

ISSN 1998-7927(print) ISSN 2664-6498 (online)

DOI: <https://doi.org/10.33216/1998-7927-2026-304-6-88-101>

УДК 004.8:519.246.8:620.9

ПРОГНОЗУВАННЯ ГАЗОСПОЖИВАННЯ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ: MLP, ДЕРЕВА РІШЕНЬ ТА ГРАДІЄНТНОГО БУСТИНГУ

Шимчук Г.В., Литвиненко Я.В.

NATURAL GAS CONSUMPTION FORECASTING USING MACHINE LEARNING METHODS: MLP, DECISION TREE, AND GRADIENT BOOSTING

Shymchuk G.V., Lytvynenko I.V.

У роботі досліджено точність прогнозування споживання природного газу за допомогою методів машинного навчання: багатошарового перцептрона (MLP), дерева рішень і градієнтного бустингу. Прогноз формувалася для послідовних трьохгодинних інтервалів, а точність обраних моделей оцінювалася на повному річному тестовому інтервалі – 2019 року. Основну увагу зосереджено не лише на виборі алгоритму, а й на способі подання інформації про попередню динаміку газоспоживання. Для всіх моделей використано однаковий набір із 47 вхідних ознак. До нього входили 16 метеорологічних, шість циклічних часових, 11 лагових, дев'ять ковзних і п'ять різницеових ознак. Це дало змогу порівнювати прогнозні можливості алгоритмів за однакового складу вхідних ознак. Дані попередніх років використано для побудови та налаштування моделей, а 2019 рік залишено для незалежної перевірки. Результати показали, що найвищу точність прогнозування на річному тестовому інтервалі формує дерево рішень. Градієнтний бустинг і MLP також достатньо точно відтворюють загальну динаміку газоспоживання, однак частіше помиляються на ділянках зі швидкою зміною режиму. Найбільші розбіжності між моделями спостерігаються восени, коли газоспоживання переходить від низького літнього рівня до рівня, характерного для опалювального сезону. Аналіз важливості вхідних ознак показав, що прогноз найбільшою мірою визначається попередніми значеннями газоспоживання, тоді як метеорологічні та календарні характеристики дають змогу врахувати вплив погодних умов, сезонності та часових закономірностей процесу. Отримані результати підтверджують, що точність прогнозування залежить не лише від вибору моделі, а й від того, наскільки повно у вхідних даних відображено часову структуру процесу. Статистична перевірка показала, що дерево рішень має статистично підтверджену перевагу над MLP і градієнтним бустингом, тобто виявлені відмінності в точності прогнозування не можна пояснити лише випадковою варіативністю похибок на тестовій вибірці. Практичне значення роботи полягає у можливості використання отриманих

результатів для побудови систем оперативного прогнозування газоспоживання в умовах сезонної мінливості попиту та погодних змін.

Ключові слова: газоспоживання, прогнозування, методи машинного навчання, MLP, дерево рішень, градієнтний бустинг, часові ряди.

Вступ. Процес газоспоживання має виразну сезонність, але саме сезонність робить задачу прогнозування оманливо простою. Узимку рівень газоспоживання є високим, улітку знижується до мінімуму, а восени швидко й нерівномірно змінюється. За тригодинного кроку даних до цього додаються добові піки, тижнева повторюваність, інерційність навантаження та вплив погодних умов. Тому модель, яка добре відтворює середню річну форму споживання газу, часто робить помилки під час прогнозування на перехідних ділянках [1, 2].

Для прогнозування газоспоживання застосовують нейронні мережі, зокрема моделі LSTM, ансамблеві методи на основі дерев рішень, а також гібридні підходи, що передбачають попереднє розкладання часового ряду газоспоживання на окремі складові [3–6]. Останні публікації присвячені підходам від простого нарощування глибини мережі до побудови інформативного набору ознак і поєднання часових масштабів. У роботі [7] підходи до прогнозування газоспоживання систематизовано за методами обробки вхідних ознак, декомпозиції даних, типами прогнозних моделей та алгоритмами оптимізації. У багатьох сучасних роботах запропоновано низку комбінованих підходів до прогнозування газоспоживання. Зокрема, використано варіаційне модальне розкладання (Variational Mode Decomposition, VMD) у поєднанні з

Transformer і LSTM [8], двоетапне виділення ознак [9], оптимізацію вхідних ознак для моделі LSTM [10], використання мультимодальних даних у моделі на основі Transformer [11], а також часово-частотне представлення даних із залученням апіорних знань про процес газоспоживання [12]. У цих роботах підвищення точності прогнозування досягається шляхом удосконалення як прогнозних моделей, так і способів попереднього оброблення та представлення даних. Водночас залишається актуальним порівняння різних алгоритмів прогнозування за однакового набору вхідних ознак.

Аналіз останніх досліджень і публікацій. У роботі [3] показано можливість застосування багатоваріантного перцептрона для прогнозування газоспоживання з урахуванням метеорологічних і календарних чинників. У дослідженні [4] для короткострокового прогнозування газоспоживання використано глибоку нейронну мережу, яка продемонструвала підвищення точності порівнянні з традиційними підходами. У роботі [5] застосовано модель LSTM, здатну враховувати залежності між поточними та попередніми значеннями часового ряду газоспоживання. У праці [6] досліджено моделі на основі градієнтного бустингу та штучних нейронних мереж, а також їх комбінування; отримані результати показали, що поєднання окремих моделей прогнозів може зменшувати загальну похибку. Дослідження, описані в працях [8–13] свідчать, що точність прогнозування залежить не лише від вибору прогнозної моделі, а й від способу підготовки вхідних даних, відбору ознак та врахування часових, метеорологічних та інших чинників, що впливають на газоспоживання.

Адитивну математичну модель газоспоживання з циклічним випадковим процесом, трендом і стохастичним залишком сформовано у праці [14], а розвиток даної моделі з урахуванням масштабних коефіцієнтів на сегментах-циклах подано у праці [15]. У роботі [14] авторами запропоновано метод прогнозування за допомогою градієнтного бустингу та на основі моделі дерева рішень з врахуванням метеорологічних факторів (було враховано лише чотири показники: температура, вологість, атмосферний тиск та сила вітру). У роботі показано, що кращий результат прогнозування було отримано на основі дерев рішень.

Метою роботи є порівняння точності прогнозування газоспоживання за допомогою MLP-5, дерева рішень і градієнтного бустингу за однакового набору вхідних ознак та однакових періодів навчання, валідації і тестування, а також аналіз відповідності отриманих прогнозів фактичному газоспоживанню на незалежному тестовому інтервалі 2019 року.

Вихідні дані та схема дослідження. У дослідженні використано два ряди даних за 2007–2019 рр.: обсяги газоспоживання та метеорологічні спостереження. Обидва ряди приведено до однакового часового кроку три години: 00:00, 03:00,

..., 21:00. Загалом сформовано 37 984 спостереження. Для кожного значення газоспоживання використовувалися метеорологічні дані, що відповідали тому самому часовому інтервалу.

Перед формуванням вхідних ознак метеорологічні дані перевірили на наявність пропущених значень. Після приведення даних до єдиного тригодинного часового кроку пропущених значень у використаних метеорологічних показниках не виявлено. Тому заповнення пропусків або вилучення спостережень через відсутність метеорологічних даних не виконували.

Оскільки дослідження виконувалося на історичних даних, фактичні метеорологічні показники для кожного прогнозованого інтервалу були відомі. У практичній системі прогнозування замість фактичних метеорологічних даних необхідно використовувати прогноз погоди, похибка якого може додатково впливати на точність прогнозу газоспоживання.

Дані поділено на навчальний, валідаційний і тестовий періоди з дотриманням їх хронологічної послідовності [17, 18]. Дані 2007–2017 рр. використовували для побудови моделей, дані 2018 року – для вибору їхніх параметрів і конфігурацій, а дані 2019 року не використовували на етапах навчання та налаштування й залишили для остаточної перевірки моделей. Після вибору конфігурацій моделі повторно навчали на даних 2007–2018 рр.

Максимальний використаний лаг становив 57 тригодинних інтервалів. Тому перші 57 спостережень 2007 року, для яких неможливо було сформувати повний набір лагових, ковзних і різницевих ознак, вилучили з вибірки. Після формування вхідних ознак загальна кількість придатних для моделювання спостережень становила 37 927: навчальний період містив 32 087 спостережень, валідаційний – 2920, тестовий – 2920.

Для спостережень 2018 і 2019 років вхідні ознаки обчислювали з використанням відомих попередніх значень газоспоживання за 2017 і 2018 роки відповідно. Тому на початку валідаційного і тестового періодів додаткові спостереження не вилучали. Таким чином, під час формування вхідних ознак інформацію з майбутніх спостережень не використовували.

