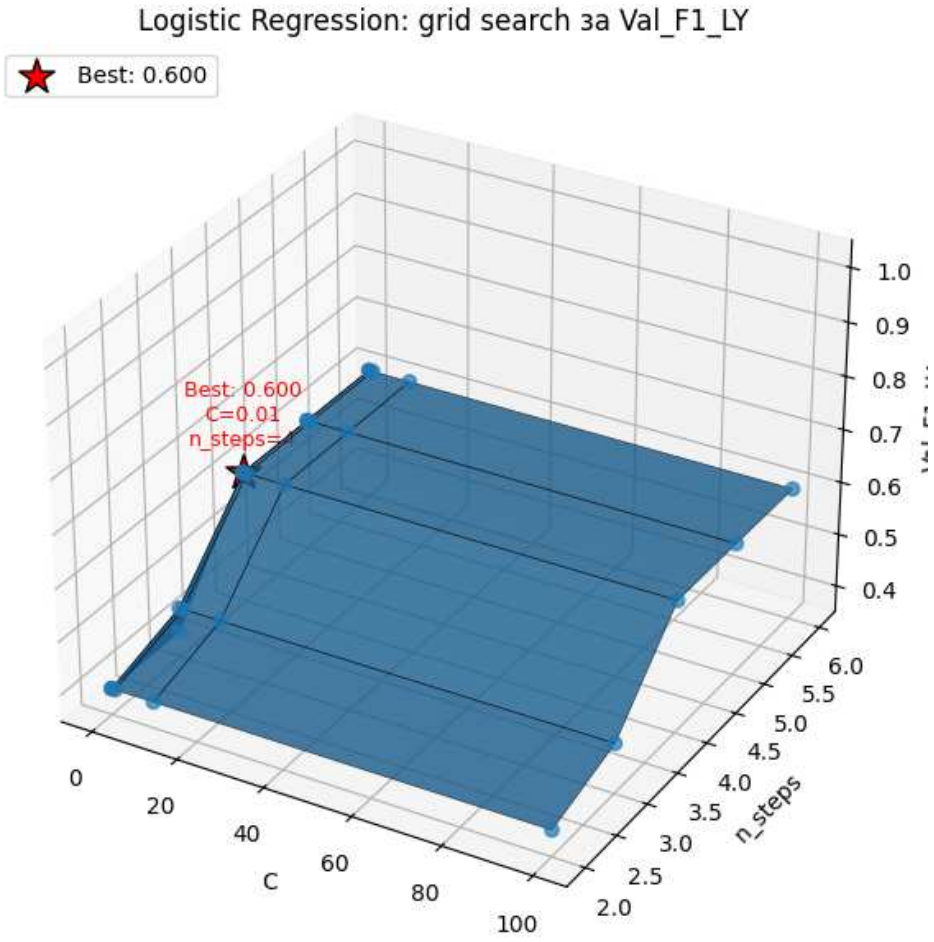

Рис. 3.6. Решітковий пошук гіперпараметрів моделі Logistic Regression

Суттєвий вплив на ефективність моделі Support Vector Machine з RBF-ядром має параметр $gamma$, який визначає, наскільки далеко поширюється вплив одного навчального спостереження: мале значення $gamma$ означає ширшу зону впливу, велике значення $gamma$ — більш локалізований вплив окремих точок. Зазвичай значення параметра $gamma$ задається вручну: $gamma$ = 0,01; 0,1; 1,0. Інколи використовують значення $gamma$=$scale$, за якого параметр γ автоматично обчислюється як величина, обернена до добутку кількості ознак і дисперсії вхідних

даних. Такий підхід дозволяє адаптувати ширину ядра до масштабу сформованого набору даних. Для більш детального аналізу ефективності моделі Support Vector Machine проведемо комп'ютерні експерименти з декількома значеннями параметра $gamma$.

Результати решіткового пошуку гіперпараметрів моделі Support Vector Machine (RBF kernel) наведені у табл. 3.6–3.7.

Таблиця 3.6

Значення метрики $F1_{LY}$ для моделі Support Vector Machine (RBF kernel), розраховані на решітці гіперпараметрів при значенні $gamma = scale$

n_steps \ C	**0,01**	**0,10**	**1,00**	**10,00**	**100,00**
2	0,5714	0,5714	0,4000	0,4000	0,4000
3	0,0000	0,5714	0,5714	0,5714	0,4444
4	0,0000	0,3333	0,3333	0,5000	0,5000
5	0,0000	0,0000	0,7500	0,5714	0,5714
6	0,6154	0,5455	0,5714	0,5714	0,5714

У табл. 3.6 подано результати решіткового пошуку при значенні параметра $gamma = scale$. Найвище значення метрики $F1_{LY}$ = 0,7500 отримано для n_{steps} = 5; C = 1,00. При значенні гіперпараметра C = 0,01 результати є дуже нестійкими. Оптимальним значенням параметра C є C = 1,00. Метрики оптимальної моделі мають такі значення: $Accuracy$ = 0,7778; $Precision$ = 0,7500; $Recall$ = 0,7500; $Specificity$ = 0,8000; $F1$ = 0,7500; ROC_AUC = 0,6000.

Таблиця 3.7

Значення метрики $F1_{LY}$ для моделі Support Vector Machine (RBF kernel), розраховані на решітці гіперпараметрів при значенні $gamma$ = 0,1

n_steps \ C	**0,01**	**0,10**	**1,00**	**10,00**	**100,00**
2	0,5714	0,5714	0,2500	0,4444	0,4444
3	0,0000	0,5714	0,2222	0,6000	0,4000
4	0,0000	0,3333	0,7500	0,5000	0,5000
5	0,0000	0,0000	0,5455	0,3333	0,5714
6	0,6154	0,2500	0,2222	0,5714	0,5714

У табл. 3.7 подано результати решіткового пошуку при значенні параметра $gamma$ = 0,1. Найвище значення метрики $F1_{LY}$ = 0,7500 отримано для n_{steps} = 4;

C = 1,00. При значенні гіперпараметра C = 0,01 результати є дуже нестійкими. Метрики оптимальної моделі мають такі значення: $Accuracy$ = 0,7778; $Precision$ = 0,7500; $Recall$ = 0,7500; $Specificity$ = 0,8000; $F1$ = 0,7500; ROC_AUC = 0,6500. Як бачимо, параметри цієї конфігурації моделі Support Vector Machine практично збігаються з параметрами попередньої моделі. Графічну ілюстрацію решіткового пошуку наведено на рис. 3.7.

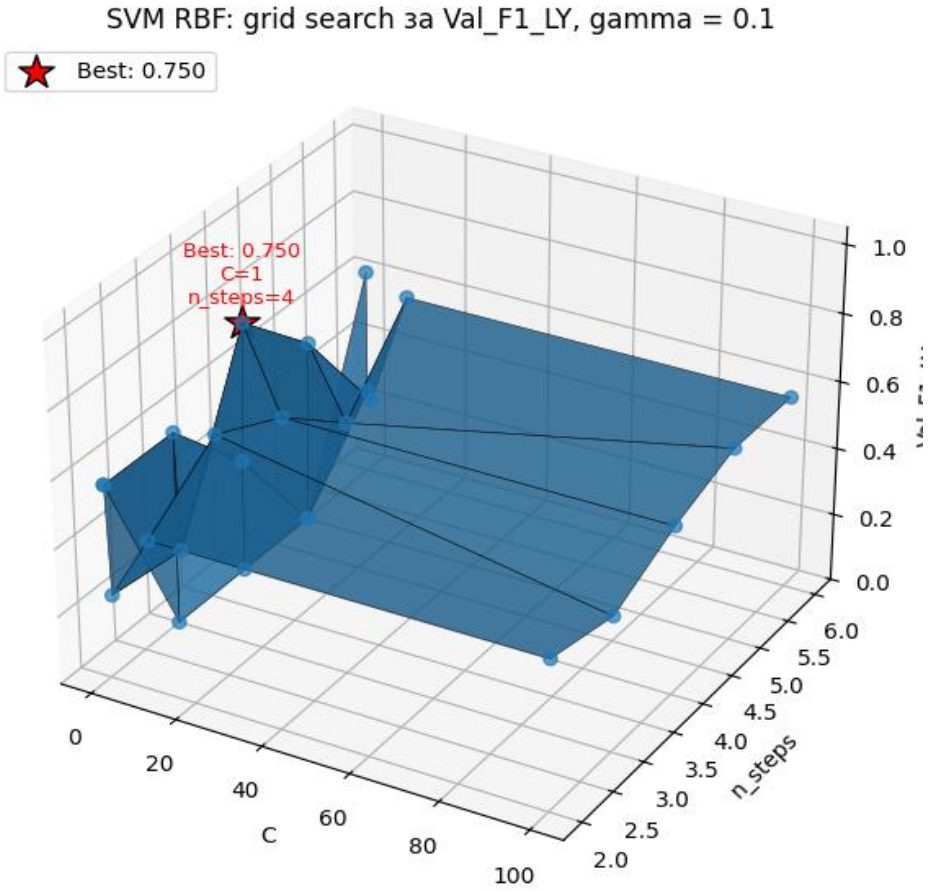

Рис. 3.7. Решітковий пошук гіперпараметрів моделі Support Vector Machine

Якість побудованих ендогенних класифікаційних моделей перевірено на тестовій вибірці. З огляду на малий обсяг вихідного часового ряду та можливу нестійкість результатів для окремої конфігурації моделі, для перевірки використано не одну, а три найкращі конфігурації кожного типу моделі за значенням критерію $F1$ на валідаційній вибірці. Такий підхід дає змогу підвищити стійкість інтерпретації результатів, оскільки висновки щодо якості прогнозування формуються не за однією випадково найкращою конфігурацією, а за групою близьких за якістю моделей.

Порівняння усереднених значень метрик для валідаційної та тестової вибірок дозволяє оцінити, наскільки прогнозні властивості моделі зберігаються під час переходу від етапу налаштування гіперпараметрів до незалежного тестування. Якщо середні значення метрик на тестовій вибірці істотно знижуються порівняно з валідаційною, це свідчить про недостатню стійкість моделі або про її чутливість до конкретного розбиття короткого часового ряду.

Таблиця 3.8

Порівняння метрик моделі Random Forest для валідаційної та тестової вибірки

Валідаційна вибірка							
n_steps	**n_estimators**	**Accuracy**	**Precision_LY**	**Recall_LY**	**F1_LY**	**Specificity_HY**	**ROC_AUC**
4	120	0,8889	1,0000	0,7500	0,8571	1,0000	0,8500
4	80	0,8889	1,0000	0,7500	0,8571	1,0000	0,8500
4	160	0,8889	1,0000	0,7500	0,8571	1,0000	0,8500
Середнє		**0,8889**	**1,0000**	**0,7500**	**0,8571**	**1,0000**	**0,8500**
Тестова вибірка							
n_steps	**n_estimators**	**Accuracy**	**Precision_LY**	**Recall_LY**	**F1_LY**	**Specificity_HY**	**ROC_AUC**
4	120	0,5000	0,6000	0,5000	0,5455	0,5000	0,4583
4	80	0,5000	0,6000	0,5000	0,5455	0,5000	0,5417
4	160	0,5000	0,6000	0,5000	0,5455	0,5000	0,5000
Середнє		**0,5000**	**0,6000**	**0,5000**	**0,5455**	**0,5000**	**0,5000**
Різниця середніх		**0,3889**	**0,4000**	**0,2500**	**0,3116**	**0,5000**	**0,3500**

У табл. 3.8 наведено результати для трьох найкращих конфігурацій моделі Random Forest. На валідаційній вибірці всі три конфігурації мають однакові значення основних метрик: середнє значення $Accuracy$ = 0,8889; $Precision$ = 1,0000; $Recall$ = 0,7500; $Specificity$ = 1,0000; $F1$ = 0,8571; ROC_AUC = 0,8500. Такі результати свідчать про високу якість класифікації на етапі валідації, зокрема про відсутність помилкових віднесень років високої врожайності до класу низької врожайності.

Однак на тестовій вибірці якість моделі Random Forest істотно знижується. Середнє значення $Accuracy$ становить лише 0,5000, $F1$ = 0,5455, $Specificity$ = 0,5000, а ROC_AUC = 0,5000. Це означає, що висока якість, отримана на валідаційній вибірці, не зберігається на незалежних тестових даних. Зниження метрик свідчить про нестійкість моделі Random Forest в ендогенній постановці задачі та про можливу чутливість результатів до конкретного розбиття короткого часового ряду.

Таблиця 3.9

Порівняння метрик моделі Support Vector Machine для валідаційної та тестової вибірки

Валідаційна вибірка								
n_steps	**C**	**gamma**	**Accuracy**	**Precision_LY**	**Recall_LY**	**F1_LY**	**Specificity_HY**	**ROC_AUC**
5	1	scale	0,7778	0,7500	0,7500	0,7500	0,8000	0,6000
4	1	0,1	0,7778	0,7500	0,7500	0,7500	0,8000	0,6500
6	0,01	scale	0,4444	0,4444	1,0000	0,6154	0,0000	0,5000
Середнє			**0,6667**	**0,6481**	**0,8333**	**0,7051**	**0,5333**	**0,5833**
Тестова вибірка								
n_steps	**C**	**gamma**	**Accuracy**	**Precision_LY**	**Recall_LY**	**F1_LY**	**Specificity_HY**	**ROC_AUC**
5	1	scale	0,6000	0,6250	0,8333	0,7143	0,2500	0,5417
4	1	0,1	0,3000	0,4000	0,3333	0,3636	0,2500	0,3750
6	0,01	scale	0,6000	0,6000	1,0000	0,7500	0,0000	0,5000
Середнє			**0,5000**	**0,5417**	**0,7222**	**0,6093**	**0,1667**	**0,4722**
Різниця середніх			**0,1667**	**0,1064**	**0,1111**	**0,0958**	**0,3666**	**0,1111**

У табл. 3.9 подано результати для моделі Support Vector Machine. На валідаційній вибірці середнє значення $F1$ для трьох найкращих конфігурацій становить 0,7051. Це нижче, ніж для Random Forest, проте на тестовій вибірці середнє значення цієї метрики знижується лише до 0,6093. Отже, у порівнянні з Random Forest модель SVM демонструє помірнішу втрату якості під час переходу від валідаційної до тестової вибірки.

Водночас результати SVM потребують обережної інтерпретації. На тестовій вибірці середнє значення $Recall$ становить 0,7222, що вказує на здатність моделі виявляти значну частину років низької врожайності. Проте середнє значення $Specificity$ дорівнює лише 0,1667. Це означає, що модель має схильність частіше відносити спостереження до класу низької врожайності й гірше розпізнає роки високої врожайності. Тому SVM забезпечує відносно краще виявлення несприятливих років, але ціною збільшення кількості помилкових попереджень.

У табл. 3.10 наведено результати для моделі Logistic Regression. На валідаційній вибірці три найкращі конфігурації мають однакові значення метрик: $Accuracy$ = 0,5556; $Precision$ = 0,5000; $Recall$ = 0,7500; $Specificity$ = 0,4000; $F1$ = 0,6000; ROC_AUC = 0,6500. Це свідчить про помірну якість моделі на етапі налаштування гіперпараметрів.

Таблиця 3.10

Порівняння метрик моделі Logistic Regression для валідаційної та тестової вибірки

Валідаційна вибірка							
n_steps	**C**	**Accuracy**	**Precision_LY**	**Recall_LY**	**F1_LY**	**Specificity_HY**	**ROC_AUC**
6	0,01	0,5556	0,5000	0,7500	0,6000	0,4000	0,6500
6	0,1	0,5556	0,5000	0,7500	0,6000	0,4000	0,6500
6	1	0,5556	0,5000	0,7500	0,6000	0,4000	0,6500
Середнє		**0,5556**	**0,5000**	**0,7500**	**0,6000**	**0,4000**	**0,6500**
Тестова вибірка							
C	**C**	**Accuracy**	**Precision_LY**	**Recall_LY**	**F1_LY**	**Specificity_HY**	**ROC_AUC**
6	0,01	0,3000	0,4000	0,3333	0,3636	0,2500	0,2083
6	0,1	0,3000	0,4000	0,3333	0,3636	0,2500	0,1667
6	1	0,3000	0,4000	0,3333	0,3636	0,2500	0,1667
Середнє		**0,3000**	**0,4000**	**0,3333**	**0,3636**	**0,2500**	**0,1806**
Різниця середніх		**0,2556**	**0,1000**	**0,4167**	**0,2364**	**0,1500**	**0,4694**

На тестовій вибірці якість Logistic Regression суттєво погіршується. Середнє значення $Accuracy$ становить 0,3000, $F1 = 0{,}3636$, $Specificity = 0{,}2500$, а $ROC\_AUC = 0{,}1806$. Такі значення свідчать про слабку здатність моделі узагальнювати закономірності на незалежній тестовій вибірці. Особливо низьке значення ROC_AUC вказує на те, що модель недостатньо добре розділяє класи низької та високої врожайності у межах ендогенного підходу.

Порівняння трьох типів моделей показує, що найвищі середні валідаційні метрики отримано для Random Forest. Проте саме ця модель демонструє суттєве зниження якості на тестовій вибірці. Модель SVM має нижчі валідаційні показники, але за середнім значенням $F1$ на тестовій вибірці вона є найкращою серед трьох розглянутих моделей. Logistic Regression показує найнижчу якість тестового прогнозування, що свідчить про недостатню ефективність лінійної класифікаційної моделі для цієї ендогенної задачі.

Отже, серед розглянутих ендогенних моделей машинного навчання найкращий результат на тестовій вибірці за середнім значенням $F1$ отримано для Support Vector Machine. Водночас низьке значення специфічності цієї моделі свідчить про незбалансованість класифікації та її схильність до надмірного прогнозування класу низької врожайності. Random Forest показав високі результати

на валідаційній вибірці, але виявився недостатньо стійким на тестовій вибірці. Logistic Regression продемонструвала найнижчу якість незалежного тестування.

Попередні результати підрозділу свідчать, що ендогенні моделі машинного навчання, побудовані лише на основі попередніх значень трендових залишків урожайності, мають обмежену прогнозну здатність. Незважаючи на те, що окремі конфігурації можуть демонструвати прийнятні або високі значення метрик на валідаційній вибірці, під час незалежного тестування якість прогнозування часто знижується. Це підтверджує, що внутрішня динаміка короткого ряду трендових залишків не містить достатньо стійкої інформації для надійного прогнозування класу врожайності.

Таким чином, ендогенний підхід до класифікаційного прогнозування врожайності може бути використаний як базовий або порівняльний варіант моделювання. Однак отримані результати показують, що для підвищення якості та стійкості прогнозування доцільно залучати екзогенні агрокліматичні фактори, які безпосередньо характеризують умови формування врожайності.

3.3. Застосування рекурентних нейронних мереж для прогнозування врожайності

Інформація про врожайність найчастіше представлена у вигляді часових рядів. Для прогнозування часових рядів традиційно використовуються методи екстраполяції тренду, ковзного середнього, експоненційного згладжування, авторегресійні моделі, моделі ARMA та ARIMA [161]. Сучасні підходи до прогнозування часових рядів активно використовують нейронні мережі, зокрема рекурентні нейронні мережі RNN, які здатні враховувати залежності між послідовними значеннями даних [62; 151; 152].

Особливо ефективними є модифікації RNN — LSTM і GRU. Класичні RNN мають проблему: їм важко зберігати інформацію про віддалені попередні стани ряду. LSTM і GRU частково розв’язують цю проблему завдяки спеціальним механізмам керування пам’яттю. Вони можуть «вирішувати», яку інформацію зберігати, яку послаблювати, а яку використовувати для прогнозу. Прикладом застосування рекурентних нейронних мереж до задачі прогнозування врожайності

пшениці є роботи [151; 152]. На відміну від лінійних і класичних класифікаційних моделей, рекурентні нейронні мережі орієнтовані на виявлення часових залежностей у послідовностях спостережень і тому є природним інструментом для аналізу динаміки врожайності та її трендових залишків [62].

Основна ідея моделі LSTM полягає у використанні спеціальної комірки пам'яті та системи керуючих механізмів, які визначають, яку інформацію необхідно зберегти, яку — оновити, а яку — відкинути. У загальному випадку LSTM-комірка містить три основні типи воріт: ворота забування, вхідні ворота та вихідні ворота. Ворота забування регулюють, яка частина попереднього стану пам'яті втрачає актуальність; вхідні ворота визначають, яка нова інформація буде додана до стану комірки; вихідні ворота формують поточний вихід моделі [153].

Модель GRU (Gated Recurrent Unit) є компактнішою рекурентною архітектурою, яка має спрощену структуру порівняно з LSTM. У GRU відсутня окрема комірка пам'яті, а основним носієм інформації про попередню динаміку є прихований стан. Для керування передаванням інформації використовуються два основні механізми: ворота оновлення та ворота скидання. Ворота оновлення визначають, яку частину попереднього прихованого стану доцільно зберегти для наступного кроку, а ворота скидання регулюють, яку частину минулої інформації потрібно тимчасово ігнорувати під час формування нового стану [154–156]. Структурну схему комірки GRU наведено на рис. 3.8.

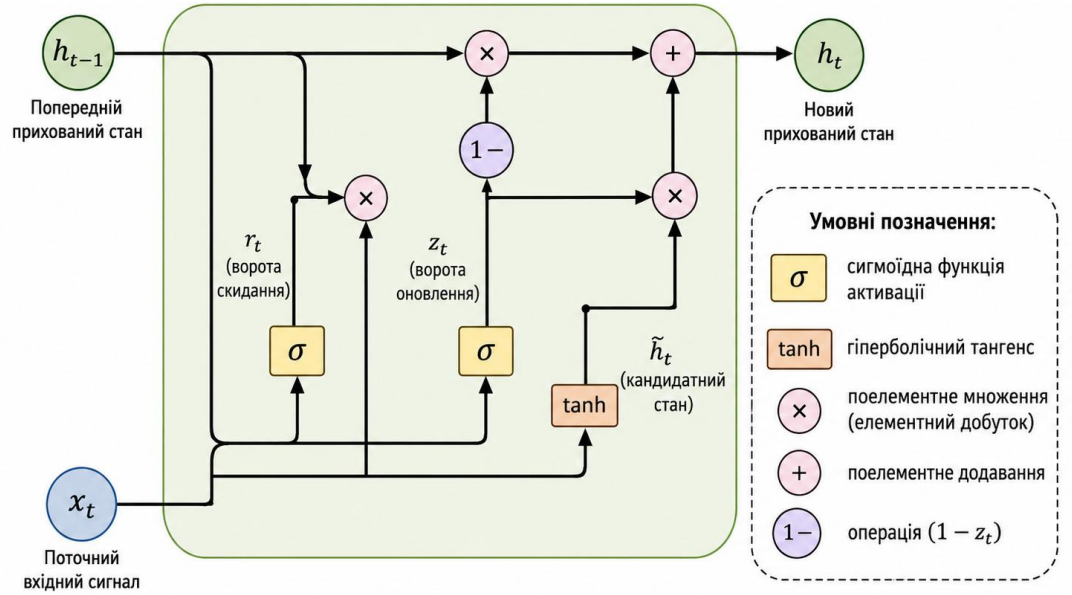

Рис. 3.8. Структурна схема комірки GRU

Порівняно з LSTM, модель GRU має меншу кількість параметрів, простішу внутрішню структуру та зазвичай швидше навчається. Це є важливою перевагою для задач, у яких обсяг навчальних даних є обмеженим. Для аграрних часових рядів така особливість має суттєве значення, оскільки кількість спостережень визначається кількістю років статистичного обліку і є значно меншою, ніж у задачах технічного, фінансового або текстового моделювання. У дисертації проведено дослідження прогнозних можливостей обох архітектур мережі RNN з урахуванням кількості вхідних ознак, кількості нейронів прихованого шару та способу формування навчальної і контрольної вибірки.

Особливістю використання рекурентних мереж у нашому дослідженні є те, що вони застосовуються до відносно коротких часових рядів, побудованих для врожайності пшениці або для трендових залишків урожайності [62; 151; 152]. На відміну від фінансових чи технічних часових рядів, де кількість спостережень є набагато більшою, в аграрних задачах довжина ряду обмежується кількістю років спостереження. За таких умов ефективність рекурентних мереж істотно залежить від того, наскільки вираженими є часові залежності в ряді, а також від того, чи містить ряд достатньо інформації для навчання складної параметричної моделі. Саме тому в дисертаційному дослідженні окремо досліджується питання про межі застосовності моделей GRU та LSTM у задачах прогнозування врожайності пшениці.

У найпростішому підході рекурентні прогнозні моделі розглядають у межах ендогенного підходу, коли на вхід мережі подається лише часовий ряд врожайності або трендових залишків. Такий підхід дозволяє оцінити, чи містить сам ряд достатню внутрішню часову структуру для побудови корисного прогнозу. Водночас аналіз, відображений у вступі дисертації, показує, що ендогенний підхід до прогнозування коротких часових рядів трендових залишків із низькими значеннями автокореляційної та часткової автокореляційної функцій може мати обмежену прогностичну здатність. Це пов’язано з тим, що сам по собі ряд трендових залишків не завжди містить достатньо інформації для стабільного нейромережевого прогнозування. Тому, однією із задач, яку ми повинні вирішити, є дослідження

прогнозної ефективності та виявлення обмежень на застосування рекурентних нейронних мереж до прогнозування врожайності на основі ендогенного підходу.

З алгоритмічної точки зору архітектура LSTM дозволяє моделювати тривалі часові залежності за рахунок використання механізму пам'яті та керованого оновлення стану комірки. Архітектура GRU є компактнішою, має меншу кількість параметрів і тому може виявитися більш стійкою на коротких вибірках. У задачах прогнозування врожайності це є важливим, оскільки коротка довжина ряду часто обмежує можливість ефективного навчання складних моделей. Саме тому в межах дослідження доцільно розглядати обидві архітектури як два різні варіанти рекурентного моделювання, чутливі до структури ряду, кількості вхідних ознак і способу формування вибірки.

Архітектуру досліджуваної GRU-моделі наведено на рис. 3.9. Вона включає вхідну послідовність лагових значень трендових залишків, рекурентний шар GRU, повнозв'язний вихідний шар Dense з одним нейроном та сигмоїдальною функцією активації. Вихід моделі інтерпретується як імовірність належності наступного року до класу низької врожайності. Якщо ця ймовірність перевищує порогове значення 0,5, рік класифікується як рік із низькою врожайністю; в іншому випадку — як рік із високою врожайністю.

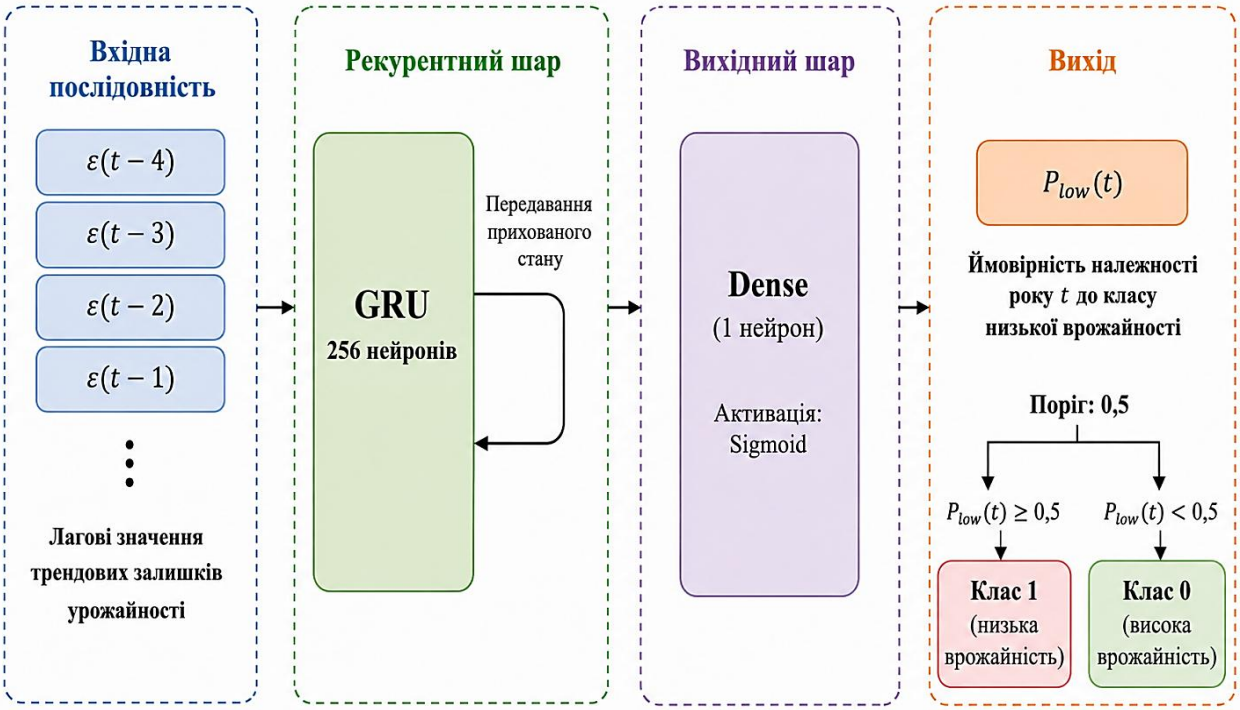

Рис. 3.9. Загальна архітектура GRU-моделі для прогнозування класу врожайності

Під час побудови рекурентних моделей принципово важливим є збереження часової структури даних. На відміну від стандартних табличних постановок машинного навчання, де можливе випадкове розбиття вибірки на навчальну та тестову частини, у випадку часових рядів необхідно дотримуватися хронологічного порядку спостережень. Це дозволяє уникнути витоку інформації з майбутніх періодів у навчання моделі та забезпечує коректність експериментальної оцінки якості прогнозу. У зв'язку з цим при використанні моделей GRU та LSTM доцільно застосовувати схеми розбиття, що зберігають часову послідовність, а також окремо контролювати вплив довжини вхідного вікна, кількості зовнішніх ознак і параметрів мережі на стабільність результатів.

Для реалізації прогнозних моделей врожайності ми використали розширені часові ряди врожайності пшениці для областей України за період 1955–2021 рр. Основним об'єктом дослідження є ряд трендових залишків врожайності, який після вилучення довгострокового тренду розглядається як стаціонарний ряд із середнім значенням, близьким до 0 [6; 14; 20; 78]. Використання трендових залишків урожайності як цільової змінної у подальших моделях прогнозування дозволяє відокремити довгострокову технологічно зумовлену складову врожайності від міжрічної кліматично зумовленої мінливості і вирішити задачу ідентифікації, оцінювання та моделювання зовнішніх впливів на коливання врожайності пшениці. Апробація моделей здійснювалася на прикладі Херсонської області. Такий вибір пояснюється тим, що врожайність пшениці в областях степової агрокліматичної зони найбільш сильно відчуває вплив коливань кліматичних факторів і це дозволяє відобразити цей вплив у межах реалізації прогнозних моделей врожайності.

У цьому підрозділі ми використовуємо ендогенні моделі для прогнозування часового ряду трендових залишків врожайності пшениці для Херсонської області. Термін «ендогенна модель» тут означає, що у ролі впливаючих факторів використовуються лише минулі значення ряду врожайності (чи ряду трендових залишків врожайності). Результуючою змінною виступає поточне значення врожайності. Ендогенна модель використовує лише внутрішні закономірності, які присутні у часовому ряду врожайності і не враховує впливу зовнішніх економічних

та кліматичних факторів. В минулому підрозділі проаналізовані класичні моделі машинного навчання Random Forest, Support Vector Machine та Logistic Regression. У даному підрозділі розглянуті удосконалені рекурентні мережі LSTM та GRU, які представляють групу моделей глибинного навчання. Дані моделі є значно потужнішими від звичайних моделей машинного навчання, але їх можливості можуть бути повністю розкриті лише для довгих часових рядів.

В ендогенному підході ширина вікна розглядалася як гіперпараметр. Ми протестували значення n_steps від 2 до 6, і вибрали найкращі значення для кожної моделі. Іншим гіперпараметром виступає $units$ — кількість нейронів у прихованому шарі. Ми використовували такі значення гіперпараметра $units$: 16, 32, 64, 128 (табл. 3.11).

Таблиця 3.11

Гіперпараметри рекурентних нейронних мереж

LSTM	n_steps	Довжина вхідної послідовності ковзного часового вікна	2, 3, 4, 5, 6
GRU	units	Кількість нейронів у прихованому шарі	16, 32, 64, 128

Модель LSTM (Long Short-Term Memory) є різновидом рекурентної нейронної мережі, спеціально розробленим для опрацювання послідовних даних і часових рядів. Її особливістю є наявність внутрішнього стану пам’яті та спеціальних керувальних механізмів — вхідного, вихідного та забувального вентилів, які регулюють збереження, оновлення і вилучення інформації з пам’яті моделі. Завдяки цьому LSTM здатна враховувати не лише короткострокові, а й потенційно довгострокові залежності між елементами часового ряду. У задачі ендогенного прогнозування врожайності така модель використовується для виявлення прихованих закономірностей у послідовності попередніх значень трендових залишків урожайності та прогнозування класу майбутнього стану — низької або високої врожайності.

У задачі ендогенного прогнозування врожайності така модель використовується для виявлення прихованих закономірностей у послідовності попередніх значень трендових залишків урожайності та прогнозування класу майбутнього стану — низької або високої врожайності.

Алгоритмічну послідовність решіткового пошуку гіперпараметрів моделі LSTM наведено на рис. 3.10. У цьому випадку зовнішній цикл відповідає перебору довжини вхідної послідовності n_steps, а внутрішній цикл — перебору кількості нейронів $units$ у рекурентному шарі. Критерієм вибору найкращої конфігурації є максимальне значення метрики $F1\_LY$ на валідаційній вибірці.

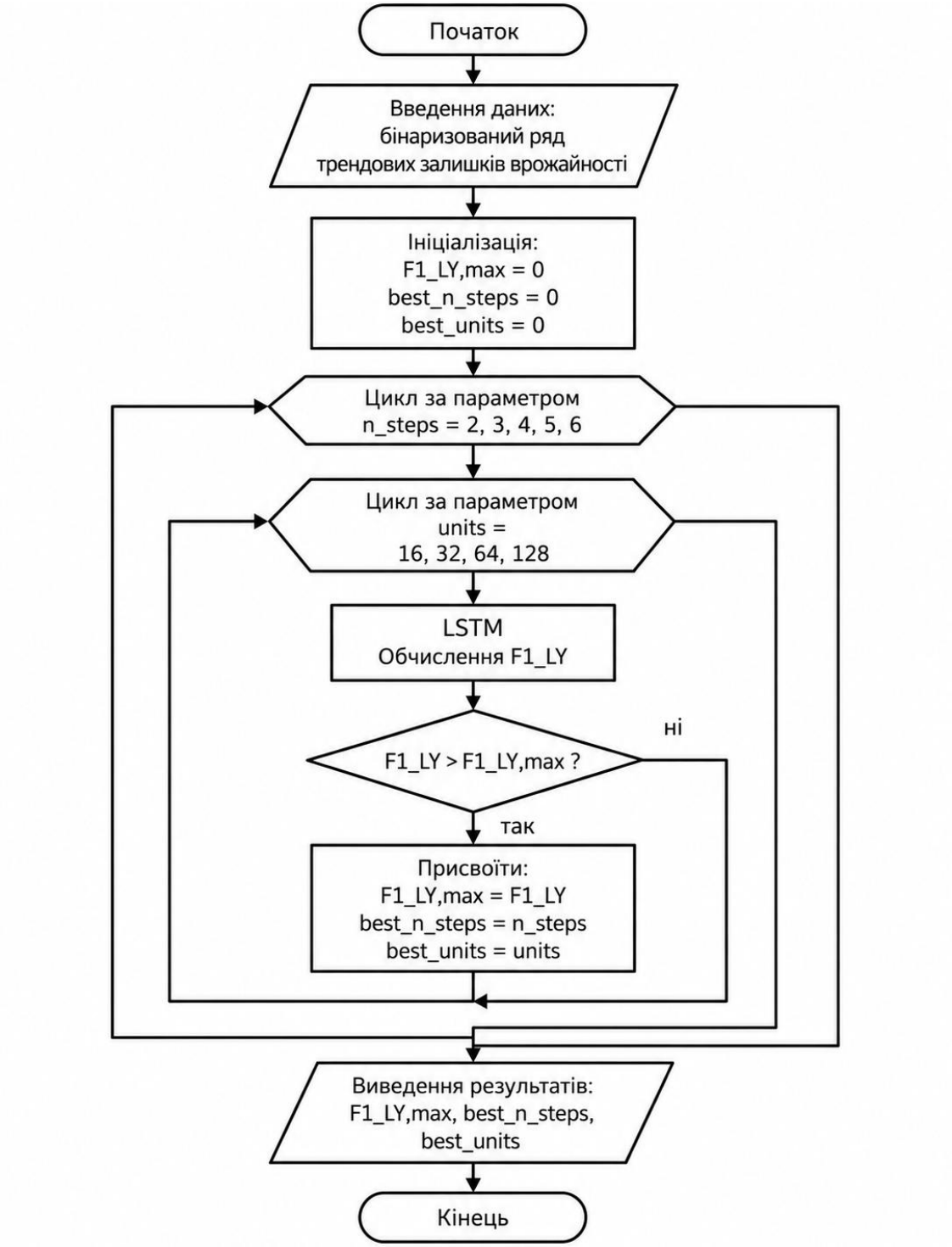

Рис. 3.10. Блок-схема алгоритму решіткового пошуку гіперпараметрів моделі LSTM

Згідно з описаною вище методикою ми здійснюємо решітковий пошук на валідаційній вибірці для пошуку найкращих гіперпараметрів прогнозних моделей. Для кожного класу моделей машинного навчання будуть відібрані три моделі з різною конфігурацією та максимальним значенням цільового показника $F1$. Метрика $F1$ була нами обрана як найбільш об'єктивний показник якості прогнозної моделі, який узгоджує точність позитивного прогнозу з повнотою виявлення об'єктів позитивного класу. Результати решіткового пошуку гіперпараметрів моделі LSTM наведені у табл. 3.12.

Таблиця 3.12

Значення метрики $F1_{LY}$ для моделі Long Short-Term Memory, розраховані на решітці гіперпараметрів

n_steps \ units	**16**	**32**	**64**	**128**
2	0,8000	0,5714	0,5714	0,2500
3	0,7500	0,5455	0,5000	0,7500
4	0,8000	0,6667	0,5714	0,6667
5	0,7500	0,6154	0,5000	0,0000
6	0,8000	0,6154	0,6154	0,5000

Як видно з табл. 3.12, найвище значення метрики $F1$ = 0,8000 отримано для набору гіперпараметрів $units$ = 16; n_{steps} = 2, 4, 6. Метрики оптимальної моделі мають такі значення: $Accuracy$ = 0,8000; $Precision$ = 0,6667; $Recall$ = 1,0000; $Specificity$ = 0,6667; ROC_AUC = 0,7500. Усі метрики розраховані на валідаційній вибірці. Графічну ілюстрацію решіткового пошуку наведено на рис. 3.11.

Модель GRU (Gated Recurrent Unit) також належить до класу рекурентних нейронних мереж із вентильними механізмами, однак має простішу архітектуру порівняно з LSTM. У GRU використовуються два основні вентилі — оновлення та скидання, які визначають, яку частину попередньої інформації потрібно зберегти, а яку — замінити новою. Завдяки меншій кількості параметрів GRU часто навчається швидше та може бути ефективнішою на невеликих наборах даних. У цьому дослідженні GRU застосовується як альтернативна рекурентна модель для ендогенної класифікації врожайності, що дозволяє перевірити, чи здатна

компактніша нейромережева архітектура краще узагальнювати інформацію, наявну в короткому часовому ряді трендових залишків.

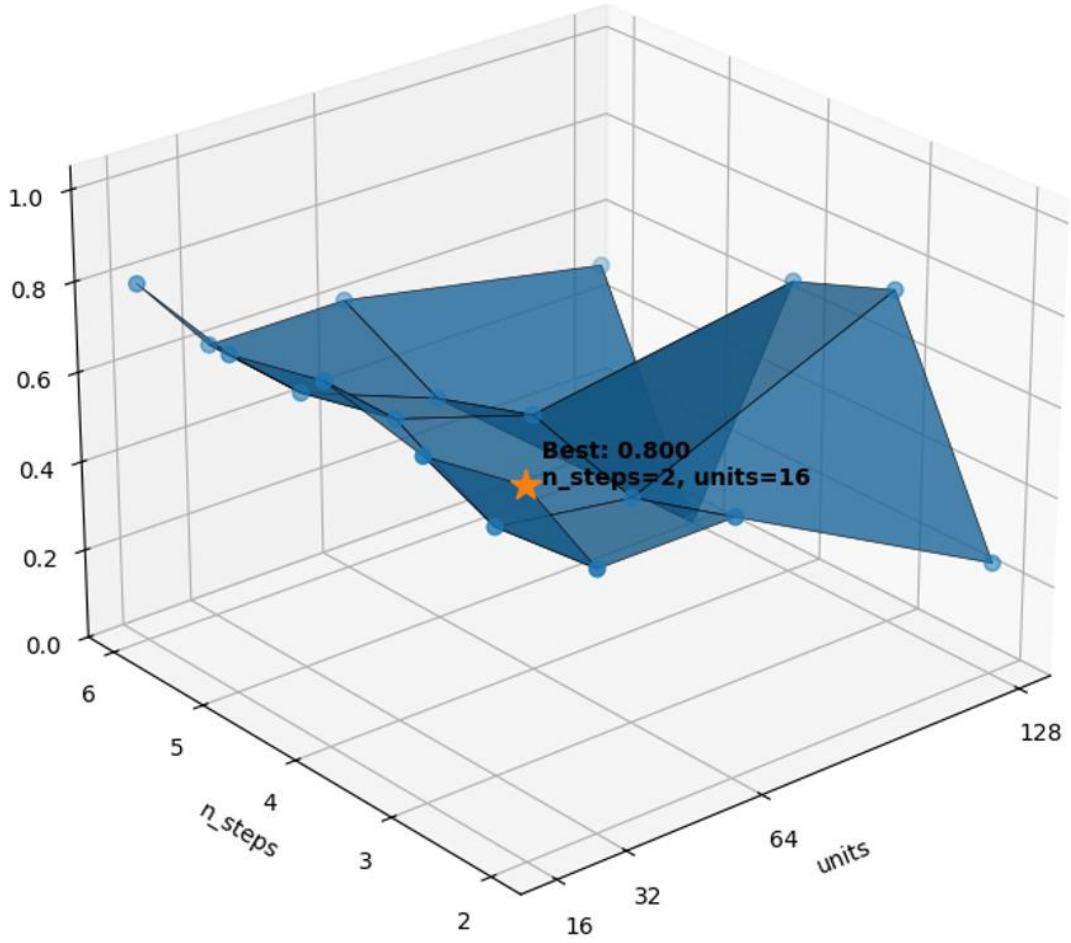

Рис. 3.11. Решітковий пошук гіперпараметрів моделі LSTM

Решітковий пошук на валідаційній вибірці оптимальних гіперпараметрів моделі GRU за цільовою метрикою $F1$ представлений у табл. 3.13.

