«Системні технології» 4 (165) 2026 «System technologies»
DOI 10.34185/1562-9945-4-165-2026-01
УДК 004.8:622.7

Б.Г. Кісельов, А.О. Сенько, А.І. Купін, Д.К. Балик

ЕМПІРИЧНЕ ВИЗНАЧЕННЯ МІНІМАЛЬНО ДОСТАТНЬОГО ОБСЯГУ НАВЧАЛЬНОЇ ВИБІРКИ ДЛЯ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ЗА ЗАДАНОГО РІВНЯ ПОХИБКИ

Анотація. Статтю присвячено задачі емпіричного визначення мінімально достатнього обсягу навчальної вибірки для регресійних моделей машинного навчання у системах сенсорного сортування руд. Актуальність теми зумовлена значними витратами на формування репрезентативних наборів даних у гірничодобувній промисловості. Постановка проблеми полягає в необхідності переходу від якісного аналізу кривої навчання до кількісної оцінки достатнього розміру вибірки за заданого рівня похибки. Метою дослідження є перевірка ієрархії підходів: крива навчання – параметрична степенева екстраполяція – GP-based learning-type curve. Методи: 10-fold GroupKFold крос-валідація, байєсівський підбір гіперпараметрів (Optuna), нелінійна регресія, гауссівські процеси. Отримано: $R^2 = 0{,}93$ на тестових фолдах; мінімально достатній обсяг вибірки для RMSE ≤ 12 оцінено в діапазоні 559–810 спостережень. Ключовий висновок: запропонована методика дозволяє обґрунтовано визначати поріг, після якого подальше розширення вибірки перестає давати практичний ефект.

Ключові слова: машинне навчання; крива навчання; мінімальний обсяг вибірки; степенева апроксимація; гауссівський процес; сортування руд; крос-валідація; HistGradientBoosting; Neural Scaling Laws; екстраполяція.

Постановка проблеми. У задачах машинного навчання не існує універсальної відповіді на питання про те, яким має бути достатній обсяг навчальної вибірки для досягнення належної якості моделі. Необхідний розмір вибірки визначається не лише видом моделі, а й складністю задачі, структурою даних, співвідношенням між корисним сигналом і шумом, вибраною метрикою якості, а також наявністю неусувної похибки. У прикладних умовах проблема додатково ускладнюється тим, що формування нових даних, їх підготовка, верифікація та подальший аналіз потребують значних часових, фінансових і організаційних витрат. З цієї причини питання визначення достатнього обсягу навчальної вибірки має не лише теоретичне, а й безпосереднє практичне значення.

Для задачі сортування руд ця проблема є особливо актуальною. Формування репрезентативної навчальної вибірки пов’язане з необхідністю отримання значної кількості спостережень для різних типів матеріалу, проведення супровідних лабораторних та інструментальних досліджень, а також використання спеціалізованих засобів вимірювання. До складу таких даних можуть входити: результати хімічного аналізу,

© Кісельов Б.Г., Сенько А.О., Купін А.І., Балик Д.К., 2026
Copyright for this paper by its authors. Use permitted under License CC BY 4.0.
ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

магнітні характеристики, геометричні параметри, температурні показники сенсорів та інші спостережувані величини. Саме тому в цій предметній області недостатньо виходити з формального припущення, що збільшення кількості даних завжди є безумовно корисним: необхідно обґрунтовано визначити, на якому етапі подальше нарощування вибірки перестає давати суттєвий практичний ефект.

Недостатній обсяг навчальної вибірки може призводити до нестійкого навчання, підвищеної варіативності результатів та погіршення узагальнювальної здатності моделі навіть у тих випадках, коли сам клас моделей є загалом придатним до поставленої задачі. Разом із тим просте збільшення кількості спостережень не гарантує пропорційного покращення якості. Якщо основне обмеження пов'язане з неусувною похибкою, шумом вимірювань або слабкою інформативністю ознак, то додаткові приклади можуть забезпечувати лише незначний виграш. Отже, практичний інтерес становить визначення мінімально достатнього розміру вибірки за заданого рівня похибки або за критерієм малопомітного приросту якості при подальшому розширенні даних [1].