Формування системи вхідних ознак. Кожна з трьох моделей отримувала однаковий 47-вимірний вектор вхідних ознак. До нього входили 16 метеорологічних показників: температура повітря (T), атмосферний тиск на рівні метеостанції (P_0), атмосферний тиск, приведений до рівня моря (P), зміна атмосферного тиску за попередні три години (P_a), відносна вологість повітря (U), середня швидкість вітру (F_f), максимальна швидкість пориву вітру протягом 10 хв перед спостереженням ($ff10$), максимальна швидкість пориву вітру між послідовними спостереженнями ($ff3$), мінімальна (T_n) і максимальна (T_x) температура повітря за

попередній період, горизонтальна видимість (VV), температура точки роси (Td), кількість опадів (RRR), тривалість періоду накопичення опадів (tR), мінімальна температура поверхні ґрунту вночі (Tg) та висота снігового покриву (sss). Крім них, використовувалися шість циклічних часових ознак, 11 лагових значень газоспоживання, дев'ять ковзних характеристик та п'ять різницевої ознак газоспоживання. Різницеві ознаки визначали зміну між вибраними попередніми значеннями ряду: $D_{1,2}$, $D_{2,3}$, $D_{8,16}$, $D_{16,56}$, $D_{55,57}$. Усі лагові, ковзні та різницеві ознаки подавалися на вхід кожної з трьох моделей, а не лише MLP-5. Номер тригодинного інтервалу доби, номер дня тижня та номер дня року кодували синусно-косинусними парами відповідно до формули (1).

При такому кодуванні кінець циклу залишається близьким до його початку, а не подається як штучно віддалена точка.

Циклічні часові ознаки відображали повторюваність газоспоживання протягом доби, тижня та року. Лагові ознаки містили попередні значення газоспоживання і давали змогу враховувати короточасну, добову та тижневу залежність поточного рівня споживання від його попередніх значень. Ковзні характеристики описували поведінку ряду на певному попередньому часовому інтервалі та включали середні значення, стандартні відхилення, мінімум, максимум та медіану. Різницеві ознаки характеризували зміну газоспоживання між вибраними попередніми моментами часу та давали змогу враховувати не лише рівень споживання, а й напрям і величину його зміни. Таким чином, сформований набір ознак містив інформацію про погодні умови, добові, тижневі та річні закономірності газоспоживання, його попередні значення та характер зміни.

Для відображення циклічного характеру календарних і часових ознак їх перетворено за допомогою функцій синуса та косинуса:

$$z_{\sin}(t) = \sin\left(\frac{2\pi z(t)}{L_z}\right), z_{\cos}(t) = \cos\left(\frac{2\pi z(t)}{L_z}\right), \quad (1)$$

де t – номер тригодинного часового інтервалу; $z(t)$ – значення відповідної часової ознаки в момент t , наприклад номер тригодинного інтервалу доби, номер дня тижня або номер дня року; L_z – тривалість повного циклу відповідної часової ознаки. Для номера тригодинного інтервалу доби $L_z = 8$, для номера дня тижня $L_z = 7$, а для номера дня року $L_z = 365$ у звичайному році та $L_z = 366$ у високосному році; π – математична стала; $z_{\sin}(t)$ і $z_{\cos}(t)$ – синусна та косинусна компоненти відповідної часової ознаки. Таке перетворення дає змогу подати циклічні часові характеристики у вигляді двох числових ознак і врахувати, що кінець

одного часового циклу безпосередньо переходить у початок наступного.

Для врахування попередніх значень газоспоживання використано 11 лагових ознак: $Q(t-1)$, $Q(t-2)$, $Q(t-3)$, $Q(t-8)$, $Q(t-9)$, $Q(t-15)$, $Q(t-16)$, $Q(t-17)$, $Q(t-55)$, $Q(t-56)$, $Q(t-57)$.

За тригодинного кроку даних лаги 1, 2 і 3 відповідають значенням газоспоживання 3, 6 і 9 годин тому; лаги 8 і 9 – 24 і 27 годин тому; лаги 15, 16 і 17 – 45, 48 і 51 годину тому; лаги 55, 56 і 57 – 165, 168 і 171 годину тому. Використання таких ознак дає змогу врахувати короткострокові зміни, а також добову, дводобову і тижневу повторюваність газоспоживання.

Лагову ознаку визначали за формулою:

$$Q_{lag,k}(t) = Q(t-k), \quad (2)$$

де $Q_{lag,k}(t)$ – значення лагової ознаки, сформованої для тригодинного інтервалу t ; $Q(t-k)$ – фактичний обсяг газоспоживання у попередньому інтервалі з номером $t-k$; Q – фактичний обсяг газоспоживання в m^3 ; t – порядковий номер тригодинного інтервалу, для якого формується прогноз; k – кількість тригодинних інтервалів між прогнозованим і попереднім значеннями газоспоживання.

Крім окремих попередніх значень газоспоживання, використовували дев'ять ковзних ознак. Середні значення та стандартне відхилення обчислювали за 8, 16 і 56 попередніми відліками, що відповідає одній добі, двом добам та одному тижню. Для восьми попередніх відліків додатково визначали мінімальне, максимальне та медіанне значення. Таким чином, до набору входили три ковзні середні, три стандартні відхилення, мінімум, максимум і медіана.

Ковзне середнє визначали за формулою:

$$\bar{Q}_w(t) = \frac{1}{w} \sum_{j=1}^w Q(t-j), \quad (3)$$

де $\bar{Q}_w(t)$ – середнє значення газоспоживання за w попередніми тригодинними інтервалами, обчислене для прогнозованого моменту часу t ; w – кількість попередніх інтервалів у ковзному вікні, причому w набуває значень 8, 16 або 56; j – порядковий номер попереднього інтервалу в межах ковзного вікна; $Q(t-j)$ – фактичний обсяг газоспоживання, зафіксований за j тригодинних інтервалів до прогнозованого моменту, Q – фактичний обсяг газоспоживання в m^3 .

Окрім ковзного середнього, для кожного з вибраних часових вікон визначали стандартне

відхилення, яке характеризує розкид попередніх значень газоспоживання відносно їхнього середнього значення:

$$S_w(t) = \sqrt{\frac{1}{w-1} \sum_{j=1}^w [Q(t-j) - \bar{Q}_w(t)]^2}, \quad (4)$$

де $S_w(t)$ – стандартне відхилення значень газоспоживання за w тригодинними інтервалами, що передують інтервалу t ; w – кількість попередніх інтервалів у ковзному вікні, причому w набуває значень 8, 16 або 56; j – порядковий номер попереднього інтервалу в межах ковзного вікна; $Q(t-j)$ – фактичний обсяг газоспоживання в інтервалі з номером $t-j$; $\bar{Q}_w(t)$ – ковзне середнє за w попередніми інтервалами, визначене за формулою (3).

Більше значення $S_w(t)$ свідчить про значніші коливання газоспоживання в межах відповідного часового вікна, а менше – про стабільніші попередні значення.

Для врахування характеру зміни газоспоживання між вибраними інтервалами використано також п'ять різницевих ознак. У загальному вигляді різницеву ознаку визначали за формулою:

$$D_{a,b}(t) = Q(t-a) - Q(t-b), \quad (5)$$

де $D_{a,b}(t)$ – різницева ознака, сформована для інтервалу t ; $Q(t-a)$ і $Q(t-b)$ – фактичні обсяги газоспоживання в попередніх інтервалах з номерами $t-a$ і $t-b$ відповідно; t – порядковий номер тригодинного інтервалу, для якого формується прогноз; a і b – величини лагів, виражені кількістю тригодинних інтервалів.

У дослідженні використано такі різницеві ознаки: $D_{1,2}(t)$, $D_{2,3}(t)$, $D_{8,16}(t)$, $D_{16,56}(t)$ і $D_{55,57}(t)$. Перші дві ознаки характеризують короткострокові зміни газоспоживання впродовж трьох попередніх інтервалів. Ознака $D_{8,16}(t)$ відображає різницю між значеннями газоспоживання за одну і дві доби до прогнозованого інтервалу, $D_{16,56}(t)$ – різницю між значеннями за дві доби й один тиждень, $D_{55,57}(t)$ – короткострокову зміну газоспоживання поблизу тижневого лагу.

Під час формування лагових, ковзних і різницевих ознак використовували лише значення газоспоживання, зафіксовані до початку інтервалу t . Фактичне значення $Q(t)$, яке прогнозувалося моделлю, до набору вхідних ознак не включали.

Порівнювані моделі. У дослідженні порівнювали модель багатошарового перцептрона, дерево рішень і модель на основі градієнтного

бустингу. Для багатошарового перцептрона використано конфігурацію MLP-5, отриману в результаті попереднього поетапного розширення системи вхідних ознак.

На першому етапі модель MLP-1 отримувала чотири метеорологічні ознаки: температуру повітря (T), атмосферний тиск (P_0), максимальну швидкість пориву вітру ($ff10$) та відносну вологість повітря (U). У моделі MLP-2 використано повний набір із 16 метеорологічних ознак. У моделі MLP-3 до них додано шість циклічних часових ознак, які описували номер тригодинного інтервалу доби, день тижня та день року за допомогою синусно-косинусного кодування.

У моделі MLP-4 систему вхідних даних доповнено лаговими значеннями газоспоживання $Q_{lag,1}$, $Q_{lag,2}$, $Q_{lag,8}$, $Q_{lag,16}$, і $Q_{lag,56}$, а також ковзними характеристиками. Фінальна конфігурація MLP-5 отримувала повний набір із 47 вхідних ознак: 16 метеорологічних, шість циклічних часових, 11 лагових, дев'ять ковзних і п'ять різницевих ознак. Саме цей набір ознак надалі використовували також для дерева рішень і градієнтного бустингу, що забезпечило однакові умови порівняння трьох моделей.