Таблиця 3.13

Значення метрики $F1_{LY}$ для моделі Gated Recurrent Unit, розраховані на решітці гіперпараметрів

n_steps \ units	**16**	**32**	**64**	**128**
2	0,7273	0,4615	0,6154	0,2500
3	0,7273	0,5714	0,7500	0,8571
4	0,8000	0,6667	0,8571	0,8571
5	0,8000	0,4000	0,4000	0,4000
6	0,8000	0,6154	0,4000	0,4444

Як видно з табл. 3.13, найвище значення метрики $F1_{LY}$ = 0,8571 отримано для набору гіперпараметрів $units$ = 128; n_{steps} = 3, 4. Метрики оптимальної моделі мають такі значення: $Accuracy$ = 0,9000; $Precision$ = 1,0000; $Recall$ = 0,7500; $Specificity$ = 1,0000; ROC_AUC = 0,7917. Усі метрики розраховані на валідаційній вибірці. Графічну ілюстрацію решіткового пошуку наведено на рис. 3.12.

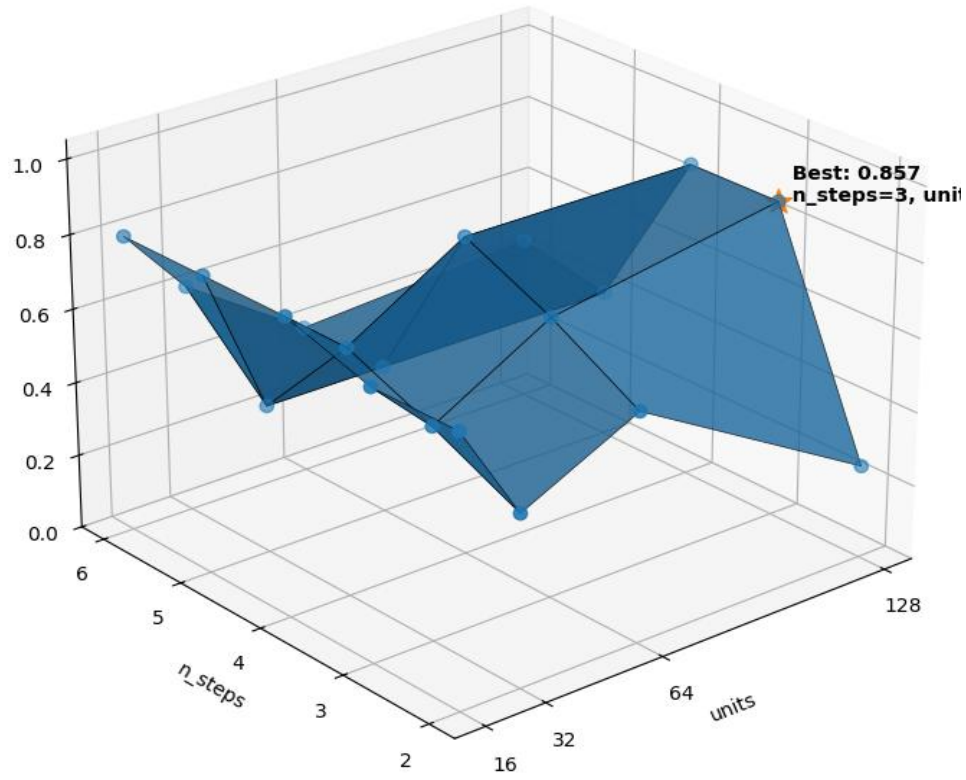

Рис. 3.12. Решітковий пошук гіперпараметрів моделі GRU

Для перевірки якості ендогенних моделей глибинного навчання використано тестову вибірку, яка не застосовувалася під час навчання моделей і налаштування їх гіперпараметрів. З огляду на малий обсяг вихідного часового ряду, оцінювання виконано не лише для однієї найкращої конфігурації, а для трьох найкращих конфігурацій кожної архітектури, відібраних за значенням метрики $F1$ на валідаційній вибірці. Це дозволяє оцінити не тільки максимальну якість окремої моделі, а й стійкість результатів для групи близьких за якістю конфігурацій. Порівняння усереднених значень метрик для валідаційної та тестової вибірок дозволяє оцінити, наскільки прогнозні властивості моделі зберігаються під час переходу від етапу налаштування гіперпараметрів до незалежного тестування. Якщо середні значення метрик на тестовій вибірці істотно знижуються порівняно з

валідаційною, це свідчить про недостатню стійкість моделі або про її чутливість до конкретного розбиття короткого часового ряду.

Таблиця 3.14

Порівняння метрик моделі LSTM для валідаційної та тестової вибірки

Валідаційна вибірка							
n_steps	**units**	**Accuracy**	**Precision**	**Recall**	**F1_LY**	**Specificity**	**ROC_AUC**
2	16	0,8000	0,6667	1,0000	0,8000	0,6667	0,7500
6	16	0,7778	0,6667	1,0000	0,8000	0,6000	0,9500
4	16	0,7778	0,6667	1,0000	0,8000	0,6000	0,9000
Середнє		**0,7852**	**0,6667**	**1,0000**	**0,8000**	**0,6222**	**0,8667**
Тестова вибірка							
n_steps	**units**	**Accuracy**	**Precision**	**Recall**	**F1_LY**	**Specificity**	**ROC_AUC**
2	16	0,5000	0,6000	0,5000	0,5455	0,5000	0,3750
6	16	0,7000	0,8000	0,6667	0,7273	0,7500	0,6667
4	16	0,7000	0,8000	0,6667	0,7273	0,7500	0,7083
Середнє		**0,6333**	**0,7333**	**0,6111**	**0,6667**	**0,6667**	**0,5833**
Різниця середніх		**0,1519**	**–0,0666**	**0,3889**	**0,1333**	**–0,0444**	**0,2833**

У табл. 3.14 наведено результати для моделі LSTM. На валідаційній вибірці три найкращі конфігурації мають однакове значення $F1$ = 0,8000, що свідчить про достатньо високу якість виявлення років низької врожайності на етапі налаштування гіперпараметрів. Середнє значення $Recall$ дорівнює 1,0000, тобто на валідаційній вибірці модель правильно виявляє всі спостереження класу низької врожайності. Водночас середнє значення $Precision$ = 0,6667 показує, що частина прогнозів класу низької врожайності є помилковими. Середнє значення ROC_AUC = 0,8667 додатково вказує на достатньо добру здатність моделі розділяти класи на валідаційній вибірці.

На тестовій вибірці якість LSTM знижується, однак це зниження не є критичним за всіма метриками. Середнє значення $F1$ = 0,6667 лише на 0,1333 менше, ніж на валідаційній вибірці. Середнє значення $Accuracy$ знижується з 0,7852 до 0,6333, а ROC_AUC — з 0,8667 до 0,5833. Найбільше зниження спостерігається для $Recall$: з 1,0000 на валідаційній вибірці до 0,6111 на тестовій. Це означає, що під час незалежного тестування модель уже не виявляє всі роки низької врожайності. Водночас середні значення $Precision$ і $Specificity$ на

тестовій вибірці навіть дещо вищі, ніж на валідаційній, що відображено від'ємними значеннями різниці середніх для цих метрик. Отже, модель LSTM демонструє помірне зниження якості на тестовій вибірці, але загалом зберігає прийнятну прогнозну здатність порівняно з іншими ендогенними моделями.

До таблиці введено рядок «Різниця середніх». Цей рядок показує зниження метрик моделі при переході від валідаційної до тестової вибірки. Чим більше значення цієї різниці, тим сильніше знижується оцінка моделі при переході від валідаційної вибірки до тестової вибірки і це означає нестійкість прогнозної сили моделі при переході до нових невідомих даних. Якщо у цьому рядку є від'ємне значення, це означає не погіршення, а навпаки — підвищення метрики на тестовій вибірці.

У табл. 3.15 наведено результати для моделі GRU. На валідаційній вибірці ця модель демонструє дуже високі значення основних метрик. Середнє значення $Accuracy$ становить 0,8926, $Precision$ — 1,0000, $F1$ — 0,8571, $Specificity$ — 1,0000, а ROC_AUC — 0,8139. Такі результати свідчать про високу якість моделі на етапі валідації, зокрема про відсутність помилкових віднесень років високої врожайності до класу низької врожайності. Проте всі три найкращі конфігурації GRU мають порівняно велику кількість нейронів у прихованому шарі — 64 або 128, що для короткого ендогенного часового ряду може підвищувати ризик перенавчання.

На тестовій вибірці якість GRU істотно погіршується. Середнє значення $Accuracy$ знижується до 0,3333, $Precision$ — до 0,4333, $Recall$ — до 0,3333, $F1$ — до 0,3757, а ROC_AUC — до 0,2778. Різниця між середніми значеннями валідаційних і тестових метрик є значною для всіх основних показників: для $F1$ вона становить 0,4814, для $Accuracy$ — 0,5593, для $Specificity$ — 0,6667, для ROC_AUC — 0,5361. Таке різке падіння якості свідчить про недостатню стійкість моделі GRU на незалежній тестовій вибірці. Отже, попри високі валідаційні показники, GRU у цій ендогенній постановці задачі не забезпечує надійного узагальнення результатів.

Таблиця 3.15

Порівняння метрик моделі GRU для валідаційної та тестової вибірки

Валідаційна вибірка							
n_steps	**units**	**Accuracy**	**Precision_LY**	**Recall_LY**	**F1_LY**	**Specificity_HY**	**ROC_AUC**
3	128	0,9000	1,0000	0,7500	0,8571	1,0000	0,7917
4	128	0,8889	1,0000	0,7500	0,8571	1,0000	0,8500
4	64	0,8889	1,0000	0,7500	0,8571	1,0000	0,8000
Середнє		**0,8926**	**1,0000**	**0,7500**	**0,8571**	**1,0000**	**0,8139**
Тестова вибірка							
n_steps	**units**	**Accuracy**	**Precision_LY**	**Recall_LY**	**F1_LY**	**Specificity_HY**	**ROC_AUC**
3	128	0,3000	0,4000	0,3333	0,3636	0,2500	0,3750
4	128	0,4000	0,5000	0,3333	0,4000	0,5000	0,2083
4	64	0,3000	0,4000	0,3333	0,3636	0,2500	0,2500
Середнє		**0,3333**	**0,4333**	**0,3333**	**0,3757**	**0,3333**	**0,2778**
Різниця середніх		**0,5593**	**0,5667**	**0,4167**	**0,4814**	**0,6667**	**0,5361**

Порівняння таблиць 3.14 і 3.15 показує, що GRU має вищі середні метрики на валідаційній вибірці, ніж LSTM, однак значно гірше переносить ці результати на тестову вибірку. LSTM, навпаки, має дещо нижчі валідаційні показники, але демонструє помітно кращу стійкість під час незалежного тестування. За середнім значенням $F1$ на тестовій вибірці LSTM досягає 0,6667, тоді як GRU — лише 0,3757. Аналогічно, середнє значення ROC_AUC для LSTM на тестовій вибірці становить 0,5833, а для GRU — 0,2778. Це дозволяє зробити висновок, що серед двох розглянутих рекурентних архітектур LSTM є більш стійкою для ендогенної класифікації врожайності на основі короткого часового ряду трендових залишків.

Отримані результати також показують, що високе значення метрик на валідаційній вибірці саме по собі не гарантує високої якості моделі на незалежних даних. Особливо це помітно для GRU, де спостерігається різкий розрив між валідаційними та тестовими показниками. Така ситуація може бути наслідком малого обсягу навчальної вибірки, слабкої автокореляційної структури ряду трендових залишків і надмірної гнучкості нейромережевої моделі порівняно з обсягом доступної інформації.

Додатковим індикатором стійкості нейромережевої прогнозної моделі може бути складність її архітектури, зокрема кількість нейронів у прихованому шарі. Якщо модель демонструє прийнятні прогнозні властивості на валідаційній вибірці

за відносно невеликої кількості нейронів, це може свідчити про кращу здатність до узагальнення та менший ризик перенавчання. Натомість, якщо найкращі валідаційні результати досягаються лише за умови використання великої кількості нейронів, така модель може надмірно адаптуватися до особливостей навчальної та валідаційної вибірок. У цьому випадку під час переходу до нових незалежних даних імовірним є істотне зниження значень тестових метрик, що свідчить про недостатню стійкість моделі. У даному дослідженні така ситуація спостерігається для моделі GRU: найкращі валідаційні результати отримано для конфігурацій із 64 та 128 нейронами, однак на тестовій вибірці якість моделі істотно знижується. Натомість для LSTM найкращі результати досягнуто за меншої кількості нейронів, що частково пояснює вищу стійкість цієї моделі під час незалежного тестування.

Попередні висновки підрозділу свідчать, що ендогенні моделі глибинного навчання мають обмежену прогнозну здатність у задачі класифікації врожайності за коротким рядом трендових залишків. Модель LSTM показала прийнятну, але не високу стійкість результатів, тоді як GRU продемонструвала ознаки нестабільності та можливого перенавчання. Це підтверджує, що лише внутрішньої динаміки трендових залишків урожайності недостатньо для побудови надійної прогнозної моделі. Тому ендогенний підхід доцільно розглядати як базовий варіант прогнозних моделей, а для підвищення точності та стійкості прогнозування необхідно залучати екзогенні агрокліматичні чинники, які безпосередньо характеризують умови формування врожайності.

3.4. Порівняльне оцінювання ефективності моделей машинного та глибинного навчання у ендогенній постановці

У підрозділах 3.2 і 3.3 було побудовано ендогенні класифікаційні моделі прогнозування врожайності на основі методів машинного та глибинного навчання. Для кожного типу моделі було виконано решітковий пошук гіперпараметрів, після чого відібрано три найкращі конфігурації за значенням метрики $F1$ на валідаційній вибірці. Підсумкова порівняльна таблиця 3.16 містить усереднені метрики для цих трьох конфігурацій на валідаційній і тестовій вибірках. Такий підхід дає змогу

оцінити не лише максимальну якість окремої моделі, а й стійкість результатів під час переходу від етапу налаштування гіперпараметрів до незалежного тестування.

Таблиця 3.16

Порівняння середніх значень $F1$ для моделей машинного та глибинного навчання на валідаційній і тестовій вибірках

Модель	Вибірка	Accuracy	Precision	Recall	F1_LY	Specificity	ROC_AUC
Random Forest	валідаційна	0,8889	1,0000	0,7500	0,8571	1,0000	0,8500
Random Forest	тестова	0,5000	0,6000	0,5000	0,5455	0,5000	0,5000
Support Vector Machine RBF	валідаційна	0,6667	0,6481	0,8333	0,7051	0,5333	0,5833
Support Vector Machine RBF	тестова	0,5000	0,5417	0,7222	0,6093	0,1667	0,4722
Logistic Regression	валідаційна	0,5556	0,5000	0,7500	0,6000	0,4000	0,6500
Logistic Regression	тестова	0,3000	0,4000	0,3333	0,3636	0,2500	0,1806
LSTM	валідаційна	0,7852	0,6667	1,0000	0,8000	0,6222	0,8667
LSTM	тестова	0,6333	0,7333	0,6111	0,6667	0,6667	0,5833
GRU	валідаційна	0,8926	1,0000	0,7500	0,8571	1,0000	0,8139
GRU	тестова	0,3333	0,4333	0,3333	0,3757	0,3333	0,2778

Порівняння результатів на валідаційній вибірці показує, що найвищі середні значення метрики $F1$ отримано для моделей Random Forest і GRU. Для обох моделей цей показник становить 0,8571. Дещо нижче значення має модель LSTM — 0,8000. Модель Support Vector Machine забезпечує середнє значення F_1 = 0,7051, а Logistic Regression — 0,6000. Отже, на валідаційній вибірці найкращі результати демонструють Random Forest і GRU, що може створювати враження їхньої переваги над іншими моделями.

Однак результати на тестовій вибірці істотно змінюють таку оцінку. Найвище середнє значення $F1$ на тестовій вибірці отримано для моделі LSTM — 0,6667. Другий результат має Support Vector Machine — 0,6093. Модель Random Forest на тестовій вибірці знижує значення $F1$ до 0,5455. Найнижчі тестові значення цього показника отримано для GRU — 0,3757 та Logistic Regression — 0,3636. Таким чином, модель GRU, яка була однією з найкращих на валідаційній вибірці, виявилася нестійкою під час незалежного тестування.

Важливим критерієм порівняння є не лише абсолютне значення тестових метрик, а й різниця між валідаційними та тестовими результатами. Чим більшою є

ця різниця, тим менш стійкою можна вважати прогнозну модель. За зниженням метрики $F1$ найменший розрив спостерігається для Support Vector Machine: середнє значення $F1$ зменшується з 0,7051 до 0,6093, тобто лише на 0,0958. Для моделі LSTM різниця становить 0,1333, що також свідчить про відносно прийнятну стійкість. Для Logistic Regression зниження становить 0,2364, для Random Forest — 0,3116, а для GRU — 0,4814. Отже, найбільшу нестабільність результатів демонструє GRU, тоді як SVM і LSTM мають більш стійкі значення цільової метрики. Графічна ілюстрація значення метрики $F1$ для різних моделей представлена на рис. 3.13.

Разом з тим стійкість SVM потребує обережної інтерпретації. Хоча ця модель має найменше зниження $F1$, на тестовій вибірці вона характеризується низьким середнім значенням $Specificity$ = 0,1667. Це означає, що модель краще виявляє роки низької врожайності, але гірше розпізнає роки високої врожайності. Така властивість може бути прийнятною для задач ризикового попередження, де важливішим є виявлення несприятливих років, однак для збалансованої класифікації вона є обмеженням.

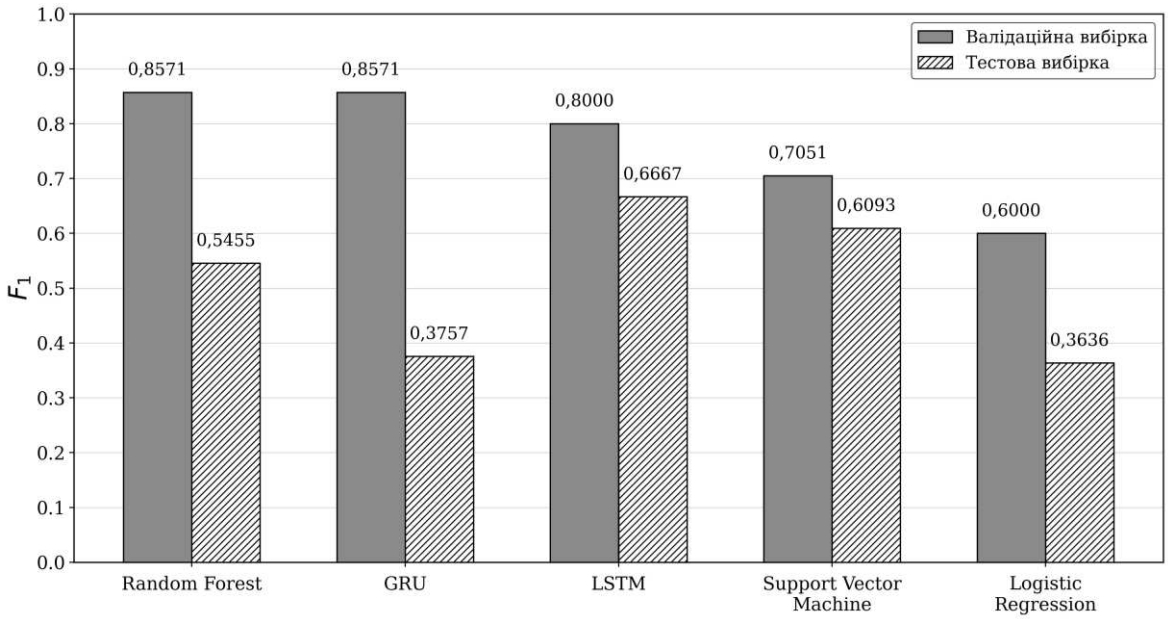

Рис. 3.13. Порівняння середніх значень $F1$ для моделей машинного та глибинного навчання

Модель LSTM демонструє найкращий баланс між якістю та стійкістю. На тестовій вибірці вона має найвище середнє значення $F1$ = 0,6667, найвище середнє значення $Accuracy$ = 0,6333 і найкраще середнє значення ROC_AUC = 0,5833 серед усіх розглянутих моделей. Хоча порівняно з валідаційною вибіркою спостерігається зниження $Recall$ і ROC_AUC, модель LSTM загалом краще зберігає прогнозні властивості на незалежних даних, ніж GRU та Random Forest. Додатково слід зазначити, що найкращі конфігурації LSTM отримано за відносно малої кількості нейронів у прихованому шарі, що знижує ризик перенавчання.

Модель GRU, навпаки, має дуже високі валідаційні показники, але різке погіршення якості на тестовій вибірці. Середнє значення $F1$ зменшується з 0,8571 до 0,3757, $Accuracy$ — з 0,8926 до 0,3333, а ROC_AUC — з 0,8139 до 0,2778. Такий розрив свідчить про недостатню здатність моделі до узагальнення. Однією з можливих причин є надмірна складність архітектури порівняно з обсягом доступного часового ряду, оскільки найкращі конфігурації GRU були отримані для 64 і 128 нейронів у прихованому шарі.

Порівняння моделей машинного та глибинного навчання показує, що глибинне навчання не забезпечує автоматичної переваги в ендогенній постановці задачі. Серед моделей глибинного навчання LSTM показала найкращий результат, тоді як GRU виявилася нестійкою. Серед моделей машинного навчання найкращим варіантом за тестовим $F1$ є Support Vector Machine, однак ця модель має низьку специфічність. Random Forest має дуже високі валідаційні показники, але суттєво втрачає якість на тестовій вибірці, що може свідчити про чутливість до конкретного розбиття даних. Logistic Regression демонструє найнижчу тестову якість серед усіх моделей.

Узагальнюючи результати, можна зробити висновок, що серед розглянутих ендогенних моделей найкращий компроміс між якістю прогнозування та стійкістю результатів забезпечує LSTM. Вона має найвище середнє значення $F1$ на тестовій вибірці та найкращі тестові значення $Accuracy$ і ROC_AUC. Support Vector Machine можна розглядати як конкурентну модель з погляду стійкості метрики $F1$, однак її низька специфічність обмежує застосування для збалансованої класифікації. GRU

і Random Forest демонструють ознаки нестабільності, оскільки їх високі валідаційні результати не підтверджуються на тестовій вибірці. Варто зауважити, що такий детальний аналіз точності, класифікаційної здатності та прогнозної стійкості моделей став можливим завдяки використанню методики решіткового пошуку гіперпараметрів. Саме ця методика дала змогу не обмежуватися оцінюванням окремої випадково вибраної конфігурації, а порівняти групи найкращих моделей, відібраних за валідаційною вибіркою, та перевірити збереження їхніх прогнозних властивостей на незалежній тестовій вибірці. Отже, решітковий пошук у цьому дослідженні виконує не лише технічну функцію підбору гіперпараметрів, а й методичну функцію порівняльного оцінювання стійкості прогнозних моделей.

Отримані результати підтверджують загальне обмеження ендогенного підходу до прогнозування врожайності. Короткий часовий ряд трендових залишків містить недостатньо стійкої внутрішньої інформації для побудови надійної класифікаційної моделі. Тому ендогенні моделі доцільно розглядати як базовий рівень прогнозного моделювання. Для підвищення точності, стійкості та інтерпретованості прогнозування необхідно переходити до екзогенного підходу із залученням агрокліматичних факторів, які безпосередньо характеризують умови формування врожайності.

Висновки до розділу 3

У розділі 3 досліджено ендогенний підхід до прогнозування врожайності, за якого як вхідна інформація для моделей використовується лише внутрішня динаміка часового ряду врожайності або його трендових залишків. Така постановка дозволяє оцінити, наскільки сам часовий ряд містить достатньо інформації для прогнозування майбутнього класу врожайності без залучення зовнішніх агрокліматичних або економічних факторів.

Сформовано методичну схему ендогенного класифікаційного прогнозування врожайності. Вона передбачає побудову тренду врожайності, розрахунок трендових залишків, їх бінаризацію на класи низької та високої врожайності, формування лагового простору ознак методом ковзного вікна, хронологічний поділ даних на

навчальну, валідаційну та тестову вибірки, решітковий пошук гіперпараметрів і незалежне оцінювання якості моделей на тестовій вибірці.

Обґрунтовано використання класифікаційної постановки задачі, у якій позитивним класом визначено клас низької врожайності. Такий підхід має прикладне значення для задач раннього виявлення несприятливих років, оцінювання ризиків і підтримки управлінських рішень в аграрному виробництві. Як основну цільову метрику для підбору гіперпараметрів використано $F1$, оскільки вона поєднує точність позитивного прогнозу та повноту виявлення років низької врожайності.

Для класичних моделей машинного навчання досліджено Random Forest, Support Vector Machine з RBF-ядром та Logistic Regression. Результати решіткового пошуку показали, що окремі конфігурації цих моделей можуть демонструвати прийнятні або високі значення метрик на валідаційній вибірці. Зокрема, Random Forest забезпечив високе середнє значення $F1$ на валідаційній вибірці, однак під час незалежного тестування якість моделі суттєво знизилася. Модель SVM показала кращу стійкість за метрикою $F1$, проте характеризувалася низькою специфічністю на тестовій вибірці, що свідчить про схильність до надмірного прогнозування класу низької врожайності. Logistic Regression продемонструвала найнижчу тестову якість серед класичних моделей машинного навчання.

Для моделей глибинного навчання досліджено рекурентні архітектури LSTM і GRU. Встановлено, що LSTM забезпечила кращу стійкість результатів на тестовій вибірці порівняно з GRU. Хоча GRU мала високі валідаційні метрики, під час переходу до тестової вибірки спостерігалося різке погіршення якості, що може свідчити про перенавчання або надмірну складність моделі для короткого часового ряду. LSTM показала найкращий компроміс між якістю прогнозування та стійкістю результатів серед усіх розглянутих ендогенних моделей.

Порівняльне оцінювання моделей машинного та глибинного навчання показало, що високі значення метрик на валідаційній вибірці не гарантують збереження якості на незалежній тестовій вибірці. Найвищі середні значення $F1$ на валідаційній вибірці отримано для Random Forest і GRU, однак саме ці моделі

продемонстрували суттєве зниження якості на тестових даних. Найкраще середнє значення $F1$ на тестовій вибірці отримано для LSTM, а найменше зниження цієї метрики — для SVM. Це підтверджує необхідність оцінювати моделі не лише за максимальним валідаційним результатом, а й за стійкістю прогнозних властивостей на незалежних даних.

Використання решіткового пошуку гіперпараметрів дало змогу не обмежуватися оцінюванням окремої конфігурації моделі, а порівняти групи найкращих моделей за валідаційною вибіркою та перевірити збереження їхніх прогнозних властивостей на тестовій вибірці. Таким чином, решітковий пошук у цьому розділі виконує не лише технічну функцію підбору гіперпараметрів, а й методичну функцію оцінювання стійкості прогнозних моделей.

Отримані результати свідчать про обмежену прогнозну здатність ендогенного підходу для короткого часового ряду трендових залишків урожайності. Внутрішня динаміка такого ряду не містить достатньо стійкої інформації для побудови надійної класифікаційної моделі. Тому ендогенні моделі доцільно розглядати як базовий або порівняльний рівень прогнозного моделювання. Для підвищення точності, стійкості та інтерпретованості прогнозування необхідно перейти до екзогенного підходу із залученням агрокліматичних факторів, які безпосередньо характеризують умови формування врожайності.

РОЗДІЛ 4. ЕКЗОГЕННІ МОДЕЛІ МАШИННОГО ТА ГЛИБИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ

4.1. Екзогенні класифікаційні моделі машинного навчання

Екзогенний підхід до прогнозного моделювання врожайності передбачає використання не лише внутрішньої динаміки самого ряду врожайності, а й зовнішніх факторів, які безпосередньо характеризують умови її формування. У цьому дослідженні такими факторами є агрокліматичні показники: декадні температурні характеристики та місячні суми опадів за період активної вегетації. На відміну від ендогенного підходу, де модель використовує лише попередні значення трендових залишків урожайності, екзогенна постановка дає змогу врахувати фізично змістовні причини міжрічних коливань урожайності.

Схема експериментальних досліджень у цьому підрозділі будується з урахуванням регіональної агрокліматичної неоднорідності України. У розділі 2 була запропонована методика агрокліматичної кластеризації областей та подальшої регіональної агрегації даних у межах однорідних зон. Це дозволяє збільшити обсяг навчальних вибірок, підвищити статистичну надійність моделей і зменшити вплив випадкових локальних коливань. Саме тому при проведенні комп'ютерних експериментів у цьому підрозділі моделі будуються не на рівні окремої області, а на рівні степової, лісостепової або західної агрокліматичної зони.

Для екзогенного моделювання врожайності у роботі використано об'єднані вибірки агрокліматичних даних, сформовані окремо для кожної з виділених раніше агрокліматичних зон. Кожна така вибірка містить 132 спостереження типу «область–рік» і охоплює шість областей відповідної зони за період 2000–2021 рр. Для кожного спостереження задано 12 кліматичних ознак періоду вегетації поточного року: дев'ять декадних температурних показників t1–t9 та три місячні суми опадів R10, R20, R30. Крім того, у вихідному наборі даних містяться фактичне та трендове значення врожайності, на основі яких розраховується трендовий залишок ($\varepsilon = y - y_{trend}$) і формується бінарна цільова змінна $eps1$. У матрицю вхідних ознак прогнозної моделі включаються лише агрокліматичні показники, тоді

як фактична врожайність, тренд, трендовий залишок і сформована на його основі змінна $eps1$ використовуються лише для побудови цільового класу та не подаються на вхід класифікатора.

У межах класифікаційної методики прогнозування цільова змінна формується на основі трендового залишку врожайності ε, який визначається як різниця між фактичною врожайністю та її трендовим рівнем. Якщо $\varepsilon > 0$, спостереження належить до класу високої врожайності ($eps1$=0), а якщо $\varepsilon \leq 0$ — до класу низької врожайності ($eps1$=1). Таким чином, задача прогнозування формулюється як задача бінарної класифікації, у якій позитивним класом є клас низької врожайності. Саме цей клас має найбільше прикладне значення, оскільки він пов'язаний з ризиком несприятливого агровиробничого результату.

У подальших комп'ютерних експериментах для верифікації екзогенних класифікаційних моделей використано поділ даних на навчальну, валідаційну та тестову вибірки. Такий підхід є доцільним, оскільки в цьому підрозділі виконується решітковий пошук гіперпараметрів моделей машинного навчання. Навчальна вибірка використовується для побудови моделей, валідаційна — для вибору найкращих конфігурацій гіперпараметрів, а тестова — для незалежного оцінювання якості вибраних моделей [64].

У кожній агрокліматичній зоні вихідний набір даних було поділено у співвідношенні 70 % / 15 % / 15 %. Навчальна вибірка містить 92 спостереження, валідаційна — 20 спостережень, тестова — 20 спостережень. Під час поділу зберігалося співвідношення класів низької та високої врожайності, що забезпечує коректніше порівняння метрик на різних етапах оцінювання.

Вибір оптимальних гіперпараметрів здійснювався за значенням метрики $F1$ на валідаційній вибірці. Після завершення решіткового пошуку для кожної моделі було відібрано три найкращі конфігурації за цією метрикою. Далі ці конфігурації оцінювалися на тестовій вибірці, яка не використовувалася ні для навчання, ні для вибору гіперпараметрів. Це дало змогу порівняти не лише максимальні валідаційні результати, а й стійкість прогнозних властивостей моделей на незалежних даних.

Степова зона. Перед побудовою моделей було проаналізовано розподіл класів у сформованій вибірці. Загальний обсяг даних становить 132 спостереження, з яких 75 спостережень належать до класу високої врожайності, а 57 — до класу низької врожайності. Частка класу низької врожайності становить 43,18 %, а класу високої врожайності — 56,82 %. Отже, вибірка не є різко незбалансованою, що створює сприятливіші умови для побудови класифікаційних моделей порівняно з випадком суттєвого домінування одного класу.

Для верифікації моделей дані були поділені на навчальну, валідаційну та тестову вибірки. Навчальна вибірка містить 92 спостереження, валідаційна та тестова — по 20 спостережень. При цьому співвідношення класів у валідаційній і тестовій вибірках є однаковим: 9 спостережень класу низької врожайності та 11 спостережень класу високої врожайності. Це важливо, оскільки забезпечує коректніше порівняння метрик між валідаційною та тестовою вибірками.

Для вибору оптимальних гіперпараметрів було виконано решітковий пошук для трьох моделей машинного навчання: Logistic Regression, Random Forest та Support Vector Machine з RBF-ядром. Основним критерієм вибору гіперпараметрів була метрика $F1$ на валідаційній вибірці. Такий вибір є методично доцільним, оскільки F1 одночасно враховує точність прогнозування класу низької врожайності та повноту його виявлення. Усереднені значення гіперпараметрів моделей машинного навчання, визначені за результатами решіткового пошуку, наведені у табл. 4.1.

Таблиця 4.1

Оптимізація гіперпараметрів машинного навчання за методом решіткового пошуку

Модель	**Вибірка**	**Accuracy**	**Precision**	**Recall**	**F1_LY**	**Specificity**	**ROC_AUC**
Log Regression	Валідаційна	0,9000	0,8889	0,8889	0,8889	0,9091	0,9024
Log Regression	тестова	0,8000	0,8571	0,6667	0,7500	0,9091	0,8822
	Різниця	**0,1000**	**0,0318**	**0,2222**	**0,1389**	**0,0000**	**0,0202**
Random Forest	Валідаційна	0,8000	0,7778	0,7778	0,7778	0,8182	0,8620
Random Forest	Тестова	0,7500	0,7500	0,6667	0,7059	0,8182	0,9024
	Різниця	**0,0500**	**0,0278**	**0,1111**	**0,0719**	**0,0000**	**–0,0404**
SVM RBF	Валідаційна	0,7500	0,7500	0,6667	0,7059	0,8182	0,8384
SVM RBF	Тестова	0,8500	0,8000	0,8889	0,8421	0,8182	0,9091
	Різниця	**–0,1000**	**–0,0500**	**–0,2222**	**–0,1362**	**0,0000**	**–0,0707**

Для моделі Logistic Regression найкращі результати на валідаційній вибірці отримано для значень параметра регуляризації C = 0,1; C = 0,2; C = 0,3. Усі три конфігурації мають однакове значення $F1$ = 0,8889, $Accuracy$ = 0,9000, $Recall$ = 0,8889 та $Specificity$ = 0,9091. Це свідчить про те, що модель є досить стійкою до невеликих змін параметра регуляризації в цьому діапазоні. На тестовій вибірці середнє значення $F1$ для трьох найкращих конфігурацій становить 0,7500, $Accuracy$ = 0,8000, $Precision$ = 0,8571, $Recall$ = 0,6667, $Specificity$ = 0,9091, а ROC_AUC = 0,8822). Отже, модель Logistic Regression демонструє добру якість і достатню стійкість, хоча під час переходу до тестової вибірки спостерігається певне зниження повноти виявлення класу низької врожайності.

Для моделі Random Forest найкращі валідаційні результати отримано для конфігурацій з невеликою глибиною дерев. До трьох найкращих увійшли варіанти з параметрами $n\_estimators$ = 40, max_depth = 2, $n\_estimators$ = 40, max_depth = 3 та $n\_estimators$ = 80, max_depth = 2. Для цих конфігурацій середнє значення $F1$ на валідаційній вибірці становить 0,7778, а на тестовій — 0,7059. Зниження метрики $F1$ становить лише 0,0719, що свідчить про відносно добру стійкість моделі. Водночас за абсолютним значенням тестової метрики $F1$ Random Forest поступається Logistic Regression і SVM RBF. Це означає, що в екзогенній постановці для даного набору ознак ансамблева модель дерев рішень не забезпечила найвищої якості класифікації, хоча її результати є стабільними.

Для моделі Support Vector Machine з RBF-ядром найкращі валідаційні результати отримано для конфігурацій C = 1, $gamma$ = 0,125; C = 1, $gamma$ = 0,15; C = 1,5, $gamma$ = 0,125. На валідаційній вибірці середнє значення $F1$ становить 0,7059, що є нижчим, ніж для Logistic Regression і Random Forest. Проте на тестовій вибірці модель SVM RBF демонструє найкращий результат серед усіх розглянутих моделей: середнє значення метрики $Accuracy$ становить 0,8500, $Precision$ = 0,8000, $Recall$ = 0,8889, $F1$ = 0,8421, $Specificity$ = 0,8182, а ROC_AUC = 0,9091. Важливо, що значення тестових метрик для SVM RBF навіть перевищують відповідні валідаційні показники. Зокрема, різниця усереднених метрик Val – Test

для F1 дорівнює –0,1362, що означає не погіршення, а покращення якості на тестовій вибірці.

Порівняння трьох моделей за результатами тестової вибірки показує, що найвищу якість класифікації забезпечила модель SVM RBF. Вона має найкраще середнє значення $F1$ = 0,8421, найвище значення $Recall$ = 0,8889 і найбільше значення ROC_AUC = 0,9091. Це означає, що модель добре виявляє роки низької врожайності та має високу здатність розділяти класи високої й низької врожайності. Logistic Regression посідає друге місце за тестовим значенням $F1$ = 0,7500, але має найвище значення $Specificity$ = 0,9091, тобто краще за інші моделі розпізнає роки високої врожайності. Random Forest показав третій результат за $F1$ = 0,7059, але продемонстрував достатньо стабільну поведінку між валідаційною та тестовою вибірками (рис. 4.1).

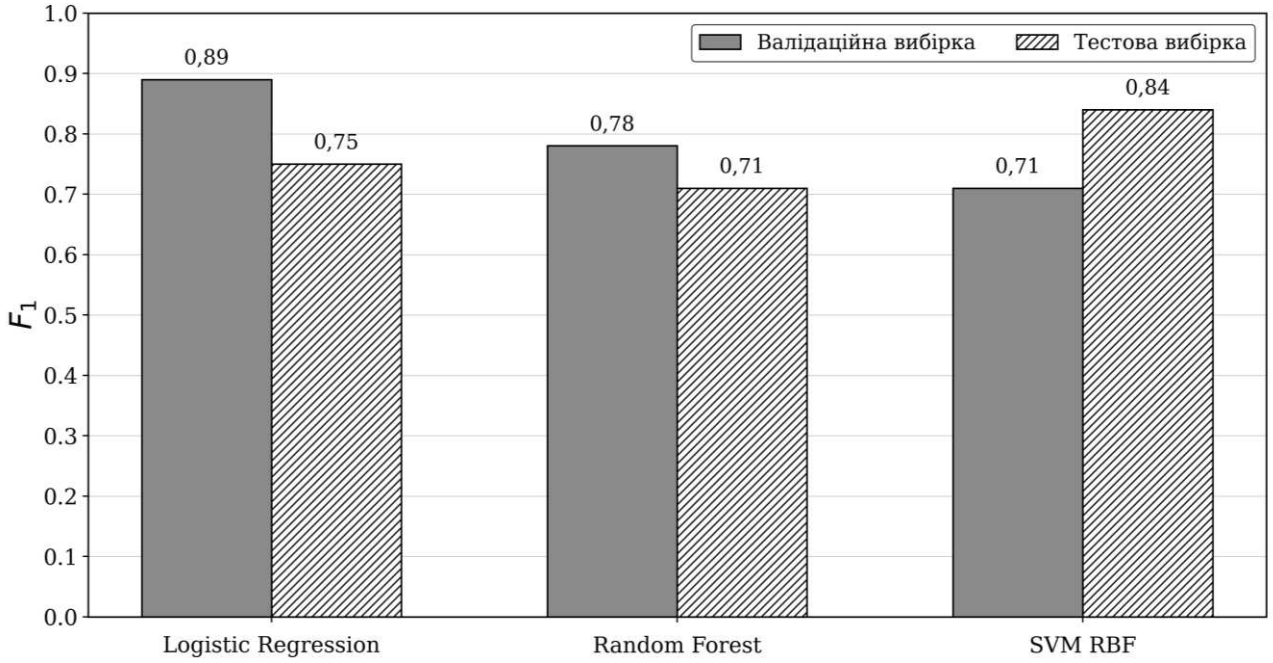

Рис. 4.1. Усереднене значення метрики $F1$ для валідаційної вибірки (суцільний фон) та тестової вибірки (штриховий фон). Степовий регіон

Отримані результати свідчать, що екзогенний підхід суттєво підвищує якість класифікаційного прогнозування порівняно з ендогенною постановкою, оскільки моделі використовують не лише внутрішню динаміку ряду, а й агрокліматичні фактори, які безпосередньо впливають на формування врожайності. Особливо важливим є те, що на тестовій вибірці всі три моделі забезпечили значення $F1$ вище

0,70, а модель SVM RBF досягла значення 0,8421. Це підтверджує інформативність сформованого простору агрокліматичних ознак для задачі прогнозування класу врожайності.

З погляду практичного використання, модель SVM RBF можна розглядати як найкращу серед досліджених моделей машинного навчання в екзогенній постановці, оскільки вона забезпечує найвищу збалансовану якість виявлення класу низької врожайності. Logistic Regression є важливою альтернативою завдяки простоті, інтерпретованості та високій специфічності. Random Forest демонструє стабільність, однак за підсумковим значенням $F1$ поступається двом іншим моделям. Загалом результати решіткового пошуку підтверджують доцільність використання екзогенних агрокліматичних факторів для прогнозного моделювання врожайності та створюють основу для подальшого порівняння моделей машинного й глибинного навчання в екзогенній постановці.

Лісостепова зона. Для лісостепової агрокліматичної зони було виконано решітковий пошук гіперпараметрів для трьох моделей машинного навчання: Logistic Regression, Random Forest та SVM RBF. Як і в попередніх експериментах, основним критерієм вибору найкращих конфігурацій була метрика $F1$, оскільки вона характеризує якість виявлення класу низької врожайності. Усереднені значення гіперпараметрів моделей машинного навчання, визначені за результатами решіткового пошуку, наведені у табл. 4.2.