У зв'язку з цим виникає потреба в підході, який би дозволяв на основі вже наявної, відносно невеликої вибірки емпірично оцінити, як змінюється похибка моделі зі зростанням кількості навчальних прикладів, і на цій основі прогнозувати доцільність подальшого розширення набору даних. Такий підхід має бути орієнтований не лише на констатацію поточного рівня якості, а й на виявлення тієї межі, після якої приріст ефективності стає малопомітним. Саме така постановка є найбільш релевантною для прикладних систем сортування руд, у яких вартість отримання додаткових спостережень є суттєвою, а рішення щодо необхідності розширення вибірки повинно бути методично обґрунтованим, а не інтуїтивним.

Аналіз останніх досліджень і публікацій. Для дослідження залежності якості моделі від обсягу навчальної вибірки традиційно використовують криві навчання (learning curves). Комплексний огляд форм кривих навчання та методів їх аналізу представлено у роботі [2], де показано, що для більшості прикладних задач спадання похибки зі збільшенням N задовільно описується степеневою залежністю.

Питання практичного планування розміру вибірки для задач класифікації та прогнозування досліджувалось у ряді робіт. Figueroa et al. [3] запропонували методику екстраполяції кривих навчання для бінарної класифікації медичних даних. Beleites et al. [4] провели систематичне дослідження залежності між розміром вибірки та якістю класифікатора у хімічній аналітиці. Vabalas et al. [5] проаналізували вплив обмеженого розміру вибірки на достовірність оцінок якості моделей машинного навчання та показали небезпеку завищення якості при недостатній кількості даних.

Сучасним розвитком параметричної екстраполяції є GP-based learning-type curves, у яких детермінований скелет (степенева функція) доповнюється гауссівським процесом. Такий підхід дозволяє не лише оцінити центральну траєкторію зміни похибки, а й кількісно охарактеризувати невизначеність прогнозу при виході за межі наявного діапазону обсягів вибірки [6].

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Важливим теоретичним підґрунтям є концепція Neural Scaling Laws [7], згідно з якою похибка нейромережевих моделей змінюється зі зростанням обсягу даних за гладкою степеневою залежністю. У межах даної роботи ця концепція використовується не для прямого перенесення висновків, а для обґрунтування гіпотези про наявність степеневої масштабної закономірності у задачі сортування руд.

Слід відзначити, що сучасні методи автоматичного підбору гіперпараметрів [8] суттєво спрощують процес оптимізації моделі та дозволяють зосередити дослідницькі зусилля на аналізі поведінки якості залежно від обсягу даних. Алгоритм HistGradientBoosting [9], використаний у даному дослідженні, демонструє стабільну поведінку на різних обсягах вибірки завдяки вбудованій регуляризації та нативній обробці пропущених значень.

Разом із тим застосування зазначених підходів безпосередньо до задачі сортування руд на основі сенсорних даних у наявній літературі не розглядалось. Це визначає актуальність та наукову новизну даного дослідження.

Мета дослідження. Метою дослідження є розробка та практична перевірка методики емпіричного визначення мінімально достатнього обсягу навчальної вибірки для регресійних моделей машинного навчання у задачі сенсорного сортування руд.

Для досягнення поставленої мети вирішуються такі задачі:

– формування та попередня обробка навчального набору реальних сенсорних даних;

– навчання та оцінювання моделі в умовах групової крос-валідації;

– побудова кривої навчання та параметрична степенева екстраполяція;

– застосування GP-based learning-type curve для статистичного оцінювання невизначеності прогнозу;

– формулювання кількісних рекомендацій щодо достатнього обсягу вибірки.

Викладення основного матеріалу дослідження. Дослідження базується на реальних даних системи сенсорного сортування руд. Нижче послідовно описано набір даних, обрану модель, схему валідації, побудову кривої навчання та методи екстраполяції.