На кожному етапі перевіряли декілька архітектур багатошарового перцептрона. За результатами валідації для моделей MLP-1, MLP-4 і MLP-5 вибрано архітектуру з двома прихованими шарами, що містили 256 і 128 нейронів, а для MLP-2 і MLP-3 – архітектуру з трьома прихованими шарами, що містили 128, 64 і 32 нейрони. У подальшому порівнянні з деревом рішень і градієнтним бустингом використано лише фінальну конфігурацію MLP-5 з архітектурою 47→256→128→1.

Перед навчанням MLP-5 усі вхідні ознаки стандартизували за допомогою StandardScaler: від кожного значення віднімали середнє значення відповідної ознаки та отриману різницю ділили на її стандартне відхилення. Аналогічне перетворення застосовували до цільового значення газоспоживання $Q(t)$. На етапі налаштування параметри стандартизації визначали лише за навчальними даними 2007–2017 рр., після чого без змін застосовували до валідаційних даних 2018 року. Після вибору конфігурації параметри стандартизації повторно визначали за об'єднаними даними 2007–2018 рр. і застосовували до тестових даних 2019 року. Отримані прогнозовані значення повертали до початкової шкали та подавали в кубічних метрах.

Для забезпечення відтворюваності результатів у процедурах, що містили випадкову складову, використовували фіксоване значення параметра `random_state = 42`. Це значення задавали під час побудови моделей і проведення перестановочної оцінки важливості ознак.

Модель MLP-5 має два приховані шари з 256 і 128 нейронами та функцію активації ReLU[19]. Для навчання використано оптимізатор Adam (Adaptive Moment Estimation)[18]. Також застосовано L2-

регуляризацию з коефіцієнтом $\alpha = 10^{-4}$ і розмір пакета 128.

Кількість епох навчання MLP-5 визначали за результатами прогнозування на валідаційних даних 2018 року. Перевіряли від 1 до 40 епох. Найменше значення $WAPE$ на валідаційному періоді становило 5,01 % і було отримане після 33 епох. Після визначення архітектури та кількості епох модель повторно навчали на об'єднаних даних 2007–2018 рр. упродовж 33 епох. Випадкове виділення частини цих даних для внутрішньої валідації не застосовували.

Архітектура MLP-5 має вигляд $47 \rightarrow 256 \rightarrow 128 \rightarrow 1$, де 47 – кількість вхідних ознак, 256 і 128 – кількість нейронів у першому та другому прихованих шарах відповідно, а 1 – одне прогнозоване значення газоспоживання.

Нехай $\mathbf{x}(t)$ – вектор із 47 вхідних ознак, сформованих для тригодинного інтервалу t . Вектор вихідних значень першого прихованого шару визначали за формулою:

$$\mathbf{h}^{(1)}(t) = \text{ReLU}(\mathbf{W}^{(1)}\mathbf{x}(t) + \mathbf{b}^{(1)}), \quad (6)$$

де $\mathbf{x}(t)$ – вектор із 47 вхідних ознак, сформованих для тригодинного інтервалу t ; $\mathbf{W}^{(1)}$ – матриця вагових коефіцієнтів між вхідним і першим прихованим шарами розміром 256×47 ; $\mathbf{b}^{(1)}$ – вектор зміщень першого прихованого шару, що містить 256 значень; $\mathbf{h}^{(1)}(t)$ – вектор вихідних значень першого прихованого шару, що містить 256 значень; ReLU – функція активації, яку застосовують до кожного елемента отриманого вектора.

Функція ReLU залишає додатні значення без змін, а від'ємні замінює нулем: $\text{ReLU}(u) = \max(0, u)$.

Вектор вихідних значень другого прихованого шару визначали за формулою:

$$\mathbf{h}^{(2)}(t) = \text{ReLU}(\mathbf{W}^{(2)}\mathbf{h}^{(1)}(t) + \mathbf{b}^{(2)}), \quad (7)$$

де $\mathbf{h}^{(1)}(t)$ – вектор вихідних значень першого прихованого шару; $\mathbf{W}^{(2)}$ – матриця вагових коефіцієнтів між першим і другим прихованими шарами розміром 128×256 ; $\mathbf{b}^{(2)}$ – вектор зміщень другого прихованого шару, що містить 128 значень; $\mathbf{h}^{(2)}(t)$ – вектор вихідних значень другого прихованого шару, що містить 128 значень.

Вихідний шар формував прогнозоване значення газоспоживання для тригодинного інтервалу t ; за формулою:

$$\hat{Q}_{MLP}(t) = \sigma Q[\mathbf{W}^{(3)}\mathbf{h}^{(2)}(t) + \mathbf{b}^{(3)}] + \mu Q, \quad (8)$$

де $\hat{Q}_{MLP}(t)$ – прогнозований моделлю MLP-5 обсяг газоспоживання для тригодинного інтервалу t ; м^3 ;

$\mathbf{W}^{(3)}$ – матриця вагових коефіцієнтів вихідного шару; $\mathbf{h}^{(2)}(t)$ – вектор вихідних значень другого прихованого шару; $\mathbf{b}^{(3)}$ – зміщення вихідного шару, σQ і μQ – відповідно середнє значення та стандартне відхилення цільової змінної, визначені за навчальними даними. Вираз у квадратних дужках є прогнозом у стандартизованій шкалі.

Модель на основі дерева рішень. Регресійне дерево побудовано за алгоритмом CART [26]. На етапі налаштування перевіряли максимальну глибину дерева 8, 12, 16, 20 та варіант без обмеження глибини, а також мінімальну кількість спостережень у кінцевому вузлі – 5, 10, 20 і 40. За критерієм мінімального $WAPE$ на валідаційних даних 2018 року обрано дерево з максимальною глибиною 16 і мінімальною кількістю 20 спостережень у кінцевому вузлі.

Прогноз дерева рішень для тригодинного інтервалу t визначається значенням у кінцевому вузлі, до якого потрапляє вектор вхідних ознак:

$$\hat{Q}_{DT}(t) = c_r, \quad \mathbf{x}(t) \in R_r, \quad (9)$$

де $\hat{Q}_{DT}(t)$ – прогнозований деревом рішень обсяг газоспоживання для тригодинного інтервалу t , м^3 ; $\mathbf{x}(t)$ – вектор із 47 вхідних ознак, сформованих для інтервалу t ; R_r – область значень вхідних ознак, що відповідає кінцевому вузлу r ; c_r – прогнозоване значення газоспоживання в цьому кінцевому вузлі; r – номер кінцевого вузла дерева.

Модель на основі градієнтного бустингу. Для третього підходу використано модель HistGradientBoostingRegressor із бібліотеки scikit-learn [17], яка належить до методів гістограмного градієнтного бустингу дерев рішень [25]. У цій моделі значення вхідних ознак попередньо розподіляються за інтервалами, а прогноз формується ансамблем послідовно побудованих дерев, кожне з яких коригує похибки попередніх дерев. За результатами налаштування на валідаційних даних 2018 року встановлено 300 ітерацій, швидкість навчання 0,1, максимальну кількість 63 кінцевих вузлів дерева та мінімальну кількість 10 спостережень у кінцевому вузлі. Автоматичну ранню зупинку під час навчання не застосовували, тому було виконано всі 300 ітерацій і побудовано 300 дерев ансамблю.

Прогноз моделі градієнтного бустингу для тригодинного інтервалу t , визначали за формулою:

$$\hat{Q}_{GB}(t) = F_0(\mathbf{x}(t)) + \eta \sum_{m=1}^M g_m(\mathbf{x}(t)), \quad (10)$$

де $\hat{Q}_{GB}(t)$ – прогнозований моделлю градієнтного бустингу обсяг газоспоживання для тригодинного

інтервалу t , m^3 ; $x(t)$ – вектор із 47 вхідних ознак, сформованих для інтервалу t ; $F_0(x(t))$ – початковий прогноз моделі; η – швидкість навчання, яка в цій моделі дорівнює 0,1; M – кількість послідовно побудованих дерев, що дорівнює 300; m – порядковий номер дерева в ансамблі; $g_m(x(t))$ – поправка до прогнозу, сформована m -м деревом.

Кожне наступне дерево формує поправку до поточного прогнозу, спрямовану на зменшення похибок попередньо побудованого ансамблю. Швидкість навчання η визначає внесок кожного нового дерева в остаточний прогноз.

Вибрані за результатами валідації на даних 2018 року конфігурації порівнюваних моделей наведено в табл. 1. Для MLP-5 визначено архітектуру мережі та параметри навчання, для дерева рішень – максимальну глибину і мінімальну кількість спостережень у кінцевому вузлі, для градієнтного бустингу – кількість ітерацій, швидкість навчання та параметри побудови дерев ансамблю.

Таблиця 1

Конфігурації моделей, вибрані за результатами валідації на даних 2018 року

Модель	Архітектура / метод	Вибрані параметри
MLP-5	47→256→128→1; ReLU	Adam; $\alpha=10^{-4}$; пакет 128; 33 епохи
Дерево рішень	CART	максимальна глибина – 16, мінімальна кількість спостережень у кінцевому вузлі – 20
Градієнтний бустинг	гістограмний GBDT	кількість ітерацій – 300; $\eta = 0,1$; максимальна кількість кінцевих вузлів одного дерева – 63; мінімальна кількість спостережень у кінцевому вузлі – 10

Наведені конфігурації використовували без змін під час остаточного навчання моделей на даних 2007–2018 рр. та оцінювання їхньої точності на тестових даних 2019 року.