Таблиця 4.2

Оптимізація гіперпараметрів машинного навчання за методом решіткового пошуку

Модель	**Вибірка**	**Accuracy**	**Precision**	**Recall**	**F1_LY**	**Specificity**	**ROC_AUC**
Log Regression	Валідаційна	0,7000	0,6154	0,8889	0,7273	0,5455	0,7946
Log Regression	Тестова	0,6667	0,5636	0,7500	0,6433	0,6111	0,7014
	Різниця	**0,0333**	**0,0517**	**0,1389**	**0,0840**	**–0,0657**	**0,0932**
Random Forest	Валідаційна	0,7167	0,6576	0,7778	0,7123	0,6667	0,7576
Random Forest	Тестова	0,6500	0,5455	0,7500	0,6316	0,5833	0,7153
	Різниця	**0,0667**	**0,1121**	**0,0278**	**0,0807**	**0,0833**	**0,0423**
SVM RBF	Валідаційна	0,7000	0,7143	0,5556	0,6250	0,8182	0,6667
SVM RBF	Тестова	0,6833	0,6476	0,5000	0,5607	0,8056	0,7049
	Різниця	**0,0167**	**0,0667**	**0,0556**	**0,0643**	**0,0126**	**–0,0382**

На валідаційній вибірці найкращий середній результат за метрикою F1 отримано для моделі Logistic Regression — 0,7273. Ця модель також має найвище значення $Recall$ = 0,8889, що свідчить про її здатність виявляти більшість років низької врожайності. Водночас значення $Specificity$ = 0,5455 є порівняно невисоким, тобто модель частіше помилково відносить частину років високої врожайності до класу низької.

Модель Random Forest на валідаційній вибірці показала близький результат: середнє значення $F1$ становить 0,7123. Порівняно з Logistic Regression вона має нижчу повноту виявлення класу низької врожайності $Recall$ = 0,7778, але кращу специфічність $Specificity$ = 0,6667. Це означає, що Random Forest дещо збалансованіше розпізнає обидва класи, хоча за основною цільовою метрикою поступається Logistic Regression.

Модель SVM RBF на валідаційній вибірці має нижче значення $F1$ = 0,6250, проте характеризується найвищою специфічністю $Specificity$ = 0,8182. Отже, ця модель краще розпізнає роки високої врожайності, але гірше виявляє роки низької врожайності, що підтверджується нижчим значенням $Recall$ = 0,5556.

На тестовій вибірці найвище середнє значення $F1$ також отримано для моделі Logistic Regression — 0,6433. Модель Random Forest має дуже близький результат $F1$ = 0,6316, тоді як SVM RBF поступається двом іншим моделям за цією метрикою $F1$ = 0,5607. Таким чином, у лісостеповій зоні Logistic Regression показала найкращий результат за основною цільовою метрикою як на валідаційній, так і на тестовій вибірках.

Порівняння валідаційних і тестових результатів показує, що для всіх моделей спостерігається помірне зниження якості на тестовій вибірці (рис. 4.2). Для Logistic Regression зниження $F1$ становить 0,0840, для Random Forest — 0,0807, для SVM RBF — 0,0643. Найменше зниження цієї метрики має SVM RBF, однак її абсолютне значення $F1$ залишається найнижчим серед трьох моделей. Тому стабільність SVM RBF не компенсує її слабшу здатність виявляти клас низької врожайності.

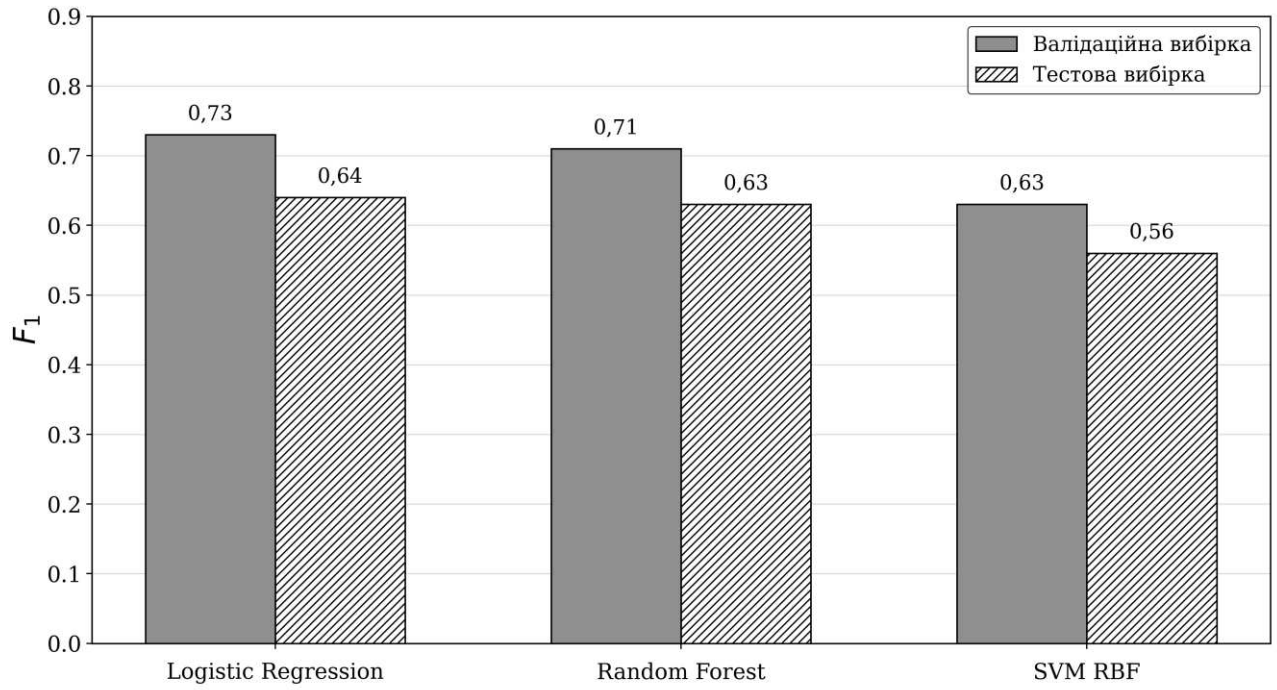

Рис. 4.2. Усереднене значення метрики $F1$ для валідаційної вибірки (суцільний фон) та тестової вибірки (штриховий фон). Лісостеповий регіон

Отже, для лісостепової зони результати класифікаційного прогнозування є помірними. Найкращою моделлю за метрикою $F1$ можна вважати Logistic Regression, яка забезпечує найвищу повноту виявлення років низької врожайності та найкраще середнє значення $F1$ на тестовій вибірці. Random Forest демонструє близький результат і може розглядатися як альтернативна модель із дещо кращим балансом між розпізнаванням низької та високої врожайності. Модель SVM RBF у цьому експерименті краще розпізнає роки високої врожайності, але менш ефективна для виявлення класу низької врожайності, який є основним з погляду оцінювання агровиробничого ризику.

Західна зона. Для західної агрокліматичної зони результати решіткового пошуку свідчать про вищу ефективність моделей машинного навчання порівняно з лісостеповою зоною. Усі три моделі забезпечили на тестовій вибірці значення $F1$ вище 0,66, а найкращий результат отримано для моделі SVM RBF (табл. 4.3).

Таблиця 4.3

Оптимізація гіперпараметрів машинного навчання за методом решіткового пошуку

Модель	Вибірка	Accuracy	Precision	Recall	F1_LY	Specificity	ROC_AUC
Log Regression	Валідаційна	0,6500	0,5714	0,8889	0,6957	0,4545	0,6263
Log Regression	Тестова	0,6000	0,5333	0,8889	0,6667	0,3636	0,6095
	Різниця	**0,0500**	**0,0381**	**0,0000**	**0,0290**	**0,0909**	**0,0168**
Random Forest	Валідаційна	0,7667	0,7593	0,7037	0,7299	0,8182	0,8350
Random Forest	Тестова	0,6667	0,6010	0,7778	0,6778	0,5758	0,8384
	Різниця	**0,1000**	**0,1583**	**–0,0741**	**0,0521**	**0,2424**	**–0,0034**
SVM RBF	Валідаційна	0,8500	0,8750	0,7778	0,8235	0,9091	0,8687
SVM RBF	Тестова	0,7500	0,6667	0,8889	0,7619	0,6364	0,7845
	Різниця	**0,1000**	**0,2083**	**–0,1111**	**0,0616**	**0,2727**	**0,0842**

На валідаційній вибірці найвищу якість класифікації продемонструвала модель SVM RBF. Для трьох найкращих конфігурацій її середнє значення $F1$ становить 0,8235, $Accuracy$ = 0,8500, $Precision$ = 0,8750, $Recall$ = 0,7778, $Specificity$ = 0,9091, а ROC_AUC = 0,8687. Це свідчить про добру здатність моделі розпізнавати як роки низької, так і роки високої врожайності.

Модель Random Forest на валідаційній вибірці показала другий результат за метрикою $F1$ = 0,7299. Вона має достатньо збалансовані показники точності, повноти та специфічності: $Precision$ = 0,7593, $Recall$ = 0,7037, $Specificity$ = 0,8182. Значення ROC_AUC = 0,8350 також підтверджує добру роздільну здатність цієї моделі.

Модель Logistic Regression має нижче валідаційне значення $F1$ = 0,6957. Її особливістю є висока повнота виявлення класу низької врожайності $Recall$ = 0,8889, однак специфічність є низькою $Specificity$ = 0,4545. Це означає, що модель добре виявляє роки низької врожайності, але водночас часто помилково відносить роки високої врожайності до класу низької.

На тестовій вибірці найкращий результат також отримано для моделі SVM RBF. Її середнє значення $F1$ становить 0,7619, що є найвищим серед усіх розглянутих моделей. Крім того, SVM RBF має найвище значення $Accuracy$ = 0,7500 і високу повноту виявлення класу низької врожайності $Recall$ = 0,8889. Водночас порівняно з валідаційною вибіркою помітно знижується специфічність: з

0,9091 до 0,6364. Це свідчить про те, що на тестових даних модель частіше помилково класифікує роки високої врожайності як низьковрожайні.

Random Forest на тестовій вибірці посідає друге місце за метрикою $F1$ = 0,6778. Порівняно з валідаційною вибіркою у цієї моделі знижується точність і специфічність, але повнота виявлення класу низької врожайності навіть зростає з 0,7037 до 0,7778. Значення ROC_AUC залишається практично незмінним і навіть трохи зростає: з 0,8350 до 0,8384, що свідчить про достатню стійкість роздільної здатності моделі.

Logistic Regression на тестовій вибірці має $F1$ = 0,6667, що є нижчим, ніж у SVM RBF та Random Forest. Разом з тим ця модель демонструє найменше зниження $F1$ між валідаційною і тестовою вибірками — лише 0,0290. Це свідчить про відносну стабільність моделі, хоча її абсолютна якість є нижчою. Основним недоліком Logistic Regression залишається низька специфічність $Specificity$ = 0,3636, тобто недостатньо надійне розпізнавання років високої врожайності.

Порівняння моделей показує, що для західної агрокліматичної зони найкращою за основною цільовою метрикою $F1$ є модель SVM RBF (рис. 4.3). Вона має найвище значення цієї метрики як на валідаційній, так і на тестовій вибірках. Random Forest забезпечує дещо нижчий, але достатньо стабільний результат і характеризується високим значенням ROC_AUC. Logistic Regression є найстабільнішою за різницею між валідаційним і тестовим $F1$, однак поступається іншим моделям за загальною якістю класифікації.

Отже, для західного регіону модель SVM RBF можна розглядати як найефективнішу серед досліджених моделей машинного навчання в екзогенній постановці. Вона найкраще виявляє роки низької врожайності й забезпечує найвищу збалансовану якість класифікації. Водночас під час інтерпретації результатів слід враховувати зниження специфічності на тестовій вибірці, що вказує на певну схильність моделі до надмірного прогнозування класу низької врожайності.

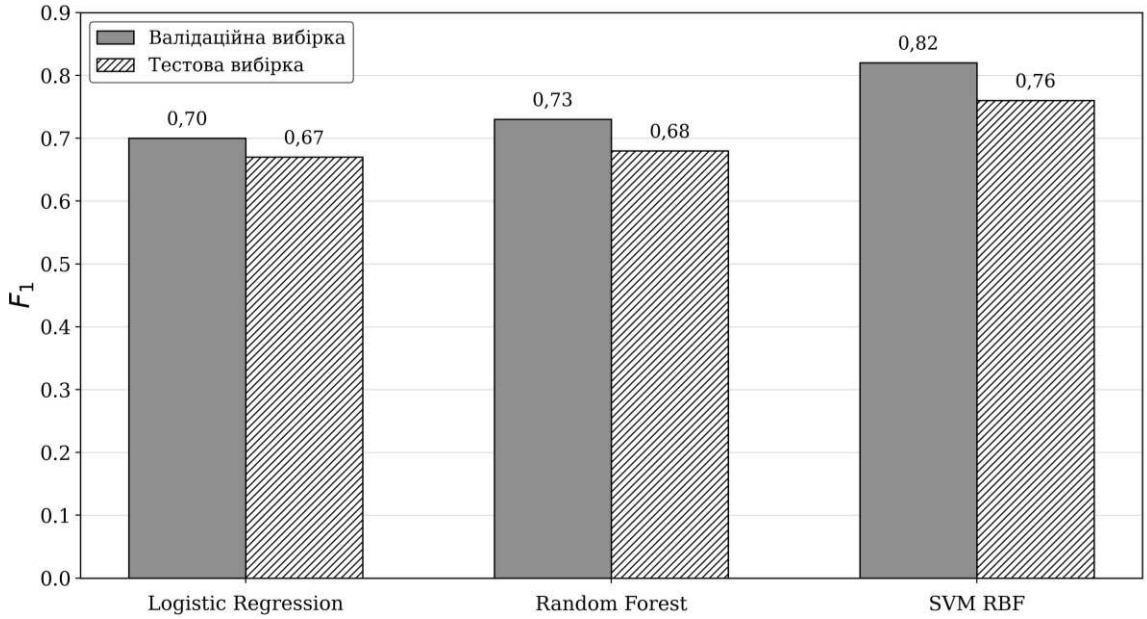

Рис. 4.3. Усереднене значення метрики $F1$ для валідаційної вибірки (суцільний фон) та тестової вибірки (штриховий фон). Західний регіон

Узагальнююче порівняння результатів екзогенного класифікаційного моделювання для трьох агрокліматичних зон показує, що якість моделей машинного навчання істотно залежить від регіональної структури даних. Хоча для всіх зон використовувалася однакова методика формування цільової змінної, однаковий набір кліматичних ознак і спільна схема решіткового пошуку, отримані значення метрики F1 відрізняються як між моделями, так і між зонами.

За результатами тестової вибірки найвищу якість класифікації отримано для степової зони. Середнє значення $F1$ для трьох моделей у цій зоні становить приблизно 0,7660. Найкращий результат забезпечила модель SVM RBF, для якої $F1$ = 0,8421. Це свідчить про високу інформативність агрокліматичних ознак у степовій зоні, де міжрічні коливання врожайності значною мірою пов’язані з температурно-вологісним режимом періоду вегетації.

Для західної агрокліматичної зони результати також є достатньо високими. Середнє тестове значення $F1$ для трьох моделей становить приблизно 0,7021. Найкращу якість знову забезпечила модель SVM RBF, для якої $F1$ = 0,7619. Random Forest і Logistic Regression показали нижчі, але прийнятні результати: відповідно 0,6778 і 0,6667. Отже, для західної зони нелінійна модель SVM RBF найкраще використовує інформацію, наявну в агрокліматичних ознаках.

Найнижчі результати отримано для лісостепової зони. Середнє тестове значення F1 для трьох моделей становить приблизно 0,6119. Найкращою моделлю у цій зоні є Logistic Regression з $F1$ = 0,6433, тоді як Random Forest має дуже близький результат $F1$ = 0,6316. Модель SVM RBF у лісостеповій зоні показала найнижче значення $F1$ = 0,5607, що свідчить про її слабшу здатність виявляти роки низької врожайності для цього регіону.

Порівняння моделей у розрізі всіх агрокліматичних зон (табл. 4.4) показує, що найвищий середній тестовий результат за метрикою $F1$ має SVM RBF. Середнє значення цієї метрики для трьох зон становить приблизно 0,7216. Модель Logistic Regression має середнє тестове значення $F1$ близько 0,6867, а Random Forest — близько 0,6718. Отже, за узагальненим критерієм якості SVM RBF є найефективнішою моделлю серед розглянутих, хоча її перевага не є однаковою для всіх зон.

Таблиця 4.4

Порівняння тестових значень метрики $F1$ для моделей машинного навчання у трьох агрокліматичних зонах

Зона	**Logistic Regression**	**Random Forest**	**SVM RBF**	**Найкраща модель**
Степова	0,7500	0,7059	0,8421	SVM RBF
Лісостепова	0,6433	0,6316	0,5607	Logistic Regression
Західна	0,6667	0,6778	0,7619	SVM RBF
Середнє	0,6867	0,6718	0,7216	SVM RBF

Важливо, що SVM RBF забезпечила найкращий результат у двох із трьох агрокліматичних зон — степовій та західній. Проте у лісостеповій зоні найкращою виявилася Logistic Regression. Це означає, що універсальної моделі, яка була б безумовно найкращою для всіх регіонів, не виявлено. Ефективність конкретного алгоритму залежить від характеру регіональних агрокліматичних залежностей і структури розподілу класів у відповідній зональній вибірці.

Random Forest не показала найкращого результату в жодній із трьох зон, однак у всіх випадках забезпечила проміжну або близьку до найкращої якість. Особливо важливо, що ця модель демонструє відносно стабільну поведінку під час

переходу від валідаційної до тестової вибірки. Тому Random Forest можна розглядати як надійну альтернативну модель, яка не завжди дає максимальне значення $F1$, але забезпечує прийнятний баланс між якістю та стійкістю.

Logistic Regression продемонструвала найкращий результат у лісостеповій зоні та достатньо високі результати у степовій і західній зонах. Її перевагою є простота, інтерпретованість і відносно стабільна поведінка. Водночас у західній зоні ця модель має низьку специфічність, що свідчить про схильність до надмірного віднесення частини років високої врожайності до класу низької врожайності.

Загалом отримані результати підтверджують доцільність використання екзогенного підходу до класифікаційного прогнозування врожайності. У всіх трьох агрокліматичних зонах моделі машинного навчання змогли використати інформацію, наявну в кліматичних ознаках поточного року, для виявлення ризику низької врожайності. Найкращі результати отримано для степової та західної зон, тоді як для лісостепової зони якість класифікації є помірною. Це свідчить про необхідність зонального підходу до прогнозного моделювання врожайності та підтверджує, що агрокліматична неоднорідність території має бути врахована під час побудови інтелектуальних моделей прогнозування.

Таким чином, за підсумками трьох агрокліматичних зон модель SVM RBF можна вважати найрезультативнішою за середнім значенням тестової метрики $F1$. Logistic Regression є найбільш інтерпретованою та ефективною для лісостепової зони, а Random Forest може розглядатися як стійка альтернативна модель. Отримані результати створюють підстави для подальшого порівняння класичних моделей машинного навчання з моделями глибинного навчання в екзогенній постановці.

4.2. Екзогенні моделі глибинного навчання

У підрозділі 3.3 проведено дослідження ефективності застосування рекурентних нейронних мереж до прогнозування врожайності на основі ендогенного підходу. При цьому результативність моделі суттєво залежить від довжини часового ряду врожайності та наявності внутрішніх зв’язків між елементами цього ряду. Для коротких часових рядів трендових залишків

врожайності із слабко вираженими внутрішніми зв’язками такий підхід має обмежену ефективність. У такому випадку виникає необхідність переходу від ендогенних до екзогенних рекурентних моделей. У такому випадку на вхід мережі подаються не лише значення врожайності або трендових залишків за попередні періоди, а й додаткові зовнішні змінні, що характеризують агрокліматичні та режимні умови. До них можуть належати температурні та опадові показники, сформовані в розділі 1, а також окремі економічні чи режимні змінні, якщо вони є доступними для конкретної постановки задачі. Такий підхід дає змогу поєднати часову динаміку врожайності з інформацією про зовнішні чинники, що потенційно підвищує точність і стійкість прогнозу.

Екзогенний підхід до прогнозного моделювання врожайності ґрунтується на залученні зовнішніх пояснювальних змінних, які відображають умови формування врожаю. У межах цього дослідження такими змінними є агрокліматичні показники поточного вегетаційного періоду, зокрема декадні температурні характеристики та місячні суми опадів. На відміну від ендогенної постановки, у якій прогноз будується лише на основі попередніх значень трендових залишків врожайності, екзогенний підхід дає змогу врахувати вплив температурно-вологісного режиму на міжрічні коливання врожайності та підвищити змістовну інтерпретованість прогнозної моделі.

У цьому підрозділі ми здійснимо апробацію моделей глибинного навчання на прикладі Херсонської області. Ця область відноситься до степової зони України. Врожайність пшениці в областях степової агрокліматичної зони найбільш сильно відчуває вплив коливань кліматичних факторів і це дозволяє відобразити цей вплив у межах реалізації прогнозних моделей врожайності. Для комп’ютерних експериментів ми використали розширений набір агрокліматичних даних для Херсонської області України за період 1955–2021 рр. Кожен зразок включає кліматичні показники t1–t9, R10, R20, R30, трендовий залишок врожайності ε, та категоріальну економічну змінну $econ$, яка відображає тип економічних відносин, який був характерним для даного часового відрізку спостережень. Змінна $econ$ приймає значення 0 для періоду колективного землеробства (1955–1989), $econ$=1

для перехідного періоду (1990–1999) і значення $econ$=2 для періоду ринково-орієнтованої системи сільськогосподарського виробництва (2000–2021). Результуюча змінна — це позначка класу врожайності: 0 позначає високий урожай, а 1 — низький урожай.

Для вибору найкращої конфігурації моделі ми застосували методику решіткового пошуку. У цьому випадку решітка є одновимірною, оскільки оптимізації піддається лише один гіперпараметр units — кількість нейронів у прихованому шарі нейронної мережі RNN. Результати решіткового пошуку наведені у табл. 4.5. До таблиці включені лише три найкращі конфігурації за критерієм $F1$ для моделей глибинного навчання LSTM та GRU. Порівняння метрик цих моделей на даних валідаційної та тестової вибірки дозволяє зробити висновок про стійкість прогнозних характеристик моделі при переході від валідаційної вибірки до нових невідомих даних.

Таблиця 4.5

Метрики найкращих конфігурацій моделей LSTM та GRU для валідаційної та тестової вибірки

Model	**units**	**Accuracy**	**Precision**	**Recall**	**F1_LY**	**Specificity**	**ROC_AUC**
LSTM	16	0,9000	0,8333	1,0000	0,9091	0,8000	0,9600
LSTM	32	0,8000	0,7143	1,0000	0,8333	0,6000	1,0000
LSTM	128	0,8000	0,7143	1,0000	0,8333	0,6000	0,9200
Середнє		**0,8333**	**0,7540**	**1,0000**	**0,8586**	**0,6667**	**0,9600**
LSTM	16	0,6364	0,6667	0,6667	0,6667	0,6000	0,5667
LSTM	32	0,6364	0,6667	0,6667	0,6667	0,6000	0,6667
LSTM	128	0,7273	0,7143	0,8333	0,7692	0,6000	0,7000
Середнє		**0,6667**	**0,6826**	**0,7222**	**0,7009**	**0,6000**	**0,6445**
Різниця середніх		**0,1666**	**0,0714**	**0,2778**	**0,1577**	**0,0667**	**0,3155**
GRU	64	0,9000	0,8333	1,0000	0,9091	0,8000	1,0000
GRU	128	0,9000	0,8333	1,0000	0,9091	0,8000	0,9600
GRU	32	0,8000	0,7143	1,0000	0,8333	0,6000	0,9200
Середнє		**0,8667**	**0,7936**	**1,0000**	**0,8838**	**0,7333**	**0,9600**
GRU	64	0,7273	0,8000	0,6667	0,7273	0,8000	0,6667
GRU	128	0,7273	0,7143	0,8333	0,7692	0,6000	0,5667
GRU	32	0,6364	0,6667	0,6667	0,6667	0,6000	0,6667
Середнє		**0,6970**	**0,7270**	**0,7222**	**0,7211**	**0,6667**	**0,6334**
Різниця середніх		**0,1697**	**0,0666**	**0,2778**	**0,1628**	**0,0667**	**0,3266**

У табл. 4.5 наведено результати оцінювання трьох найкращих конфігурацій моделей LSTM та GRU в екзогенній постановці для Херсонської області. На відміну від ендогенного підходу, у цьому випадку на вхід рекурентних нейронних мереж подаються не лише значення трендових залишків урожайності, а й зовнішні пояснювальні змінні: агрокліматичні показники та економічний фактор $econ$. Оскільки використовується інформація лише поточного року, параметр n_steps не розглядається як гіперпараметр, а основним параметром решіткового пошуку є кількість нейронів у прихованому шарі $units$.

На валідаційній вибірці обидві рекурентні архітектури показали високі результати. Для LSTM середнє значення $F1$ для трьох найкращих конфігурацій становить 0,8586, а для GRU — 0,8838. В обох випадках середнє значення $Recall$ дорівнює 1,0000, тобто на етапі валідації моделі виявляють усі роки низької врожайності. Це є важливим результатом з погляду задачі оцінювання агровиробничого ризику, оскільки позитивним класом у дослідженні є саме клас низької врожайності.

Разом з тим висока повнота виявлення класу низької врожайності супроводжується помірною специфічністю. Для LSTM середнє значення $Specificity$ на валідаційній вибірці становить 0,6667, а для GRU — 0,7333. Це означає, що частина років високої врожайності помилково відноситься моделями до класу низької врожайності. Така властивість може бути прийнятною для задач раннього попередження про ризик, однак для збалансованої класифікації її потрібно враховувати під час інтерпретації результатів.

На тестовій вибірці якість обох моделей знижується (рис. 4.4), проте падіння метрики F1 не є критичним. Для LSTM середнє тестове значення $F1$ становить 0,7009, а для GRU — 0,7211. Отже, в екзогенній постановці GRU демонструє дещо кращу якість за основною цільовою метрикою. Вона також має вищі середні значення $Accuracy$ — 0,6970 проти 0,6667 для LSTM, $Precision$ — 0,7270 проти 0,6826, а також $Specificity$ — 0,6667 проти 0,6000. Водночас LSTM має трохи вище середнє значення ROC_AUC: 0,6445 проти 0,6334 для GRU.

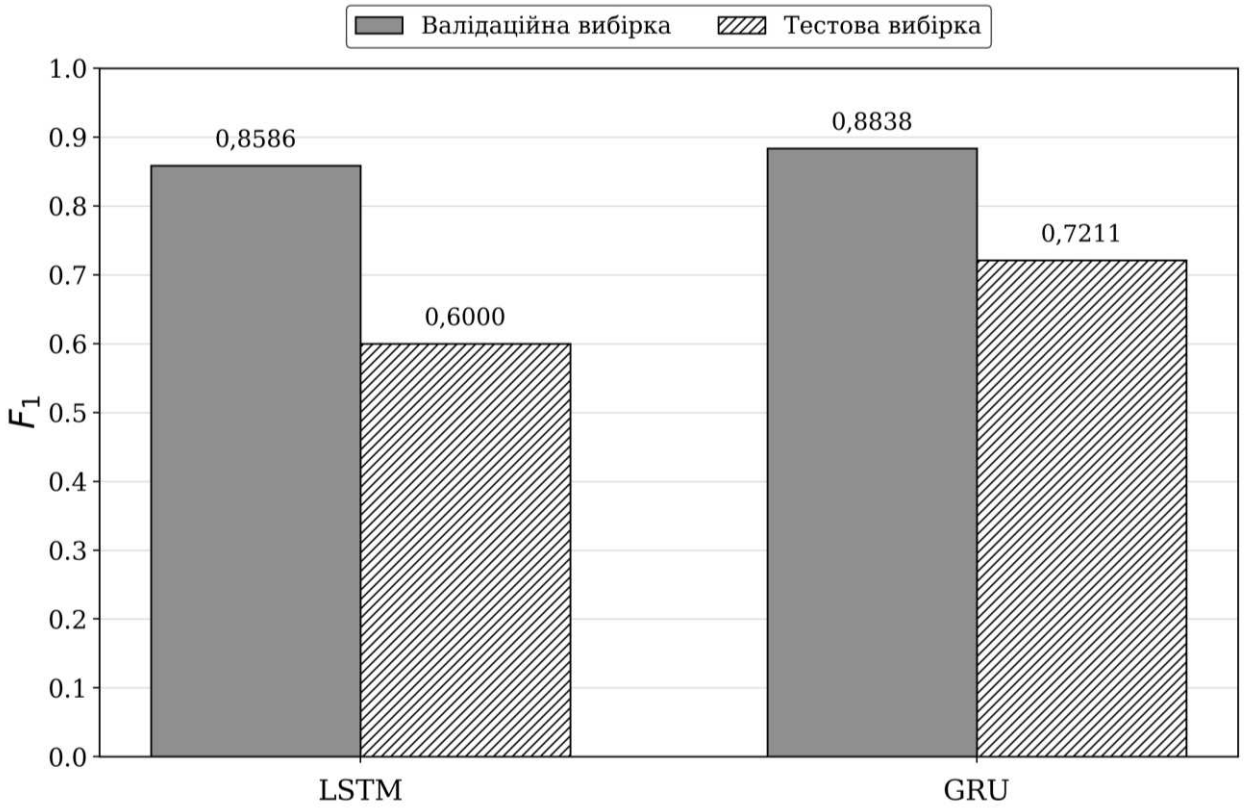

Рис. 4.4. Значення метрики $F1$ на валідаційній (суцільний фон) та тестовій вибірці (штриховий фон) для моделей LSTM та GRU

Рядок «різниця середніх» показує, наскільки знижуються метрики під час переходу від валідаційної до тестової вибірки. Для LSTM зниження $F1$ становить 0,1577, а для GRU — 0,1628. Отже, за стійкістю основної метрики обидві моделі є близькими. Найбільше погіршення спостерігається для $Recall$: в обох моделей він знижується з 1,0000 на валідаційній вибірці до 0,7222 на тестовій. Це означає, що на незалежних даних моделі вже не виявляють усі роки низької врожайності, хоча загальний рівень виявлення цього класу залишається прийнятним.

Порівняння LSTM і GRU показує, що в екзогенній постановці GRU має невелику перевагу за більшістю тестових метрик. Найкращі конфігурації GRU відповідають значенням $units$ = 64, 128 і 32, тоді як для LSTM до трьох найкращих увійшли конфігурації з $units$ = 16, 32 і 128. Це свідчить, що в екзогенній постановці висока якість не пов’язана лише з максимальною кількістю нейронів. Моделі з помірною складністю також здатні забезпечувати прийнятні результати.

Далі проведено порівняльний аналіз ефективності моделей машинного навчання на прикладі степової зони та моделей глибинного навчання для Херсонської області (табл. 4.6).

Таблиця 4.6

Порівняльна характеристика моделей машинного і глибинного навчання за метрикою $F1$ на тестовій вибірці

	Logistic Regression	**Random Forest**	**SVM RBF**	**LSTM**	**GRU**
Ендогенний підхід	0,3636	0,5455	0,6093	0,6667	0,3757
Екзогенний підхід	0,7500	0,7059	0,8421	0,7009	0,7211

Порівняльний аналіз моделей машинного та глибинного навчання за метрикою $F1$ на тестовій вибірці показує суттєву перевагу екзогенного підходу над ендогенним (рис. 4.5). У всіх розглянутих моделях, за винятком LSTM, приріст якості є суттєвим. Це підтверджує, що залучення зовнішніх агрокліматичних та економічних факторів істотно підвищує інформативність вхідного простору ознак порівняно з використанням лише внутрішньої динаміки трендових залишків урожайності [4; 7–9; 24].

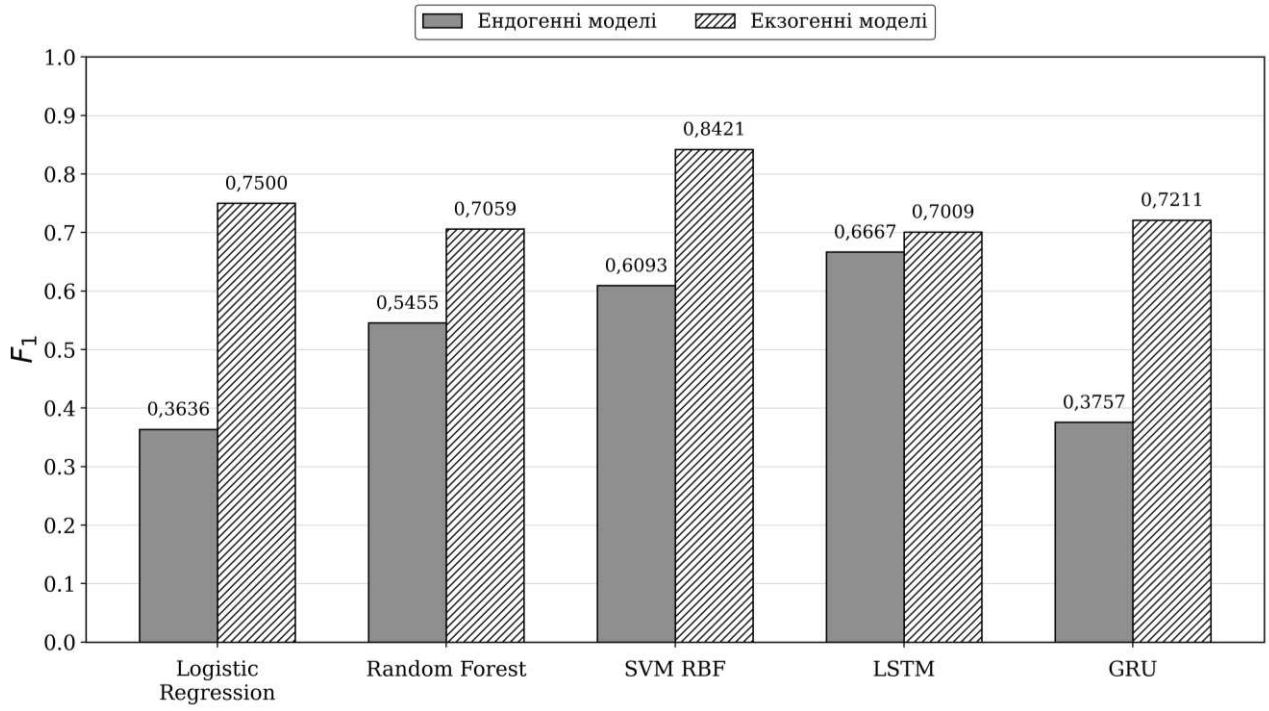

Рис. 4.5. Значення метрики $F1$ для ендогенних моделей (суцільний фон) та екзогенних моделей (штриховий фон)

Найбільше покращення спостерігається для моделі GRU. В ендогенній постановці її тестове значення $F1$ становило лише 0,3757, тоді як в екзогенній постановці воно зросло до 0,7211. Отже, приріст становить 0,3454. Це свідчить про те, що низька якість GRU в ендогенному підході була зумовлена не стільки недоліками самої архітектури, скільки недостатньою інформативністю короткого ряду трендових залишків. Після додавання зовнішніх факторів модель GRU змогла значно краще виявляти роки низької врожайності.

Суттєве покращення також отримано для Logistic Regression. Значення $F1$ зросло з 0,3636 в ендогенній постановці до 0,7500 в екзогенній, тобто приріст становить 0,3864. Це є найбільшим приростом серед усіх моделей. Такий результат показує, що навіть відносно проста лінійна модель може бути достатньо ефективною за умови використання змістовно обґрунтованого простору агрокліматичних ознак.

Для SVM RBF значення $F1$ зросло з 0,6093 до 0,8421. Ця модель показала найвище абсолютне значення цільової метрики в екзогенній постановці. Отже, саме SVM RBF забезпечила найкращу якість виявлення класу низької врожайності серед усіх розглянутих моделей. Такий результат свідчить про важливість нелінійного розділення класів у просторі агрокліматичних ознак.

Модель Random Forest також покращила результат: $F1$ зросло з 0,5455 до 0,7059. Хоча ця модель не стала найкращою в екзогенній постановці, вона забезпечила прийнятну якість класифікації і підтвердила загальну перевагу екзогенного підходу.

Найменший приріст отримано для LSTM. Значення $F1$ зросло з 0,6667 до 0,7009, тобто лише на 0,0342. Це означає, що LSTM уже в ендогенній постановці була найсильнішою серед моделей глибинного навчання, тому додавання екзогенних факторів покращило її результат лише помірно. Водночас у екзогенній постановці LSTM поступилася GRU, яка значно краще використала додаткову інформацію про агрокліматичні та економічні умови.

Загалом таблиця показує, що екзогенний підхід забезпечує вищу якість класифікаційного прогнозування для всіх розглянутих моделей. Найкращою

моделлю в ендогенній постановці була LSTM з $F1 = 0{,}6667$, тоді як в екзогенній постановці найкращий результат отримано для SVM RBF з $F1 = 0{,}8421$. Це підтверджує, що підвищення якості прогнозування досягається не лише за рахунок вибору складнішої моделі, а насамперед завдяки змістовному розширенню простору вхідних ознак.

Отримані результати свідчать, що класичні моделі машинного навчання в екзогенній постановці залишаються дуже конкурентними порівняно з моделями глибинного навчання. Зокрема, SVM RBF і Logistic Regression перевищують за метрикою $F1$ обидві рекурентні моделі. Отриманий результат можна пояснити малою довжиною часових рядів врожайності. Ця обставина не дозволила повністю розкрити прогнозні можливості потужних моделей глибинного навчання LSTM та GRU.

4.3. Нелінійний підхід до прогнозування врожайності

Лінійний підхід до прогнозування врожайності передбачає адитивний характер впливу кліматичних факторів на врожайність, тобто кожний температурний або опадовий показник розглядається незалежно від інших. Проте формування врожаю відбувається послідовно в часі, а вплив погодних умов однієї фази вегетації може залежати від умов попередньої або наступної фази. Для врахування такого кумулятивного ефекту до вхідного простору моделей включено добутки кліматичних факторів, які відповідають суміжним часовим інтервалам.

У регресійній постановці цільовою змінною є трендовий залишок врожайності ε, тоді як вхідними змінними виступають агрокліматичні ознаки, сформовані в підрозділі 2.1. До базового набору вхідних факторів належать дев’ять декадних температурних показників квітня–червня t1–t9 та три місячні суми опадів R10, R20, R30. У даному підрозділі цей базовий простір може бути розширений шляхом синтезу нелінійних ознак, а також залученням додаткових зовнішніх змінних у нейромережевих постановках задачі прогнозування. Таким чином, сформований у другому розділі набір даних виступає основою для побудови різних типів прогнозних моделей.

Для оцінювання якості прогнозних моделей використовують різні групи метрик залежно від постановки задачі. Для регресійних моделей основними показниками виступають коефіцієнт детермінації R^2, середня абсолютна похибка MAE та корінь із середньоквадратичної похибки $RMSE$ [160; 170]. Коефіцієнт детермінації характеризує частку варіації цільової змінної, пояснену моделлю, і може бути поданий у вигляді

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{4.1}$$

де y_i — фактичне значення цільової змінної, $\hat{y}_i$ — її прогнозоване значення, $\bar{y}$ — середнє значення цільової змінної у вибірці.

Середня абсолютна похибка визначається як

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{4.2}$$

а корінь із середньоквадратичної похибки — як

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4.3}$$

Ці показники дозволяють оцінити точність регресійного прогнозу як у середньому, так і з урахуванням впливу великих помилок.

4.3.1. Введення нелінійних кліматичних ознак

Точність моделювання врожайності засобами лінійних моделей зазвичай є низькою. У зв’язку з цим виникає потреба дослідити нелінійні моделі прогнозування врожайності та підходи до синтезу інформативних кліматичних ознак, які здатні розширити лінійний ознаковий простір. Необхідність такого підходу зумовлена тим, що вплив температури та вологозабезпечення на врожайність має неадитивний характер і залежить від поєднання факторів у суміжних фазах вегетації. Крім того, відомо, що вегетація рослин відбувається лише у межах зони екологічної толерантності (зона оптимальних температур, зона оптимальних опадів). Наближення температурних або вологісних параметрів до меж цієї зони нелінійно знижує продуктивність рослин. Тому, вплив кліматичних факторів на врожайність доцільно описувати не лише лінійними залежностями, а і ознаками другого порядку.

У сучасних дослідженнях неодноразово підкреслюється, що залежність урожайності від погодних умов є нелінійною, а екстремальні погодні значення можуть справляти значно сильніший вплив, ніж це випливає з лінійних моделей. У праці [2] для України прямо показано, що врахування нелінійних факторів дозволяє значно підвищити якість регресійного опису врожайності. Зокрема, для трьох виділених агрокліматичних зон — степової, лісостепової та західної — побудовано нелінійні регресійні моделі, які істотно перевищують за точністю відповідні лінійні аналоги.

У межах дисертаційного дослідження як базові кліматичні предиктори використовуються середньодекадні температури квітня–червня t1–t9 та місячні суми опадів R10, R20, R30. Саме ці показники було обґрунтовано у другому розділі як достатній набір ознак для аналізу лінійного впливу погодно-кліматичних умов на врожайність. Для переходу до нелінійного моделювання до базового набору факторів додаються квадрати змінних та добутки послідовних (у часі) факторів.

Зокрема, для температурних показників сформовано ознаки t1t2, t2t3, …, t8t9, що характеризують мультиплікативну взаємодію температурного режиму двох послідовних декад. Для показників опадів використано добутки R10R20 та R20R30, які відображають спільний вплив вологозабезпечення в послідовні місяці критичного періоду вегетації. Такий підхід можна пояснити на прикладі. Нехай, наприклад посуха на часовому відрізку 10 днів зменшує врожайність на 3 ц/га. Якщо посуха триває двадцять днів підряд, то врожайність зменшиться не на 6 ц/га (6 = 3 + 3) а на 9 ц/га (9 = 3 × 3).