Критерії оцінювання. Точність прогнозування оцінювали за коефіцієнтом детермінації R^2 , середньою абсолютною похибкою MAE , коренем із середньої квадратичної похибки $RMSE$, середньою абсолютною відсотковою похибкою $MAPE$ та зваженою абсолютною відсотковою похибкою $WAPE$. Коефіцієнт R^2 характеризує, яку частину варіації фактичних значень відтворює прогнозна модель. Чим ближче його значення до одиниці, тим точніше прогноз відтворює фактичні дані. MAE визначає середню абсолютну величину похибки, а $RMSE$ є чутливішою до великих відхилень унаслідок піднесення похибок до квадрата. $MAPE$ характеризує середню відносну похибку у відсотках, однак може набувати нестійких значень за малих фактичних

обсягів газоспоживання. $WAPE$ визначає відношення сумарної абсолютної похибки прогнозу до сумарного фактичного газоспоживання. Для порівняння моделей на річному тестовому інтервалі основною відносною метрикою обрано $WAPE$ [21].

Нехай $Q_i = Q(t_i)$ – фактичний обсяг газоспоживання в i -му тестовому спостереженні; $\hat{Q}_i$ – відповідне прогнозоване значення $\bar{Q}$ – середнє фактичне значення газоспоживання на тестовому інтервалі; i – порядковий номер тестового спостереження; n – загальна кількість тестових спостережень. Для тестового періоду 2019 року $n = 2920$. Середнє фактичне значення визначається як $\bar{Q} = \frac{1}{n} \sum_{i=1}^n Q_i$.

Коефіцієнт детермінації. Коефіцієнт детермінації обчислювали за формулою:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^n (Q_i - \bar{Q})^2}, \quad (11)$$

де R^2 – коефіцієнт детермінації; Q_i – фактичний обсяг газоспоживання; $\hat{Q}_i$ – прогнозований обсяг газоспоживання; $\bar{Q}$ – середнє фактичне значення газоспоживання; n – кількість тестових спостережень.

Середня абсолютна похибка. Середню абсолютну похибку визначали за формулою:

$$MAE = \frac{1}{n} \sum_{i=1}^n |Q_i - \hat{Q}_i|, \quad (12)$$

де MAE – середня абсолютна похибка прогнозування, m^3 . Менше значення MAE відповідає точнішому прогнозу.

Корінь із середньої квадратичної похибки. $RMSE$ обчислювали за формулою:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Q_i - \hat{Q}_i)^2}, \quad (13)$$

де $RMSE$ – корінь із середньої квадратичної похибки прогнозування, m^3 . Через піднесення похибок до квадрата $RMSE$ сильніше реагує на окремі великі відхилення, ніж MAE .

Середня абсолютна відсоткова похибка. $MAPE$ визначали за формулою:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \frac{|Q_i - \hat{Q}_i|}{|Q_i|}, \quad (14)$$

де $MAPE$ – середня абсолютна відсоткова похибка прогнозування, %. За малих значень Q_i ця метрика може непропорційно зростати.

Зважена абсолютна відсоткова похибка. $WAPE$ обчислювали за формулою:

$$WAPE = 100\% \frac{\sum_{i=1}^n |Q_i - \hat{Q}_i|}{\sum_{i=1}^n |Q_i|}, \quad (15)$$

де $WAPE$ – зважена абсолютна відсоткова похибка прогнозування, %. На відміну від $MAPE$ у $WAPE$ сумарну абсолютну похибку ділять на сумарний фактичний обсяг газоспоживання, тому окремі малі значення Q_i мають менший вплив на підсумковий результат.

Статистичне порівняння моделей.

Відмінності між похибками прогнозування моделей попарно перевіряли за допомогою двостороннього тесту Дібольда-Маріано [25] для квадратичної та абсолютної функцій втрат. Для кожного моменту часу різницю функцій втрат обчислювали як значення функції втрат першої моделі мінус значення функції втрат другої. Тому додатне значення статистики DM у поєднанні з $p < 0,05$ свідчить про статистично значущу перевагу другої моделі.

Для врахування часової залежності між похибками довгострокову дисперсію різниці функцій втрат оцінювали за допомогою стійкого до гетероскедастичності й автокореляції НАС-оцінювача Ньюї-Веста з максимальним лагом 56. За тригодинного кроку даних цей лаг відповідає одному тижню.

Для різниці $WAPE$ між моделями також будували 95 % довірчі інтервали методом рухомого блокового бутстрепу. Цей метод передбачає багаторазове формування нових вибірок із послідовних блоків початкового часового ряду, що дає змогу частково зберегти часову залежність між сусідніми спостереженнями. У кожному з 2000 повторень із блоків довжиною 56 спостережень формували нову вибірку та повторно обчислювали різницю $WAPE$. Для обох моделей у кожній порівнюваній парі використовували однакові блоки спостережень.

Для кожної пари моделей різницю $WAPE$ визначали за формулою:

$$\Delta WAPE = WAPE_A - WAPE_B, \quad (16)$$

де $WAPE_A$ і $WAPE_B$ – значення $WAPE$ першої та другої моделей у відповідній парі. Додатне значення $\Delta WAPE$ означає, що перша модель має більшу похибку. Якщо 95 % довірчий інтервал не містить нуля, різницю між моделями вважають статистично підтвердженою. Якщо інтервал містить нуль, наявних даних недостатньо для підтвердження відмінності між моделями.

Методика оцінювання важливості вхідних ознак. Оскільки дерево рішень забезпечило найвищу точність прогнозування, перестановочну оцінку важливості вхідних ознак проводили лише для цієї моделі на незалежних тестових даних 2019 року. Спочатку обчислювали базове значення коефіцієнта R^2 за незмінними тестовими даними. Потім значення кожної ознаки випадково перемішували між тестовими спостереженнями, залишаючи інші ознаки незмінними, після чого повторно обчислювали R^2 [26]. Перемішування кожної ознаки виконували п'ять разів, а її важливість визначали як середнє зменшення R^2 порівняно з базовим значенням. Чим більшим було зменшення R^2 , тим сильніше точність прогнозу дерева рішень залежала від інформації, що містилася у відповідній ознаці. Значення, близьке до нуля або від'ємне, свідчило про відсутність помітного позитивного внеску ознаки в точність цієї моделі.

Результати прогнозування на незалежному тестовому інтервалі. Значення основних показників точності для базових прогнозів і трьох досліджуваних моделей на тестових даних 2019 року наведено в табл. 2. Базові прогнози не передбачали навчання моделей: прогнозоване значення газоспоживання для інтервалу t прирівнювали до фактичного значення, зафіксованого за k тригодинних інтервалів до нього:

$$\hat{Q}_{basek}(t) = Q(t-k), k \in \{1, 8, 56\}, \quad (17)$$

де $\hat{Q}_{basek}(t)$ – базовий прогноз газоспоживання для інтервалу t ; $Q(t-k)$ – фактичний обсяг газоспоживання за k тригодинних інтервалів до прогнозованого інтервалу; k – величина лагу. Значення $k=1, k=8, k=56$ відповідають прогнозам за значеннями газоспоживання, зафіксованими відповідно три години, одну добу й один тиждень тому.

Серед базових варіантів найкращий результат отримано для добового лагового прогнозу:

$$\hat{Q}_{base8}(t) = Q(t-8) \quad (18)$$

для якого $R^2 = 0,9500$, а $WAPE$ становить 11,74 %. Це підтверджує виражену добову повторюваність газоспоживання та визначає базовий рівень точності для порівняння з досліджуваними моделями.

Серед досліджуваних моделей MLP-5, дерева рішень і градієнтного бустингу найвищу точність показало дерево рішень. Для нього отримано $R^2 = 0,9704$, $MAE = 4233,03 \text{ м}^3$, $RMSE = 7380,45 \text{ м}^3$, $MAPE = 12,99\%$ і $WAPE = 8,48\%$. Отже, дерево рішень має найбільше значення R^2 та найменші значення MAE , $RMSE$, $MAPE$, $WAPE$ і серед усіх порівнюваних моделей.

Таблиця 2

Якість прогнозування на незалежному тестовому інтервалі 2019 року

Модель	R^2	MAE , м ³	$RMSE$ м ³	$MAPE$ %	$WAPE$ %
Q(t-1)	0,8602	10882,15	16044,79	34,23	21,81
Q(t-8)	0,9500	5857,79	9600,06	13,92	11,74
Q(t-56)	0,8480	10087,73	16732,69	20,58	20,22
MLP-5	0,9514	5458,08	9460,61	18,53	10,94
Дерево рішень	0,9704	4233,03	7380,45	12,99	8,48
Гradientний бустинг	0,9612	5114,19	8451,45	13,10	10,25

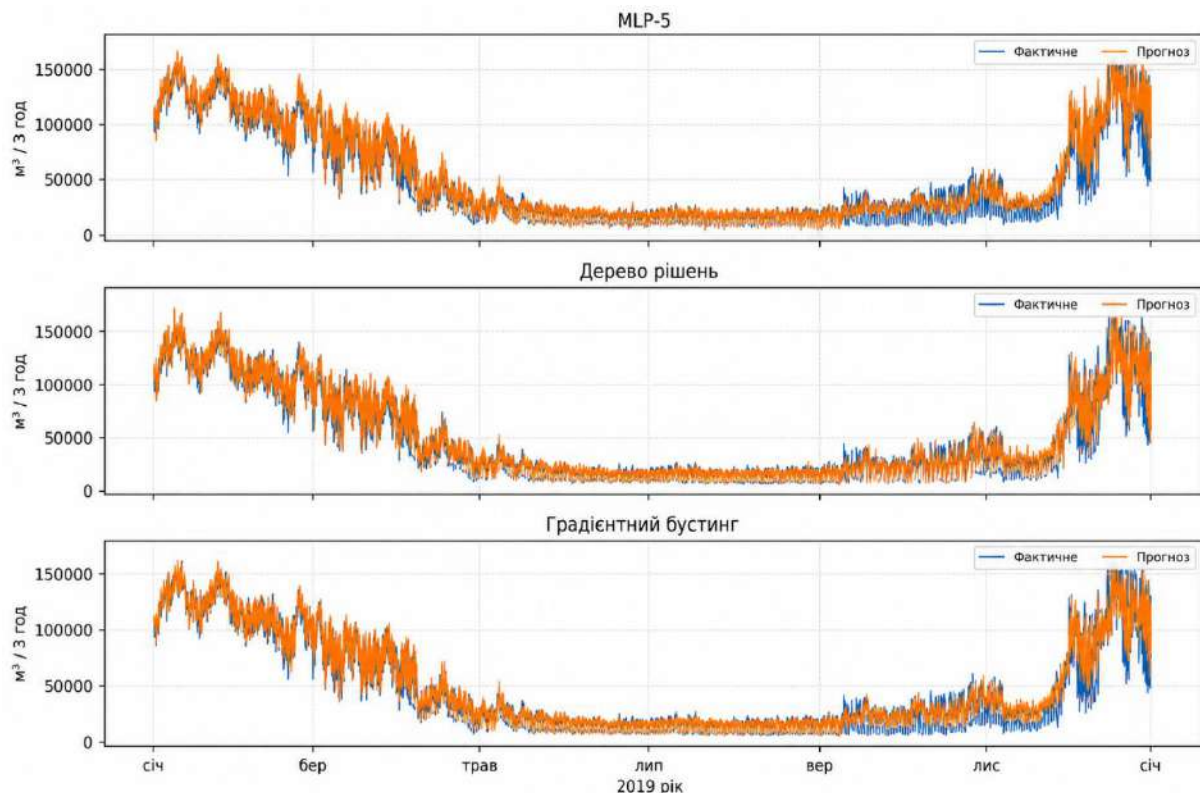


Рис. 1. Фактичний і прогнозований ряд газоспоживання на незалежному тестовому інтервалі 2019 року

Порівняно з MLP-5 значення MAE для дерева рішень є меншим на 22,44 %, $RMSE$ – на 21,99 %, а $WAPE$ – на 22,49 %. Коефіцієнт детермінації при цьому зростає з 0,9514 до 0,9704. Порівняно з gradientним бустингом дерево рішень також має менші похибки: MAE – на 17,23 %, $RMSE$ – на 12,67 %, а $WAPE$ – на 17,27 %. Значення $MAPE$ для цих двох моделей є близькими: 12,99 % для дерева рішень і 13,10 % для gradientного бустингу.

Фактичні та прогнозовані значення газоспоживання у 2019 році. На рис. 1 наведено фактичні та прогнозовані значення газоспоживання для всіх 2920 тригодинних інтервалів тестового періоду 2019 року.

На початку року всі три моделі загалом відтворюють високі значення газоспоживання, характерні для зимового періоду, а також їхні внутрішньодобові коливання. Навесні

газоспоживання поступово зменшується, а влітку залишається відносно низьким.

Найбільші розбіжності між фактичними та прогнозованими значеннями спостерігаються під час переходу від літнього до осінньо-зимового періоду. Починаючи з вересня фактичне газоспоживання зростає, а амплітуда його коливань збільшується. У цей період похибка моделі MLP-5 помітно зростає. Дерево рішень і gradientний бустинг загалом точніше відтворюють початок осіннього зростання газоспоживання. У грудні похибки знову збільшуються для всіх трьох моделей, що збігається з підвищенням газоспоживання та збільшенням амплітуди його коливань.

На рис. 2 кожна точка відповідає одному тригодинному інтервалу 2019 року. На горизонтальній осі відкладено фактичний обсяг газоспоживання, а на вертикальній – прогнозований.

Пряма лінія на кожній діаграмі є лінією ідеального прогнозу, для якої виконується рівність:

$$\hat{Q}(t) = Q(t) \quad (19)$$

де $Q(t)$ – фактичний обсяг газоспоживання у тригодинному інтервалі t , а $\hat{Q}(t)$ – відповідне прогнозоване значення. У кожній точці цієї лінії прогноз повністю збігається з фактичним значенням. Чим ближче точки розташовані до лінії ідеального прогнозу, тим меншою є похибка моделі.

Для дерева рішень точки загалом розташовані ближче до лінії ідеального прогнозу, ніж для MLP-5 і градієнтного бустингу. Це узгоджується з результатами, наведеними в табл. 2, згідно з якими дерево рішень має найбільше значення R^2 і найменші значення MAE , $RMSE$, $MAPE$ та $WAPE$.

За великих фактичних обсягів газоспоживання розкид точок збільшується для всіх трьох моделей. Це свідчить про наявність окремих тригодинних інтервалів зі значними абсолютними похибками. Саме тому значення $RMSE$ для всіх моделей перевищують MAE , під час обчислення $RMSE$ великі

відхилення мають більший вплив на підсумкове значення показника.

Зміна точності протягом року та за різних температур. Річні показники характеризують загальну точність моделей, але не відображають зміни похибки в окремі періоди року. Тому для кожного місяця 2019 року окремо обчислено $WAPE$ за всіма тригодинними інтервалами відповідного місяця. Отримані результати наведено на рис. 3.

Для MLP-5 значення $WAPE$ становило 3,19 % у січні та 10,56 % у серпні, після чого різко збільшилося: до 26,06 % у вересні та 29,70 % у жовтні. Для градієнтного бустингу у вересні та жовтні отримано відповідно 19,03 і 15,90 %, а для дерева рішень – 18,89 і 17,20 %. Отже, у вересні дерево рішень мало дещо меншу похибку, ніж градієнтний бустинг, тоді як у жовтні точнішим був градієнтний бустинг. Згідно з рис. 3, градієнтний бустинг також мав меншу похибку в листопаді. Проте протягом більшої частини року найменші значення $WAPE$ отримано для дерева рішень, що узгоджується з його найкращим річним значенням $WAPE$ на тестових даних.

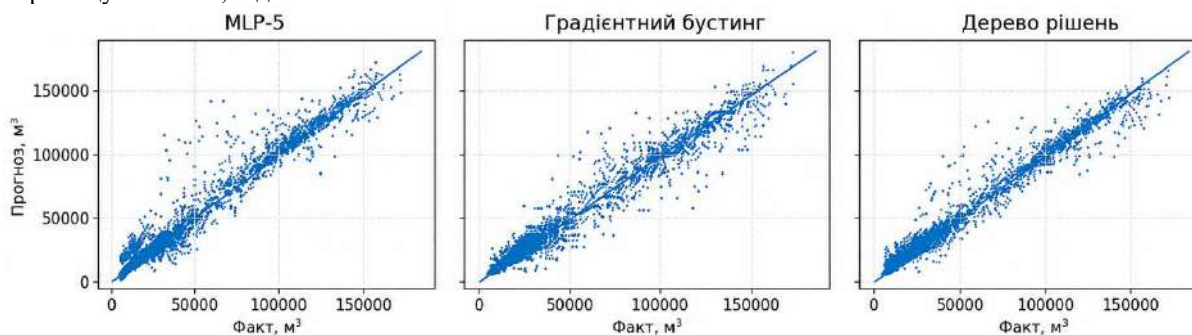


Рис. 2. Порівняння фактичних і прогнозованих значень газоспоживання на незалежному тестовому інтервалі 2019 року