Включення добутків послідовних ознак дозволяє описувати ситуації, коли несприятливий вплив одного кліматичного чинника посилюється або послаблюється умовами наступного часового інтервалу. Наприклад, висока температура наприкінці травня може мати особливо негативний вплив на врожайність за умови збереження високої температури на початку червня, тоді як достатня кількість опадів у суміжні місяці може частково компенсувати температурний стрес. Отже, добутки суміжних кліматичних факторів використовуються як ознаки другого порядку, призначені для моделювання

неадитивного та кумулятивного характеру кліматичного впливу на врожайність [2; 15; 17]. Таким чином, ми реалізуємо, поставлене раніше завдання щодо синтезу нелінійних кліматичних ознак для врахування неадитивного та кумулятивного впливу погодних умов на врожайність.

Загальний вигляд нелінійної регресійної моделі можна подати як

$$\varepsilon = \beta_0 + \sum_{i=1}^{12} \beta_i x_i + \sum_{i=1}^{12} \gamma_i {x_i}^2 + \sum_{j=1}^{10} \delta_j x_j x_{j+1} + \eta, \qquad (4.4)$$

де x_i — базові кліматичні фактори t1–t9, R10, R20, R30, β_i — коефіцієнти при лінійних членах, γ_i — коефіцієнти при квадратичних членах, δ_j — коефіцієнти при добутках послідовних факторів, η — випадкова похибка. На практиці до фінальної моделі типу (4.4) включаються лише статистично значущі члени, що забезпечує її інтерпретованість і знижує ризик перенавчання.

У праці [2] показано, що структура нелінійних моделей суттєво відрізняється для різних агрокліматичних зон України. Для степової зони фінальна регресійна модель містить 4 лінійні фактори, 7 квадратів факторів і 3 добутки послідовних факторів (табл. 4.7).

Таблиця 4.7

Нелінійна OLS регресійна модель врожайності пшениці для степової зони

Factor	Coefficient	Std. Error	t-statistic	p-value	CI 2,5%	CI 97,5%
const	0,8485	0,096	8,820	0,000	0,658	1,039
t1	1,3093	0,336	3,900	0,000	0,644	1,974
t6	1,7673	0,571	3,096	0,002	0,637	2,898
t7	–3,2870	0,652	–5,045	0,000	–4,577	–1,997
R10	0,2727	0,042	6,420	0,000	0,189	0,357
t_1^2	–1,2376	0,338	–3,663	0,000	–1,907	–0,568
t_3^2	–3,3253	0,578	–5,756	0,000	–4,469	–2,181
t_4^2	–3,4301	0,559	–6,133	0,000	–4,538	–2,322
t_5^2	–4,4963	0,744	–6,044	0,000	–5,970	–3,023
t_6^2	–2,4313	0,824	–2,952	0,004	–4,062	–0,800
t_7^2	5,8134	0,924	6,292	0,000	3,984	7,643
t_8^2	–0,2197	0,079	–2,790	0,006	–0,376	–0,064
t3*t4	6,8607	1,108	6,190	0,000	4,666	9,056
t5*t6	8,3899	1,487	5,640	0,000	5,444	11,336
t6*t7	–5,8699	1,407	–4,172	0,000	–8,657	–3,083

Параметри OLS-моделі: R^2 = 0,734; скоригований R^2 = 0,702; F = 23,06; $p < 0{,}001$; кількість спостережень — 132.

Для лісостепової зони модель включає 3 лінійні фактори, 8 квадратів та 2 добутки. Для західної зони — 5 лінійних факторів, 3 квадрати та 8 добутків послідовних факторів. Такий результат свідчить, що для степу й лісостепу більш важливу роль відіграє саме нелінійна чутливість до окремих факторів, тоді як для західної зони більш істотним є тривалий сумарний вплив кліматичних умов на окремих етапах вегетаційного періоду. Структуру нелінійних моделей врожайності для різних агрокліматичних зон наведено у табл. 4.8.

Таблиця 4.8

Кількість значущих лінійних і нелінійних факторів у регресійних моделях врожайності

Агрокліматична зона	Лінійні фактори	Квадрати факторів	Добутки послідовних факторів
Степова	4	7	3
Лісостепова	3	8	2
Західна	5	3	8

Дані табл. 4.8 показують, що в усіх агрокліматичних зонах кількість нелінійних компонентів перевищує або є співмірною з кількістю лінійних факторів. Це підтверджує важливість урахування квадратичних ефектів і мультиплікативних взаємодій між погодними чинниками при моделюванні врожайності. Порівняння ефективності лінійних і нелінійних моделей наведено на рис. 4.6.

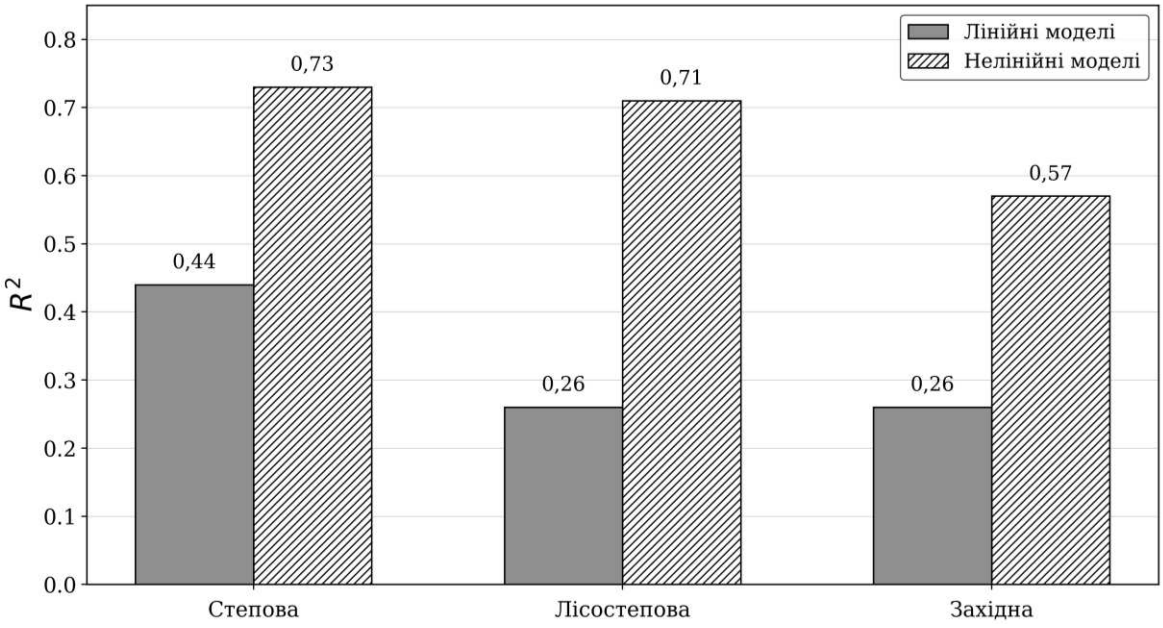

Рис. 4.6. Порівняння коефіцієнта детермінації лінійних і нелінійних регресійних моделей врожайності для різних агрокліматичних зон України

Одержані результати мають важливе значення. По-перше, вони підтверджують, що урожайність пшениці в Україні визначається не лише прямим лінійним впливом температури й опадів, а й нелінійними режимними ефектами, коли вплив фактора змінюється залежно від його рівня. По-друге, взаємодія послідовних факторів свідчить про те, що погодні умови окремих декад не можна розглядати ізольовано: важливим є їх поєднаний вплив у межах усієї вегетації. По-третє, відмінності між зонами означають, що для різних регіонів України не існує єдиної універсальної моделі врожайності; замість цього доцільно використовувати зональний підхід, обґрунтований у підрозділі 2.3. Таким чином, розроблений нами метод розширення вхідного простору ознак, дозволив урахувати нелінійні та кумулятивні ефекти кліматичного впливу в межах регресійної моделі, зберігаючи можливість інтерпретації внеску окремих факторів.

Практичне значення нелінійних моделей полягає в тому, що вони забезпечують більш точне раннє прогнозування врожайності за кілька місяців до збирання врожаю. У праці [2] прямо зазначено, що погодні умови квітня, травня та червня є критичними для формування майбутнього врожаю, а використання середньодекадних температур і місячних сум опадів є достатнім для оцінювання впливу погодно-кліматичних умов на врожайність.

Нелінійні моделі прогнозування врожайності завдяки своїй ефективності можуть бути використані не лише для прогнозування врожайності, але і для формування оптимальних кліматичних траєкторій та сценарного аналізу і можуть служити основою для створення системи прийняття рішень у аграрному бізнесі. Описана вище методологія може також слугувати основою для інструментів раннього попередження про ймовірність низької врожайності. Для сільськогосподарських виробників така інформація може допомогти прийняти рішення щодо адаптації агрономічних заходів та управління виробничими ризиками. Для страхових компаній і регіональних сільськогосподарських органів ця модель може бути корисною для попередньої оцінки ризиків, планування потреб у зернових ресурсах та розробки заходів реагування на потенційно несприятливі умови вирощування. Завдяки простоті реалізації такі моделі можуть бути

використані як великими агровиробниками, так і фермерами при плануванні інвестиційних і маркетингових рішень.

4.3.2. Оптимізація температурної траєкторії вегетації

Одним із практично важливих наслідків побудови нелінійних моделей є можливість переходу від опису впливу кліматичних факторів до задачі їх цільового оптимізаційного аналізу. Розглянемо задачу визначення оптимальної температурної траєкторії вегетації, за якої досягається найбільш сприятливий прогнозований рівень урожайності [3; 18]. Така постановка дозволяє інтерпретувати результати моделювання не лише в пояснювальному, а і в прикладному аспекті, пов’язаному з підтримкою агротехнологічних рішень.

На відміну від класичної задачі прогнозування врожайності, де шукається значення функції відгуку за відомими кліматичними факторами, при оптимізаційній постановці задача полягає у визначенні такого набору середньодекадних температур t1–t9, якому відповідає максимальне позитивне відхилення врожайності від тренду. Під оптимальною температурною траєкторією вегетації будемо розуміти сукупність середньодекадних температур квітня, травня і червня t1–t9 за яких цільова функція детрендованої врожайності ε досягає найбільшого значення. Саме таке трактування оптимальної температурної траєкторії, прийняте у праці [3], буде покладено в основу цього параграфа. Слід зазначити, що отримана температурна траєкторія інтерпретується не як керований агротехнологічний параметр, а як модельний температурний сценарій, асоційований із високими позитивними відхиленнями врожайності від тренду. Нашим завданням буде пошук температурного режиму, який можна розглядати як орієнтир максимально сприятливої вегетації. Таким чином, ми переходимо до задачі оптимізації зерновиробництва в межах реально спостережуваного температурного коридору для вегетації пшениці у степовій зоні.

Вихідною інформаційною основою нашого аналізу є сукупність детрендованих значень урожайності ε та відповідних їм температурних факторів t1–t9, сформованих для степової зони України. Повний масив даних містить 150

спостережень, отриманих для шести областей степового регіону: Херсонської, Миколаївської, Одеської, Дніпропетровської, Запорізької та Кіровоградської за період 2000–2024 роки. Як і в попередніх підрозділах, детрендована врожайність використовується для відокремлення кліматичної складової від довгострокового тренду, зумовленого розвитком технологій, інвестицій і структурними змінами в аграрному секторі.

Для побудови оптимізаційної моделі не використовувався весь масив спостережень. Спочатку дані було впорядковано за спаданням значень ε, після чого з повної вибірки виокремлено верхню частину спостережень, що відповідає режимам максимальної детрендованої врожайності. Такий підхід ґрунтується на тому, що саме верхній сегмент вибірки є найбільш інформативним для пошуку температурної траєкторії, пов’язаної з найвищими врожаями. У результаті серії обчислювальних експериментів для побудови моделі було обрано підвибірку з 35 спостережень [3]. Збільшення обсягу цієї підвибірки призводило до зниження статистичної значущості коефіцієнтів регресії, що ускладнювало використання методів диференціального числення для пошуку екстремуму цільової функції. Потрібно зауважити, що оскільки підвибірка формувалася на основі найбільших значень цільової змінної, отримана модель використовується не для загального прогнозування врожайності, а для дослідження температурних умов, характерних для режимів високої детрендованої врожайності.

До сформованої підвибірки увійшли вісім спостережень з Одеської області, по шість — з Дніпропетровської, Кіровоградської та Херсонської областей, п’ять — з Миколаївської області та чотири — із Запорізької. Підвибірка зберігає просторове представлення всіх основних областей степової зони, але при цьому концентрується на тих роках і регіонах, де було досягнуто найбільших позитивних відхилень урожайності від тренду. Це дозволяє інтерпретувати модель саме як модель оптимальних температурних режимів, а не як модель середньої поведінки системи.

Для узагальнення процедури формування вибірки основні її характеристики наведено в табл. 4.9.

Таблиця 4.9

Формування вибірки для задачі визначення оптимальної температурної траєкторії

Показник	**Значення**	**Пояснення**
Повний обсяг вибірки	150	Спостереження для шести областей степової зони
Критерій упорядкування	спадання ε	Відбір спостережень із найбільшими позитивними відхиленнями від тренду
Обсяг підвибірки для моделі	35	Верхня частина вибірки, що відповідає режимам максимальної врожайності
Мета відбору	оптимізаційне моделювання	Виділення однорідної групи спостережень для пошуку максимуму цільової функції

Для опису зв’язку між детрендованою врожайністю та температурними факторами вегетації було використано квадратичну регресійну модель. Такий вибір обумовлений двома причинами. По-перше, результати попередніх досліджень для степової зони показали, що нелінійна регресія дає істотно кращий опис залежності врожайності від кліматичних факторів порівняно з лінійною моделлю [2; 15; 17; 19; 21]. По-друге, саме квадратична форма функції є зручною для подальшої оптимізаційної постановки, оскільки дозволяє використовувати апарат аналітичного пошуку стаціонарної точки та дослідження максимуму в багатовимірному просторі факторів. Таким чином, у задачі визначення оптимальної температурної траєкторії нами використано скорочену квадратичну модель без перехресних добутків факторів, що пояснюється зручністю математичного аналізу відповідної цільової функції.

Початкову квадратичну модель було побудовано на підвибірці з 35 спостережень. Її основні характеристики є такими: коефіцієнт детермінації становив R^2 = 0,979, скоригований коефіцієнт детермінації — 0,957, а загальна статистична значущість моделі відповідала рівню $Prob(F)$ = 7,48·10^{-11}. Після усунення незначущих предикторів було отримано вдосконалену OLS-модель, для якої статистична значущість ще більше підвищилася: $Prob(F)$ = 4,80·10^{-14}, тоді як коефіцієнт детермінації залишився практично незмінним (R^2 = 0,977, R^2_{adj} = 0,962). Саме ця вдосконалена модель надалі використовується як цільова функція задачі оптимізації.

З урахуванням статистично значущих предикторів цільова функція може бути записана у вигляді

$$\varepsilon = a_2 t_2 + a_3 t_3 + a_5 t_5 + a_7 t_7 + a_8 t_8 + a_9 t_9 + b_1 t_1^2 + b_2 t_2^2 + \quad (4.5)$$

$$+ b_3 t_3^2 + b_5 t_5^2 + b_6 t_6^2 + b_7 t_7^2 + b_8 t_8^2 + b_9 t_9^2,$$

де a_i — коефіцієнти при лінійних членах, b_i — коефіцієнти при квадратичних членах. У цій функції відсутні лінійні або квадратичні члени для тих факторів, які було вилучено на етапі статистичного відбору як незначущі. Функція (4.5) використовується надалі для пошуку точки максимуму в межах допустимої області значень температур. Різні знаки коефіцієнтів при лінійних і квадратичних членах можуть свідчити про неоднорідний і нелінійний характер впливу температури на врожайність у різні декади вегетації.

Значення коефіцієнтів моделі були оцінені за методом найменших квадратів на сформованій вище вибірці. З метою перевірки стійкості побудованої моделі було використано чотирикратну перехресну перевірку з випадковим поділом на навчальні й тестові підвибірки. Середнє значення коефіцієнта детермінації за результатами такої перевірки становило 0,9850, а середні значення RMSE для навчальної та тестової підвибірок — відповідно 1,0517 і 3,4482. Отримані результати свідчать про стабільність апроксимації в межах виділеного сегмента спостережень із високими значеннями детрендованої врожайності.

Приймемо гіпотезу про те, що температура в кожній декаді вегетаційного періоду може змінюватися лише в межах фактично спостережуваного для степової зони інтервалу. Тоді задача максимізації цільової функції повинна розв’язуватися не на всьому дев’ятимірному просторі, а в межах допустимої області T, заданої обмеженнями на кожен із температурних факторів t1–t9. Така постановка дозволяє інтерпретувати знайдену траєкторію не як абстрактний математичний екстремум, а як температурний режим, що лежить у коридорі реально спостережуваних кліматичних умов степової зони України.

Область допустимих значень температур має наступний вигляд

$$T = \left\{ (t_1, \dots, t_9) \middle| t_i^{min} \le t_i \le t_i^{max}, i = 1, \dots, 9 \right\}, \quad (4.6)$$

де t_i^{min} і t_i^{max} — відповідно мінімальне та максимальне значення i-го температурного фактора у вихідній вибірці. Межі цих факторів наведено в табл. 4.10. Оптимальна температурна траєкторія не повинна виходити за межі реально спостережуваних для степової зони значень.

Таблиця 4.10

Граничні значення температурних факторів t1–t9

Показник	**t1**	**t2**	**t3**	**t4**	**t5**	**t6**	**t7**	**t8**	**t9**
t_i^{min}	4,27	8,08	8,51	10,49	11,50	13,05	14,80	17,19	17,37
t_i^{max}	12,05	14,95	16,07	19,03	19,50	26,32	22,51	24,25	26,25

Для знаходження точки максимуму спочатку визначається стаціонарна точка цільової функції. Із цією метою перші частинні похідні функції ε(t1–t9) за всіма аргументами прирівнюються до нуля:

$$\frac{\partial \varepsilon}{\partial t_i} = 0, i = 1, \dots, 9. \tag{4.7}$$

У більш детальному вигляді система (4.7) записується так:

$$\begin{gathered}
\frac{\partial(eps)}{\partial t_1} = \ 2b_1 t_1 = 0; \ \frac{\partial(eps)}{\partial t_2} = \ 2b_2 t_2 + a_2 = 0; \\
\frac{\partial(eps)}{\partial t_3} = \ 2b_3 t_3 + a_3 = 0; \ \frac{\partial(eps)}{\partial t_5} = \ 2b_5 t_5 + a_5 = 0; \\
\frac{\partial(eps)}{\partial t_6} = \ 2b_6 t_6 = 0; \frac{\partial(eps)}{\partial t_7} = \ 2b_7 t_7 + a_7 = 0; \\
\frac{\partial(eps)}{\partial t_8} = \ 2b_8 t_8 + a_8 = 0; \frac{\partial(eps)}{\partial t_9} = \ 2b_9 t_9 + a_9 = 0.
\end{gathered} \tag{4.8}$$

Розв’язання цієї системи дає координати стаціонарної точки, які надалі зіставляються з допустимим температурним коридором. Варто окремо зазначити, що цільова функція не містить змінної t4, тому для неї стаціонарна умова не формується, а значення визначається додатково на етапі пошуку максимуму в межах інтервалу допустимих значень.

Наявність стаціонарної точки сама по собі ще не означає, що вона є точкою максимуму в межах допустимої області. Тому в роботі використано спеціальний алгоритм, який дозволяє для кожного температурного фактора вибрати координату

максимуму з урахуванням знака квадратичного коефіцієнта, положення стаціонарної точки відносно допустимого інтервалу та статистичної визначеності коефіцієнтів моделі. Якщо квадратичний коефіцієнт для відповідного фактора є від'ємним, максимум досягається в стаціонарній точці за умови, що вона належить допустимому інтервалу. Якщо стаціонарна точка лежить за межами інтервалу, максимум обирається на найближчій до неї межі. Якщо квадратичний коефіцієнт є додатним, максимум досягається на тій межі допустимого інтервалу, що розташована далі від стаціонарної точки. Для слабо визначених коефіцієнтів із недостатньою статистичною значущістю доцільно використовувати середину допустимого інтервалу.

Логіка, яка була застосована для переходу від стаціонарної точки до точки максимуму цільової функції eps, може бути описана наступним алгоритмом:

1. Якщо $b_i < 0$ (парабола вітками донизу):

   • Якщо $t_{imin} \le t_{istac} \le t_{imax}$, то максимум функції eps збігається із стаціонарною точкою, визначеною за умовою (4.8);

   • Якщо стаціонарна точка лежить за межами інтервалу $[t_{imin}, t_{imax}]$, точка максимуму співпадає із тією границею інтервалу, яка є ближчою до стаціонарної точки:

$$t_{iopt} = \begin{cases} t_{imax},\ t_{istac} > t_{imax}; \\ t_{imin},\ t_{istac} \le t_{imin}. \end{cases} \quad (4.9)$$

2. Якщо $b_i > 0$ (парабола вітками доверху):

   • Якщо $t_{imin} \le t_{istac} \le t_{imax}$, то максимум функції eps співпадає із тією границею інтервалу, яка є більш віддаленою від стаціонарної точки:

$$t_{iopt} = \begin{cases} t_{imin},\ t_{istac} - t_{imin} > t_{imax} - t_{istac}; \\ t_{imax},\ t_{istac} - t_{imin} \le t_{imax} - t_{istac}. \end{cases} \quad (4.10)$$

   • Якщо стаціонарна точка лежить за межами інтервалу $[t_{imin}, t_{imax}]$, точка максимуму eps співпадає із тією границею інтервалу, яка більш віддалена від стаціонарної точки.

3. Якщо $b_i = 0$ (відсутній квадратичний член):

   • Оптимальна точка визначається знаком коефіцієнта a_i:

$$t_{iopt} = \begin{cases} t_{imin},\ a_i \leq 0; \\ t_{imax},\ a_i > 0. \end{cases} \quad (4.11)$$

4. Якщо маємо випадок погано означених коефіцієнтів регресії (*p-value* $\gg 0{,}05$), визначаємо точку максимуму з рівності:

$$t_{iopt} = \frac{t_{imin} + t_{imax}}{2}. \quad (4.12)$$

У результаті застосування наведеного алгоритму було отримано координати максимуму цільової функції, які наведено в табл. 4.11. Оптимальну температурну траєкторію графічно наведено на рис. 4.7.

Таблиця 4.11

Оптимальні значення температурних факторів t_{opt}

Показник	**t1**	**t2**	**t3**	**t4**	**t5**	**t6**	**t7**	**t8**	**t9**
t_{opt}	12,05	11,75	16,07	14,76	19,50	13,05	22,51	17,19	24,37

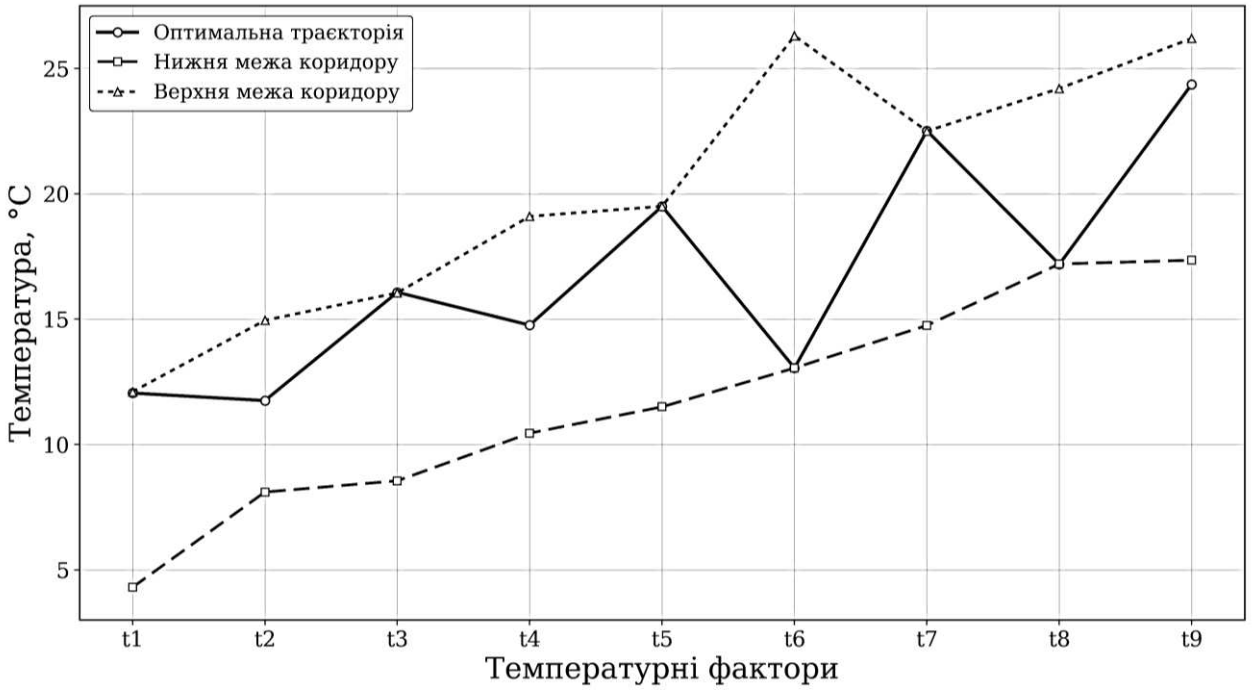

Рис. 4.7. Оптимальна температурна траєкторія вегетації пшениці для степової зони України

Квадратична регресійна модель має важливу перевагу: вона забезпечує явну інтерпретацію впливу кожного фактора на цільову функцію та дозволяє застосувати строгий математичний апарат для пошуку максимуму. Водночас її головне обмеження полягає в тому, що коефіцієнти оцінюються лише за підвибіркою з 35

спостережень. Якщо оцінки коефіцієнтів є недостатньо точними, знайдена траєкторія може бути далекою від глобального оптимуму. У праці [3] цей ефект прямо ілюструється тим, що отримана квадратичною регресією траєкторія добре узгоджується з очікуваним оптимумом у перших чотирьох точках, але відхиляється від нижньої межі коридору в точках 5, 7 і 9.

З метою уточнення оптимальної температурної траєкторії в роботі додатково застосовано методи штучного інтелекту — Random Forest (RF) і Support Vector Machine (SVM) — у поєднанні з алгоритмом Differential Evolution (DE). У такій гібридній схемі модель машинного навчання використовується як предиктор значення цільової функції ε, а Differential Evolution виконує роль метаевристичного оптимізатора, який шукає в просторі допустимих температурних значень такий вектор параметрів, що максимізує прогнозовану врожайність. Такий підхід добре підходить до задач оптимізації складних багатовимірних функцій, особливо за наявності кількох локальних екстремумів [64; 165; 166].

Основні кроки гібридної процедури є такими: спочатку на наявних даних навчається модель RF або SVM, далі задаються допустимі межі температурних факторів, після чого алгоритм DE генерує популяцію векторів параметрів, яка еволюціонує через операції відбору, мутації та кросоверу. Кожен кандидатний вектор оцінюється за значенням цільової функції, а найкращі розв’язки, що забезпечують максимум ε, зберігаються. Перевагою описаної методики є те, що відбір зразків для навчання здійснювався на всьому масиві даних, а не лише на 35 найкращих зразках.

Порівняння температурних траєкторій, отриманих квадратичною регресією, та інтелектуальними методами SVM і RF, зображено на рис. 4.8. На підставі виконаного порівняльного аналізу можна стверджувати, що оптимальні температурні траєкторії, побудовані методами SVM і RF, є досить близькими між собою. Це дає підстави розглядати гібридні AI-підходи не як альтернативу аналітичному методу, а як засіб уточнення оптимальної траєкторії.

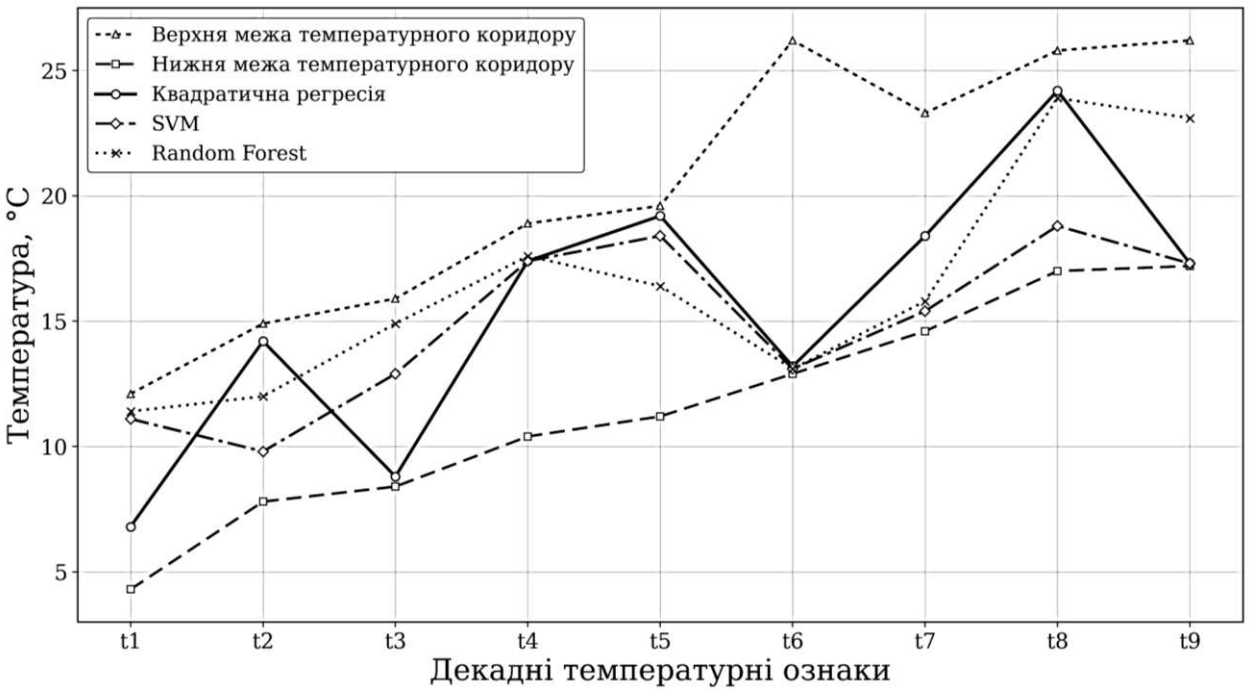

Рис. 4.8. Порівняння оптимальних температурних траєкторій, отриманих квадратичною регресією, SVM та Random Forest

Отже, побудова оптимальної температурної траєкторії вегетації є прикладом застосування нелінійної регресійної моделі. Одержані результати конкретизують роль температурних факторів у різні фази розвитку культури та можуть бути використані як аналітична основа для оцінювання адаптивних сценаріїв зерновиробництва. Оптимальна температурна траєкторія та нелінійна модель врожайності можуть бути використані для сценарного аналізу, який включений до розробленої у наступному розділі інформаційно-аналітичної системи «Пшениця України».

Висновки до розділу 4

У розділі 4 досліджено екзогенний підхід до прогнозування врожайності, за якого до вхідного простору моделей включаються зовнішні фактори, що безпосередньо характеризують умови формування врожаю. На відміну від ендогенного підходу, який спирається лише на внутрішню динаміку ряду врожайності або його трендових залишків, екзогенна постановка дає змогу врахувати агрокліматичні та, в окремих експериментах, економічні чинники. Це

дозволило підвищити інформативність вхідних даних і покращити якість класифікаційного прогнозування років низької врожайності.

Для екзогенних класифікаційних моделей машинного навчання використано об’єднані вибірки агрокліматичних даних, сформовані окремо для степової, лісостепової та західної агрокліматичних зон. Кожна вибірка містила 132 спостереження типу «область–рік» за період 2000–2021 рр. До вхідного простору моделей було включено 12 кліматичних ознак: дев’ять декадних температурних показників квітня–червня та три місячні суми опадів. Цільова змінна формувалася на основі трендового залишку врожайності, що дозволило перейти до задачі бінарної класифікації з позитивним класом низької врожайності.

Для вибору гіперпараметрів моделей Logistic Regression, Random Forest і SVM RBF застосовано решітковий пошук із поділом даних на навчальну, валідаційну та тестову вибірки. Найкращі конфігурації відбиралися за значенням метрики F1 на валідаційній вибірці, після чого три найкращі конфігурації кожної моделі оцінювалися на незалежній тестовій вибірці. Такий підхід дозволив оцінити не лише максимальну якість окремої конфігурації, а й стійкість результатів для групи близьких за якістю моделей.

Результати комп’ютерних експериментів показали, що якість екзогенних моделей машинного навчання істотно залежить від агрокліматичної зони. Для степової зони найкращий результат отримано для моделі SVM RBF, яка на тестовій вибірці досягла значення $F1 = 0{,}8421$. Для західної зони найефективнішою також виявилася SVM RBF із тестовим значенням $F1 = 0{,}7619$. Для лісостепової зони найкращий результат показала Logistic Regression із $F1 = 0{,}6433$. Отже, універсальної моделі, яка була б найкращою для всіх агрокліматичних зон, не виявлено.

Узагальнення результатів для трьох агрокліматичних зон показало, що за середнім тестовим значенням метрики $F1$ найкращою серед моделей машинного навчання є SVM RBF. Її середнє значення для трьох зон становить 0,7216. Logistic Regression має середнє тестове значення $F1 = 0{,}6867$, а Random Forest — 0,6718. Водночас Logistic Regression продемонструвала найкращий результат у

лісостеповій зоні, що підтверджує необхідність зонального підходу до прогнозного моделювання врожайності.

Дослідження екзогенних моделей глибинного навчання виконано на прикладі Херсонської області. До вхідного простору моделей LSTM і GRU було включено агрокліматичні показники та економічний фактор $econ$, який відображає зміну типу економічних відносин у різні історичні періоди. У цій постановці основним гіперпараметром була кількість нейронів у прихованому шарі $units$.

Результати екзогенних RNN-моделей показали, що GRU має невелику перевагу над LSTM за основною тестовою метрикою. Для LSTM середнє тестове значення $F1$ становить 0,7009, а для GRU — 0,7211. Обидві моделі продемонстрували прийнятний рівень якості, однак не забезпечили однозначної переваги над класичними моделями машинного навчання. Зокрема, SVM RBF у степовій зоні мала вище тестове значення $F1$ = 0,8421, а Logistic Regression — 0,7500.

Порівняння ендогенного та екзогенного підходів засвідчило суттєву перевагу екзогенної постановки. Значення $F1$ зросло для всіх розглянутих моделей: для Logistic Regression — з 0,3636 до 0,7500, для Random Forest — з 0,5455 до 0,7059, для SVM RBF — з 0,6093 до 0,8421, для LSTM — з 0,6667 до 0,7009, для GRU — з 0,3757 до 0,7211. Найбільше покращення отримано для Logistic Regression та GRU. Це підтверджує, що підвищення якості прогнозування досягається не лише за рахунок ускладнення архітектури моделі, а насамперед завдяки змістовному розширенню простору вхідних ознак.

У розділі також досліджено нелінійний підхід до моделювання врожайності, який передбачає розширення базового простору кліматичних факторів квадратами змінних і добутками послідовних у часі ознак. Такий підхід дозволяє врахувати неадитивний і кумулятивний характер впливу погодних умов на формування врожаю. Встановлено, що структура нелінійних моделей суттєво відрізняється між агрокліматичними зонами. Для степової зони важливу роль відіграють квадратичні ефекти окремих факторів, тоді як для західної зони більшу роль мають добутки

послідовних факторів, що відображають тривалий сумарний вплив погодних умов упродовж вегетаційного періоду.

Побудова нелінійних регресійних моделей підтвердила, що врожайність не може бути достатньо повно описана лише лінійною адитивною дією температури та опадів. Важливими є також режимні ефекти, коли вплив фактора змінюється залежно від його рівня, а також взаємодія погодних умов суміжних часових інтервалів. Це обґрунтовує доцільність використання ознак другого порядку для підвищення точності та інтерпретованості моделей прогнозування врожайності.

На основі нелінійної квадратичної моделі для степової зони розглянуто задачу оптимізації температурної траєкторії вегетації. Побудована квадратична модель використана як цільова функція для пошуку температурного режиму, асоційованого з максимально сприятливими умовами формування врожаю.

З метою уточнення оптимальної температурної траєкторії додатково застосовано гібридний підхід, що поєднує моделі Random Forest і SVM з алгоритмом Differential Evolution. Моделі машинного навчання виконували роль предикторів цільової функції, а алгоритм Differential Evolution використовувався для пошуку максимуму в межах допустимого температурного коридору. Отримані за допомогою Random Forest і SVM температурні траєкторії виявилися близькими між собою, що підтверджує стійкість результатів оптимізаційного аналізу.

Узагальнюючи, можна підтвердити, що екзогенний підхід є більш ефективним для прогнозування врожайності порівняно з ендогенним підходом. Використання агрокліматичних та економічних факторів дозволяє підвищити якість класифікаційного прогнозування, а врахування нелінійних ефектів і взаємодій між погодними чинниками поглиблює змістовну інтерпретацію моделей. Отримані результати створюють методичну основу для побудови інформаційно-аналітичної системи підтримки прийняття рішень у зерновиробництві, зокрема для оцінювання ризику низької врожайності, сценарного аналізу та визначення сприятливих температурних режимів вегетації.

РОЗДІЛ 5. ІНФОРМАЦІЙНО-АНАЛІТИЧНА СИСТЕМА «ПШЕНИЦЯ УКРАЇНИ» ЯК ЗАСІБ ПРОГРАМНОЇ РЕАЛІЗАЦІЇ, АПРОБАЦІЇ ТА ВІДТВОРЮВАНОГО ЗАСТОСУВАННЯ ІНТЕЛЕКТУАЛЬНИХ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ВРОЖАЙНОСТІ

5.1. Призначення, позиціонування та вимоги до інформаційно-аналітичної системи

Результати, отримані в попередніх розділах дисертації, охоплюють формування агрокліматичного ознакового простору, побудову ендогенних та екзогенних моделей прогнозування врожайності, класифікацію рівня врожайності, оцінювання ризику низької врожайності, сценарно-оптимізаційний аналіз кліматичних умов і просторову інтерпретацію результатів. Подання цих результатів лише у вигляді окремих таблиць, графіків, моделей або програмних фрагментів не забезпечує їх повторного використання в межах єдиного дослідницького процесу. Тому завершальним етапом дисертаційного дослідження є програмна реалізація запропонованих моделей і методів в інформаційно-аналітичній системі (ІАС) «Пшениця України» [12; 13; 23].

Інформаційно-аналітична система «Пшениця України» розроблена як предметно орієнтоване вебсередовище для програмної реалізації, апробації та відтворюваного застосування інтелектуальних моделей прогнозування врожайності за агрокліматичними факторами. Її призначення полягає в забезпеченні повного циклу роботи з даними та моделями: від формування агрокліматичної вибірки й попередньої аналітичної перевірки даних до побудови прогнозних оцінок, класифікації ризикових років, оцінювання ризику, сценарного аналізу, просторової візуалізації та формування звітних матеріалів.

На відміну від окремих розрахункових модулів або ізольованих програмних реалізацій моделей, розроблена система орієнтована на збереження не тільки кінцевих результатів, а й конфігурацій дослідницького процесу. У системі можуть фіксуватися параметри сформованої вибірки, налаштування графічного подання, параметри прогнозного експерименту, сценарні припущення, результати

порівняння альтернатив, просторові відображення та звітні матеріали. Такий підхід забезпечує повторне відкриття, перевірку, коригування та документування отриманих результатів.

Окремі алгоритмічно-програмні напрацювання, пов’язані з моделюванням впливу кліматичних факторів на врожайність методами машинного навчання, були реалізовані в комп’ютерній програмі [23]. У межах дисертаційного дослідження ці напрацювання розвинуто та інтегровано в інформаційно-аналітичну систему «Пшениця України», яка забезпечує не лише виконання окремих розрахункових процедур, а й повний відтворюваний цикл роботи з агрокліматичними даними, прогнозними моделями, ризиковими оцінками, сценаріями, картографічним поданням і звітністю.

Основними користувачами системи є дослідники, аналітики, фахівці аграрного профілю та користувачі, які працюють із задачами прогнозування врожайності, оцінювання агрокліматичних ризиків і підготовки аналітичних матеріалів. Для дослідника система забезпечує відтворюване середовище перевірки моделей і сценаріїв; для аналітика — засоби порівняння результатів, виявлення ризикових станів і формування звітів; для прикладного користувача — інструменти візуального подання прогнозних та ризикових оцінок у регіональному розрізі.

Функціональна логіка системи ґрунтується на поєднанні кількох взаємопов’язаних контурів. Перший контур забезпечує формування вибірок і підготовку агрокліматичних даних. Другий контур відповідає за візуальну та просторову аналітику. Третій контур реалізує прогнозування врожайності, класифікацію ризикових років і оцінювання ризику. Четвертий контур забезпечує сценарний аналіз, порівняння альтернатив, збереження сценаріїв, формування колекцій і підготовку звітних матеріалів. У сукупності ці контури утворюють дослідницький процес «дані→ознаки→моделі→ризики→сценарії→звітність».

Функціональні вимоги до системи охоплюють формування вибірок за просторовими, часовими та атрибутивними критеріями; попередній перегляд і перевірку даних; побудову графіків і карт; запуск прогнозних процедур; класифікацію ризикових років; оцінювання ризику на основі трендових залишків;

сценарне порівняння результатів; збереження конфігурацій дослідження; формування звітів і експорт результатів. Нефункціональні вимоги включають відтворюваність результатів, узгодженість між модулями, розширюваність архітектури, структурованість інтерфейсу, керованість доступом і придатність результатів до документування та повторного використання.

Зв'язок інформаційно-аналітичної системи «Пшениця України» з результатами попередніх розділів дисертації подано на рис. 5.1. Схема відображає, що система узагальнює результати формування агрокліматичного ознакового простору, побудови ендогенних та екзогенних моделей прогнозування, класифікації ризикових років, оцінювання ризику та сценарно-оптимізаційного аналізу, забезпечуючи їх програмну реалізацію в єдиному дослідницькому середовищі.

Рис. 5.1. Позиціонування ІАС «Пшениця України» у структурі дисертаційного дослідження

Як видно з рис. 5.1, ІАС «Пшениця України» виконує інтегрувальну функцію щодо результатів дисертаційного дослідження. Результати формування ознакового

простору, побудови прогнозних моделей, оцінювання ризику та сценарного аналізу реалізуються в системі через модулі підготовки даних, прогнозування, ризик-аналізу, просторової візуалізації, сценаріїв і звітності. Це забезпечує перехід від окремих моделей і розрахункових процедур до відтворюваного програмного середовища їх застосування.