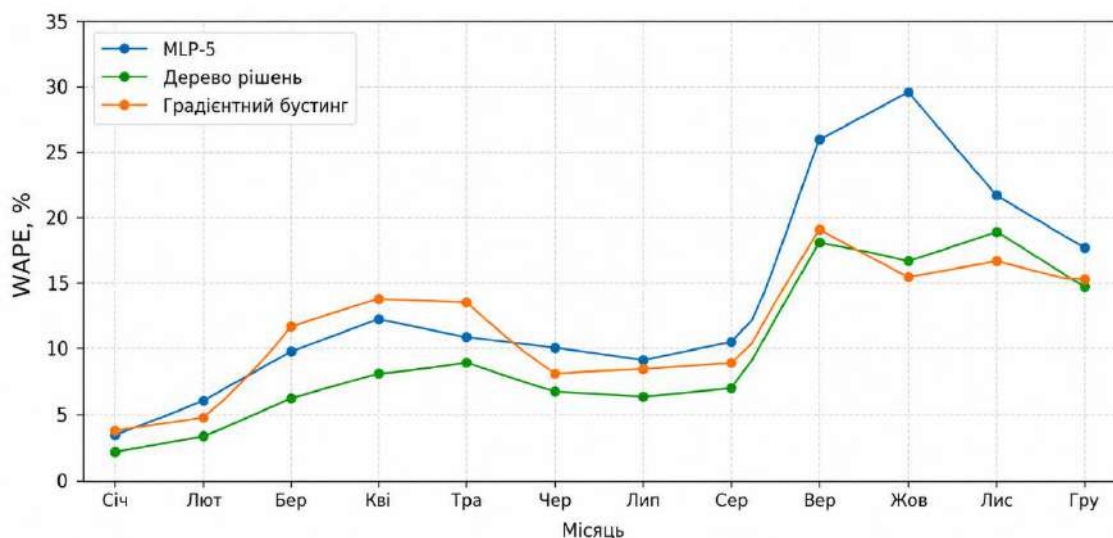
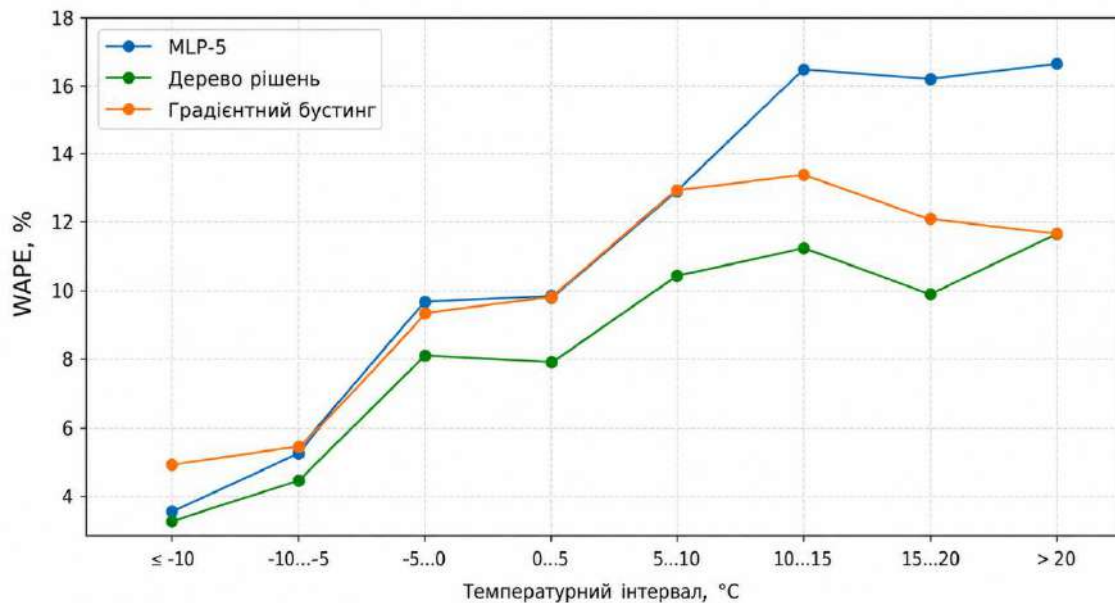


Рис. 3. Помісячні значення $WAPE$ моделей на тестових даних 2019 року

Рис. 4. Значення *WAPE* моделей за температурними інтервалами на тестових даних 2019 року

Найбільше зростання *WAPE* для всіх трьох моделей спостерігалося під час переходу від серпня до вересня. Це свідчить про те, що перехід від відносно низького й стабільного літнього газоспоживання до його осіннього зростання був складнішим для прогнозування, ніж періоди з менш різкими змінами обсягів споживання.

Для доповнення помісячного аналізу тестові спостереження також згруповано за температурними інтервалами. До кожної групи входили тригодинні спостереження з відповідного діапазону температури незалежно від місяця, у якому вони були зафіксовані. Для кожної температурної групи окремо обчислено *WAPE* трьох моделей. Результати наведено на рис. 4.

На відміну від помісячного порівняння на рис. 3, рис. 4 показує зміну похибки прогнозування залежно від температури повітря. Таке групування дає змогу порівняти точність моделей за однакових температурних умов, об'єднавши спостереження з різних місяців.

У діапазоні температур від 10 до 15 °C значення *WAPE* становило 16,47 % для MLP-5, 13,40 % для градієнтного бустингу та 11,31 % для дерева рішень. За температур від 15 до 20 °C відповідні значення становили 16,18, 12,12 і 9,91 %. Отже, у цих температурних інтервалах найменшу похибку мало дерево рішень, а найбільшу – MLP-5. Температури від 10 до 20 °C часто спостерігаються у перехідні періоди року, коли характер газоспоживання змінюється. За таких умов відмінності між точністю моделей стають особливо помітними.

Водночас результати на рис. 4 не доводять, що перевага дерева рішень зумовлена лише температурою повітря.

Температура є однією із 47 вхідних ознак, тому точність прогнозування визначається спільним використанням метеорологічних, часових, лагових, ковзних і різницевих ознак.

Результати статистичного порівняння моделей. Значення *MAE*, *RMSE* і *WAPE* характеризують величину похибок, але самі по собі не дають змоги встановити, чи є відмінності між точністю моделей статистично значущими. Для цього використано результати попарного двостороннього тесту Дібольда-Маріано [25], методику якого описано вище.

Для кожної пари моделей тест виконували окремо для квадратичної та абсолютної функції втрат. У табл. 3 використано такі позначення: DM_{kv} – значення статистики Дібольда-Маріано для квадратичної функції втрат; p_{kv} – відповідне p -значення; DM_{abs} – значення статистики Дібольда-Маріано для абсолютної функції втрат; p_{abs} – відповідне p -значення.

Оскільки різницю функцій втрат обчислювали як значення функції втрат першої моделі в парі мінус значення функції втрат другої, додатне значення статистики DM означає, що перша модель має більше середнє значення функції втрат. За додатного значення статистики DM і $p < 0,05$ перевагу другої моделі вважали статистично підтвердженою.

Останній стовпець таблиці містить 95 % довірчий інтервал різниці *WAPE* між моделями. Різницю наведено у відсоткових пунктах, що в таблиці позначено скороченням «п.п.».

За прийнятого рівня значущості 0,05 значення $p < 0,05$ вважали підставою для відхилення припущення про однакову прогнозу точність моделей.

Для пари «MLP-5 – дерево рішень» отримано додатні значення статистики Дібольда-Маріано: $DM_{kv}=3,47$ для квадратичної та $DM_{abs}=6,47$ для абсолютної функції втрат. В обох випадках отримано $p < 0,001$. Крім того, 95 % довірчий інтервал різниці *WAPE* становить [1,69; 3,45] п.п. і повністю розташований вище нуля. Оскільки $\Delta WAPE$

Таблиця 3

Попарна статистична перевірка прогнозової точності

Пара моделей	$DM_{кв}$	$p_{кв}$	$DM_{абс}$	$p_{абс}$	95 % довірчий інтервал $\Delta WAPE$ п.п.
MLP-5 – дерево рішень	3,47	<0,001	6,47	<0,001	[1,69; 3,45]
Гradientний бустинг – дерево рішень	3,96	<0,001	4,86	<0,001	[1,25; 2,44]
MLP-5 – gradientний бустинг	1,54	0,123	1,24	0,215	[-0,35; 1,87]

визначено як $WAPE$ першої моделі мінус $WAPE$ другої, отримані результати підтверджують меншу похибку дерева рішень порівняно з MLP-5.

Для пари «gradientний бустинг - дерево рішень» також отримано додатні значення статистики Дібольда-Маріано: $DM_{кв}=3,96$ і $DM_{абс}=4,86$. Для обох функцій втрат значення $p<0,001$ а 95 % довірчий інтервал різниці $WAPE$ становить [1,25; 2,44] п.п. і не містить нуля. Отже, дерево рішень має статистично підтверджену перевагу над gradientним бустингом.

Інший результат отримано для пари «MLP-5 – gradientний бустинг». Значення статистики становлять $DM_{кв}=1,54$ і $DM_{абс}=1,24$ а відповідні p -значення – $p_{кв}=0,123$ і $p_{абс}=0,215$. Обидва p -значення перевищують 0,05. Крім того, 95 % довірчий інтервал різниці $WAPE$ [-0,35; 1,87] п.п. містить нуль. Отже, за використаною процедурою статистичної перевірки немає достатніх підстав стверджувати про перевагу MLP-5 або gradientного бустингу одна над одною.

Результати оцінювання важливості вхідних ознак. Оскільки дерево рішень показало найвищу точність на незалежних тестових даних, результати перестановочної оцінки важливості наведено саме для цієї моделі. Ранжування ознак за середнім зменшенням коефіцієнта R^2 представлено на рис. 5.

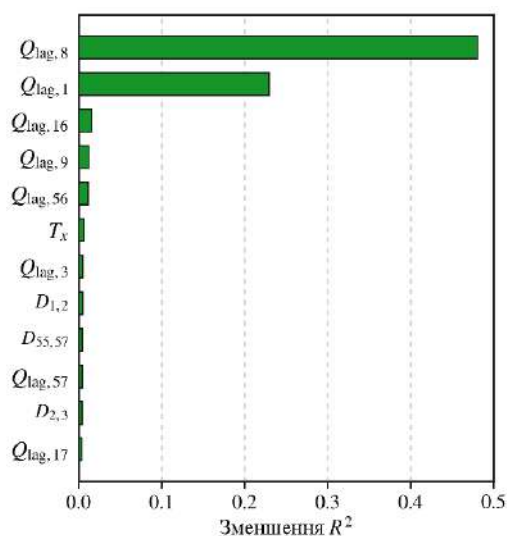


Рис. 5. Перестановочна важливість вхідних ознак для моделі дерева рішень

Найбільше середнє зменшення R^2 отримано після перестановки ознак $Q_{lag,8}$ і $Q_{lag,1}$: відповідно на 0,48059 і 0,22948. Ознака $Q_{lag,8}$ відповідає обсягу газоспоживання в тому самому тригодинному інтервалі попередньої доби, а $Q_{lag,1}$ – обсягу газоспоживання за попередні три години.

Значно менші значення перестановочної важливості отримано для ознак $Q_{lag,16}$, $Q_{lag,9}$ і $Q_{lag,56}$ відповідно 0,01583, 0,01202 і 0,01150. Вони відповідають значенням газоспоживання, зафіксованим відповідно 48, 27 і 168 годин до прогнозованого інтервалу. Серед метеорологічних ознак найбільшу перестановочну важливість мала максимальна температура повітря T_x , однак її перестановочна важливість є меншою, ніж важливість п'яти провідних лагових ознак.

Отже, прогноз дерева рішень найбільше залежить від попередніх значень самого газоспоживання, насамперед від значення за одну добу та за три години до прогнозованого інтервалу.

Обговорення результатів. Проведений експеримент підтвердив отриманий у роботі [16] результат щодо високої точності дерева рішень при прогнозуванні газоспоживання. У цьому дослідженні всі три моделі отримували однаковий набір із 47 вхідних ознак, а їх навчання, валідацію та тестування виконували на однакових часових інтервалах. За таких умов найвище значення R^2 та найнижчі значення MAE , $RMSE$, $MAPE$ і $WAPE$ отримано для окремого налаштованого дерева рішень, а не для ансамблевого методу gradientного бустингу.

На тестових даних 2019 року дерево рішень забезпечило $R^2=0,9704$, $MAE=4233,03\text{ м}^3$, $RMSE=7380,45\text{ м}^3$, $MAPE=12,99\%$ і $WAPE=8,48\%$. Для gradientного бустингу значення $WAPE$ становило 10,25 %, а для MLP-5 – 10,94 %. Порівняно з MLP-5 дерево рішень зменшило MAE на 22,44%, $RMSE$ – на 21,99 %, а $WAPE$ – на 22,49 %. Порівняно з gradientним бустингом зменшення цих показників становило відповідно 17,23%, 12,67% і 17,27%.

Отриманий результат не означає, що модель на основі дерева рішень загалом є точнішою за gradientний бустинг або MLP у будь-якій задачі прогнозування газоспоживання. Сформульований висновок стосується конкретного набору даних,

сформованого набору з 47 вхідних ознак, вибраних конфігурацій моделей і застосованої процедури навчання, валідації та тестування.

Важливим результатом є те, що відносно просте дерево рішень змогло ефективно використати сформований набір вхідних ознак, зокрема попередні значення газоспоживання, ознаки його добової і тижневої повторюваності, різниці між попередніми значеннями та метеорологічні характеристики. На відміну від градієнтного бустингу, дерево рішень не є ансамблевою моделлю: прогноз формується одним деревом унаслідок послідовного поділу простору вхідних ознак за визначеними пороговими значеннями. Завдяки цьому структуру моделі та вплив окремих ознак можна інтерпретувати безпосередніше, ніж у випадку ансамблю з великої кількості дерев.

Градієнтний бустинг, навпаки, є ансамблевим методом, у якому кожне наступне дерево доповнює попередні дерева та спрямоване на зменшення залишкової похибки [23]. Проте в проведеному експерименті така складніша структура не забезпечила підвищення точності порівняно з одним налаштованим деревом рішень.

Отримані результати також підтверджують важливість обґрунтованого формування вхідних ознак. Сучасні гібридні моделі прогнозування [8–13] використовують складніші способи оброблення даних та архітектури. Однак їх доцільно порівнювати не лише з простими базовими прогнозами, а й із налаштованими моделями машинного навчання, які отримують однаковий набір вхідної інформації. Таке порівняння дає змогу відокремити підвищення точності, зумовлене складнішою структурою моделі, від ефекту належно сформованих вхідних ознак.

Обмеження дослідження. Дослідження має кілька обмежень. Тестування запропонованого підходу виконано для одного локального процесу газоспоживання та одного календарного року, що певною мірою обмежує можливість узагальнення отриманих результатів. Розширення експериментальної бази за рахунок даних інших регіонів України в сучасних умовах є ускладненим, оскільки інформація про газоспоживання належить до стратегічно важливих даних і має обмежений доступ. У подальших дослідженнях планується перевірити запропонований підхід на даних інших регіонів, а також, за можливості, використати відкриті дані про газоспоживання інших країн і провести тестування для кількох послідовних років. Це дасть змогу оцінити стійкість, узагальнюваність та переносимість моделей за різних кліматичних, територіальних і споживчих умов.

Крім того, у ретроспективному експерименті для кожного прогнозованого тригодинного інтервалу використовували фактичні метеорологічні спостереження. У реальній системі прогнозування такі значення наперед невідомі, тому на практиці необхідно використовувати прогнозні метеорологічні дані. Похибка прогнозу температури,

вологості, швидкості вітру, атмосферного тиску та інших метеорологічних параметрів може додатково впливати на точність прогнозування газоспоживання. Тому одним із напрямів подальших досліджень є оцінювання точності моделей із використанням саме прогнозних, а не фактичних метеорологічних даних. Для річного горизонту прогнозування газоспоживання доцільно або окремо формувати прогноз відповідних кліматичних показників, або використовувати сезонні та довгострокові прогнозні метеорологічні дані з авторитетних відкритих джерел, зокрема Європейського центру середньострокових прогнозів погоди (ECMWF).

Окремим обмеженням є застосування перестановочної оцінки важливості ознак. За наявності взаємопов'язаних або корельованих факторів, які містять подібну інформацію, їхня важливість може частково розподілятися між кількома ознаками. Тому отримане ранжування слід розглядати як характеристику конкретної навченої моделі та її реакції на зміну вхідних ознак, а не як безпосередню оцінку причинного впливу окремих факторів на обсяг газоспоживання.

Висновки. За однакового набору з 47 вхідних ознак та однакового поділу даних на навчальний, валідаційний і тестовий періоди найвищу точність на незалежних даних 2019 року забезпечило дерево рішень. Для цієї моделі отримано $R^2=0,9704$, $MAE=4233,03\text{ м}^3$, $RMSE=7380,45\text{ м}^3$, $MAPE=12,99\%$ і $WAPE=8,48\%$. Для градієнтного бустингу значення $WAPE$ становило 10,25 %, а для MLP-5 – 10,94 %.

Статистична перевірка підтвердила перевагу дерева рішень над MLP-5 і градієнтним бустингом. Для обох порівнянь у тесті Дібольда-Маріано отримано $p<0,001$ як для квадратичної, так і для абсолютної функції втрат. Відповідні 95 % довірчі інтервали різниці $WAPE$ не містили нуля. Водночас статистично значущої різниці між MLP-5 і градієнтним бустингом не виявлено: $p=0,123$ для квадратичної та $p=0,215$ для абсолютної функції втрат, а 95 % довірчий інтервал $\Delta WAPE[-0,35;1,87]$ процентного пункту містив нуль.

Найбільше зростання похибок усіх моделей спостерігалось в осінній період, коли після літнього мінімуму газоспоживання починало швидко збільшуватися. Аналіз за температурними інтервалами також показав підвищення похибок у перехідних температурних умовах, особливо для моделі MLP-5.

Перестановочна оцінка важливості показала, що найбільший вплив на прогноз дерева рішень мали ознаки $Q_{lag,8}$ і $Q_{lag,1}$. Перша ознака відповідає значенню газоспоживання у тому самому тригодинному інтервалі попередньої доби, а друга – значенню за попередні три години. Отриманий результат підтверджує важливість урахування попередніх значень газоспоживання при формуванні вхідних даних прогнозної моделі.

Подальші дослідження доцільно спрямувати на перевірку результатів прогнозування на даних інших регіонів і для кількох тестових років, використання прогнозованих метеорологічних даних замість фактичних, а також порівняння прогнозування вихідного ряду газоспоживання з прогнозуванням його циклічної компоненти, отриманої на основі моделі циклічного випадкового процесу. Це дасть змогу обґрунтовано визначити, чи забезпечують складніші гібридні моделі додаткове підвищення точності порівняно з простішими моделями, зокрема з налаштованим деревом рішень.