Таким чином, ІАС «Пшениця України» у межах дисертаційного дослідження розглядається не як допоміжний програмний засіб, а як інтегроване інформаційно-аналітичне середовище, у якому реалізовано запропоновані інтелектуальні моделі прогнозування врожайності, засоби оцінювання ризиків, сценарного аналізу та відтворюваного представлення результатів.

5.2. Архітектура системи, структура даних та інформаційні потоки

Архітектура інформаційно-аналітичної системи «Пшениця України» сформована відповідно до логіки повного циклу роботи з агрокліматичними даними та моделями прогнозування врожайності. Система має забезпечувати узгоджене виконання кількох взаємопов’язаних процедур: формування вибірки, попередню аналітичну перевірку даних, побудову графічних і картографічних представлень, запуск прогнозних моделей, оцінювання ризику, сценарне порівняння та формування звітних матеріалів. Тому її архітектура побудована як модульна веборієнтована структура, у якій функціональні блоки пов’язані спільною базою даних, збереженими конфігураціями дослідження та сценарним контуром [23; 25; 26].

Вибір такої архітектурної організації зумовлений потребою поєднати в одному середовищі роботу з предметними агрокліматичними даними, запуск моделей машинного навчання, просторове подання результатів, сценарне порівняння та документування отриманих висновків. За таких умов ізольоване виконання окремих розрахункових процедур було б недостатнім, оскільки не забезпечувало б узгодженого переходу від сформованої вибірки до прогнозних, ризикових і звітних результатів.

У структурі системи виділено кілька взаємопов’язаних рівнів. Рівень збереження даних містить предметні агрокліматичні та врожайнісні дані, а також службові об’єкти, необхідні для фіксації конфігурацій дослідження, сценаріїв, колекцій і журналу подій. Прикладний серверний рівень забезпечує маршрутизацію запитів, обробку форм, керування правами доступу та взаємодію між модулями. Аналітико-модельний рівень відповідає за підготовку вибірок, розвідувальний аналіз даних, прогнозування, класифікацію, оцінювання ризику та сценарний аналіз. Рівень користувацького інтерфейсу забезпечує доступ до функціональних модулів, а контур документування результатів формує графічні, картографічні, прогнозні, ризикові та звітні матеріали.

Ключовою архітектурною особливістю системи є використання збережених конфігурацій дослідження. Центральним об’єктом такого підходу є конфігурація вибірки даних, яка фіксує параметри сформованої вибірки: просторовий рівень, перелік областей або зон, часовий інтервал і набір агрокліматичних та врожайнісних показників. Після збереження така конфігурація може повторно використовуватися в модулях графічної аналітики, прогнозування, ризик-аналізу, сценарного порівняння й звітності. Це усуває розрив між етапом підготовки даних і подальшими аналітичними процедурами, оскільки всі модулі працюють із формалізованою конфігурацією вхідних даних.

Аналогічну роль у системі виконують конфігурація графічного подання та конфігурація прогнозного експерименту. Конфігурація графічного подання зберігає параметри побудованої візуалізації, що дає змогу повторно відкрити графік із тими самими налаштуваннями. Конфігурація прогнозного експерименту фіксує вибрану базову вибірку, модель, цільовий показник, набір предикторів, сценарні зміни та параметри оцінювання. У сукупності ці конфігурації формують основу повторного використання результатів дослідження, оскільки пов’язують кінцевий результат з умовами його отримання.

Узагальнену архітектуру інформаційно-аналітичної системи подано на рис. 5.2. На схемі відображено логічний взаємозв’язок основних рівнів ІАС:

прикладного серверного рівня, рівня збереження даних, аналітико-модельного контуру, рівня користувацького інтерфейсу та контуру документування результатів.

Рис. 5.2. Архітектура інформаційно-аналітичної системи «Пшениця України»

Рис. 5.2 відображає логічну організацію ІАС «Пшениця України», тобто взаємозв’язок основних рівнів системи та послідовність переходу від збереження й опрацювання даних до їх аналітичного подання, інтерпретації та документування. Рівень збереження даних і прикладний серверний рівень належать до серверної частини системи, тоді як аналітико-модельний контур, користувацький інтерфейс і контур документування результатів забезпечують прикладне опрацювання та представлення результатів дослідження.

У межах такої логічної архітектури інформаційні потоки системи мають послідовно-циклічний характер. Початковим етапом є імпорт або вибір наявних агрокліматичних і врожайнісних даних. Далі користувач формує вибірку за просторовими, часовими й атрибутивними критеріями, після чого система створює конфігурацію вибірки даних. На її основі можуть виконуватися попередній

перегляд таблиці, розвідувальний аналіз, побудова графіків, просторове відображення показників, регресійне або класифікаційне прогнозування, оцінювання ризику та сценарне порівняння результатів. Отримані результати можуть бути збережені у складі сценарію, об’єднані в колекції, наведені на карті або включені до звітних матеріалів. Таким чином, інформаційний потік не завершується одноразовим прогнозом, а переходить у відтворюваний сценарний контур.

Структура даних системи підпорядкована логіці повторного використання результатів дослідження. У ній виділено три взаємопов’язані групи сутностей: предметні агрокліматичні дані, конфігурації дослідницького процесу та об’єкти організації сценаріїв. До першої групи належать сутності, що описують агрокліматичні зони, області, часові ряди врожайності та кліматичні показники. Вони формують інформаційну основу для побудови ознакового простору, аналізу трендових залишків, прогнозування та оцінювання ризику.

До другої групи належать сутності, які забезпечують фіксацію конфігурацій дослідження. У програмній моделі системи для цього використовується службова сутність збереженої конфігурації дослідження, що може містити параметри вибірки даних, графічного подання або прогнозного експерименту. Така структура дає змогу застосовувати один механізм збереження для різних типів дослідницьких конфігурацій: вибірок, графічних представлень і прогнозних експериментів. Крім того, вона створює умови для повторного запуску моделей, порівняння результатів і документування умов отримання прогнозних та ризикових оцінок.

До третьої групи належать об’єкти сценарного контуру: дослідницький сценарій, папка сценаріїв, колекція сценаріїв і журнал подій сценарію. Дослідницький сценарій фіксує логіку окремого аналітичного експерименту, у якому можуть бути пов’язані вибірка, графічне представлення, прогнозна конфігурація, сценарні параметри, статус і версія. Папка сценаріїв забезпечує групування сценаріїв за темами або етапами дослідження. Колекція сценаріїв використовується для об’єднання кількох сценаріїв у логічно завершену добірку, придатну для порівняння або звітності. Журнал подій сценарію фіксує історію змін,

запусків і дій користувача, що підсилює відтворюваність і контрольованість дослідницького процесу.

Логічну модель даних та об’єктів відтворюваного дослідницького процесу подано на рис. 5.3. Схема відображає не лише предметні таблиці, а й зв’язок між даними, збереженими конфігураціями, сценаріями та результатами. У центральній частині схеми розміщено збережену конфігурацію дослідження та дослідницький сценарій, оскільки саме ці об’єкти забезпечують перехід від збереження даних до організації відтворюваного дослідницького процесу. Предметні сутності агрокліматичної зони, області, даних урожайності, місячних і декадних кліматичних показників виступають джерелом формування конфігурації вибірки даних, а папка сценаріїв, колекція сценаріїв і журнал подій сценарію забезпечують організацію, порівняння та документування результатів.

Рис. 5.3. Логічна модель даних та об’єктів відтворюваного дослідницького процесу в ІАС «Пшениця України»

Програмна реалізація системи виконана з використанням веборієнтованого підходу. Серверна частина реалізована на основі Django, що забезпечує опис моделей даних, маршрутизацію, роботу з формами, шаблонами, автентифікацією та

правами доступу. Для реалізації прикладного програмного інтерфейсу використано Django REST Framework, а для документування API — засоби автоматизованого опису сервісних ендпоінтів. Збереження предметних даних, конфігурацій дослідження, сценаріїв, колекцій і службових об’єктів здійснюється в реляційній базі даних MySQL. Інтерактивність користувацького інтерфейсу підтримується засобами шаблонів Django та HTMX, що дає змогу оновлювати окремі частини сторінок без повного перезавантаження інтерфейсу.

Для побудови графічних представлень використано бібліотеку Plotly, що забезпечує інтерактивне відображення часових рядів, порівнянь і аналітичних графіків. Просторова візуалізація реалізована із застосуванням картографічних засобів бібліотеки Leaflet, що дає змогу подавати врожайність, прогнозні оцінки, ризики та сценарні результати в регіональному розрізі. Таблична обробка, підготовка вибірок, експорт і службові перетворення даних виконуються засобами Python-бібліотек для роботи з табличними структурами. Формування звітів і експорт результатів забезпечують перехід від інтерактивного аналізу до документованого представлення результатів.

Використання зазначених програмних засобів відповідає модульній структурі ІАС і характеру опрацьовуваних даних. Веборієнтована реалізація забезпечує централізоване збереження даних і доступ до результатів через єдиний інтерфейс, реляційна база даних узгоджується зі структурованим поданням областей, років, кліматичних показників, сценаріїв і конфігурацій дослідження, а засоби інтерактивної графічної та картографічної візуалізації дають змогу подавати результати прогнозування й ризик-аналізу у формі, придатній для інтерпретації та звітності.

Важливою складовою програмної реалізації є рольова організація доступу. У системі передбачено розмежування користувачів за ролями адміністратора, редактора та переглядача. Це дає змогу відокремити функції адміністрування, редагування даних і конфігурацій, а також перегляду результатів. Такий підхід є важливим для системи, що використовується не лише як демонстраційний

застосунок, а як дослідницьке середовище з можливістю збереження, повторного відкриття та документування аналітичних результатів.

Отже, архітектура ІАС «Пшениця України» поєднує модульну вебструктуру, предметно орієнтовану базу агрокліматичних даних, механізми збереження дослідницьких конфігурацій, сценарний контур і засоби звітності. Така побудова забезпечує зв’язок між даними, моделями, прогнозними результатами, ризиковими оцінками, сценаріями та документованими матеріалами, що відповідає призначенню системи як засобу програмної реалізації, апробації та відтворюваного застосування інтелектуальних моделей прогнозування врожайності.

5.3. Реалізація модулів підготовки даних, аналітики, прогнозування та ризик-аналізу

Функціональні модулі інформаційно-аналітичної системи «Пшениця України» реалізують прикладний контур використання інтелектуальних моделей прогнозування врожайності. Їх призначення полягає не лише в окремому відображенні даних або запуску розрахункових процедур, а в забезпеченні послідовного переходу від підготовки агрокліматичної вибірки до побудови прогнозних оцінок, класифікації ризикових років, оцінювання ризику, просторового подання результатів і подальшого включення отриманих матеріалів до сценаріїв та звітів.

Початковим компонентом цього контуру є модуль формування вибірок і підготовки даних. Він забезпечує відбір даних за просторовими, часовими та атрибутивними критеріями: агрокліматичною зоною, областями, періодом спостереження та набором показників. У межах цього модуля користувач формує базовий набір даних, який може містити показники врожайності, трендові залишки, температурні предиктори, показники опадів і похідні агрокліматичні ознаки [79]. Такий підхід відповідає логіці формування ознакового простору, розглянутій у попередніх розділах дисертації, де основною одиницею аналізу виступає узгоджене спостереження типу «регіон–рік».

Важливою функцією модуля підготовки даних є збереження сформованої вибірки у вигляді конфігурації вибірки даних. Це дає змогу перетворити вибірку з одноразового результату фільтрації на відтворювану конфігурацію дослідження. Збережена конфігурація вибірки може використовуватися в модулях графічної аналітики, прогнозування, ризик-аналізу, сценарного порівняння та звітності. У такий спосіб система забезпечує єдність вхідних даних для різних етапів аналізу й унеможливлює ситуацію, коли графіки, прогнози та сценарії будуються на різних незафіксованих наборах даних.

У модулі формування вибірки передбачено попередній перегляд сформованих даних та їх первинну аналітичну перевірку. Користувач може оцінити структуру таблиці, перелік включених показників, часові межі, кількість спостережень і наявність необхідних предикторів. Додатково система підтримує формування звіту розвідувального аналізу даних, що дає змогу перевірити повноту даних, виявити пропуски, оцінити розподіли показників і підготувати вибірку до використання в прогнозних процедурах. Така перевірка є важливою, оскільки якість моделей машинного навчання значною мірою залежить від узгодженості, повноти та інформативності вхідного ознакового простору.

Після формування вибірки користувач може перейти до модуля візуальної аналітики. Цей модуль призначений для побудови графічних представлень часових рядів, порівняння регіонів, аналізу динаміки врожайності, температурних показників, опадів і трендових залишків. Його завдання полягає в перетворенні табличних даних на наочні аналітичні представлення, що дають змогу виявляти тенденції, міжрічні коливання, просторові відмінності та потенційні аномальні значення. Графічні представлення можуть використовуватися як самостійні результати розвідувального аналізу, а також як допоміжні матеріали для інтерпретації прогнозних і ризикових оцінок.

Конфігурації побудованих графіків можуть зберігатися як конфігурації графічного подання. Це забезпечує повторне відкриття візуалізації з тими самими параметрами: вибіркою, показниками, типом графіка, режимом порівняння та стилем відображення. Завдяки цьому графік у системі є не лише зображенням, а

відтворюваним аналітичним об’єктом, пов’язаним із конкретною вибіркою даних. Такий підхід важливий для дисертаційного дослідження, оскільки дає змогу зберігати не тільки результати, а й умови їх отримання.

Окремий модуль системи забезпечує просторову візуалізацію результатів. Картографічне подання використовується для відображення врожайності, трендових залишків, прогнозних оцінок, ризикових показників і сценарних змін у розрізі областей України. На відміну від звичайного графічного представлення, карта дає змогу інтерпретувати результати з урахуванням регіональної неоднорідності та агрокліматичного зонування. Це особливо важливо для задач прогнозування врожайності, оскільки одна й та сама модельна оцінка може мати різне практичне значення для різних зон і областей.

Прогнозний модуль системи реалізує використання сформованих агрокліматичних вибірок для побудови прогнозних оцінок урожайності. Його робота ґрунтується на конфігурації вибірки даних, яка визначає склад вхідних даних, перелік предикторів, часовий інтервал і територіальне охоплення. У межах прогнозного контуру можуть використовуватися регресійні, класифікаційні та нейромережеві постановки, що відповідає логіці дисертаційного дослідження. Регресійний підхід орієнтований на отримання числової прогнозної оцінки або відхилення від очікуваного рівня, тоді як класифікаційний підхід дає змогу визначати належність року або регіону до категорій «низька» / «висока» врожайність.

Класифікаційний контур має особливе значення, оскільки в задачах аграрної прогнозної аналітики точне числове значення не завжди є єдиним або головним результатом. Для практичного використання часто важливо своєчасно виявити ризиковий стан, тобто визначити, чи належить спостереження до класу низької врожайності. У системі така постановка реалізується через вибір моделі класифікації, цільового показника, порогового критерію та метрик оцінювання. Це дає змогу порівнювати моделі за здатністю виявляти несприятливі роки, а не лише за середньою точністю прогнозу.

Модуль прогнозування пов’язаний із механізмом конфігурації прогнозного експерименту, який фіксує параметри моделювання. До таких параметрів належать вибрана базова вибірка, тип моделі, набір предикторів, цільова змінна, сценарні зміни вхідних показників і параметри оцінювання результату [23; 180]. Збереження конфігурації прогнозного експерименту забезпечує повторний запуск або повторну інтерпретацію прогнозного експерименту без втрати інформації про умови його проведення. У подальшому така конфігурація може бути пов’язана з дослідницьким сценарієм, що переводить окремий прогноз у структуру відтворюваного сценарію.

Оцінювання ризику низької врожайності реалізовано як окремий функціональний блок, пов’язаний із трендовими залишками врожайності та нижньохвостовими характеристиками їх розподілу. Використання трендових залишків дає змогу відокремити довгострокову тенденцію зміни врожайності від несприятливої міжрічної мінливості. У межах ризикового аналізу система може використовувати квантильні пороги, від’ємні залишки, показники відхилення від очікуваного рівня та просторову інтерпретацію ризикових областей. Такий підхід узгоджується з дисертаційною логікою оцінювання ризику через нижній хвіст розподілу трендових залишків.

Ризик-аналітичний модуль доповнює прогнозування, оскільки дає змогу оцінити не тільки очікуваний рівень урожайності, а й небезпеку несприятливих відхилень. Для користувача це означає можливість виявити регіони або роки з підвищеним ризиком низької врожайності, порівняти їх між собою, подати результати на карті та включити ризикові оцінки до сценарного або звітного контуру. У такому вигляді ризик-аналіз доповнює прогнозний контур системи та формує окремий аналітичний результат, пов’язаний з інтерпретацією несприятливих відхилень урожайності.

Сценарний модуль забезпечує зміну вхідних агрокліматичних параметрів і порівняння альтернативних конфігурацій. Його використання дає змогу досліджувати, як зміна температурних показників або опадів у визначених часових інтервалах може впливати на прогнозну оцінку, клас урожайності або ризиковий стан. Такий підхід відповідає сценарно-оптимізаційній постановці, у межах якої

модель використовується не тільки для передбачення, а й для аналізу можливих умов формування сприятливого або несприятливого результату. Приклади інтерфейсної реалізації модулів формування вибірки, візуальної аналітики, картографічного подання, прогнозування та ризик-аналізу наведено в додатку Г.

Послідовність взаємодії функціональних модулів ІАС «Пшениця України» подано на рис. 5.4. Схема відображає перехід від формування агрокліматичної вибірки та збереження конфігурації вибірки даних до візуальної аналітики, прогнозування, ризик-аналізу, сценарного порівняння, формування дослідницького сценарію, об'єднання результатів у колекцію сценаріїв та підготовки звітних матеріалів. Така організація забезпечує узгоджене використання даних, моделей і результатів у межах єдиного відтворюваного дослідницького контуру.

Рис. 5.4. Функціональний контур модулів ІАС «Пшениця України»

Як видно з рис. 5.4, модулі системи не функціонують ізольовано, а утворюють послідовний аналітичний контур. Сформована вибірка є основою для графічного, картографічного, прогнозного та ризик-аналітичного опрацювання, а отримані результати можуть бути збережені у вигляді сценаріїв, об'єднані в колекції та використані для формування звітів.

Таким чином, реалізовані модулі ІАС «Пшениця України» забезпечують повний прикладний цикл роботи з інтелектуальними моделями прогнозування врожайності. Модуль формування вибірки створює інформаційну основу аналізу, модулі графіків і карти забезпечують візуальну та просторову інтерпретацію результатів, прогнозний модуль реалізує числові й класифікаційні моделі, ризик-аналітичний блок дає змогу оцінювати несприятливі відхилення, а сценарний модуль підтримує порівняння альтернативних умов. У сукупності ці компоненти формують програмно реалізований контур застосування результатів дисертаційного дослідження.

5.4. Сценарії, колекції, звітність і відтворюваність дослідницького процесу

Особливістю інформаційно-аналітичної системи «Пшениця України» є підтримка не лише окремих аналітичних процедур, а й цілісного відтворюваного дослідницького процесу. У задачах прогнозування врожайності важливо зберігати не тільки кінцеві значення прогнозу, графіки чи карти, а й умови їх отримання: склад вхідної вибірки, набір предикторів, обрану модель, параметри сценарію, критерії класифікації, ризикові пороги та спосіб подання результатів. Без фіксації цих умов повторна перевірка, порівняння або використання результатів у подальшому аналізі стає ускладненим.

Для розв’язання цієї задачі в системі реалізовано сценарний контур, у межах якого результати роботи окремих модулів можуть об’єднуватися в дослідницькі сценарії. Сценарій у системі розглядається як структурований об’єкт, що поєднує вхідні дані, аналітичні налаштування, прогнозну конфігурацію, сценарні припущення, результати моделювання та службову інформацію про стан дослідницького процесу. Такий підхід дає змогу перейти від одноразового виконання моделі до збереження повної логіки аналітичного експерименту.

Основним об’єктом сценарного контуру є дослідницький сценарій. Він забезпечує зв’язок між збереженими конфігураціями даних, графіків і прогнозування, а також фіксує параметри дослідницького сценарію. У межах

дослідницького сценарію можуть бути пов’язані конфігурація вибірки даних, конфігурація графічного подання та конфігурація прогнозного експерименту, що відображають відповідно базову вибірку, параметри графічного подання та налаштування прогнозного експерименту. Завдяки цьому сценарій містить не лише кінцевий результат, а й опис шляху його отримання.

Конфігурація вибірки даних у сценарному контурі визначає інформаційну основу дослідження. Вона фіксує, для яких зон, областей, років і показників сформовано вибірку. Конфігурація графічного подання зберігає спосіб представлення результатів, що дає змогу повторно відкрити потрібну візуалізацію без ручного відновлення параметрів. Конфігурація прогнозного експерименту описує модель, цільову змінну, набір предикторів і сценарні зміни вхідних факторів. Поєднання цих об’єктів у межах дослідницького сценарію забезпечує зв’язок між даними, моделями, візуалізацією та інтерпретацією результатів.

Сценарії можуть використовуватися для різних типів дослідницьких задач. Наприклад, один сценарій може відповідати базовому прогнозуванню врожайності за фактичними агрокліматичними ознаками, інший — класифікації низьковрожайних років, третій — оцінюванню ризику за трендовими залишками, четвертий – сценарному аналізу зміни температурних показників або опадів. Така організація дає змогу зберігати результати різних постановок задач у єдиному середовищі та порівнювати їх між собою.

Для впорядкування сценаріїв у системі передбачено папки та колекції. Папка сценаріїв використовується для групування сценаріїв за тематикою, етапом дослідження або типом аналітичної задачі. Наприклад, окремі папки можуть відповідати формуванню вибірок, прогнозуванню, ризик-аналізу, сценарному моделюванню або підготовці матеріалів для звітності. Це забезпечує структурованість роботи користувача та спрощує повторне відкриття раніше створених сценаріїв.

Колекція сценаріїв призначена для об’єднання кількох сценаріїв у логічно завершену добірку. Колекція може використовуватися для порівняння альтернативних моделей, аналізу різних агрокліматичних зон, зіставлення базового

та зміненого сценаріїв, узагальнення ризикових оцінок або підготовки підсумкового звіту. На відміну від окремого сценарію, колекція дає змогу представити не один результат, а групу пов’язаних аналітичних результатів, що мають спільну дослідницьку мету.

Важливим елементом відтворюваності є фіксація історії дій у межах сценарію. Для цього використовується журнал подій сценарію, який зберігає інформацію про події, пов’язані зі створенням, редагуванням, оновленням або використанням сценарію. Журнал подій дає змогу відстежувати послідовність змін, контролювати розвиток дослідницького сценарію та пояснювати, як саме було отримано певний результат. Це особливо важливо для досліджень, у яких одна й та сама вибірка може використовуватися в різних моделях або сценарних припущеннях.

Послідовність формування сценарію в ІАС «Пшениця України» подано на рис. 5.5. Схема відображає перехід від формування вибірки, графічного подання та прогнозного експерименту до дослідницького сценарію, його організації в папках і колекціях, фіксації подій та підготовки звітних матеріалів. У такій структурі кожен результат зберігає зв’язок із вихідними даними, параметрами моделі та способом аналітичного подання.

Рис. 5.5. Життєвий цикл дослідницького сценарію в ІАС «Пшениця України»

Як видно з рис. 5.5, сценарний контур системи забезпечує не лінійне одноразове виконання розрахунку, а повторно використовуваний цикл дослідження. Користувач може сформувати вибірку, зберегти її конфігурацію, побудувати графічне або картографічне подання, налаштувати прогнозний експеримент, зберегти його конфігурацію, сформувати дослідницький сценарій, об’єднати кілька сценаріїв у колекцію сценаріїв і підготувати звітні матеріали. За потреби сценарій може бути повторно відкритий, уточнений, доповнений або використаний як основа для нового варіанта аналізу.

Формування звітів є завершальним етапом сценарного контуру. Звіт у системі розглядається як документоване представлення результатів, отриманих у межах одного сценарію або колекції сценаріїв. До такого звіту можуть входити опис сформованої вибірки, основні параметри моделі, графіки, карти, прогнозні оцінки, ризикові показники, сценарні зміни та короткі аналітичні висновки. Це забезпечує перехід від інтерактивної роботи в системі до підсумкового матеріалу, придатного для архівування, демонстрації, порівняння або подальшого використання. Приклади інтерфейсних форм організації сценаріїв, колекцій і звітних матеріалів подано в додатку Г.

Підтримка експорту результатів у поширені формати дає змогу використовувати матеріали системи поза межами вебінтерфейсу. Експорт може застосовуватися для збереження табличних результатів, передавання сформованих даних в інші програмні середовища, підготовки аналітичних документів або формування додаткових матеріалів до дослідження. У такий спосіб звітність у системі виконує не лише презентаційну функцію, а й забезпечує документування отриманих результатів.

Відтворюваність дослідницького процесу в ІАС «Пшениця України» забезпечується поєднанням кількох механізмів: збереженням конфігурацій вибірки, графіків і прогнозування; організацією сценаріїв; групуванням сценаріїв у папки й колекції; фіксацією історії подій; формуванням звітів та експортом результатів. Така організація дає змогу повторно перевіряти отримані результати, порівнювати альтернативні постановки задач, аналізувати вплив зміни вхідних умов і

використовувати попередньо сформовані конфігурації в нових дослідницьких сценаріях.

Таким чином, сценарії, колекції та звітність у системі «Пшениця України» забезпечують організацію, повторне відкриття, порівняння та документування результатів прогнозного аналізу. Вони пов’язують вхідні дані, параметри моделей, прогнозні оцінки, ризикові показники, картографічне подання та звітні матеріали в єдину структуру дослідницького процесу. Завдяки цьому система підтримує не тільки виконання окремих розрахунків, а й збереження умов отримання результатів для їх подальшої перевірки та використання.

5.5. Апробація системи, експериментальна перевірка та напрями використання

Апробацію інформаційно-аналітичної системи «Пшениця України» проведено з метою перевірки її працездатності, узгодженості функціональних модулів і можливості повторного застосування результатів інтелектуальних моделей прогнозування врожайності. У межах перевірки оцінювалася не лише робота окремих модулів, а й цілісність наскрізного дослідницького циклу: формування агрокліматичної вибірки, побудова графічних і картографічних представлень, запуск прогнозних процедур, класифікація ризикових років, оцінювання ризику, сценарне порівняння результатів, збереження сценаріїв, формування колекцій і підготовка звітних матеріалів.

Експериментальна перевірка ґрунтувалася на використанні регіональних даних врожайності пшениці та агрокліматичних показників, сформованих у попередніх розділах дисертації. Такий підхід дав змогу перевірити систему не на умовних тестових даних, а на предметній базі, яка використовувалася для побудови та дослідження запропонованих моделей. Основною одиницею аналізу виступала узгоджена комбінація «область–рік», що містила показники врожайності, трендові залишки, температурні ознаки, показники опадів і похідні предиктори.

Перевірка працездатності системи проводилася за наскрізним сценарієм. На першому етапі в модулі підготовки даних формувалася вибірка за обраною

агрокліматичною зоною, переліком областей, часовим інтервалом і набором показників. Після попереднього перегляду та перевірки структури даних сформована конфігурація вибірки зберігалася для повторного використання в інших модулях системи без повторного ручного налаштування параметрів.

На другому етапі перевірялася робота модулів візуальної та просторової аналітики. На основі збереженої вибірки будувалися графічні представлення динаміки врожайності, трендових залишків, температурних показників та опадів. Картографічний модуль використовувався для просторового подання показників у розрізі областей України. Це дало змогу перевірити узгодженість між табличною вибіркою, графічним представленням і просторовою інтерпретацією результатів.

На третьому етапі виконувалася перевірка прогнозного контуру. Сформована конфігурація вибірки даних використовувалася як джерело вхідних даних для побудови прогнозних оцінок і класифікації рівня врожайності. У межах цього етапу перевірялася можливість використання моделей машинного навчання для числового та категоріального прогнозування, зокрема для віднесення спостережень до класів «низька» / «висока» врожайність. Додатково оцінювалася можливість збереження параметрів прогнозного експерименту у вигляді відповідної конфігурації, що забезпечує повторне відкриття та відтворення умов моделювання.

На четвертому етапі перевірявся ризик-аналітичний контур системи. Для цього використовувалися трендові залишки врожайності, квантильні пороги та показники несприятливих відхилень від очікуваного рівня. Система забезпечувала можливість виявлення низьковрожайних років, оцінювання ризикових станів, просторового подання областей із підвищеною вразливістю та включення результатів ризик-аналізу до сценарного або звітного контуру. У межах цієї перевірки ризик-аналіз розглядався як складова повного циклу прогнозної аналітики, пов’язана з інтерпретацією несприятливих відхилень урожайності.

На п’ятому етапі перевірялася робота сценарного контуру. Сформовані вибірки, графічні представлення, прогнозні конфігурації та ризикові оцінки об’єднувалися в дослідницький сценарій. Окремі сценарії групувалися в папки, об’єднувалися в колекції сценаріїв і використовувалися для порівняння

альтернативних результатів. Журнал подій давав змогу фіксувати послідовність змін і дій, пов’язаних зі сценарієм. У такий спосіб було перевірено, що система підтримує не лише виконання окремих операцій, а й організацію відтворюваного дослідницького процесу.

Завершальним етапом апробації була перевірка формування звітів та експорту результатів. Звітні матеріали формувалися на основі сценарію або колекції сценаріїв і містили опис використаної вибірки, результати графічного й картографічного подання, прогнозні оцінки, ризикові показники, сценарні порівняння та аналітичні висновки. Це забезпечило перехід від інтерактивної роботи в системі до документованого результату, придатного для архівування, демонстрації або подальшого використання.

Узагальнену послідовність експериментальної перевірки ІАС «Пшениця України» подано на рис. 5.6. Схема відображає наскрізний контроль працездатності системи від формування вибірки до звіту та показує, що кожен етап перевірки спирається на результати попереднього.

Рис. 5.6. Послідовність експериментальної перевірки працездатності та відтворюваності ІАС «Пшениця України»

Як видно з рис. 5.6, експериментальна перевірка охоплює всі основні функціональні контури системи: підготовку даних, аналітику, прогнозування,

ризик-аналіз, сценарне порівняння та звітність. Такий підхід дає змогу оцінити систему не лише за фактом наявності окремих модулів, а за здатністю забезпечувати повний відтворюваний цикл застосування інтелектуальних моделей прогнозування врожайності.

Результати апробації підтвердили, що ІАС «Пшениця України» забезпечує узгоджене використання агрокліматичних даних, моделей машинного навчання, прогнозних оцінок, ризикових показників і сценарних конфігурацій. Система дає змогу повторно відкривати сформовані вибірки, відтворювати графічні та прогнозні налаштування, порівнювати альтернативні сценарії, подавати результати в регіональному розрізі та формувати документовані аналітичні матеріали. Отримані результати підтверджують можливість використання системи як програмного середовища апробації та повторного застосування запропонованих моделей. Окремі результати проходження експериментального сценарію в інтерфейсі ІАС «Пшениця України» наведено в додатку Г.

Практичне використання системи може здійснюватися у кількох напрямах. У дослідницькій діяльності вона може застосовуватися для перевірки моделей прогнозування врожайності, порівняння різних постановок задач, аналізу впливу агрокліматичних ознак і формування відтворюваних сценаріїв. В аналітичній практиці система може використовуватися для виявлення регіонів і років із підвищеним ризиком низької врожайності, оцінювання наслідків зміни кліматичних параметрів і підготовки матеріалів для підтримки управлінських рішень. В освітньому процесі система може бути використана як приклад програмної реалізації повного циклу інтелектуального аналізу аграрних даних.

Водночас розроблена система має предметні межі застосування, зумовлені вихідною базою даних, обраним просторовим рівнем аналізу та переліком реалізованих моделей. У межах дисертаційного дослідження її апробацію виконано на даних врожайності пшениці в областях України. Адаптація системи до інших культур або територіальних рівнів потребує відповідного наповнення бази статистичними, кліматичними та просторовими даними, а також перевірки інформативності ознак і стійкості моделей для нової предметної області.

Подальший розвиток ІАС «Пшениця України» може бути пов’язаний із розширенням набору культур, підключенням додаткових джерел даних, використанням супутникових індексів, поглибленням модуля ризик-аналізу, додаванням нових моделей машинного та глибинного навчання, удосконаленням журналу запусків моделей і розширенням засобів автоматизованої звітності. Перспективним є також розвиток механізмів порівняння моделей, збереження історії експериментів і формування інтерактивних аналітичних панелей для різних категорій користувачів.

Таким чином, апробація ІАС «Пшениця України» підтвердила її працездатність і логічну узгодженість як програмного середовища для застосування інтелектуальних моделей прогнозування врожайності. Система забезпечує зв’язок між даними, ознаками, моделями, ризиковими оцінками, сценаріями, картографічним поданням і звітними матеріалами. Це дає змогу використовувати її для відтворення результатів дисертаційного дослідження, порівняння альтернативних сценаріїв та підготовки аналітичних матеріалів.

Висновки до розділу 5

1. Розроблено інформаційно-аналітичну систему «Пшениця України», призначену для програмної реалізації, апробації та відтворюваного застосування інтелектуальних моделей прогнозування врожайності за агрокліматичними факторами. Показано, що система є не окремим розрахунковим модулем, а інтегрованим дослідницьким середовищем, у якому поєднано підготовку даних, формування ознак, прогнозування, класифікацію ризикових років, оцінювання ризику, сценарний аналіз, просторове подання та звітність. Окремі алгоритмічно-програмні напрацювання, реалізовані в комп’ютерній програмі [23], розвинуто та інтегровано в межах ІАС «Пшениця України».

2. Обґрунтовано архітектуру системи, яка поєднує рівень збереження даних, прикладний серверний рівень, аналітико-модельний контур, користувацький інтерфейс і засоби формування звітних матеріалів. Структура даних системи підтримує не лише збереження агрокліматичних і врожайнісних показників, а й

фіксацію конфігурацій дослідницького процесу. Використання конфігурації вибірки даних, конфігурації графічного подання та конфігурації прогнозного експерименту забезпечує повторне застосування сформованих вибірок, графічних представлень і прогнозних експериментів у різних модулях системи.

3. Реалізовано функціональні модулі підготовки даних, візуальної аналітики, просторового подання, прогнозування, класифікації ризикових років, оцінювання ризику та сценарного аналізу. Модуль формування вибірок забезпечує створення відтворюваної агрокліматичної основи для подальшого моделювання; модулі графіків і карт дають змогу інтерпретувати динаміку та регіональні відмінності; прогнозний модуль реалізує числове й класифікаційне прогнозування; ризик-аналітичний контур забезпечує аналіз несприятливих відхилень на основі трендових залишків і квантильних порогів.

4. Розроблено сценарний контур системи, у межах якого результати роботи окремих модулів можуть бути об’єднані в дослідницький сценарій, згруповані в папці сценаріїв, порівняні в колекції сценаріїв і задокументовані через журнал подій та звітні матеріали. Така організація дає змогу зберігати не лише кінцеві результати, а й умови їх отримання: склад вибірки, параметри графічного подання, налаштування прогнозного експерименту, сценарні припущення та результати ризик-аналізу. Це забезпечує відтворюваність, повторне відкриття, порівняння й документування дослідницьких конфігурацій.

5. Проведено експериментальну перевірку працездатності ІАС «Пшениця України» на даних врожайності пшениці та агрокліматичних показниках областей України. Перевірка охопила повний цикл роботи системи: формування вибірки, збереження її конфігурації, побудову графічних і картографічних представлень, запуск прогнозних і класифікаційних процедур, оцінювання ризиків, створення сценаріїв, формування колекцій і підготовку звітів. Результати апробації підтвердили придатність системи для практичного застосування запропонованих моделей у задачах прогнозування врожайності, оцінювання агрокліматичних ризиків, сценарного аналізу та підтримки прийняття рішень у рослинництві.

ВИСНОВКИ

У дисертаційній роботі розв’язано актуальну науково-прикладну задачу розроблення, обґрунтування та експериментальної перевірки інтелектуальних моделей прогнозування врожайності на основі методів машинного навчання для підвищення точності, стійкості та інформативності прогнозних оцінок з урахуванням агрокліматичної мінливості, регіональної неоднорідності та ризику низької врожайності. При цьому отримано такі наукові та практичні результати:

1. Сформовано агрокліматичну матрицю ознак для 18 областей України за період 2000–2021 рр., яка містить декадні температурні показники квітня–червня та місячні суми опадів, просторово узгоджені з обласним рівнем статистичних даних урожайності. Обґрунтовано використання трендових залишків як інформаційного показника міжрічної мінливості врожайності, що дає змогу відокремити довгострокову технологічну складову від короткострокових коливань, зумовлених агрокліматичними та іншими чинниками. Виконано агрокліматичне групування областей України та виділено три укрупнені зони: степову, лісостепову та західну. Це дало змогу перейти від аналізу окремих коротких регіональних рядів до формування структурованих зональних вибірок, придатних для побудови, порівняння та верифікації моделей машинного навчання. Отриманий ознаковий, часовий і просторовий базис використано в подальших розділах дисертації для прогнозного моделювання, оцінювання ризику низької врожайності та сценарно-оптимізаційного аналізу.

2. Запропоновано оцінювати ризик низької врожайності на основі від’ємних трендових залишків, що дозволяє розглядати несприятливі відхилення врожайності від очікуваного трендового рівня як кількісну характеристику агровиробничого ризику. Використання трендових залишків дало змогу відокремити ризик міжрічної мінливості від довгострокових змін, пов’язаних із технологічним розвитком, зміною організації виробництва та іншими поступовими факторами. Установлено, що розподіли трендових залишків мають зональну специфіку: для степової зони характерні важчі хвости розподілу, що свідчить про вищу ймовірність значних несприятливих відхилень, тоді як для лісостепової та західної зон прийнятною

апроксимацією є нормальний розподіл. На основі квантильних мір VaR і CVaR показано, що степова зона має найвищий рівень ризику низької врожайності, а західна зона — найнижчий. Отримані результати можуть бути використані для порівняльного аналізу агрокліматичних ризиків, обґрунтування страхових і управлінських рішень, а також для подальшого сценарного аналізу несприятливих умов виробництва.

3. Обґрунтовано доцільність розгляду задачі прогнозування врожайності не лише як задачі числового передбачення її очікуваного рівня, а й як задачі класифікації років на категорії низької та високої врожайності. Такий перехід має важливе прикладне значення, оскільки в задачах підтримки прийняття рішень часто більш важливим є не точне числове значення врожайності, а своєчасне виявлення ризику несприятливого року. Для формування цільової змінної використано трендові залишки, що дозволило пов’язати класифікацію не з абсолютним рівнем урожайності, а з відхиленням фактичного результату від очікуваного трендового значення. Позитивним класом визначено клас низької врожайності, що спрямовує модель на виявлення саме несприятливих станів. Як основну метрику якості використано F1_LY, яка поєднує точність прогнозування класу низької врожайності та повноту його виявлення. Запропонована класифікаційна постановка стала методичною основою для подальшого порівняння ендогенних і екзогенних моделей машинного та глибинного навчання.

4. Побудовано та порівняно ендогенні й екзогенні класифікаційні моделі машинного навчання. Ендогенні моделі, що використовують лише лагову інформацію про трендові залишки врожайності, показали обмежену якість на тестовій вибірці: Logistic Regression — F1_LY = 0,3636, Random Forest — F1_LY = 0,5455, SVM RBF — F1_LY = 0,6093. Екзогенні моделі на основі агрокліматичних ознак забезпечили істотне покращення: для степової зони SVM RBF досягла F1_LY = 0,8421, Logistic Regression — F1_LY = 0,7500, Random Forest — F1_LY = 0,7059. Для західної зони найкращою також була SVM RBF з F1_LY = 0,7619, а для лісостепової зони — Logistic Regression з F1_LY = 0,6433. Отримані результати

підтвердили перевагу екзогенного підходу та необхідність урахування зональної специфіки даних.

5. Встановлено, що ендогенні RNN-моделі мають обмежену прогнозну здатність для короткого ряду трендових залишків урожайності, який характеризується слабкою автокореляційною структурою. У такій постановці рекурентні нейронні мережі не отримують достатньої кількості внутрішньої часової інформації для стабільного виявлення закономірностей, тому збільшення складності архітектури не гарантує підвищення якості прогнозування. В ендогенній постановці найкращий компроміс між якістю та стійкістю забезпечила LSTM з тестовим F1_LY = 0,6667, тоді як GRU продемонструвала істотне зниження якості на тестовій вибірці. В екзогенній постановці для Херсонської області залучення агрокліматичних і режимних факторів суттєво підвищило результативність RNN-моделей: якість GRU зросла до F1_LY = 0,7211, а LSTM досягла F1_LY = 0,7009. Це підтвердило, що ефективність рекурентних нейронних мереж у задачах прогнозування врожайності визначається не лише вибором архітектури, а насамперед інформативністю вхідного простору ознак. Отримані результати дозволили уточнити роль LSTM і GRU у системі інтелектуальних моделей прогнозування врожайності та обґрунтувати доцільність їх використання переважно в екзогенній постановці, коли до моделі залучаються зовнішні кліматичні та економічні предиктори.

6. Застосування решіткового пошуку з відбором трьох найкращих конфігурацій дало змогу оцінювати не лише максимальну якість окремої моделі на валідаційній вибірці, а й стійкість групи близьких за якістю конфігурацій на незалежних тестових даних. Такий підхід є особливо важливим для задач прогнозування врожайності, де обсяг доступних вибірок часто є обмеженим, а результати окремої конфігурації можуть істотно залежати від способу поділу даних, початкових параметрів і випадкових коливань у навчальній вибірці. Усереднене оцінювання TOP-3 конфігурацій дозволило зменшити ризик переінтерпретації випадково найкращого результату та забезпечити більш надійне порівняння моделей. Методика була використана для оцінювання моделей машинного та

глибинного навчання в ендогенній і екзогенній постановках, а також для порівняння результатів між агрокліматичними зонами. Це дало змогу підвищити обґрунтованість висновків щодо переваг окремих модельних підходів і забезпечити більш стійку інтерпретацію результатів прогнозування низької врожайності.

7. Показано, що врожайність не може бути достатньо повно описана лише лінійною адитивною дією температури та опадів, оскільки вплив погодних умов на рослини має нелінійний, фазовий і кумулятивний характер. Для врахування таких ефектів розроблено підхід до синтезу нелінійних кліматичних ознак другого порядку, який передбачає використання квадратів базових агрокліматичних предикторів і добутків кліматичних показників суміжних часових інтервалів. Це дало змогу формалізувати як можливу наявність оптимальних або критичних рівнів окремих факторів, так і взаємодію погодних умов у послідовних фазах вегетації. Установлено, що структура нелінійних моделей відрізняється залежно від агрокліматичної зони: у степовій і лісостеповій зонах важливішими є квадратичні ефекти окремих факторів, тоді як у західній зоні істотнішу роль відіграють добутки послідовних факторів. Отримані результати підтвердили доцільність розширення ознакового простору нелінійними кліматичними предикторами та показали, що такі ознаки підвищують інтерпретованість моделей, оскільки дозволяють змістовно описувати неадитивний і кумулятивний вплив погодних умов. Нелінійна модель урожайності була використана також як основа для побудови сприятливої температурної траєкторії вегетації.

8. Розроблені методи й моделі реалізовано в інформаційно-аналітичній системі «Пшениця України», яка забезпечує єдиний програмний контур опрацювання даних: від завантаження, очищення, просторово-часового узгодження та формування агрокліматичних ознак до побудови прогнозних моделей, оцінювання ризику низької врожайності, сценарного аналізу, візуалізації та подання результатів користувачу. Це дало змогу перейти від окремих модельних експериментів до цілісного інформаційно-аналітичного середовища, у якому реалізовано послідовний цикл підготовки даних, прогнозування, інтерпретації результатів і формування аналітичних оцінок. Програмна реалізація підтверджує

можливість практичного використання запропонованих інтелектуальних моделей у задачах аналітичної підтримки рішень в аграрному виробництві. Інтеграція моделей в інформаційно-аналітичній системі забезпечує відтворюваність розрахунків, узгоджене використання агрокліматичних даних і прогнозних процедур, а також подання результатів у формі, придатній для оцінювання ризиків, аналізу можливих сценаріїв і підтримки управлінських рішень у рослинництві.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Hrytsiuk P., Babych T., Baranovsky S., Havryliuk M. Assessing of climate impact on wheat yield using machine learning techniques. In: *Information Control Systems & Technologies* : proceedings of the 11th International Conference, ICST-2023, Odesa, Ukraine, September 21–23, 2023. *CEUR Workshop Proceedings*. 2023. Vol. 3513. P. 314–329.

2. Hrytsiuk P., Havryliuk M. Modeling of the nonlinear impact of climatic factors on wheat yield using machine learning techniques. In: Ermolayev V. et al. (eds) *Information and Communication Technologies in Education, Research, and Industrial Applications* : 19th International Conference, ICTERI 2024, Lviv, Ukraine, September 23–27, 2024, Proceedings. Communications in Computer and Information Science. Cham : Springer, 2025. Vol. 2359. P. 20–35. https://doi.org/10.1007/978-3-031-81372-6_2

3. Hrytsiuk P., Havryliuk M., Nehrey M. Constructing the optimal temperature trajectory for the wheat vegetative cycle. In: Ermolayev V. et al. (eds) *Information and Communication Technologies in Education, Research, and Industrial Applications* : 20th International Conference, ICTERI 2025, Nice, France, September 1–4, 2025, Proceedings. Communications in Computer and Information Science. Cham : Springer, 2025. Vol. 2763. P. 101–116. https://doi.org/10.1007/978-3-032-10477-9_8

4. Hrytsiuk P., Babych T., Kardash O., Havryliuk M. Application of Machine Learning for Wheat Yield Prediction. In: Potapov I. et al. (eds) *Digitalisation and Digital Transformation* : First Research Twinning Conference, RTC-Digital 2023, Liverpool, UK, March 27–30, 2023, Proceedings. Communications in Computer and Information Science. Cham : Springer, 2026. Vol. 2647. P. 3–11. https://doi.org/10.1007/978-3-032-04731-1_1

5. Грицюк П. М., Гаврилюк М. С. Моделювання впливу кліматичних факторів на врожайність пшениці методами машинного навчання. *Штучний інтелект*. 2025. Т. 30, № 1. С. 121–130. https://doi.org/10.15407/jai2025.01.121

6. Грицюк П. М., Гаврилюк М. С., Джоші О. І. Ідентифікація закону розподілу трендових залишків врожайності як інструмент моделювання ризиків

зерновиробництва. *Штучний інтелект*. 2025. Т. 30, № 2. С. 72–83. https://doi.org/10.15407/jai2025.02.072

7. Грицюк П. М., Гаврилюк М. С. Класифікаційний підхід до прогнозування врожайності. *Форум молодих економістів-кібернетиків «Моделювання економіки: проблеми, тенденції, досвід»* : тези доповідей X Всеукраїнської науково-практичної конференції, м. Львів, 25 листопада 2022 р. Львів, 2022. С. 47–49.

8. Havryliuk M., Hrytsiuk P., Kardash O., Babych T. Forecasting of wheat yield using machine learning techniques. *Digital Theme UK–Ukraine Research Twinning Conference* : UA-DIGITAL 2023, online, March 27–31, 2023.

9. Havryliuk M. S. Wheat yield forecasting using machine learning methods. *Проблеми та перспективи розвитку сучасної науки* : збірник тез доповідей Міжнародної науково-практичної конференції молодих науковців, аспірантів і здобувачів вищої освіти, м. Рівне, 11–12 травня 2023 р. Рівне : НУВГП, 2023. С. 139–142.

10. Грицюк П. М., Гаврилюк М. С. Оцінка втрат врожаю пшениці внаслідок військової агресії росії. *Інтегровані інтелектуальні робототехнічні комплекси (ІІРТК-2023)* : матеріали XVI Міжнародної науково-практичної конференції, м. Київ, 23–24 травня 2023 р. Київ : НАУ, 2023. С. 374–376.

11. Hrytsiuk P., Babych T., Baranovskii S., Havryliuk M. Assessing of climate impact on wheat yield using machine learning techniques. *Information Control Systems and Technologies (ICST-2023)* : materials of the XI International Scientific-Practical Conference, Odesa, Ukraine, September 21–23, 2023. P. 102–105.

12. Грицюк П. М., Гаврилюк М. С. Використання машинного навчання для управління процесами аграрної економіки. *Міжнародна конференція з інноваційних рішень у програмній інженерії (ICISSE 2023)*, м. Івано-Франківськ, 29–30 листопада 2023 р. Івано-Франківськ, 2023. С. 150–154. https://doi.org/10.5281/zenodo.10397356

13. Гаврилюк М. С. Використання машинного навчання для управління процесами аграрної економіки. *Науково-практична конференція студентів та*

аспірантів навчально-наукового інституту кібернетики, інформаційних технологій та інженерії, м. Рівне, 23 травня 2024 р. Рівне : НУВГП, 2024. С. 17.

14. Гаврилюк М. С., Грицюк П. М. Ідентифікація закону розподілу залишків врожайності сільськогосподарських культур. *Комп'ютерне моделювання та програмне забезпечення інформаційних систем і технологій 2024 (КМПЗ 2024)* : матеріали IV Міжнародної науково-практичної конференції, м. Чернівці, 30 травня – 1 червня 2024 р. Чернівці, 2024. С. 76–80.

15. Грицюк П. М., Гаврилюк М. С. Оцінювання нелінійного впливу кліматичних факторів на врожайність пшениці. *Штучний інтелект. Наука. Бізнес* : матеріали Всеукраїнської науково-практичної конференції, м. Одеса, 22 жовтня 2024 р. Одеса, 2024. С. 18–21.

16. Гаврилюк М. С., Грицюк П. М. Дослідження впливу кліматичних факторів на коливання врожайності пшениці. *Форум молодих економістів-кібернетиків. Моделювання економіки: проблеми, тенденції, досвід* : матеріали XII Всеукраїнської науково-практичної конференції, м. Львів, 22–23 листопада 2024 р. Львів, 2024. С. 11–14.

17. Havryliuk M., Hrytsiuk P. Modeling the nonlinear influence of climatic factors on wheat yield using machine learning methods. *Advances in Science, Technology, and Society* : proceedings of the International Scientific Conference, Oxford, United Kingdom, February 20, 2025. P. 210–214.

18. Грицюк П. М., Бабич Т. Ю., Гаврилюк М. С. Побудова оптимальної температурної траєкторії вегетації пшениці. *Математичні методи, моделі та інформаційні технології в економіці* : матеріали IX Міжнародної науково-методичної конференції, м. Чернівці, 24–25 квітня 2025 р. Чернівці, 2025. С. 69–71.

19. Гаврилюк М. С. Моделювання нелінійного впливу кліматичних факторів на врожайність пшениці методами машинного навчання. *Науково-практична конференція студентів та аспірантів навчально-наукового інституту кібернетики, інформаційних технологій та інженерії*, м. Рівне, 14 травня 2025 р. Рівне : НУВГП, 2025. С. 49.

20. Гаврилюк М. С., Грицюк П. М. Використання закону розподілу трендових залишків врожайності для моделювання ризиків виробництва зерна. *Моделювання, керування та інформаційні технології* : матеріали VIII Міжнародної науково-практичної конференції, м. Рівне, 6–8 листопада 2025 р. Рівне : НУВГП, 2025. С. 286–288. https://doi.org/10.31713/MCIT.2025.089

21. Гаврилюк М. С. Нелінійне моделювання впливу кліматичних факторів на врожайність пшениці в Україні. *Форум молодих економістів-кібернетиків. Моделювання економіки: проблеми, тенденції, досвід* : матеріали XIII Всеукраїнської науково-практичної конференції, м. Львів, 21–22 листопада 2025 р. Львів, 2025. С. 101–102.

22. Hrytsiuk P., Babych T., Hladka O., Nehrey M., Havryliuk M. Machine learning-based modeling of climate effects on wheat yield in the Ukrainian Steppe. *Journal of Lifestyle and SDGs Review*. 2026. Vol. 6, No. 1. Art. e08120. https://doi.org/10.47172/2965-730X.SDGsReview.v6.n01.pe08120

23. Гаврилюк М. С., Грицюк П. М., Бабич Т. Ю., Кардаш О. Л. Комп’ютерна програма «Математичне моделювання впливу кліматичних факторів на врожайність методами машинного навчання» : свідоцтво про реєстрацію авторського права на твір, 127859. Дата реєстрації 26 червня 2024 р. URL: https://iprop-ua.com/cr/uol4v9nc/

24. Hrytsiuk P., Havryliuk M., Kardash O. A classification approach to wheat yield forecasting using machine learning and deep learning methods. Agricultural and Resource Economics. 2026. Vol. 12, No. 2. P. 83–108. https://doi.org/10.51599/are.2026.12.02.04

25. Krisnawijaya N. N. K., Tekinerdogan B., Catal C., van der Tol R. Data analytics platforms for agricultural systems: A systematic literature review. *Computers and Electronics in Agriculture*. 2022. Vol. 195. Art. 106813. https://doi.org/10.1016/j.compag.2022.106813

26. Cravero A., Sepúlveda S., Gutiérrez F., Muñoz L. From precision agriculture to intelligent agricultural ecosystems: a systematic review of machine learning and big data applications. *Agronomy*. 2026. Vol. 16, No. 5. Art. 516. https://doi.org/10.3390/agronomy16050516

27. Put Your Data to Work. Climate FieldView. URL: https://climate.com/en-ca/resources/getting-started/fieldview-201-overview/put-your-data-to-work.html (дата звернення: 14.05.2025).

28. Data Management | Operations Center. John Deere. URL: https://www.deere.com/en/technology-products/precision-ag-technology/data-management/operations-center/ (дата звернення: 14.05.2025).

29. OneSoil Pro: Quick Start Guide. OneSoil Help Center. URL: https://help.onesoil.ai/en/articles/7007824-onesoil-pro-quick-start-guide (дата звернення: 14.05.2025).

30. Crop Monitoring Software For Remote Farm Analytics. EOS Data Analytics. URL: https://eos.com/products/crop-monitoring/ (дата звернення: 14.05.2025).

31. xarvio® Digital Farming Solutions FIELD MANAGER. xarvio Digital Farming Solutions. URL: https://ag.xarvio.com/field-manager (дата звернення: 14.05.2025).

32. CropX Agronomic Farm Management System. *CropX*. URL: https://cropx.com/ (дата звернення: 14.05.2025).

33. Granular Insights. Corteva Agriscience. URL: https://www.corteva.com/us/products-and-solutions/digital-solutions/granular-insights.html (дата звернення: 14.05.2025).

34. Snehalatha M., Patil A. AgriYieldAI an AI-driven decision support system for crop yield prediction using YieldFusionNet. *Discover Computing*. 2026. Vol. 29. Art. 222. https://doi.org/10.1007/s10791-026-10096-y

35. Aworka R., Cedric L. S., Adoni W. Y. H., Zoueu J. T., Mutombo F. K., Kimpolo C. L. M., Nahhal T., Krichen M. Agricultural decision system based on advanced machine learning models for yield prediction: case of East African countries. *Smart Agricultural Technology*. 2022. Vol. 2. Art. 100048. https://doi.org/10.1016/j.atech.2022.100048

36. Sudha S. P., Loret J. B. S. A review on machine learning-based precision agriculture techniques for crop farming monitoring with IoT. *Discover Environment*. 2026. Vol. 4. Art. 10. https://doi.org/10.1007/s44274-025-00305-8

37. Kuan Y. N., Goh K. M., Lim L. L. Systematic review on machine learning and computer vision in precision agriculture: applications, trends, and emerging techniques.

Engineering Applications of Artificial Intelligence. 2025. Vol. 148. Art. 110401. https://doi.org/10.1016/j.engappai.2025.110401

38. Rajbongshi A., Johora F. T., Hossain A., Sarker M. S., Rahman M. H., Rahman M. W., Alotaibi F. T., Moni M. A. Leveraging explainable AI for sustainable agriculture: a comprehensive review of recent advances. *Artificial Intelligence Review*. 2026. Vol. 59, No. 3. Art. 105. https://doi.org/10.1007/s10462-025-11459-5

39. Cartolano A., Cuzzocrea A., Pilato G. Analyzing and assessing explainable AI models for smart agriculture environments. *Multimedia Tools and Applications*. 2024. Vol. 83. P. 37225–37246. https://doi.org/10.1007/s11042-023-17978-z

40. Olatinwo D. D., Myburgh H. C., De Freitas A., Abu-Mahfouz A. Explainable data-driven approach for smart crop yield prediction in Sub-Saharan Africa: performance and interpretability analysis. *Agriculture*. 2026. Vol. 16, No. 8. Art. 826. https://doi.org/10.3390/agriculture16080826

41. Shams M. Y., Gamel S. A., Talaat F. M. Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making. *Neural Computing and Applications*. 2024. Vol. 36, No. 11. P. 5695–5714. https://doi.org/10.1007/s00521-023-09391-2

42. Xu H., Yin H., Liu J., Wang L., Feng W., Song H., Fan Y., Qi K., Liang Z., Li W., Zhang X., Zhang R., Wang S. Prediction of spatial winter wheat yield by combining multiscale time series of vegetation and meteorological indices. *Agronomy*. 2025. Vol. 15, No. 5. Art. 1114. https://doi.org/10.3390/agronomy15051114

43. Kheir A. M. S., Ammar K. A., Amer A., Ali M. G. M., Ding Z., Elnashar A. Machine learning-based cloud computing improved wheat yield simulation in arid regions. *Computers and Electronics in Agriculture*. 2022. Vol. 203. Art. 107457. https://doi.org/10.1016/j.compag.2022.107457

44. Laktionov I., Semenov S., Diachenko G., Rutkowski L. An end-to-end ML and XAI-based framework for crop yield prediction under agroclimatic variability. *Applied Soft Computing*. 2026. Art. 115394. https://doi.org/10.1016/j.asoc.2026.115394

45. Державна служба статистики України. Методологічні положення державного статистичного спостереження «Площі, валові збори та урожайність

сільськогосподарських культур» : затв. наказом Державної служби статистики України від 01.06.2023 № 204. Київ : Державна служба статистики України, 2023. 61 с. URL: https://ukrstat.gov.ua/norm_doc/2023/204/204.pdf (дата звернення: 14.05.2025).

46. Адаменко Т. Зміна клімату та сільське господарство в Україні: що варто знати фермерам? Німецько-український агрополітичний діалог. Київ, 2019. 34 с. URL: https://mepr.gov.ua/wp-content/uploads/2023/07/5_Zmina-klimatu-ta-silske-gospodarstvo-v-Ukrayini.pdf (дата звернення: 14.05.2025).

47. Müller D., Jungandreas A., Koch F., Schierhorn F. *Impact of climate change on wheat production in Ukraine*. Agricultural Policy Report APD/APR/02/2016. German–Ukrainian Agricultural Policy Dialogue. Kyiv, 2016. URL: https://www.apd-ukraine.de/images/2016/02-2016/APD_APR_02-2016_impact_on_wheat_eng_fin.pdf (дата звернення: 14.05.2025).

48. Грицюк П. М., Бачишина Л. Д. Вплив зміни кліматичних умов на динаміку врожайності зернових в Україні. *Економіка України*. 2016. Т. 59, № 6(655). С. 68–75.

49. Senapati N., Halford N. G., Hawkesford M. J., Shewry P. R., Semenov M. A. Extreme heat and drought at flowering could threaten global wheat yields under climate change. *Climatic Change*. 2026. Vol. 179. Art. 28. https://doi.org/10.1007/s10584-025-04054-8

50. Popescu A., Dinu T. A., Stoian E., Şerban V. Climate change and its impact on wheat, maize and sunflower yield in Romania in the period 2017–2021. *Scientific Papers Series Management, Economic Engineering in Agriculture and Rural Development*. 2023. Vol. 23, No. 1. P. 587–602. URL: https://managementjournal.usamv.ro/pdf/vol.23_1/Art63.pdf

51. Konduri V. S., Vandal T. J., Ganguly S., Ganguly A. R. Data science for weather impacts on crop yield. *Frontiers in Sustainable Food Systems*. 2020. Vol. 4. Art. 52. https://doi.org/10.3389/fsufs.2020.00052

52. IPCC. *Climate Change 2022: Mitigation of Climate Change*. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on

Climate Change / eds. P. R. Shukla et al. Cambridge, UK ; New York, NY, USA : Cambridge University Press, 2022. https://doi.org/10.1017/9781009157926

53. Liu Z., Di L., Yang R., Guo L., Zhang C., Li H., Shao B. In–season crop yield prediction: state of the art and future research direction. *International Journal of Applied Earth Observation and Geoinformation*. 2026. Vol. 146. Art. 105129. https://doi.org/10.1016/j.jag.2026.105129

54. IPCC. Climate Change 2022: *Impacts, Adaptation, and Vulnerability*. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change / eds. H.-O. Pörtner et al. Cambridge, UK ; New York, NY, USA : Cambridge University Press, 2022. 3056 p. https://doi.org/10.1017/9781009325844

55. Gibson P. B., Chapman W. E., Altinok A., Delle Monache L., DeFlorio M. J., Waliser D. E. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment*. 2021. Vol. 2. Art. 159. https://doi.org/10.1038/s43247-021-00225-4

56. Leng G., Hall J. W. Predicting spatial and temporal variability in crop yields: an inter–comparison of machine learning, regression and process-based models. *Environmental Research Letters*. 2020. Vol. 15, No. 4. Art. 044027. https://doi.org/10.1088/1748-9326/ab7b24

57. Crane–Droesch A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*. 2018. Vol. 13, No. 11. Art. 114003. https://doi.org/10.1088/1748-9326/aae159

58. Sidhu B. S., Mehrabi Z., Ramankutty N., Kandlikar M. How can machine learning help in understanding the impact of climate change on crop yields? *Environmental Research Letters*. 2023. Vol. 18, No. 2. Art. 024008. https://doi.org/10.1088/1748-9326/acb164

59. Zhou W., Zhou W., Cammarano D., Butterbach–Bahl K., Olesen J. E., Lin Z., Huang T., Cai G., Zhang J., Qiu J., Wang S. Unraveling the impact of environmental factors on wheat yield across the European Union via explainable machine learning. *Computers and Electronics in Agriculture*. 2026. Vol. 241. Art. 111268. https://doi.org/10.1016/j.compag.2025.111268

60. Wang B., Feng P., Waters C., Cleverly J., Liu D. L., Yu Q. Quantifying the impacts of pre-occurred ENSO signals on wheat yield variation using machine learning in Australia. *Agricultural and Forest Meteorology*. 2020. Vol. 291. Art. 108043. https://doi.org/10.1016/j.agrformet.2020.108043

61. van Klompenburg T., Kassahun A., Catal C. Crop yield prediction using machine learning: a systematic literature review. *Computers and Electronics in Agriculture*. 2020. Vol. 177. Art. 105709. https://doi.org/10.1016/j.compag.2020.105709

62. Jabed M. A., Azmi Murad M. A. Crop yield prediction in agriculture: a comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*. 2024. Vol. 10, No. 24. Art. e40836. https://doi.org/10.1016/j.heliyon.2024.e40836

63. Elavarasan D., Vincent D. R., Sharma V., Zomaya A. Y., Srinivasan K. Forecasting yield by integrating agrarian factors and machine learning models: a survey. *Computers and Electronics in Agriculture*. 2018. Vol. 155. P. 257–282. https://doi.org/10.1016/j.compag.2018.10.024

64. Kuhn M., Johnson K. *Applied Predictive Modeling*. New York : Springer, 2013. 600 p. https://doi.org/10.1007/978-1-4614-6849-3

65. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006. Vol. 27, No. 8. P. 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

66. Грицюк П. М., Бабич Т. Ю., Красько Б. В. Класифікаційні методи прогнозування врожайності. *Вісник Хмельницького національного університету. Серія: Технічні науки*. 2022. № 3(309). С. 209–216. https://doi.org/10.31891/2307-5732-2022-309-3-209-216

67. Zhang L., Li C., Zhang G., Wu X., Zhou L., Chen L., Jiao Y., Liu G., Hei W. Winter wheat yield estimation based on multisource remote sensing data: a dual-branch TCN-Transformer model and analysis of growth-stage feature transition mechanisms. *Computers and Electronics in Agriculture*. 2025. Vol. 239, Part B. Art. 111014. https://doi.org/10.1016/j.compag.2025.111014

68. Araghi A., Daccache A. Remote sensing and TerraClimate datasets for wheat yield prediction using machine learning. *Smart Agricultural Technology*. 2025. Vol. 11. Art. 100909. https://doi.org/10.1016/j.atech.2025.100909

69. Devkota K. P., Bouasria A., Devkota M., Nangia V. Predicting wheat yield gap and its determinants combining remote sensing, machine learning, and survey approaches in rainfed Mediterranean regions of Morocco. *European Journal of Agronomy*. 2024. Vol. 158. Art. 127195. https://doi.org/10.1016/j.eja.2024.127195

70. Khosravani Shariati S. A., Abbasi A. Machine learning–based winter wheat yield prediction using multisource data. *Agricultural Water Management*. 2025. Vol. 322. Art. 109951. https://doi.org/10.1016/j.agwat.2025.109951

71. Schierhorn F., Hofmann M., Gagalyuk T., Ostapchuk I., Müller D. Machine learning reveals complex effects of climatic means and weather extremes on wheat yields during different plant developmental stages. *Climatic Change*. 2021. Vol. 169. Art. 39. https://doi.org/10.1007/s10584-021-03272-0

72. Artzner P., Delbaen F., Eber J.-M., Heath D. Coherent Measures of Risk. *Mathematical Finance*. 1999. Vol. 9, No. 3. P. 203–228. https://doi.org/10.1111/1467-9965.00068

73. Rockafellar R. T., Uryasev S. Optimization of Conditional Value-at-Risk. *Journal of Risk*. 2000. Vol. 2, No. 3. P. 21–41. https://doi.org/10.21314/JOR.2000.038

74. Gao Y., Wallach D., Gu B., Tang L., Lin T., Chang X., Hasegawa T., Asseng S., Kahveci T., Hoogenboom G. Quantification and comparison of prediction uncertainty associated with different practices of crop modeling. *Agricultural and Forest Meteorology*. 2025. Vol. 371. Art. 110633. https://doi.org/10.1016/j.agrformet.2025.110633

75. Wijesena S., Pradhan B. Dynamic climate risk modelling for agriculture: a machine learning approach integrating crop phenology into weather index insurance. *International Journal of Disaster Risk Reduction*. 2026. Vol. 136. Art. 106063. https://doi.org/10.1016/j.ijdrr.2026.106063

76. Ma J., Zhao Y., Cui B., Liu L., Ding Y., Chen Y., Zhang X. Prediction of drought thresholds triggering winter wheat yield losses in the future based on the CNN-LSTM

model and Copula theory: a case study of Henan Province. *Agronomy*. 2025. Vol. 15, No. 4. Art. 954. https://doi.org/10.3390/agronomy15040954

77. Bozorgi M., Cristóbal J., Casadesús J. Assessing the performance of multi–timescale drought indices for monitoring agricultural drought impacts on wheat yield. *Agricultural Water Management*. 2026. Vol. 323. Art. 110092. https://doi.org/10.1016/j.agwat.2025.110092

78. He Y., Chen X., Li H., Li X., Sun S., Wu M. Comprehensive review of detrending methods for crop yields: approaches, applications, and future directions. *Agricultural and Forest Meteorology*. 2026. Vol. 378. Art. 111017. https://doi.org/10.1016/j.agrformet.2026.111017

79. McKinney W. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 2nd ed. Sebastopol : O'Reilly Media, 2017. 524 p.

80. Akcapinar M. C., Apaydin H. Early yield estimation in wheat using open access global datasets and artificial intelligence. *Computers and Electronics in Agriculture*. 2025. Vol. 237, Part C. Art. 110486. DOI: 10.1016/j.compag.2025.110486.

81. Fuentes I., Al–Shammari D. Field–aware and explainable modelling for early-season crop yield prediction using satellite-derived phenology. *Remote Sensing*. 2026. Vol. 18, No. 6. Art. 890. https://doi.org/10.3390/rs18060890

82. Qiao D., Wang T., Xu D. J., Ma R., Feng X., Ruan J. Early-season estimation of winter wheat yield: a hybrid machine learning-enabled approach. *Technological Forecasting and Social Change*. 2024. Vol. 201. Art. 123267. https://doi.org/10.1016/j.techfore.2024.123267

83. Грицюк П. М. Аналіз, моделювання та прогнозування динаміки врожайності озимої пшениці в розрізі областей України : монографія. Рівне : НУВГП, 2010. 350с.

84. Radočaj D., Jurišić M., Plaščak I., Galić L. Satellite remote sensing for crop yield prediction: a review. *Agriculture*. 2026. Vol. 16, No. 4. Art. 417. https://doi.org/10.3390/agriculture16040417

85. Ashourloo D., Brader J., Zhao Y., Jiang R., Zhang M., Chapman S., Hammer G., Potgieter A. B. Machine learning approaches for wheat yield prediction integrating biophysical modeling and remote sensing: effects of sample size, dimensionality, and

transferability. *Smart Agricultural Technology*. 2026. Vol. 13. Art. 101936. https://doi.org/10.1016/j.atech.2026.101936

86. Price K. V., Storn R. M., Lampinen J. A. D*ifferential Evolution: A Practical Approach to Global Optimization*. Berlin ; Heidelberg : Springer, 2005. XIX, 539 p. https://doi.org/10.1007/3-540-31306-0

87. Ruan G., Li X., Yuan F., Cammarano D., Ata-Ul-Karim S. T., Liu X., Tian Y., Zhu Y., Cao W., Cao Q. Improving wheat yield prediction integrating proximal sensing and weather data with machine learning. *Computers and Electronics in Agriculture*. 2022. Vol. 195. Art. 106852. https://doi.org/10.1016/j.compag.2022.106852

88. Li L., Wang B., Feng P., Liu D. L., He Q., Zhang Y., Wang Y., Li S., Lu X., Yue C., Li Y., He J., Feng H., Yang G., Yu Q. Developing machine learning models with multi-source environmental data to predict wheat yield in China. *Computers and Electronics in Agriculture*. 2022. Vol. 194. Art. 106790. https://doi.org/10.1016/j.compag.2022.106790

89. Han J., Zhang Z., Luo Y., Cao J., Zhang L., Zhang J., Li Z., Tao F. Prediction of winter wheat yield based on multi-source data and machine learning in China. *Remote Sensing*. 2020. Vol. 12, No. 2. Art. 236. https://doi.org/10.3390/rs12020236

90. Ashfaq M., Khan I., Afzal R. F., Shah D., Ali S., Tahir M. Enhanced wheat yield prediction through integrated climate and satellite data using advanced AI techniques. *Scientific Reports*. 2025. Vol. 15. Art. 18093. https://doi.org/10.1038/s41598-025-02700-w

91. Zeng X., Han D., Tansey K., Wang P., Pei M., Li Y., Li F., Du Y. An interpretable wheat yield estimation model using time series remote sensing data and considering meteorological and soil influences. *Remote Sensing*. 2025. Vol. 17, No. 18. Art. 3192. https://doi.org/10.3390/rs17183192

92. Naghdyzadegan Jahromi M., Zand-Parsa S., Razzaghi F., Jamshidi S., Didari S., Doosthosseini A., Pourghasemi H. R. Developing machine learning models for wheat yield prediction using ground-based data, satellite–based actual evapotranspiration and vegetation indices. *European Journal of Agronomy*. 2023. Vol. 146. Art. 126820. https://doi.org/10.1016/j.eja.2023.126820

93. Maranhão R. L. A., Caldas M. M., Kastens J., Watson J., Lollato R. P. Assessing NDVI, climate and management to predict winter wheat yields at field scale in Kansas,

USA. *Remote Sensing*. 2025. Vol. 17, No. 20. Art. 3500. https://doi.org/10.3390/rs17203500

94. Copernicus Climate Change Service (C3S). ERA5 post-processed daily statistics on single levels from 1940 to present. *Climate Data Store*. 2024. https://doi.org/10.24381/cds.4991cf48

95. Runfola D., Anderson A., Baier H., Crittenden M., Dowker E., Fuhrig S. et al. geoBoundaries: a global database of political administrative boundaries. PLOS ONE. 2020. Vol. 15, No. 4. Art. e0231866. https://doi.org/10.1371/journal.pone.0231866

96. Khechba K., Belgiu M., Laamrani A., Stein A., Amazirh A., Chehbouni A. The impact of spatiotemporal variability of environmental conditions on wheat yield forecasting using remote sensing data and machine learning. *International Journal of Applied Earth Observation and Geoinformation*. 2025. Vol. 136. Art. 104367. https://doi.org/10.1016/j.jag.2025.104367

97. Alam M. S. B., Esichaikul V., Lameesa A., Ahmed S. F., Gandomi A. H. An approach for crop recommendation with uncertainty quantification based on machine learning for sustainable agricultural decision–making. *Results in Engineering*. 2025. Vol. 26. Art. 105505. https://doi.org/10.1016/j.rineng.2025.105505

98. Zhang H., Zhang Y., Tong F., Li M. A novel winter wheat yield prediction framework using fused spatial-temporal-spectral (STS) information from Sentinel-2 and Landsat 8 via improved Pix2Pix network. *Computers and Electronics in Agriculture*. 2025. Vol. 231. Art. 109982. https://doi.org/10.1016/j.compag.2025.109982

99. Navidi M. N., Fazli E., Kharazmi R., Seyedmohammadi J. Wheat yield prediction using integrated optical and radar remote sensing with machine learning across key phenological stages. *Scientific Reports*. 2026. Vol. 16. Art. 10470. https://doi.org/10.1038/s41598-026-41501-7

100. Cai Y., Guan K., Lobell D., Potgieter A. B., Wang S., Peng J., Xu T., Asseng S., Zhang Y., You L., Peng B. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*. 2019. Vol. 274. P. 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010

101. Amazirh A., Chehbouni A., Kreiner A., Adeniyi O. D., Zeleke G. A., Bouras E., Chakir A., Wheeler J., Barnieh B. A., Khechba K., Zoltan S. Seasonal field-scale wheat yield forecasting using XGBoost with radar, optical, and weather data in Morocco. *International Journal of Applied Earth Observation and Geoinformation*. 2026. Vol. 149. Art. 105242. https://doi.org/10.1016/j.jag.2026.105242

102. Lou Z., Sun D. Yield estimation of winter wheat in the Huang–Huai–Hai region using MODIS and meteorological data: spatio–temporal analysis and county–level modeling. *Frontiers in Plant Science*. 2025. Vol. 16. Art. 1721972. https://doi.org/10.3389/fpls.2025.1721972

103. Raza A., Shahid M. A., Zaman M., Miao Y., Huang Y., Safdar M., Maqbool S., Muhammad N. E. Improving wheat yield prediction with multi–source remote sensing data and machine learning in arid regions. *Remote Sensing*. 2025. Vol. 17, No. 5. Art. 774. https://doi.org/10.3390/rs17050774

104. Ju S., Lim H., Ma J. W., Kim S., Lee K., Zhao S., Heo J. Optimal county-level crop yield prediction using MODIS-based variables and weather data: a comparative study on machine learning models. *Agricultural and Forest Meteorology*. 2021. Vol. 307. Art. 108530. https://doi.org/10.1016/j.agrformet.2021.108530

105. Zhang Z., Luo Y., Han J., Xu J., Tao F. Estimating global wheat yields at 4 km resolution during 1982–2020 by a spatiotemporal transferable method. *Remote Sensing*. 2024. Vol. 16, No. 13. Art. 2342. https://doi.org/10.3390/rs16132342

106. Khodjaev S., Bobojonov I., Kuhn L., Glauben T. Optimizing machine learning models for wheat yield estimation using a comprehensive UAV dataset. *Modeling Earth Systems and Environment*. 2025. Vol. 11. Art. 15. https://doi.org/10.1007/s40808-024-02188-9

107. Yan Y., Li Y., Jia S., Bai Y., Cao B., Mashori A. S., Li F., Zhang W. Integration of UAV multispectral and meteorological data to improve maize yield prediction accuracy. *Agronomy*. 2026. Vol. 16, No. 2. Art. 163. https://doi.org/10.3390/agronomy16020163

108. Cao J., Zhang Z., Luo Y., Zhang L., Zhang J., Li Z., Tao F. Wheat yield predictions at a county and field scale with deep learning, machine learning, and Google Earth

Engine. *European Journal of Agronomy*. 2021. Vol. 123. Art. 126204. https://doi.org/10.1016/j.eja.2020.126204

109. Zhu G., Zhao C., Zhou L., Li Z., Zhu H. Winter wheat yield prediction at a county scale using time series variation features of remote sensing spectra and machine learning. *European Journal of Agronomy*. 2025. Vol. 170. Art. 127751. https://doi.org/10.1016/j.eja.2025.127751

110. Åström O., Månsson S., Lazar I., Nilsson M., Ekelöf J., Oxenstierna A., Sopasakis A. Predicting intra-field yield variations for winter wheat using remote sensing and Graph Attention Networks. *Computers and Electronics in Agriculture*. 2025. Vol. 237, Part A. Art. 110499. https://doi.org/10.1016/j.compag.2025.110499

111. Kešelj K., Stamenković Z., Kostić M., Aćin V., Tekić D., Novaković T., Ivanišević M., Ivezić A., Magazin N. Machine learning (AutoML)-driven wheat yield prediction for European varieties: enhanced accuracy using multispectral UAV data. *Agriculture*. 2025. Vol. 15, No. 14. Art. 1534. https://doi.org/10.3390/agriculture15141534

112. Ishaq R. A. F., Zhou G., Jing G., Shah S. R. A., Ali A., Imran M., Jiang H., Obaid-ur-Rehman. Geospatial robust wheat yield prediction using machine learning and integrated crop growth model and time–series satellite data. *Remote Sensing*. 2025. Vol. 17, No. 7. Art. 1140. https://doi.org/10.3390/rs17071140

113. Xiao G., Niu Q., Du K., Huang H., Guan H., Li X., Huang J. Rapid 10 m winter wheat yield mapping with machine learning and APSIM-Wheat crop model. *Field Crops Research*. 2026. Vol. 343. Art. 110521. https://doi.org/10.1016/j.fcr.2026.110521

114. Ashfaq M., Khan I., Alzahrani A., Tariq M. U., Khan H., Ghani A. Accurate wheat yield prediction using machine learning and climate-NDVI data fusion. *IEEE Access*. 2024. Vol. 12. P. 40947–40961. https://doi.org/10.1109/ACCESS.2024.3376735

115. Mi Q., Huo Z., Li M., Zhang L., Kong R., Zhang F., Wang Y., Huo Y. Development of a drought monitoring system for winter wheat in the Huang-Huai–Hai Region, China, utilizing a machine learning-physical process hybrid model. *Agronomy*. 2025. Vol. 15, No. 3. Art. 696. https://doi.org/10.3390/agronomy15030696

116. Hernández Hernández G. C., Gómez Gómez J., Jiménez-Cabas J. Predictive models based on artificial intelligence to estimate crop yield: a literature review.

Agriculture. 2025. Vol. 15, No. 23. Art. 2438. https://doi.org/10.3390/agriculture15232438

117. Sciencia M. C. M., Creech C. F., Frels K. A., Easterly A. C. Estimating hard winter wheat yield with historical and novel methods. *Agronomy Journal*. 2025. Vol. 117, No. 1. Art. e270021. https://doi.org/10.1002/agj2.70021

118. Shawon S. M., Ema F. B., Mahi A. K., Niha F. L., Zubair H. T. Crop yield prediction using machine learning: an extensive and systematic literature review. *Smart Agricultural Technology*. 2025. Vol. 10. Art. 100718. https://doi.org/10.1016/j.atech.2024.100718

119. Botero–Valencia J., García–Pineda V., Valencia-Arias A., Valencia J., Reyes-Vera E., Mejia-Herrera M., Hernández-García R. Machine learning in sustainable agriculture: systematic review and research perspectives. *Agriculture*. 2025. Vol. 15, No. 4. Art. 377. https://doi.org/10.3390/agriculture15040377

120. Bai H., Xiao D., Tang J., Liu D. L. Evaluation of wheat yield in North China Plain under extreme climate by coupling crop model with machine learning. *Computers and Electronics in Agriculture*. 2024. Vol. 217. Art. 108651. https://doi.org/10.1016/j.compag.2024.108651

121. Anu C. S., Nirmala C. R., Balakrishnan K., Raghavendra L., Biradar S. H., Archana K. N. Interpretable crop yield prediction using stacked regression and explainable AI in precision agriculture. *Proceedings of the Indian National Science Academy*. 2026. https://doi.org/10.1007/s43538-026-00719-9

122. Özden C., Karadoğan N. Wheat yield prediction for Turkey using statistical machine learning and deep learning methods. *Pakistan Journal of Agricultural Sciences*. 2024. Vol. 61, No. 2. P. 429–435. https://doi.org/10.21162/PAKJAS/24.182

123. Fiorentini M., Schillaci C., Denora M., Zenobi S., Deligios P., Orsini R., Santilocchi R., Perniola M., Montanarella L., Ledda L. A machine learning modeling framework for *Triticum turgidum* subsp. *durum* Desf. yield forecasting in Italy. *Agronomy Journal*. 2024. Vol. 116, No. 3. P. 1050–1070. https://doi.org/10.1002/agj2.21279

124. Kalimuthu M., Vaishnavi P., Kishore M. Crop prediction using machine learning. In: *2020 Third International Conference on Smart Systems and Inventive Technology*

(ICSSIT), Tirunelveli, India, 20–22 August 2020. IEEE, 2020. P. 926–932. https://doi.org/10.1109/ICSSIT48917.2020.9214190

125. Veenadhari S., Misra B., Singh C. Machine learning approach for forecasting crop yield based on climatic parameters. In: *2014 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 3–5 January 2014. IEEE, 2014. P. 1–5. https://doi.org/10.1109/ICCCI.2014.6921718

126. Kuradusenge M., Hitimana E., Hanyurwimfura D., Rukundo P., Mtonga K., Mukasine A., Uwitonze C., Ngabonziza J., Uwamahoro A. Crop yield prediction using machine learning models: case of Irish potato and maize. *Agriculture*. 2023. Vol. 13, No. 1. Art. 225. https://doi.org/10.3390/agriculture13010225

127. Feng P., Wang B., Liu D. L., Waters C., Xiao D., Shi L., Yu Q. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agricultural and Forest Meteorology*. 2020. Vol. 285–286. Art. 107922. https://doi.org/10.1016/j.agrformet.2020.107922

128. Hrytsiuk P., Babych T., Hladka O., Nehrey M. Modeling of wheat yield in the steppe region of Ukraine using machine learning techniques. In: Information Control Systems & Technologies : proceedings of the 12th International Conference, ICST 2024, Odesa, Ukraine, September 23–25, 2024. *CEUR Workshop Proceedings*. 2024. Vol. 3790. P. 409–421. URL: https://ceur-ws.org/Vol-3790/paper36.pdf

129. James G., Witten D., Hastie T., Tibshirani R. *An Introduction to Statistical Learning: with Applications in R.* Springer Texts in Statistics. New York : Springer, 2013. 426 p. https://doi.org/10.1007/978-1-4614-7138-7

130. Babenko V., Panchyshyn A., Zomchak L., Nehrey M., Artym–Drohomyretska Z., Lahotskyi T. Classical machine learning methods in economics research: macro and micro level examples. *WSEAS Transactions on Business and Economics*. 2021. Vol. 18. Art. 22. P. 209–217. https://doi.org/10.37394/23207.2021.18.22

131. Köksal D. D., Bennin K. E., Tekinerdoğan B. Centralized and vertical federated learning for climate-based crop yield prediction. Computers and Electronics in Agriculture. 2026. Vol. 249. Art. 111851. https://doi.org/10.1016/j.compag.2026.111851

132. Haseeb M., Tahir Z., Mahmood S. A., Tariq A. Winter wheat yield prediction using linear and nonlinear machine learning algorithms based on climatological and remote sensing data. *Information Processing in Agriculture*. 2025. Vol. 12, No. 4. P. 431–444. https://doi.org/10.1016/j.inpa.2025.02.004

133. Pinto J. G. C. P., Balboa G. R., Mueller N. D., Slater G., Frels K., dos Santos C. L., Miguez F. E., Lollato R. P., Puntel L. A. Evaluation of APSIM next generation for simulating winter wheat growth, yield response to nitrogen, and nitrogen dynamics. *Frontiers in Agronomy*. 2026. Vol. 7. Art. 1740421. https://doi.org/10.3389/fagro.2025.1740421

134. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. *Classification and Regression Trees*. Monterey, CA : Wadsworth, 1984. 368 p.