Література

- Liu J., Wang S., Wei N., Chen X., Xie H., Wang J. Natural gas consumption forecasting: A discussion on forecasting history and future challenges // *Journal of Natural Gas Science and Engineering*. 2021. Vol. 90. Article 103930. DOI: <https://doi.org/10.1016/j.jngse.2021.103930>
- Šebalj D., Mesarić J., Dujak D. Analysis of methods and techniques for prediction of natural gas consumption: A literature review // *Journal of Information and Organizational Sciences*. 2019. Vol. 43, No. 1. P. 99–117. DOI: <https://doi.org/10.31341/jios.43.1.6>
- Szoplik J. Forecasting of natural gas consumption with artificial neural networks // *Energy*. 2015. Vol. 85. P. 208–220. DOI: <https://doi.org/10.1016/j.energy.2015.03.084>
- Merkel G. D., Povinelli R. J., Brown R. H. Short-Term Load Forecasting of Natural Gas with Deep Neural Network Regression // *Energies*. 2018. Vol. 11, No. 8. Article 2008. DOI: <https://doi.org/10.3390/en11082008>
- Laib O., Khadir M. T., Mihaylova L. Toward efficient energy systems based on natural gas consumption prediction with LSTM Recurrent Neural Networks // *Energy*. 2019. Vol. 177. P. 530–542. DOI: <https://doi.org/10.1016/j.energy.2019.04.075>
- Sharma V., Cali Ü., Sardana B., Kuzlu M., Banga D., Pipattanasomporn M. Data-driven short-term natural gas demand forecasting with machine learning techniques // *Journal of Petroleum Science and Engineering*. 2021. Vol. 206. Article 108979. DOI: <https://doi.org/10.1016/j.petrol.2021.108979>
- Tian N., Shao B., Bian G., Zeng H., Li X., Zhao W. Application of forecasting strategies and techniques to natural gas consumption: A comprehensive review and comparative study // *Engineering Applications of Artificial Intelligence*. 2024. Vol. 129. Article 107644. DOI: <https://doi.org/10.1016/j.engappai.2023.107644>
- Zhao M., Guo G., Fan L., Han L., Yu Q., Wang Z. Short-term natural gas load forecasting based on EL-VMD-Transformer-ResLSTM // *Scientific Reports*. 2024. Vol. 14. Article 20343. DOI: <https://doi.org/10.1038/s41598-024-70384-9>
- Lin Z., Xie L., Zhang S. A compound framework for short-term gas load forecasting combining time-enhanced perception transformer and two-stage feature extraction // *Energy*. 2024. Vol. 298. Article 131365. DOI: <https://doi.org/10.1016/j.energy.2024.131365>
- Wang H., Gao X., Zhang Y., Yang Y. Natural Gas Consumption Forecasting Model Based on Feature Optimization and Incremental Long Short-Term Memory // *Sensors*. 2025. Vol. 25, No. 10. Article 3079. DOI: <https://doi.org/10.3390/s25103079>
- Xu X., Xu H., Mei K., Tong L., Liu Z., Wang T., Fang K. Integrating multi-modal data into transformer model for short-term gas consumption forecasting // *Alexandria Engineering Journal*. 2025. Vol. 122. P. 655–664. DOI: <https://doi.org/10.1016/j.aej.2025.03.029>
- Niu P., Du J., Xu N., Wang B., Liao Q., Qiu R., Cai S., Liang Y. P-KTFNet: A prior knowledge enhanced time-frequency forecasting model for natural gas consumption // *Energy*. 2025. Vol. 328. Article 136559. DOI: <https://doi.org/10.1016/j.energy.2025.136559>
- Qiao W., Ma Q., Yang Y., Xi H., Huang N., Yang X. et al. Natural gas consumption forecasting using a novel two-stage model based on improved sparrow search algorithm // *Journal of Pipeline Science and Engineering*. 2025. Vol. 5, No. 1. Article 100220. DOI: <https://doi.org/10.1016/j.jpse.2024.100220>
- Lytvynenko I., Lupenko S., Nazarevych O., Shymchuk H., Hotovych V. Additive mathematical model of gas consumption process // *Scientific Journal of TNTU*. 2021. Vol. 104, No. 4. P. 87–97. DOI: https://doi.org/10.33108/visnyk_tntu2021.04.087
- Lytvynenko I., Lupenko S., Kunanets N., Nazarevych O., Shymchuk G., Hotovych V. Simulation of gas consumption process based on the mathematical model in the form of cyclic random process considering the scale factors // *CEUR Workshop Proceedings*. 2021. Vol. 3039. P. 97–106. URL: <https://ceur-ws.org/Vol-3039/paper23.pdf>
- Shymchuk G., Lytvynenko I., Hromyak R., Lytvynenko S., Hotovych V. Gas Consumption Forecasting Using Machine Learning Methods and Taking Into Account Climatic Indicators // *CEUR Workshop Proceedings*. 2023. Vol. 3468. P. 156–163. URL: <https://ceur-ws.org/Vol-3468/short8.pdf>
- Bergmeir C., Benítez J. M. On the use of cross-validation for time series predictor evaluation // *Information Sciences*. 2012. Vol. 191. P. 192–213. DOI: <https://doi.org/10.1016/j.ins.2011.12.028>
- Tashman L. J. Out-of-sample tests of forecasting accuracy: an analysis and review // *International Journal of Forecasting*. 2000. Vol. 16, No. 4. P. 437–450. DOI: [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)
- Nair V., Hinton G. E. Rectified Linear Units Improve Restricted Boltzmann Machines // *Proceedings of the 27th International Conference on Machine Learning*. 2010. P. 807–814. URL: <https://proceedings.mlr.press/v9/nair10a.html>
- Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization // *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. 2015. URL: <https://arxiv.org/abs/1412.6980>
- Breiman L., Friedman J. H., Olshen R. A., Stone C. J. *Classification and Regression Trees*. Belmont, CA : Wadsworth, 1984.
- Pedregosa F., Varoquaux G., Gramfort A. et al. *Scikit-learn: Machine Learning in Python* // *Journal of Machine*

- Learning Research. 2011. Vol. 12. P. 2825–2830. URL: <https://jmlr.org/papers/v12/pedregosa11a.html>
23. Friedman J. H. Greedy Function Approximation: A Gradient Boosting Machine // The Annals of Statistics. 2001. Vol. 29, No. 5. P. 1189–1232. DOI: <https://doi.org/10.1214/aos/1013203451>
 24. Hyndman R. J., Koehler A. B. Another look at measures of forecast accuracy // International Journal of Forecasting. 2006. Vol. 22, No. 4. P. 679–688. DOI: <https://doi.org/10.1016/j.ijforecast.2006.03.001>
 25. Diebold F. X., Mariano R. S. Comparing predictive accuracy // Journal of Business & Economic Statistics. 1995. Vol. 13, No. 3. P. 253–263. DOI: <https://doi.org/10.1080/07350015.1995.10524599>
 26. Altmann A., Toloşi L., Sander O., Lengauer T. Permutation importance: a corrected feature importance measure // Bioinformatics. 2010. Vol. 26, No. 10. P. 1340–1347. DOI: <https://doi.org/10.1093/bioinformatics/btq134>

Shymchuk G., Lytvynenko I. Natural gas consumption forecasting using machine learning methods: MLP, decision tree, and gradient boosting

The paper investigated the accuracy of natural gas consumption forecasting using machine learning methods: multilayer perceptron (MLP), decision tree, and gradient boosting. The forecast was formed for consecutive three-hour intervals, and the accuracy of the selected models was evaluated over the full annual test interval of 2019. The main focus was not only on the choice of algorithm, but also on the method of presenting information about the previous dynamics of gas consumption. The same set of 47 input features was used for all models. It included 16 meteorological, six cyclic time, 11 lag, nine moving, and five difference features. This made it possible to compare the forecasting capabilities of algorithms with the same composition of input features. Data from previous years were used to build and tune the models, and 2019 was left for independent verification. The results showed that the decision tree forms the highest forecasting accuracy over the annual test interval. Gradient boosting and MLP also reproduce the general dynamics of gas consumption quite accurately, but are more likely to make mistakes in areas with

rapid regime changes. The largest discrepancies between the models are observed in the fall, when gas consumption goes from a low summer level to a level typical of the heating season. Analysis of the importance of input features showed that the forecast is determined to the greatest extent by the previous values of gas consumption, while meteorological and calendar characteristics make it possible to take into account the influence of weather conditions, seasonality and time patterns of the process. The results obtained confirm that the forecast accuracy depends not only on the choice of model, but also on how fully the time structure of the process is reflected in the input data. Statistical testing showed that the decision tree has a statistically confirmed advantage over MLP and gradient boosting, i.e. the detected differences in forecast accuracy cannot be explained only by the random variability of errors in the test sample. The practical significance of the work lies in the possibility of using the obtained results to build systems for operational forecasting of gas consumption in conditions of seasonal variability of demand and weather changes.

Keywords: natural gas consumption, forecasting, machine learning methods, MLP, decision tree, gradient boosting, time series.

Шимчук Григорій Валерійович – старший викладач кафедри комп'ютерних наук Тернопільського національного технічного університету імені Івана Пулюя. <https://orcid.org/0000-0003-2362-7386>
e-mail: gorych@gmail.com

Литвиненко Ярослав Володимирович – д-р техн. наук, професор, професор кафедри комп'ютерних наук Тернопільського національного технічного університету імені Івана Пулюя.
<https://orcid.org/0000-0001-7311-4103>
e-mail: lytvynenko@tntu.edu.ua

Дата першого надходження статті 03.08.2026.

Дата прийняття статті до друку після рецензування 26.08.2026.

Дата публікації 30.09.2026.



Стаття з відкритим доступом,
відповідно до умов ліцензії
[Creative Commons \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)