135. Breiman L. Bagging predictors. *Machine Learning*. 1996. Vol. 24, No. 2. P. 123–140. https://doi.org/10.1007/BF00058655

136. Gharakhanlou N. M., Perez L. From data to harvest: leveraging ensemble machine learning for enhanced crop yield predictions across Canada amidst climate change. Science of The Total Environment. 2024. Vol. 951. Art. 175764. https://doi.org/10.1016/j.scitotenv.2024.175764

137. Friedman J. H. Stochastic gradient boosting. *Computational Statistics & Data Analysis*. 2002. Vol. 38, No. 4. P. 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

138. You J., Li X., Low M., Lobell D., Ermon S. Deep Gaussian process for crop yield prediction based on remote sensing data. Proceedings of the AAAI Conference on Artificial Intelligence. 2017. Vol. 31, No. 1. P. 4559–4566. https://doi.org/10.1609/aaai.v31i1.11172

139. Kuwata K., Shibasaki R. Estimating crop yields with deep learning and remotely sensed data. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015. IEEE, 2015. P. 858–861. https://doi.org/10.1109/IGARSS.2015.7325900

140. Iqbal N., Shahzad M. U., Sherif E.-S. M., Tariq M. U., Rashid J., Le T.-V., Ghani A. Analysis of Wheat-Yield Prediction Using Machine Learning Models under Climate

Change Scenarios. Sustainability. 2024. Vol. 16, No. 16. Art. 6976. https://doi.org/10.3390/su16166976

141. Rufaioglu S. B., Bilgili A. V., Ipekesen S., Ismael A. M., Kaya Y., Matos-Carvalho J. P. CNN-based wheat yield prediction using multi-source and multi-stage data integration from UAV imagery and sensors. Ecological Informatics. 2026. Vol. 93. Art. 103608. https://doi.org/10.1016/j.ecoinf.2026.103608

142. Zidi F. A., Ouafi A., Bougourzi F., Distante C., Taleb-Ahmed A. Advancing wheat crop analysis: a survey of deep learning approaches using hyperspectral imaging. *Computers and Electronics in Agriculture*. 2025. Vol. 238. Art. 110770. https://doi.org/10.1016/j.compag.2025.110770

143. Zhou X., Dang Y., Song J., Xiao Z., Yang H. A deep learning-driven spatio-temporal framework for timely corn yield estimation across multiple remote sensing scenarios. *Remote Sensing*. 2026. Vol. 18, No. 5. Art. 743. https://doi.org/10.3390/rs18050743

144. Barbosa A., Trevisan R., Hovakimyan N., Martin N. F. Modeling yield response to crop management using convolutional neural networks. *Computers and Electronics in Agriculture*. 2020. Vol. 170. Art. 105197. https://doi.org/10.1016/j.compag.2019.105197

145. Taremwa D., Ahishakiye E., Obbo A., Kisozi P. K., Kaggwa F. Prediction of maize yield in Uganda using CNN-LSTM architecture on a multimodal climate and remote sensing dataset. *Discover Artificial Intelligence*. 2026. Vol. 6. Art. 164. https://doi.org/10.1007/s44163-026-00855-7

146. Muruganantham P., Wibowo S., Grandhi S., Samrat N. H., Islam N. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*. 2022. Vol. 14, No. 9. Art. 1990. https://doi.org/10.3390/rs14091990

147. Wang D., Cheng Y., Shi L., Yin H., Yang G., Liu S., Dong Q., Ge J. Prediction of winter wheat yield and interpretable accuracy under different water and nitrogen treatments based on CNNResNet-50. *Agronomy*. 2025. Vol. 15, No. 7. Art. 1755. https://doi.org/10.3390/agronomy15071755

148. Pan S., Liu L. A framework for predicting winter wheat yield in Northern China with triple cross-attention and multi–source data fusion. *Plants*. 2025. Vol. 14, No. 14. Art. 2206. https://doi.org/10.3390/plants14142206

149. Fu H., Lu J., Li J., Zou W., Tang X., Ning X., Sun Y. Winter wheat yield prediction using satellite remote sensing data and deep learning models. *Agronomy*. 2025. Vol. 15, No. 1. Art. 205. https://doi.org/10.3390/agronomy15010205

150. Islam M. Z., Khatun M. M., Chyon M. S. A., Anzoom R. et al. Comparative analysis of explainable machine learning integrated with hyperspectral imaging for early prediction of wheat yield. *Talanta*. 2026. Vol. 308. Art. 129955. https://doi.org/10.1016/j.talanta.2026.129955

151. Zachow M., Ofori–Ampofo S., Kunstmann H., Kuzu R. S., Zhu X. X., Asseng S. Wheat yield forecasts with seasonal climate models and long short-term memory networks. *Computers and Electronics in Agriculture*. 2025. Vol. 239, Part B. Art. 110965. https://doi.org/10.1016/j.compag.2025.110965

152. Ashfaq M., Khan I., Shah D., Ali S., Tahir M. Predicting wheat yield using deep learning and multi-source environmental data. *Scientific Reports*. 2025. Vol. 15. Art. 26446. https://doi.org/10.1038/s41598-025-11780-7

153. Hochreiter S., Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997. Vol. 9, No. 8. P. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

154. Khodadadi N., Towfek S. K., Zaki A. M., Alharbi A. H., Khodadadi E., Khafaga D. S., Abualigah L., Ibrahim A., Abdelhamid A. A., Eid M. M. Predicting normalized difference vegetation index using a deep attention network with bidirectional GRU: a hybrid parametric optimization approach. *International Journal of Data Science and Analytics*. 2025. Vol. 20. P. 3119–3146. https://doi.org/10.1007/s41060-024-00640-8

155. Wan G., Wang K., Zeng J., Li L., Li J., Yan S., Tao Z., Yao D., Xiao Y. Wheat yield prediction using LSTMRFNet: a hybrid model based on time–series data. *Smart Agricultural Technology*. 2026. Vol. 14. Art. 102072. https://doi.org/10.1016/j.atech.2026.102072

156. Joshi A., Pradhan B., Chakraborty S., Varatharajoo R., Alamri A., Gite S., Lee C.-W. An explainable Bi-LSTM model for winter wheat yield prediction. *Frontiers in Plant Science*. 2025. Vol. 15. Art. 1491493. https://doi.org/10.3389/fpls.2024.1491493

157. Chen C. C., McCarl B. A., Schimmelpfennig D. E. Yield variability as influenced by climate: a statistical investigation. Climatic Change. 2004. Vol. 66, No. 1–2. P. 239–261. https://doi.org/10.1023/B:CLIM.0000043159.33816.e5

158. Meng H., Qian L. Performances of different yield-detrending methods in assessing the impacts of agricultural drought and flooding: a case study in the middle-and-lower reach of the Yangtze River, China. *Agricultural Water Management*. 2024. Vol. 296. Art. 108812. https://doi.org/10.1016/j.agwat.2024.108812

159. Li Y., Zeng H., Zhang M., Wu B., Qin X. Global de-trending significantly improves the accuracy of XGBoost–based county-level maize and soybean yield prediction in the Midwestern United States. *GIScience & Remote Sensing*. 2024. Vol. 61, No. 1. Art.

160. Hyndman R. J., Athanasopoulos G. Forecasting: Principles and Practice. 3rd ed. Australia : OTexts, 2021. URL: https://otexts.com/fpp3/ (дата звернення: 20.05.2025).

161. Box G. E. P., Jenkins G. M., Reinsel G. C. *Time Series Analysis: Forecasting and Control*. 4th ed. Hoboken, NJ : John Wiley & Sons, 2008. 746 p. https://doi.org/10.1002/9781118619193

162. Berrar D. Cross-validation. In: Ranganathan S., Gribskov M., Nakai K., Schönbach C. (eds) *Encyclopedia of Bioinformatics and Computational Biology*. Oxford : Academic Press, 2019. Vol. 1. P. 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

163. Hastie T., Tibshirani R., Friedman J. Model assessment and selection. In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. 2nd ed. New York : Springer, 2009. P. 219–259. https://doi.org/10.1007/978-0-387-84858-7_7

164. Petrovic B., Paluba D., Nofrizal A. Y., Kucera A. Analyzing meteorological effects on crop yield and yield prediction through machine learning with different data partitioning approaches: a case study from Czech Republic. *Journal of Agriculture and Food Research*. 2026. Vol. 26. Art. 102662. https://doi.org/10.1016/j.jafr.2026.102662

165. Storn R., Price K. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization*. 1997. Vol. 11. P. 341–359. https://doi.org/10.1023/A:1008202821328

166. Fares A., Valiya Veettil A., Rahman A., Awal R. A review of artificial intelligence applications for predicting crop performance and enhancing food security under drought-induced water stress. *Agricultural Water Management*. 2026. Vol. 325. Art. 110212. https://doi.org/10.1016/j.agwat.2026.110212

167. Вітлінський В. В., Великоіваненко Г. І. Ризикологія в економіці та підприємництві : монографія. Київ : КНЕУ, 2004. 480 с.

168. Державна служба статистики України. Портал офіційної статистики. URL: https://stat.gov.ua/ (дата звернення: 14.05.2025).

169. Ali M. A., Hassan M., Mehmood M., Kazmi D. H., Chishtie F. A., Shahid I. The potential impact of climate extremes on cotton and wheat crops in Southern Punjab, Pakistan. *Sustainability*. 2022. Vol. 14, No. 3. Art. 1609. https://doi.org/10.3390/su14031609

170. Draper N. R., Smith H. *Applied Regression Analysis*. 3rd ed. New York : Wiley, 1998. 736 p. https://doi.org/10.1002/9781118625590

171. Brandt P., Beyer F., Borrmann P., Möller M., Gerighausen H. Ensemble learning–based crop yield estimation: a scalable approach for supporting agricultural statistics. *GIScience & Remote Sensing*. 2024. Vol. 61, No. 1. Art. 2367808. https://doi.org/10.1080/15481603.2024.2367808

172. Li H.–Y., Lawrence J. A., Mason P. J., Ghail R. C. A framework for accurate annual regional crop yield prediction. *Remote Sensing*. 2026. Vol. 18, No. 8. Art. 1157. https://doi.org/10.3390/rs18081157

173. Everitt B. S., Landau S., Leese M., Stahl D. Cluster Analysis. 5th ed. *Chichester : Wiley*, 2011. 346 p.

174. Kozubowski T. J., Podgorski K. A Multivariate and Asymmetric Generalization of Laplace Distribution. *Computational Statistics*. 2000. Vol. 15. No. 4. P. 531–540. https://doi.org/10.1007/PL00022717

175. Greenwood P. E., Nikulin M. S. *A Guide to Chi-Squared Testing*. Wiley Series in Probability and Statistics. New York : John Wiley & Sons, 1996. 304 p.

176. Dowd K. Measuring Market Risk. 2nd ed. Chichester : John Wiley & Sons, 2005. 416 p. https://doi.org/10.1002/9781118673485

177. Uryasev S. Conditional Value-at-Risk: optimization algorithms and applications. In: *IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFEr).* New York : IEEE, 2000. P. 49–57. https://doi.org/10.1109/CIFER.2000.844598

178. Nehrey M., Klymenko N., Kravchenko V., Komar M. Ukrainian agriculture during the full–scale Russian–Ukrainian war: consequences, policy responses and recovery strategies. *Agricultural and Resource Economics: International Scientific E-Journal*. 2025. Vol. 11, No. 2. P. 148–182. https://doi.org/10.51599/are.2025.11.02.06

179. Srivastava A. K., Safaei N., Khaki S., Lopez G., Zeng W., Ewert F., Gaiser T., Rahimi J. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. Scientific Reports. 2022. Vol. 12. Art. 3215. https://doi.org/10.1038/s41598-022-06249-w

180. Géron A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2nd ed. Sebastopol : O'Reilly Media, 2019. 848 p.

ДОДАТКИ

Додаток А. Довідки про впровадження

АГРОТРЕЙД-РІВНЕ

Товариство з обмеженою відповідальністю "Агротрейд-Рівне"
код ЄДРПОУ 44757420
Адреса: м.Рівне., вул.Данила Галицького 19
р/р UA453510050000026000879144215 в АТ «УКРСИББАНК»
ІПН 447574217160
agrotrade2022@ukr.net

Від 18 травня 2026 року
Вихідний № 85/26

ДОВІДКА
про практичне використання результатів дисертаційної роботи Гаврилюка Максима Сергійовича на тему «Інтелектуальні моделі прогнозування врожайності на основі методів машинного навчання»

Довідка видана Гаврилюку М. С. про те, що результати його дисертаційного дослідження використано у практичній та аналітичній діяльності ТзОВ «Агротрейд-Рівне» для інформаційної підтримки оцінювання агрокліматичних умов, прогнозування врожайності пшениці та виявлення ризиків низької врожайності.

Застосовані підходи утворюють прикладний аналітичний ланцюг: підготовка кліматичних і виробничих даних, модельний прогноз, оцінювання ризику, сценарне порівняння та візуалізація висновків для прийняття управлінських рішень.

У діяльності «Агротрейд-Рівне» використано:

– підхід до оброблення агрокліматичних даних для Рівненської та Хмельницької областей України та аналіз їх впливу врожайності пшениці;

– моделі прогнозування врожайності на основі методів машинного навчання для порівняння очікуваних результатів за різних температурних і вологозабезпечувальних умов;

– методику оцінювання ризику зниження врожайності та визначення критичних агрометеорологічних умов у період квітень–червень;

– сценарний інструментарій інформаційно-аналітичної системи «Пшениця України» для порівняння базових і змінених умов, формування карт, графіків і звітів.

Використання результатів дисертаційного дослідження Гаврилюка М. С. сприяє підвищенню обґрунтованості агровиробничих і організаційних рішень, своєчасному виявленню несприятливих сценаріїв та поліпшенню інформаційного забезпечення планування діяльності підприємства.

Директор ТзОВ «Агротрейд-Рівне» Залужний М. М.

ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ
«МЕТИЗ КАПІТАЛ»

47751,Тернопільська обл.., с.Підгородне, вул. Стрийська, 10Б
Телефон (факс) /0352/43-07-70
р/р UA113253650000002600302435078 у Центральна філія ПАТ «Кредобанк» МФО 325365
ЄДРПОУ 37246348 ІПН 372463419180

Від 11 травня 2026 року
Вих № 147

ДОВІДКА
про впровадження результатів дисертаційної роботи
Гаврилюка Максима Сергійовича на тему
«Інтелектуальні моделі прогнозування врожайності на основі методів машинного навчання»

Результати, отримані в дисертаційній роботі Гаврилюка М. С., є важливими для аналітичного забезпечення процесів оцінювання стану зерновиробництва, прогнозування врожайності пшениці та врахування агрокліматичних ризиків у процесі підготовки управлінських рішень. Даною довідкою засвідчується, що у діяльність ТОВ «Метиз Капітал» впроваджено такі результати дисертаційного дослідження:

– методичний підхід до формування регіонально агрегованих агрокліматичних даних для аналізу врожайності пшениці;

– підходи до агрокліматичної кластеризації областей України та формування узагальнених регіональних вибірок для задач прогнозування;

– моделі прогнозування врожайності пшениці на основі методів машинного навчання, придатні для оцінювання можливих сценаріїв розвитку зерновиробництва;

– методичні засади оцінювання ризику низької врожайності та аналітичного врахування наслідків зовнішніх шоків для зернового виробництва;

– результати проектування та апробації інформаційно-аналітичної системи «Пшениця України» як інструменту підтримки аналітичної та прогнозної діяльності.

Впровадження зазначених результатів дисертаційної роботи Гаврилюка М. С. сприяє підвищенню якості аналітичного супроводу управлінських рішень, покращенню інформаційного забезпечення прогнозування врожайності пшениці, а також більш повному врахуванню агрокліматичних ризиків у практичній діяльності установи.

Директор ТОВ «Метиз Капітал» **Данильчук Т.В.**

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ВОДНОГО ГОСПОДАРСТВА ТА ПРИРОДОКОРИСТУВАННЯ
(НУВГП)
вул. Соборна, 11, м. Рівне, 33028, тел. (0362) 63 30 98, факс (0362) 63 32 09,
e-mail: mail@nuwm.edu.ua, web: https://nuwm.edu.ua
код ЄДРПОУ 02071116

Від 10.04.2026 № 011-11 На № _____ від _____

ДОВІДКА
про використання у навчальному процесі
Національного університету водного господарства та природокористування
результатів досліджень і розробок, одержаних при виконанні дисертаційної роботи
ГАВРИЛЮКА МАКСИМА СЕРГІЙОВИЧА
на здобуття ступеня доктора філософії
зі спеціальності 122 «Комп'ютерні науки»

Використання у навчальному процесі науково-методичних розробок та результатів досліджень здобувача вищої освіти третього (освітньо-наукового) рівня Гаврилюка М. С., що викладені в його дисертації, присвяченій інформаційній технології прогнозування врожайності пшениці на основі методів машинного навчання, сприяло оновленню змісту навчальних занять, посиленню прикладної складової підготовки здобувачів вищої освіти та впровадженню практико-орієнтованих прикладів застосування методів машинного навчання. У навчальному процесі Національного університету водного господарства та природокористування під час викладання наступних курсів спеціальності F6 «Інформаційні системи і технології» апробовано та впроваджено такі рекомендації:

- під час викладання навчальної дисципліни «Системний аналіз» для здобувачів першого (бакалаврського) рівня, які навчаються за ОПП «Інформаційні системи і технології», запроваджено приклади формалізації задач прогнозування врожайності, визначення факторів впливу, структурування вхідних даних і побудови моделей предметної області, що сприяє розвитку навичок системного аналізу прикладних процесів. Особливу увагу приділено встановленню взаємозв'язків між агрокліматичними показниками, виробничими умовами та результативними ознаками врожайності пшениці. Це дало змогу посилити практичну складову навчальної дисципліни та сформувати у здобувачів розуміння етапів постановки й обґрунтування задач аналізу складних соціально-економічних і природно-виробничих систем;

- під час викладання навчальної дисципліни «Методи машинного навчання» для здобувачів вищої освіти другого (магістерського) рівня, які навчаються за ОПП «Інформаційні технології в бізнесі», запроваджено практичні завдання з підготовки агрокліматичних і статистичних даних, побудови регресійних, класифікаційних і нейромережевих моделей, оцінювання точності прогнозування та визначення ризиків зниження врожайності. У межах практичної підготовки здобувачі опрацьовують приклади формування навчальних і тестових вибірок, вибору факторних ознак, інтерпретації результатів моделювання та порівняння якості різних алгоритмів машинного навчання. Це сприяє формуванню навичок застосування інтелектуальних методів аналізу даних для розв'язання прикладних задач прогнозування та підтримки прийняття рішень.

Проректор з наукової роботи
та міжнародних зв'язків НУВГП
доктор економічних наук, професор Наталія САВІНА

Петро ГРИЦЮК +380985935153

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ВОДНОГО ГОСПОДАРСТВА ТА ПРИРОДОКОРИСТУВАННЯ

вул. Соборна, 11, м. Рівне, 33028, тел. (0362) 63-30-98, факс (0362) 63-32-09, mail@nuwm.edu.ua

Від 17.04.26 № 16н-26

На № _____ від _____

ДОВІДКА

про використання досліджень, отриманих при виконанні дисертаційної роботи
Гаврилюку Максиму Сергійовичу
на здобуття наукового ступеня доктора філософії за спеціальністю
122 – Комп'ютерні науки
в науково-дослідних роботах, виконаних у науково-дослідній частині
Національного університету водного господарства та природокористування

Видана **Гаврилюку Максиму Сергійовичу** з підтвердженням того, що результати досліджень, викладених в дисертаційній роботі, яка присвячена розробці інтелектуальних моделей прогнозування врожайності на основі методів машинного навчання, використовувались при виконанні кафедральної науково-дослідної теми в науково-дослідній частині згідно з планами науково-дослідних робіт кафедри комп'ютерних технологій та економічної кібернетики Національного університету водного господарства та природокористування «Інформаційні технології моделювання екологічних, економічних та соціальних процесів» (Державний реєстраційний номер: 0121U111686, 2021-2026 рр.).

У процесі виконання науково-дослідної роботи автором теоретично обґрунтовано метод формування нелінійних ознак кумулятивного кліматичного впливу на врожайність на основі синтезу ознак другого порядку, який дає змогу формалізувати неадитивні ефекти взаємодії кліматичних чинників у суміжних фазах вегетації; удосконалено методичний підхід до побудови прогнозних моделей врожайності шляхом агрокліматичної кластеризації областей України та регіональної агрегації даних, що дало змогу збільшити обсяг навчальної вибірки, зменшити вплив регіональної неоднорідності та підвищити стійкість моделей. На основі отриманих результатів автором розроблено інформаційну технологію прогнозування врожайності, яка забезпечує автоматизацію підготовки даних, формування ознак, побудови прогнозних моделей, оцінювання ризиків, оптимізації параметрів вегетації, візуалізації результатів, сценарного аналізу та формування аналітичних звітів.

Проректор з наукової роботи
та міжнародних зв'язків,
д.е.н., професор Наталія САВІНА

Сергій Куницький 0967375013

Додаток Б. Список публікацій здобувача за темою дисертації

Публікації у виданнях, проіндексованих у наукометричній базі Scopus

1. Hrytsiuk P., Babych T., Baranovsky S., Havryliuk M. Assessing of climate impact on wheat yield using machine learning techniques. In: *Information Control Systems & Technologies* : proceedings of the 11th International Conference, ICST-2023, Odesa, Ukraine, September 21–23, 2023. *CEUR Workshop Proceedings*. 2023. Vol. 3513. P. 314–329. URL: https://ceur-ws.org/Vol-3513/paper26.pdf. **Scopus** (1,32/0,33 д.а.; *авторський внесок здобувача — підготовка та попередня обробка даних врожайності й кліматичних показників, формування експериментальної вибірки, реалізація моделей машинного навчання, проведення числових експериментів, аналіз отриманих результатів та підготовка графічних матеріалів*).

2. Hrytsiuk P., Havryliuk M. Modeling of the nonlinear impact of climatic factors on wheat yield using machine learning techniques. In: Ermolayev V. et al. (eds) *Information and Communication Technologies in Education, Research, and Industrial Applications* : 19th International Conference, ICTERI 2024, Lviv, Ukraine, September 23–27, 2024, Proceedings. Communications in Computer and Information Science. Cham : Springer, 2025. Vol. 2359. P. 20–35. https://doi.org/10.1007/978-3-031-81372-6_2. **Scopus** (1,06/0,53 д.а.; *авторський внесок здобувача — формування набору кліматичних факторів для моделювання врожайності пшениці, реалізація нелінійних моделей машинного навчання, проведення обчислювальних експериментів, оцінювання якості моделей, інтерпретація впливу кліматичних змінних на результат прогнозування*).

3. Hrytsiuk P., Havryliuk M., Nehrey M. Constructing the optimal temperature trajectory for the wheat vegetative cycle. In: Ermolayev V. et al. (eds) *Information and Communication Technologies in Education, Research, and Industrial Applications* : 20th International Conference, ICTERI 2025, Nice, France, September 1–4, 2025, Proceedings. Communications in Computer and Information Science. Cham : Springer, 2025. Vol. 2763. P. 101–116. https://doi.org/10.1007/978-3-032-10477-9_8. **Scopus** (0,91/0,3 д.а.; *авторський внесок здобувача — підготовка вихідних даних, участь у*

формалізації задачі побудови оптимальної температурної траєкторії вегетаційного циклу пшениці, виконання числових розрахунків, аналіз модельних результатів та підготовка частини ілюстративного матеріалу).

4. Hrytsiuk P., Babych T., Kardash O., Havryliuk M. Application of machine learning for wheat yield prediction. In: Potapov I. et al. (eds) *Digitalisation and Digital Transformation* : First Research Twinning Conference, RTC-Digital 2023, Liverpool, UK, March 27–30, 2023, Proceedings. Communications in Computer and Information Science. Cham : Springer, 2026. Vol. 2647. P. 3–11. https://doi.org/10.1007/978-3-032-04731-1_1. **Scopus** (0,6/0,15 д.а.; *авторський внесок здобувача — підготовка даних для прогнозування врожайності пшениці, реалізація та тестування моделей машинного навчання, участь у порівнянні результатів прогнозування, підготовка експериментальної частини дослідження*).

Статті у наукових фахових виданнях України категорії «Б»

5. Грицюк П. М., Гаврилюк М. С. Моделювання впливу кліматичних факторів на врожайність пшениці методами машинного навчання. *Штучний інтелект*. 2025. Т. 30, № 1. С. 121–130. https://doi.org/10.15407/jai2025.01.121 (0,92/0,46 д.а.; *авторський внесок здобувача — побудова інформаційної бази дослідження, підготовка кліматичних і врожайних показників, реалізація моделей машинного навчання, проведення обчислювальних експериментів, оцінювання точності моделей та інтерпретація одержаних результатів*).

6. Грицюк П. М., Гаврилюк М. С., Джоші О. І. Ідентифікація закону розподілу трендових залишків врожайності як інструмент моделювання ризиків зерновиробництва. *Штучний інтелект*. 2025. Т. 30, № 2. С. 72–83. https://doi.org/10.15407/jai2025.02.072 (1,02/0,34 д.а.; *авторський внесок здобувача — розрахунок трендових залишків врожайності, підготовка даних для статистичного аналізу, участь в ідентифікації закону розподілу залишків, виконання обчислювальних експериментів, інтерпретація результатів для задач моделювання ризиків зерновиробництва*).

Наукові праці, які засвідчують апробацію матеріалів дисертації

7. Грицюк П. М., Гаврилюк М. С. Класифікаційний підхід до прогнозування врожайності. *Форум молодих економістів-кібернетиків «Моделювання економіки: проблеми, тенденції, досвід»* : тези доповідей X Всеукраїнської науково-практичної конференції, м. Львів, 25 листопада 2022 р. Львів, 2022. С. 47–49. (0,15/0,07 д.а.)

8. Havryliuk M., Hrytsiuk P., Kardash O., Babych T. Forecasting of wheat yield using machine learning techniques. *Digital Theme UK–Ukraine Research Twinning Conference* : UA-DIGITAL 2023, online, March 27–31, 2023. (0,59/0,15 д.а.)

9. Havryliuk M. S. Wheat yield forecasting using machine learning methods. *Проблеми та перспективи розвитку сучасної науки* : збірник тез доповідей Міжнародної науково-практичної конференції молодих науковців, аспірантів і здобувачів вищої освіти, м. Рівне, 11–12 травня 2023 р. Рівне : НУВГП, 2023. С. 139–142. (0,17/0,17 д.а.)

10. Грицюк П. М., Гаврилюк М. С. Оцінка втрат врожаю пшениці внаслідок військової агресії росії. *Інтегровані інтелектуальні робототехнічні комплекси (ІІРТК-2023)* : матеріали XVI Міжнародної науково-практичної конференції, м. Київ, 23–24 травня 2023 р. Київ : НАУ, 2023. С. 374–376. (0,18/0,09 д.а.)

11. Hrytsiuk P., Babych T., Baranovskii S., Havryliuk M. Assessing of climate impact on wheat yield using machine learning techniques. *Information Control Systems and Technologies (ICST-2023)* : materials of the XI International Scientific-Practical Conference, Odesa, Ukraine, September 21–23, 2023. P. 102–105. (0,16/0,04 д.а.)

12. Грицюк П. М., Гаврилюк М. С. Використання машинного навчання для управління процесами аграрної економіки. *Міжнародна конференція з інноваційних рішень у програмній інженерії (ICISSE 2023)*, м. Івано-Франківськ, 29–30 листопада 2023 р. Івано-Франківськ, 2023. С. 150–154. https://doi.org/10.5281/zenodo.10397356 (0,43/0,21 д.а.)

13. Гаврилюк М. С. Використання машинного навчання для управління процесами аграрної економіки. *Науково-практична конференція студентів та аспірантів навчально-наукового інституту кібернетики, інформаційних*

технологій та інженерії, м. Рівне, 23 травня 2024 р. Рівне : НУВГП, 2024. С. 17. (0,08/0,08 д.а.)

14. Гаврилюк М. С., Грицюк П. М. Ідентифікація закону розподілу залишків врожайності сільськогосподарських культур. *Комп'ютерне моделювання та програмне забезпечення інформаційних систем і технологій 2024 (КМПЗ 2024)* : матеріали IV Міжнародної науково-практичної конференції, м. Чернівці, 30 травня – 1 червня 2024 р. Чернівці, 2024. С. 76–80. (0,2/0,1 д.а.)

15. Грицюк П. М., Гаврилюк М. С. Оцінювання нелінійного впливу кліматичних факторів на врожайність пшениці. *Штучний інтелект. Наука. Бізнес* : матеріали Всеукраїнської науково-практичної конференції, м. Одеса, 22 жовтня 2024 р. Одеса, 2024. С. 18–21. (0,19/0,1 д.а.)

16. Гаврилюк М. С., Грицюк П. М. Дослідження впливу кліматичних факторів на коливання врожайності пшениці. *Форум молодих економістів-кібернетиків. Моделювання економіки: проблеми, тенденції, досвід* : матеріали XII Всеукраїнської науково-практичної конференції, м. Львів, 22–23 листопада 2024 р. Львів, 2024. С. 11–14. (0,21/0,1 д.а.)

17. Havryliuk M., Hrytsiuk P. Modeling the nonlinear influence of climatic factors on wheat yield using machine learning methods. *Advances in Science, Technology, and Society* : proceedings of the International Scientific Conference, Oxford, United Kingdom, February 20, 2025. P. 210–214. (0,28/0,14 д.а.)

18. Грицюк П. М., Бабич Т. Ю., Гаврилюк М. С. Побудова оптимальної температурної траєкторії вегетації пшениці. *Математичні методи, моделі та інформаційні технології в економіці* : матеріали IX Міжнародної науково-методичної конференції, м. Чернівці, 24–25 квітня 2025 р. Чернівці, 2025. С. 69–71. (0,1/0,03 д.а.)

19. Гаврилюк М. С. Моделювання нелінійного впливу кліматичних факторів на врожайність пшениці методами машинного навчання. *Науково-практична конференція студентів та аспірантів навчально-наукового інституту кібернетики, інформаційних технологій та інженерії*, м. Рівне, 14 травня 2025 р. Рівне : НУВГП, 2025. С. 49. (0,07/0,07 д.а.)

20. Гаврилюк М. С., Грицюк П. М. Використання закону розподілу трендових залишків врожайності для моделювання ризиків виробництва зерна. *Моделювання, керування та інформаційні технології* : матеріали VIII Міжнародної науково-практичної конференції, м. Рівне, 6–8 листопада 2025 р. Рівне : НУВГП, 2025. С. 286–288. https://doi.org/10.31713/MCIT.2025.089 (0,26/0,13 д.а.)

21. Гаврилюк М. С. Нелінійне моделювання впливу кліматичних факторів на врожайність пшениці в Україні. *Форум молодих економістів-кібернетиків. Моделювання економіки: проблеми, тенденції, досвід* : матеріали XIII Всеукраїнської науково-практичної конференції, м. Львів, 21–22 листопада 2025 р. Львів, 2025. С. 101–102. (0,1/0,1 д.а.)

Праці, які додатково відображають наукові результати дисертації

22. Hrytsiuk P., Babych T., Hladka O., Nehrey M., Havryliuk M. Machine learning-based modeling of climate effects on wheat yield in the Ukrainian Steppe. *Journal of Lifestyle and SDGs Review*. 2026. Vol. 6, No. 1. Art. e08120. https://doi.org/10.47172/2965-730X.SDGsReview.v6.n01.pe08120. (0,88/0,18 д.а.; *авторський внесок здобувача — формування вибірки для степової зони України, підготовка кліматичних ознак, реалізація обчислювальних експериментів із використанням методів машинного навчання, аналіз впливу кліматичних факторів на врожайність пшениці, підготовка таблиць і графіків*).

23. Свідоцтво про реєстрацію авторського права на твір № 127859. Комп’ютерна програма «Математичне моделювання впливу кліматичних факторів на врожайність методами машинного навчання». Автори: Гаврилюк М. С., Грицюк П. М., Бабич Т. Ю., Кардаш О. Л. Дата реєстрації 26 червня 2024 р.

Додаток В. Свідоцтво про реєстрацію авторського права на твір

УКРАЇНА

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 127859

Комп'ютерна програма «Математичне моделювання впливу кліматичних факторів на врожайність методами машинного навчання» («Клімат – врожайність»)

(вид, назва твору)

Автор (співавтори) **Гаврилюк Максим Сергійович, Грицюк Петро Михайлович, Бабич Тетяна Юріївна, Кардаш Оксана Любомирівна**

(прізвище, ім'я, по батькові (за наявності), псевдонім (за наявності))

Авторські майнові права належать повністю **Національний університет водного господарства та природокористування, вул. Соборна, 11, м. Рівне, 33028**

(прізвище, ім'я, по батькові (за наявності) фізичної особи / найменування юридичної особи, адреса)

Дата реєстрації 26 червня 2024 р.

Директор Державної організації «Український національний офіс інтелектуальної власності та інновацій» **Олена ОРЛЮК**

УКРНОІВІ

Додаток Г. Інтерфейсні форми та результати роботи інформаційно-аналітичної системи «Пшениця України»

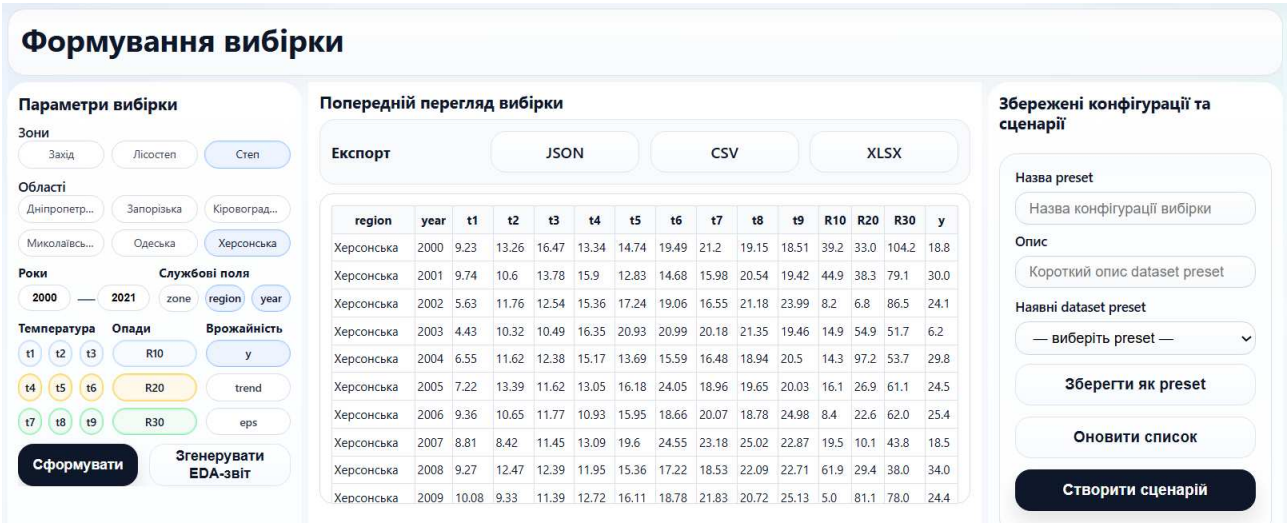

Рис. Г.1. Формування агрокліматичної вибірки та збереження конфігурації вибірки даних

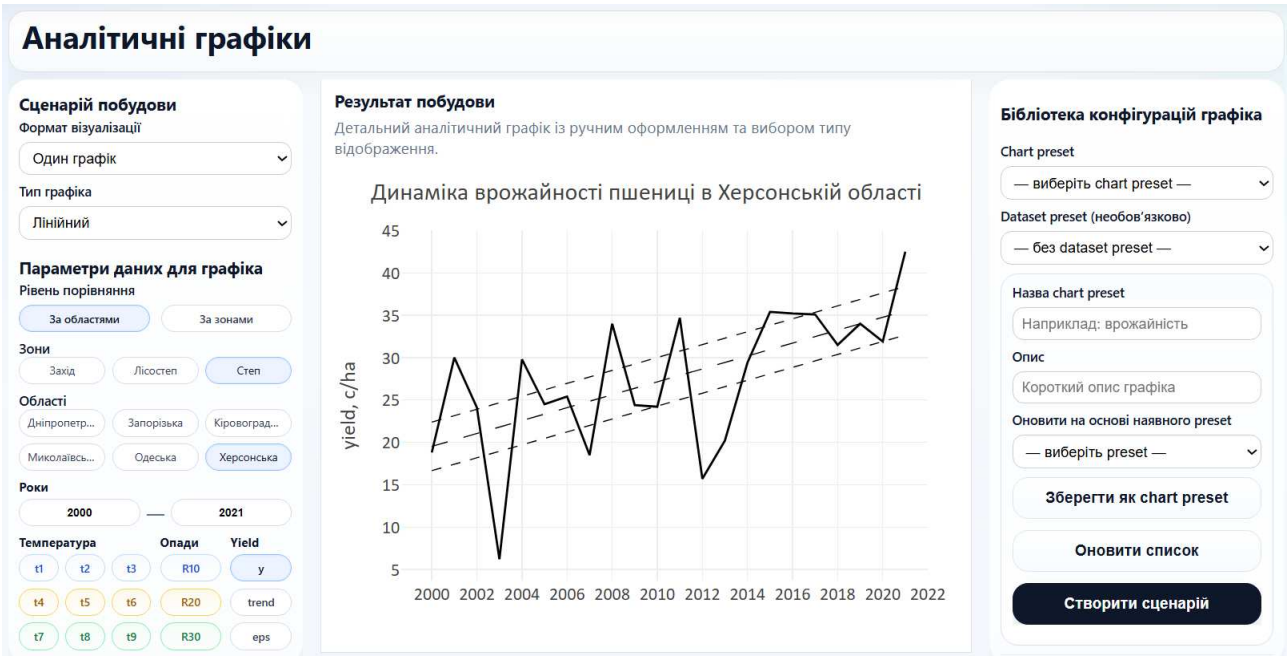

Рис. Г.2. Побудова аналітичного графіка на основі сформованої вибірки та збереження конфігурації графічного подання

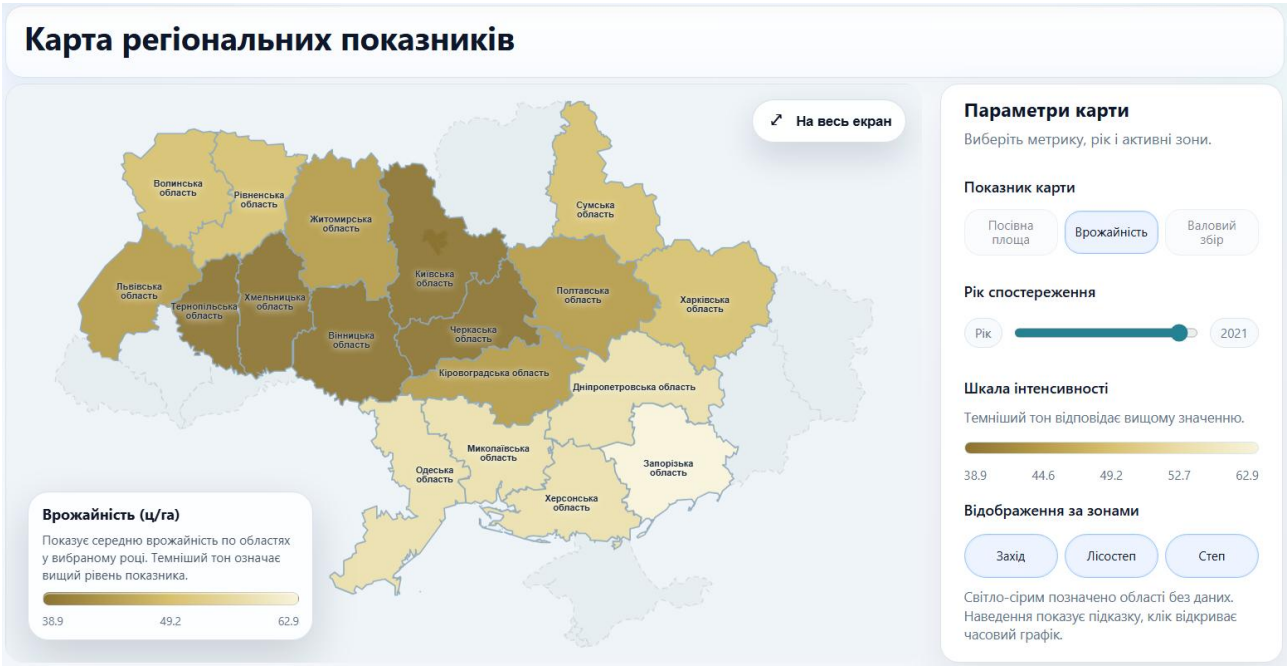

Рис. Г.3. Картографічне подання показників у розрізі областей України

Рис. Г.4. Результат прогнозування врожайності в ІАС

Рис. Г.5. Класифікація ризикових років у прогнозно-аналітичному модулі системи

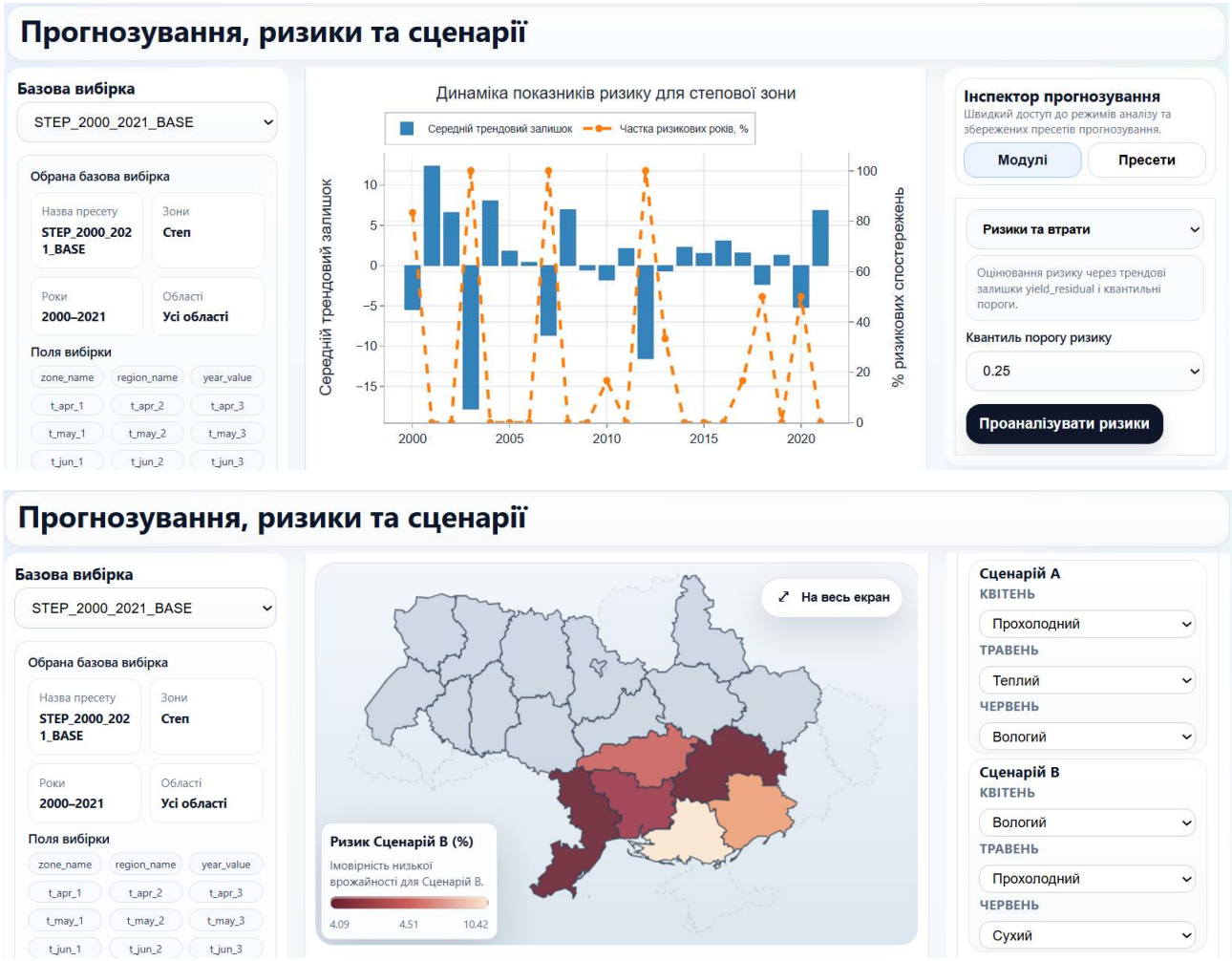

Рис. Г.6. Оцінювання ризику низької врожайності та просторове подання ризикових областей на прикладі степової зони України

Рис. Г.7. Просторове подання прогнозованої врожайності для одного з альтернативних сценаріїв на прикладі степової зони України

Рис. Г.8. Порівняння альтернативних сценаріїв прогнозування на прикладі степової зони України

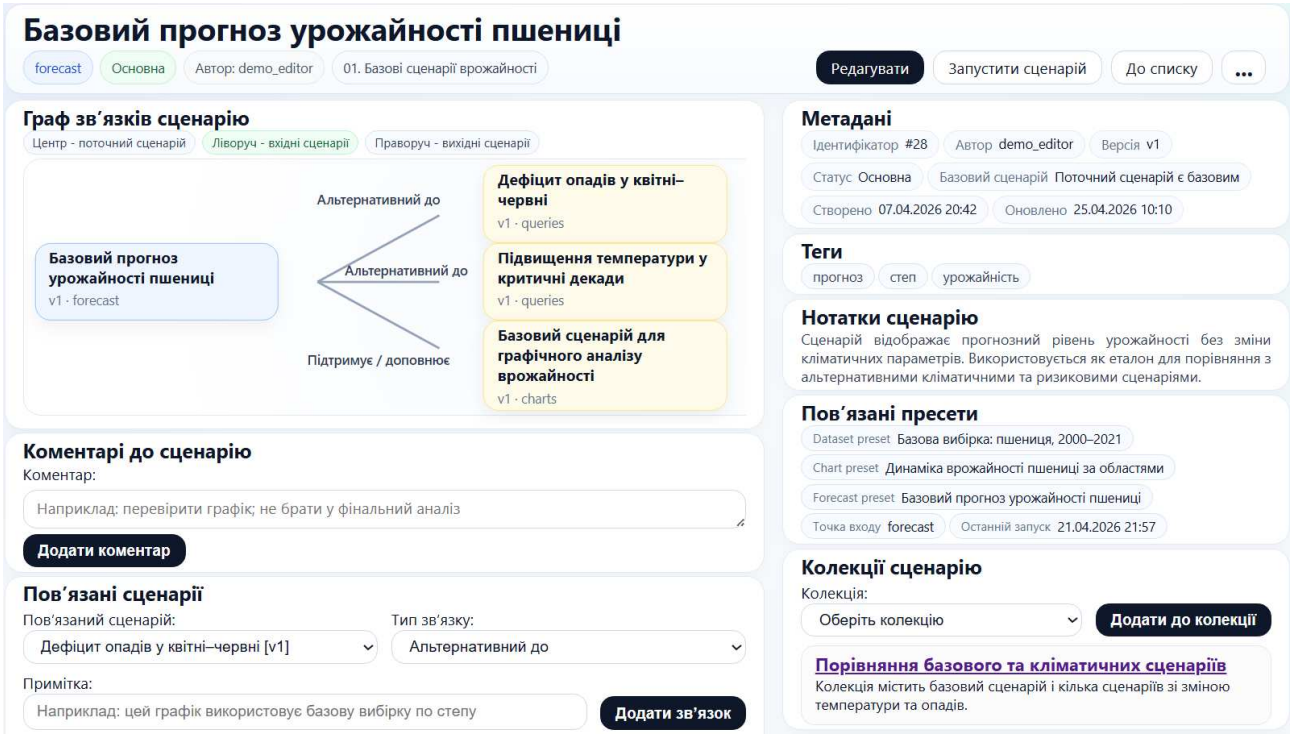

Рис. Г.9. Сторінка окремого дослідницького сценарію в ІАС «Пшениця України»

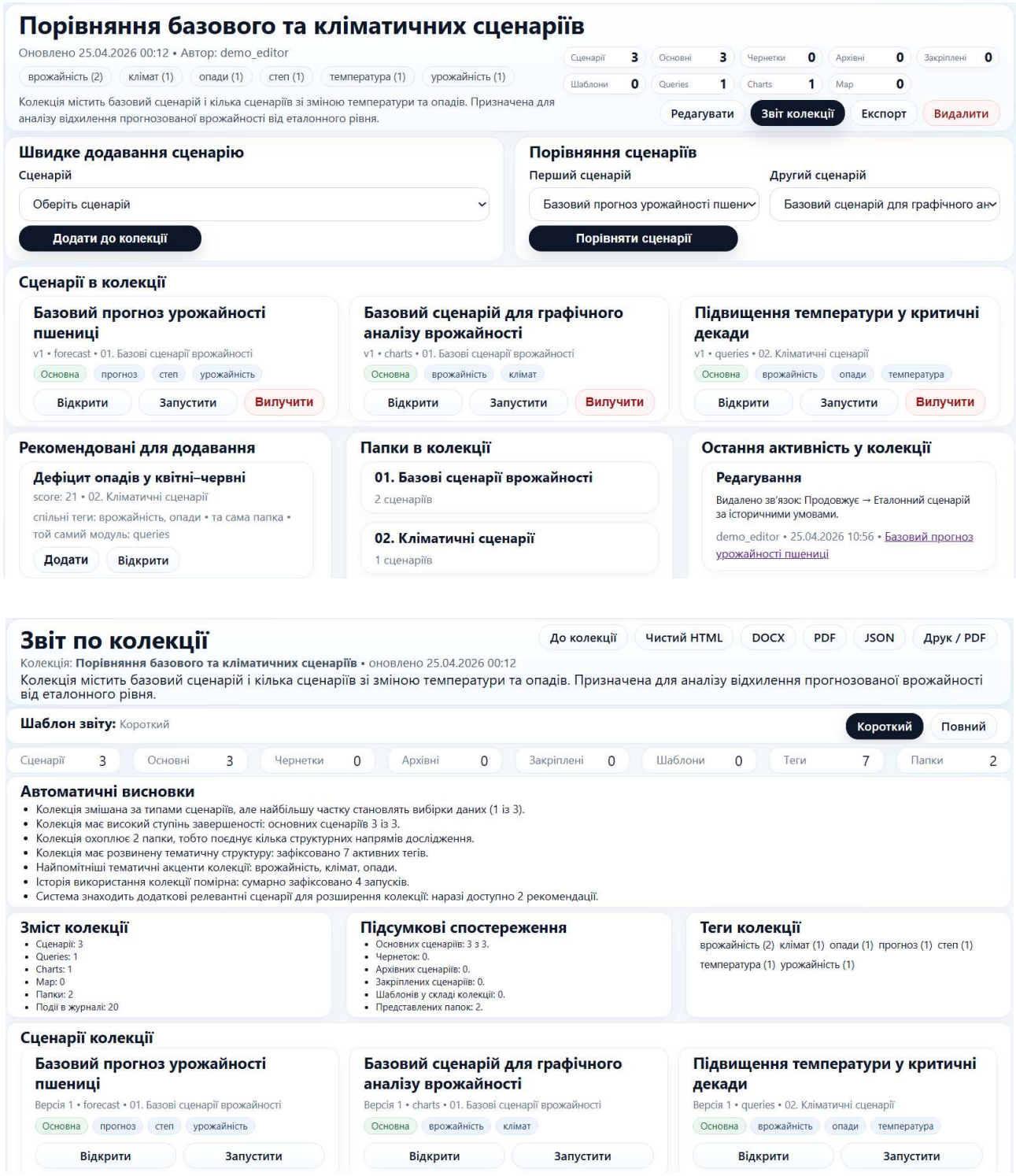

Рис. Г.10. Сторінка колекції сценаріїв і формування звітних матеріалів

Додаток Д. Програмні матеріали та відтворювані обчислювальні експерименти

У додатку наведено програмні матеріали, що забезпечують відтворення основних етапів дисертаційного дослідження: формування агрокліматичної матриці ознак, детрендування часових рядів урожайності, формування цільових класів, виконання обчислювальних експериментів з рекурентними нейронними мережами, а також оцінювання нижньохвостового ризику та сценарних змін агрокліматичних факторів.

Наведені лістинги мають допоміжний характер і доповнюють опис алгоритмів, моделей та обчислювальних процедур, поданий в основній частині дисертації. Для зручності сприйняття у додатку подано укрупнені фрагменти програмної реалізації без службових файлів середовища виконання, ключів доступу, кешованих проміжних файлів та допоміжних налаштувань, які не впливають на наукову логіку обчислень.

Д.1. Формування агрокліматичної матриці ознак

Метою програмного блоку є одержання фінального набору даних виду «область–рік–y–t1–t9–R10, R20, R30». Для цього статистичні дані врожайності поєднуються з просторово агрегованими кліматичними показниками ERA5, розрахованими в межах полігонів областей України рівня ADM1.

Лістинг Д.1 – Завантаження вихідних даних і підготовка просторового узгодження

```
# Лістинг Д.1. Формування агрокліматичної матриці ознак
# Призначення: просторово-часове узгодження статистики врожайності,
# даних ERA5 та цифрових меж областей України рівня ADM1.

import os
import zipfile
from pathlib import Path

import numpy as np
import pandas as pd
import geopandas as gpd
import xarray as xr
import rioxarray
from shapely.geometry import mapping

# За потреби у Google Colab монтується Google Drive:
try:
    from google.colab import drive
    drive.mount('/content/drive')
except Exception:
    pass

BASE_DIR = Path('/content/drive/MyDrive/wheat_yield_project')
DATA_DIR = BASE_DIR / 'data'
```

```
RAW_DIR = DATA_DIR / 'raw'
OUT_DIR = DATA_DIR / 'processed'
OUT_DIR.mkdir(parents=True, exist_ok=True)

YIELD_FILE = RAW_DIR / 'yield_regions_2000_2021.xlsx'
ADM1_FILE = RAW_DIR / 'geoBoundaries-UKR-ADM1.geojson'
ERA5_TMAX_FILE = RAW_DIR / 'era5_tmax_daily_2000_2021.nc'
ERA5_TMIN_FILE = RAW_DIR / 'era5_tmin_daily_2000_2021.nc'
ERA5_PREC_FILE = RAW_DIR / 'era5_prec_daily_2000_2021.nc'

YEARS = range(2000, 2022)
TARGET_REGIONS = [
    'Вінницька', 'Волинська', 'Дніпропетровська', 'Житомирська',
    'Запорізька', 'Київська', 'Кіровоградська', 'Львівська',
    'Миколаївська', 'Одеська', 'Полтавська', 'Рівненська',
    'Сумська', 'Тернопільська', 'Харківська', 'Херсонська',
    'Хмельницька', 'Черкаська'
]

REGION_NAME_MAP = {
    'Vinnytsia': 'Вінницька', 'Volyn': 'Волинська',
    'Dnipropetrovsk': 'Дніпропетровська', 'Zhytomyr': 'Житомирська',
    'Zaporizhia': 'Запорізька', 'Kyiv': 'Київська',
    'Kirovohrad': 'Кіровоградська', 'Lviv': 'Львівська',
    'Mykolaiv': 'Миколаївська', 'Odesa': 'Одеська',
    'Poltava': 'Полтавська', 'Rivne': 'Рівненська',
    'Sumy': 'Сумська', 'Ternopil': 'Тернопільська',
    'Kharkiv': 'Харківська', 'Kherson': 'Херсонська',
    'Khmelnytskyi': 'Хмельницька', 'Cherkasy': 'Черкаська'
}

# 1. Статистичні дані врожайності
# Очікувана структура: region, year, yield_value або відповідні українські назви.
yield_df = pd.read_excel(YIELD_FILE)
yield_df = yield_df.rename(columns={
    'Область': 'region', 'Рік': 'year', 'Врожайність': 'yield_value',
    'region_name': 'region', 'year_value': 'year', 'y': 'yield_value'
})
yield_df['region'] = yield_df['region'].astype(str).str.replace(' область', '', regex=False).str.strip()
yield_df = yield_df[yield_df['region'].isin(TARGET_REGIONS)]
yield_df = yield_df[yield_df['year'].isin(YEARS)]

# 2. Межі областей ADM1
regions_gdf = gpd.read_file(ADM1_FILE).to_crs('EPSG:4326')
name_col = 'shapeName' if 'shapeName' in regions_gdf.columns else 'shapeName_en'
regions_gdf['region'] = regions_gdf[name_col].replace(REGION_NAME_MAP)
regions_gdf = regions_gdf[regions_gdf['region'].isin(TARGET_REGIONS)]
regions_gdf = regions_gdf[['region', 'geometry']].copy()

# 3. Відкриття ERA5-масивів і приведення до CRS
# Температура ERA5 часто подається у K, тому далі виконується перехід до C.
tmax = xr.open_dataset(ERA5_TMAX_FILE)
tmin = xr.open_dataset(ERA5_TMIN_FILE)
prec = xr.open_dataset(ERA5_PREC_FILE)

# Назви змінних залежать від способу експорту з CDS.
TMAX_VAR = 'mx2t' if 'mx2t' in tmax.data_vars else list(tmax.data_vars)[0]
TMIN_VAR = 'mn2t' if 'mn2t' in tmin.data_vars else list(tmin.data_vars)[0]
PREC_VAR = 'tp' if 'tp' in prec.data_vars else list(prec.data_vars)[0]

def prepare_da(ds, var_name):
    da = ds[var_name]
    if 'longitude' in da.dims:
        da = da.rename({'longitude': 'x', 'latitude': 'y'})
    elif 'lon' in da.dims:
        da = da.rename({'lon': 'x', 'lat': 'y'})
    da = da.rio.write_crs('EPSG:4326')
    return da

tmax_da = prepare_da(tmax, TMAX_VAR)
tmin_da = prepare_da(tmin, TMIN_VAR)
prec_da = prepare_da(prec, PREC_VAR)
```

```
# 4. Допоміжні функції просторового усереднення
# Для кожної області вирізається кліматичне поле в межах полігона,
# після чого обчислюється середнє по всіх пікселях.
def clip_mean_by_region(da, geom):
    clipped = da.rio.clip([mapping(geom)], crs='EPSG:4326', drop=True)
    dims = [d for d in clipped.dims if d not in ('time', 'valid_time')]
    return clipped.mean(dim=dims, skipna=True)

def to_celsius(da):
    return da - 273.15 if float(da.mean()) > 100 else da

def precip_to_mm(da):
    # У ERA5 total precipitation зазвичай подається у метрах.
    return da * 1000.0 if float(da.mean()) < 10 else da
```

Лістинг Д.2 – Формування агрокліматичних предикторів t1-t9, R10-R30

```
# Лістинг Д.2. Агрегування добових кліматичних даних у предиктори t1-t9, R10-R30
# Призначення: перехід від добових просторових полів до регіонального набору
# ознак виду region-year-y-t1-...-t9-R10-R20-R30.

DECADS = [
    ('t1', 4,  1, 10), ('t2', 4, 11, 20), ('t3', 4, 21, 30),
    ('t4', 5,  1, 10), ('t5', 5, 11, 20), ('t6', 5, 21, 31),
    ('t7', 6,  1, 10), ('t8', 6, 11, 20), ('t9', 6, 21, 30),
]
PREC_MONTHS = [('R10', 4), ('R20', 5), ('R30', 6)]

rows = []

for _, reg in regions_gdf.iterrows():
    region_name = reg['region']
    geom = reg.geometry

    # Просторове усереднення добових рядів у межах області.
    reg_tmax = to_celsius(clip_mean_by_region(tmax_da, geom))
    reg_tmin = to_celsius(clip_mean_by_region(tmin_da, geom))
    reg_prec = precip_to_mm(clip_mean_by_region(prec_da, geom))

    # Середньодобова температура як середнє Tmax і Tmin.
    reg_temp = (reg_tmax + reg_tmin) / 2.0
    time_dim = 'time' if 'time' in reg_temp.dims else 'valid_time'

    for year in YEARS:
        record = {'region': region_name, 'year': year}

        # Температурні ознаки t1-t9: середня температура за декади квітня-червня.
        for feature, month, day_from, day_to in DECADS:
            start = f'{year}-{month:02d}-{day_from:02d}'
            end = f'{year}-{month:02d}-{day_to:02d}'
            subset = reg_temp.sel({time_dim: slice(start, end)})
            record[feature] = float(subset.mean(skipna=True).values)

        # Опадові ознаки R10, R20, R30: місячні суми опадів за квітень-червень.
        for feature, month in PREC_MONTHS:
            start = f'{year}-{month:02d}-01'
            end = f'{year}-{month:02d}-30' if month in (4, 6) else f'{year}-{month:02d}-31'
            subset = reg_prec.sel({time_dim: slice(start, end)})
            record[feature] = float(subset.sum(skipna=True).values)

        rows.append(record)

climate_df = pd.DataFrame(rows)
feature_df = yield_df.merge(climate_df, on=['region', 'year'], how='inner')
feature_df = feature_df.sort_values(['region', 'year']).reset_index(drop=True)

columns = ['region', 'year', 'yield_value'] + [f't{i}' for i in range(1, 10)] + ['R10', 'R20', 'R30']
feature_df = feature_df[columns]

# Контроль повноти: 18 областей x 22 роки = 396 спостережень.
expected_n = len(TARGET_REGIONS) * len(list(YEARS))
print('Кількість спостережень:', len(feature_df), 'очікувано:', expected_n)
```

```
print(feature_df.isna().sum())

feature_df.to_csv(OUT_DIR / 'agroclimatic_feature_matrix.csv', index=False, encoding='utf-8-sig')
feature_df.to_excel(OUT_DIR / 'agroclimatic_feature_matrix.xlsx', index=False)

display(feature_df.head())
```

Д.2. Детрендування часових рядів урожайності та формування цільових класів

Для зменшення впливу довгострокового технологічного тренду фактичну врожайність подано через трендові залишки. На основі знака залишку формується бінарний клас: 1 — низька врожайність, 0 — висока або нормальна врожайність.

Лістинг Д.3 – Детрендування та підготовка набору даних для класифікаційних моделей

```
# Лістинг Д.3. Детрендування врожайності та формування цільових класів
# Призначення: вилучення довгострокового технологічного тренду та перехід
# до залишків eps, що використовуються у класифікаційних моделях.

import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

feature_df = pd.read_csv('agroclimatic_feature_matrix.csv', encoding='utf-8-sig')

def add_linear_trend_residuals(group):
    group = group.sort_values('year').copy()
    X = group[['year']].to_numpy()
    y = group['yield_value'].to_numpy()

    trend_model = LinearRegression()
    trend_model.fit(X, y)

    group['yield_trend'] = trend_model.predict(X)
    group['eps'] = group['yield_value'] - group['yield_trend']
    group['yield_class'] = np.where(group['eps'] <= 0, 1, 0)
    group['yield_class_label'] = np.where(group['yield_class'] == 1, 'low_yield', 'high_yield')
    return group

model_df = (
    feature_df
    .groupby('region', group_keys=False)
    .apply(add_linear_trend_residuals)
    .reset_index(drop=True)
)

# Ознаки, які використовуються в моделях машинного навчання.
feature_cols = [f't{i}' for i in range(1, 10)] + ['R10', 'R20', 'R30']
model_cols = ['region', 'year', 'yield_value', 'yield_trend', 'eps',
              'yield_class', 'yield_class_label'] + feature_cols
model_df = model_df[model_cols]

model_df.to_csv('model_dataset_with_residuals.csv', index=False, encoding='utf-8-sig')
model_df.to_excel('model_dataset_with_residuals.xlsx', index=False)

print(model_df.groupby('region')['yield_class'].value_counts().unstack(fill_value=0))
                                        display(model_df.head())
```

Д.3. Обчислювальні експерименти з GRU-моделлю

Цей програмний блок використовується для відтворюваного підбору параметрів рекурентної нейронної мережі GRU. У межах експерименту змінюються ширина часового вікна n_steps, кількість нейронів GRU та кількість епох навчання. Найкраща конфігурація визначається за F1-score з урахуванням $Accuracy$, $Precision$ та $Recall$.

Лістинг Д.4 – Добір параметрів GRU-моделі та розрахунок класифікаційних метрик

```
# Лістинг Д.4. GRU-експерименти для класифікаційного прогнозування
# Призначення: добір параметрів n_steps, units, epochs і вибір конфігурації
# за F1-score, Accuracy, Precision та Recall.

import os
import time
import random
import warnings
warnings.filterwarnings('ignore')

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.preprocessing import MaxAbsScaler
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix

import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Input, GRU, Dense

SEED = 42

def reset_seeds(seed=SEED):
    random.seed(seed)
    np.random.seed(seed)
    tf.random.set_seed(seed)
    tf.keras.backend.clear_session()
    try:
        tf.config.experimental.enable_op_determinism()
    except Exception:
        pass

reset_seeds()

# Вхідний ряд: трендові залишки eps1 для Херсонської області.
file = 'HersonEps6.csv'
data = pd.read_csv(file, sep=',', decimal='.', encoding='utf-8-sig')
data = data.sort_values('Year').reset_index(drop=True)

df = data[['Year', 'eps1']].copy()
df = df.rename(columns={'eps1': 'eps'})

def create_yield_class_dataset(series_scaled, series_original, years, n_steps):
    X, y, y_value, target_years = [], [], [], []
    for i in range(len(series_scaled) - n_steps):
        X.append(series_scaled[i:i + n_steps])
        next_eps = series_original[i + n_steps]
        y.append(1 if next_eps <= 0 else 0)      # 1 = low_yield
        y_value.append(next_eps)
        target_years.append(years[i + n_steps])
    return np.array(X), np.array(y), np.array(y_value), np.array(target_years)
```

```
def train_test_split_time_order(X, y, y_value, years, test_ratio=0.2):
    n_test = int(np.ceil(len(X) * test_ratio))
    n_train = len(X) - n_test
    return (
        X[:n_train], X[n_train:],
        y[:n_train], y[n_train:],
        y_value[:n_train], y_value[n_train:],
        years[:n_train], years[n_train:]
    )

def build_gru_model(n_steps, units):
    model = Sequential([
        Input(shape=(n_steps, 1)),
        GRU(units),
        Dense(1, activation='sigmoid')
    ])
    model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
    return model

def run_experiment(n_steps, units, epochs, seed=SEED, verbose=0):
    reset_seeds(seed)
    scaler = MaxAbsScaler()
    eps_scaled = scaler.fit_transform(df[['eps']]).astype('float32')

    X, y, y_value, years = create_yield_class_dataset(
        eps_scaled, df['eps'].to_numpy(), df['Year'].to_numpy(), n_steps
    )

    X_train, X_test, y_train, y_test, yv_train, yv_test, years_train, years_test = \
        train_test_split_time_order(X, y, y_value, years, test_ratio=0.2)

    model = build_gru_model(n_steps=n_steps, units=units)
    t0 = time.time()
    history = model.fit(X_train, y_train, epochs=epochs, batch_size=8, verbose=verbose)
    time_sec = time.time() - t0

    prob = model.predict(X_test, verbose=0).ravel()
    y_pred = (prob >= 0.5).astype(int)

    tn, fp, fn, tp = confusion_matrix(y_test, y_pred, labels=[0, 1]).ravel()
    result = {
        'n_steps': n_steps,
        'units': units,
        'epochs': epochs,
        'train_size': len(y_train),
        'test_size': len(y_test),
        'test_year_start': int(years_test[0]),
        'test_year_end': int(years_test[-1]),
        'Accuracy': accuracy_score(y_test, y_pred),
        'Precision': precision_score(y_test, y_pred, zero_division=0),
        'Recall': recall_score(y_test, y_pred, zero_division=0),
        'F1_score': f1_score(y_test, y_pred, zero_division=0),
        'TN': int(tn), 'FP': int(fp), 'FN': int(fn), 'TP': int(tp),
        'final_loss': float(history.history['loss'][-1]),
        'final_accuracy': float(history.history['accuracy'][-1]),
        'time_sec': time_sec,
        'seed': seed
    }
    test_results = pd.DataFrame({
        'Year': years_test,
        'eps_true': yv_test,
        'yield_class_true': y_test,
        'prob_low_yield': prob,
        'yield_class_pred': y_pred
    })
    return result, history.history, test_results

cache = {}
def get_result(n_steps, units, epochs, step_name):
    key = (n_steps, units, epochs, SEED)
    if key not in cache:
        result, history, test_results = run_experiment(n_steps, units, epochs)
        cache[key] = {'result': result, 'history': history, 'test_results': test_results}
```

```
    row = cache[key]['result'].copy()
    row['Step'] = step_name
    return row

# Крок 1: фіксуємо units=256, epochs=250; змінюємо ширину вікна n_steps.
step1_rows = []
for n_steps in [3, 4, 5, 6]:
    step1_rows.append(get_result(n_steps=n_steps, units=256, epochs=250, step_name='Step1_nsteps'))
step1_df = pd.DataFrame(step1_rows)

# Крок 2: фіксуємо n_steps=4, epochs=250; змінюємо кількість нейронів GRU.
step2_rows = []
for units in [16, 32, 64, 128, 256, 512]:
    step2_rows.append(get_result(n_steps=4, units=units, epochs=250, step_name='Step2_units'))
step2_df = pd.DataFrame(step2_rows)

# Крок 3: фіксуємо units=256; змінюємо epochs для n_steps=3 і n_steps=4.
epochs_list = [100, 150, 200, 250, 259, 300, 400, 500, 600, 800, 1000]
step3_rows = []
for n_steps in [3, 4]:
    for epochs in epochs_list:
        step3_rows.append(get_result(n_steps=n_steps, units=256, epochs=epochs,
step_name='Step3_epochs'))
step3_df = pd.DataFrame(step3_rows)

# Вибір найкращої конфігурації за F1-score, далі за Accuracy, Precision і Recall.
all_results_df = pd.concat([step1_df, step2_df, step3_df], ignore_index=True)
all_results_df = all_results_df.drop_duplicates(
    subset=['n_steps', 'units', 'epochs', 'seed']
).reset_index(drop=True)

best_df = all_results_df.sort_values(
    by=['F1_score', 'Accuracy', 'Precision', 'Recall'],
    ascending=[False, False, False, False]
).reset_index(drop=True)

best = best_df.iloc[0]
best_key = (int(best['n_steps']), int(best['units']), int(best['epochs']), int(best['seed']))
print('Найкраща конфігурація:', best_key)
print(best[['F1_score', 'Accuracy', 'Precision', 'Recall']])

all_results_df.to_csv('gru_experiment_results.csv', index=False, encoding='utf-8-sig')
best_df.head(10).to_excel('gru_best_configurations.xlsx', index=False)
```

Д.4. Оцінювання нижньохвостового ризику та сценарний аналіз

Ризиковий блок призначений для оцінювання несприятливих відхилень урожайності від тренду. Для цього розраховуються нижні квантилі, VaR, CVaR та частка від’ємних залишків. Сценарний блок дає змогу оцінювати зміну прогнозної компоненти за заданими змінами температурних і опадових ознак.

Лістинг Д.5 – Розрахунок показників ризику та сценарних змін

```
# Лістинг Д.5. Оцінювання нижньохвостового ризику та сценарних змін
# Призначення: розрахунок VaR, CVaR, квантилів трендових залишків і
# сценарне оцінювання зміни очікуваної врожайності за кліматичними ознаками.

import numpy as np
import pandas as pd

model_df = pd.read_csv('model_dataset_with_residuals.csv', encoding='utf-8-sig')

def lower_tail_risk(group, alpha=0.10):
    eps = group['eps'].dropna().to_numpy()
    var_alpha = np.quantile(eps, alpha)
    cvar_alpha = eps[eps <= var_alpha].mean() if np.any(eps <= var_alpha) else np.nan
```

```
    return pd.Series({
        'n': len(eps),
        'q10': np.quantile(eps, 0.10),
        'q25': np.quantile(eps, 0.25),
        'median': np.quantile(eps, 0.50),
        'VaR_0_10': var_alpha,
        'CVaR_0_10': cvar_alpha,
        'negative_share': float(np.mean(eps < 0))
    })

risk_by_region = model_df.groupby('region').apply(lower_tail_risk).reset_index()
risk_by_region.to_excel('risk_indicators_by_region.xlsx', index=False)

# Приклад сценарної функції для зональної лінійної моделі.
# Коефіцієнти зберігаються в окремій таблиці, а зміни ознак задаються сценарієм.
zone_coefficients = {
    'Степова': {'const': 32.173, 't5': -1.1833, 't7': -0.6660, 'R10': 0.0320},
    'Лісостепова': {'const': -18.394, 't3': 1.1670, 't4': 0.9540, 't7': -0.5860},
    'Західна': {'const': -8.045, 't3': 0.9000, 't5': 0.7870, 'R30': -0.0450}
}

def predict_eps_from_coefficients(row, coefs):
    value = coefs.get('const', 0.0)
    for feature, beta in coefs.items():
        if feature == 'const':
            continue
        value += beta * row.get(feature, 0.0)
    return value

def apply_climate_scenario(row, scenario):
    row = row.copy()
    for feature, delta in scenario.items():
        if feature in row.index:
            row[feature] = row[feature] + delta
    return row

# Приклад: підвищення температури другої декади травня на 1 C і зменшення
# опадів квітня на 10 мм для степової зони.
scenario = {'t5': 1.0, 'R10': -10.0}

steppe_df = model_df[model_df['zone'] == 'Степова'].copy() if 'zone' in model_df.columns else
model_df.copy()
coefs = zone_coefficients['Степова']

scenario_rows = []
for _, row in steppe_df.iterrows():
    base_eps = predict_eps_from_coefficients(row, coefs)
    scenario_row = apply_climate_scenario(row, scenario)
    scenario_eps = predict_eps_from_coefficients(scenario_row, coefs)
    scenario_rows.append({
        'region': row['region'],
        'year': row['year'],
        'base_eps_pred': base_eps,
        'scenario_eps_pred': scenario_eps,
        'delta_eps': scenario_eps - base_eps
    })

scenario_df = pd.DataFrame(scenario_rows)
scenario_df.to_excel('scenario_analysis_steppe.xlsx', index=False)

print(risk_by_region.head())
print(scenario_df.head())
```

Додаток Е. Фрагменти програмної реалізації інформаційно-аналітичної системи «Пшениця України»

У додатку наведено фрагменти серверної реалізації інформаційно-аналітичної системи «Пшениця України», що демонструють організацію аналітико-модельного контуру, інтеграцію сервісів прогнозування, сценарного аналізу, ризикових розрахунків, формування інтерактивних візуалізацій та експорту звітів.

Фрагменти коду подано вибірково, оскільки повна реалізація інформаційної системи містить службові модулі, шаблони інтерфейсу, налаштування доступу, маршрутизацію, статичні файли та інші компоненти, що не є необхідними для розкриття наукової логіки роботи системи.

Е.1. Підключення серверних компонентів та аналітичних сервісів

Серверний модуль IAC організовано навколо моделей предметної області, користувацьких сценаріїв, графічних конфігурацій, сервісів прогнозування та інструментів експорту. Така структура забезпечує зв’язок між базою даних, інтерфейсом користувача та аналітичними процедурами.

Лістинг Е.1 – Підключення моделей даних, бібліотек візуалізації та сервісів прогнозування

```
# Лістинг Е.1. Фрагмент серверного модуля IAC «Пшениця України»
# Призначення: підключення моделей даних, механізмів візуалізації,
# сервісів прогнозування, сценарного аналізу та експорту звітів.

import json, io, re, math, csv, time, statistics
import pandas as pd
from io import BytesIO
from pathlib import Path
from statistics import NormalDist
from datetime import timedelta

from django.conf import settings
from django.http import JsonResponse, HttpResponse, HttpResponseBadRequest, HttpResponseForbidden
from django.shortcuts import render, redirect, get_object_or_404
from django.contrib import messages
from django.urls import reverse
from django.contrib.auth.decorators import login_required
from django.db.models import Q, Min, Max, Count
from django.views.decorators.http import require_POST, require_GET
from django.template.loader import render_to_string
from django.utils import timezone
from django.utils.http import urlencode
from django.db import connection

from plotly.graph_objs import Figure, Scatter, Bar, Heatmap
from plotly.io import to_html

from .models import (
    Zone, Region, YieldData, WeatherMonthly, WeatherDecadal,
    MapRegionAlias, ChartStyleConfig, ChartSeriesStyle, SavedPreset,
```

```
    SqlQueryLog, ResearchScenario, ResearchScenarioFolder,
    ResearchScenarioTag, ResearchScenarioEvent, ResearchScenarioComment,
    ResearchScenarioRelation, ResearchScenarioCollection,
)

from .services.forecast_service import (
    run_regression_analysis,
    run_classification_analysis,
    run_risk_analysis,
    run_scenario_analysis,
    build_forecast_report,
    build_forecast_map_payload,
    build_forecast_region_detail,
    build_dual_scenario_comparison,
    build_dual_scenario_map_payload,
)
```

Е.2. Формування результатів регресійного прогнозування

Наведений фрагмент показує логіку виклику сервісу регресійного прогнозування, обробку можливих помилок, побудову інтерактивного графіка та повернення HTML-фрагмента до користувацького інтерфейсу.

Лістинг Е.2 – Серверне формування візуального результату регресійного прогнозу

```
# Лістинг Е.2. Побудова HTML-представлення регресійного прогнозу
# Призначення: виклик аналітичного сервісу, побудова Plotly-графіка
# та повернення результату в інтерфейс IAC.

def _forecast_plot_legend():
    return {
        'orientation': 'h',
        'x': 0,
        'xanchor': 'left',
        'y': 1.02,
        'yanchor': 'bottom',
    }

def _render_forecast_regression_html(request, scope, model_key='linear'):
    model_key = (model_key or 'linear').strip()
    if model_key not in ('linear', 'rf'):
        model_key = 'linear'

    try:
        result = run_regression_analysis(
            zone_id=scope['zone_id'],
            region_ids=scope['region_ids'],
            year_from=scope['year_from'],
            year_to=scope['year_to'],
            model_key=model_key,
            test_ratio=0.25,
        )
    except ValueError as e:
        return render_to_string('wheatapp/_forecast_error.html', {'msg': str(e)}, request=request)
    except Exception as e:
        return render_to_string(
            'wheatapp/_forecast_error.html',
            {'msg': f'Помилка під час обчислення прогнозу: {e}'},
            request=request,
        )

    chart = result['agg_chart']

    fig = Figure()
    fig.add_trace(Scatter(
```

```
        x=chart['x'], y=chart['actual'], mode='lines+markers',
        name='Фактична врожайність'
    ))
    fig.add_trace(Scatter(
        x=chart['x'], y=chart['predicted'], mode='lines+markers',
        name='Прогноз моделі'
    ))
    fig.update_layout(
        template='simple_white',
        title=f"Фактична та прогнозована врожайність за роками ({result['model_label']})",
        xaxis_title='Рік',
        yaxis_title='Врожайність',
        hovermode='x unified',
        title_x=0.5,
        margin=dict(l=40, r=20, t=85, b=40),
        legend=_forecast_plot_legend(),
    )

    plot_html = to_html(fig, include_plotlyjs=False, full_html=False,
                        config={'displayModeBar': False})
    return render_to_string(
        'wheatapp/_forecast_regression.html',
        {'plot_html': plot_html, 'result': result, 'scope': scope},
        request=request,
    )
```

Е.3. Сценарний аналіз у веборієнтованому середовищі

Фрагмент демонструє приймання параметрів сценарію з форми користувача, перевірку вхідних даних, запуск сценарного розрахунку та підготовку результату у вигляді інтерактивної порівняльної діаграми.

Лістинг Е.3 – Обробка сценарного запиту та формування порівняльного результату

```
# Лістинг Е.3. Обробка сценарного прогнозу та ризикового аналізу в ІАС
# Призначення: формування обчислювального запиту, запуск аналітичного сервісу
# і підготовка структурованого результату для інтерфейсу.

@workspace_required
@require_POST
def forecast_scenario_panel(request):
    scope = _forecast_scope_from_request(request)
    if not scope['region_ids']:
        return render(request, 'wheatapp/_forecast_error.html', {
            'msg': 'Оберіть хоча б одну область для сценарного аналізу.'
        })

    scenario = {
        't3': _float_param(request.POST.get('delta_t3'), 0.0),
        't4': _float_param(request.POST.get('delta_t4'), 0.0),
        't5': _float_param(request.POST.get('delta_t5'), 0.0),
        't7': _float_param(request.POST.get('delta_t7'), 0.0),
        'R10': _float_param(request.POST.get('delta_R10'), 0.0),
        'R30': _float_param(request.POST.get('delta_R30'), 0.0),
    }
    scenario = {k: v for k, v in scenario.items() if abs(v) > 1e-12}

    try:
        result = run_scenario_analysis(
            zone_id=scope['zone_id'],
            region_ids=scope['region_ids'],
            year_from=scope['year_from'],
            year_to=scope['year_to'],
            scenario=scenario,
        )
```

```
    except ValueError as e:
        return render(request, 'wheatapp/_forecast_error.html', {'msg': str(e)})

    fig = Figure()
    fig.add_trace(Bar(
        x=result['regions'],
        y=result['base_prediction'],
        name='Базовий прогноз'
    ))
    fig.add_trace(Bar(
        x=result['regions'],
        y=result['scenario_prediction'],
        name='Сценарний прогноз'
    ))
    fig.update_layout(
        template='simple_white',
        barmode='group',
        title='Порівняння базового та сценарного прогнозу',
        xaxis_title='Область',
        yaxis_title='Врожайність',
        margin=dict(l=40, r=20, t=70, b=80),
    )

    plot_html = to_html(fig, include_plotlyjs=False, full_html=False,
                        config={'displayModeBar': False})
    return render(request, 'wheatapp/_forecast_scenario.html', {
        'plot_html': plot_html,
        'result': result,
        'scenario': scenario,
        'scope': scope,
   })
```

Е.4. Експорт результатів прогнозування та сценарного аналізу

Експорт звітів забезпечує документування результатів роботи системи, збереження підсумкових таблиць і подальше використання отриманих результатів у наукових, навчальних або прикладних матеріалах.

Лістинг Е.4 – Формування звітних документів DOCX/PDF

```
# Лістинг Е.4. Узагальнена схема експорту результатів аналітичного сценарію
# Призначення: документування результатів моделювання у форматах DOCX/PDF.

from docx import Document
from docx.shared import Pt
from reportlab.lib.pagesizes import A4
from reportlab.platypus import SimpleDocTemplate, Paragraph, Spacer, Table, TableStyle
from reportlab.lib.styles import getSampleStyleSheet
from reportlab.lib import colors

def export_forecast_report_docx(result, filename='forecast_report.docx'):
    document = Document()
    document.add_heading('Звіт за результатами прогнозного сценарію', level=1)

    p = document.add_paragraph()
    p.add_run('Модель: ').bold = True
    p.add_run(str(result.get('model_label', '')))

    p = document.add_paragraph()
    p.add_run('Період аналізу: ').bold = True
    p.add_run(f"{result.get('year_from')} - {result.get('year_to')}")

    table = document.add_table(rows=1, cols=4)
    hdr = table.rows[0].cells
    hdr[0].text = 'Область'
    hdr[1].text = 'Фактичне значення'
    hdr[2].text = 'Прогноз'
```

```
    hdr[3].text = 'Похибка'

    for row in result.get('rows', []):
        cells = table.add_row().cells
        cells[0].text = str(row.get('region', ''))
        cells[1].text = f"{row.get('actual', 0):.3f}"
        cells[2].text = f"{row.get('predicted', 0):.3f}"
        cells[3].text = f"{row.get('error', 0):.3f}"

    document.add_paragraph('Звіт сформовано автоматизовано засобами ІАС «Пшениця України».')
    document.save(filename)
    return filename

def export_forecast_report_pdf(result, filename='forecast_report.pdf'):
    doc = SimpleDocTemplate(filename, pagesize=A4)
    styles = getSampleStyleSheet()
    flow = []
    flow.append(Paragraph('Звіт за результатами прогнозного сценарію', styles['Title']))
    flow.append(Spacer(1, 12))

    data = [['Область', 'Факт', 'Прогноз', 'Похибка']]
    for row in result.get('rows', []):
        data.append([
            row.get('region', ''),
            f"{row.get('actual', 0):.3f}",
            f"{row.get('predicted', 0):.3f}",
            f"{row.get('error', 0):.3f}",
        ])

    table = Table(data)
    table.setStyle(TableStyle([
        ('GRID', (0, 0), (-1, -1), 0.5, colors.black),
        ('BACKGROUND', (0, 0), (-1, 0), colors.lightgrey),
        ('ALIGN', (1, 1), (-1, -1), 'CENTER'),
    ]))
    flow.append(table)
    doc.build(flow)
    return filename
```