Набір даних та ознакова інженерія. Набір містить 699 спостережень без пропущених значень (Рисунок 1). Цільова змінна – KT10valueMax (максимальне значення вихідного сигналу котушки KT10), яка корелює з вмістом магнітного матеріалу у зразку руди.

Первинні вхідні ознаки, отримані безпосередньо із сенсорної системи:

– Volume – об'єм зразка;

– Area – площа проекції зразка;

– MaxHeight – максимальна висота зразка;

– CenterDiviation – відхилення центру мас;

– UsedCoilArea – частка задіяної площі індуктивної котушки;

– SensorTemp – температура сенсора;

– SensorValueRAW – сирий вихідний сигнал сенсора;

– SensorId – ідентифікатор сенсора.

Для збагачення ознакового простору синтезовано три похідні ознаки:

– AspectRatio = MaxHeight / (√Area + ε) – відношення висоти до кореня з площі, характеризує форму зразка;

– Density = Volume / (Area + ε) – об'ємна щільність зразка;

– NormDeviation = √(CenterDiviation / (√Area + ε)) – нормалізоване відхилення центру мас.

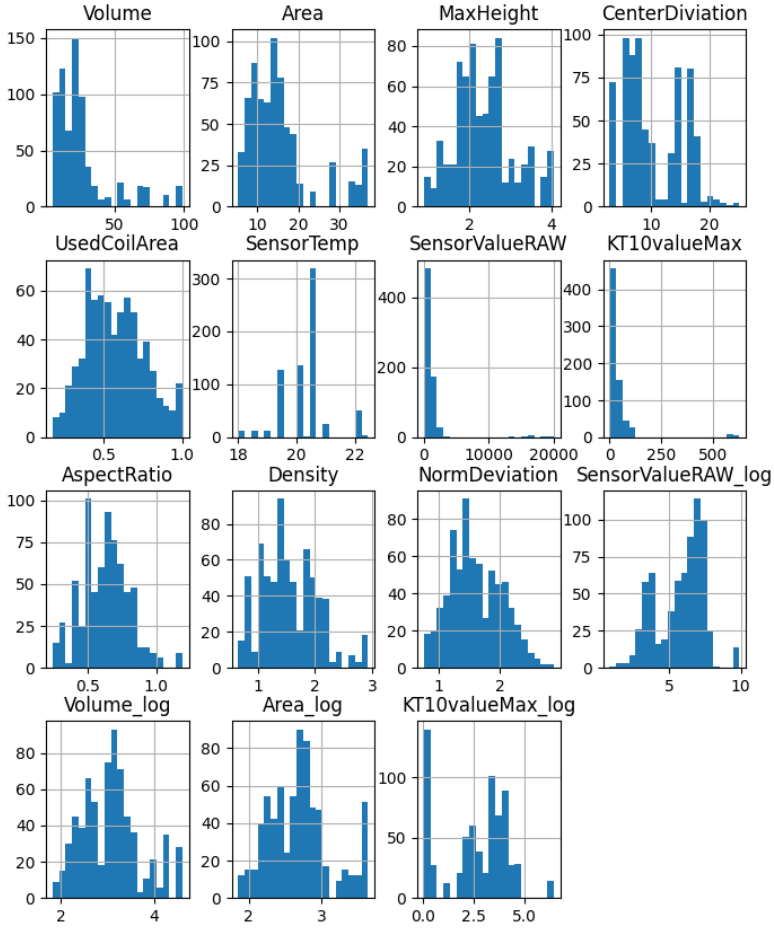


Рисунок 1 – Гістограми розподілу первинних та похідних ознак набору даних

Аналіз кореляційної матриці (Рисунок 2) виявив, що найбільш інформативною вхідною ознакою є SensorValueRAW (r = 0,985), далі – Volume (r = 0,358), Area (r = 0,313) та UsedCoilArea (r = 0,286). Температура сенсора та геометричні похідні ознаки мають низьку лінійну кореляцію з цільовою змінною.

Модель та підбір гіперпараметрів. Як основну модель обрано HistGradientBoostingRegressor у складі конвеєра з попереднім заповненням пропущених значень середнім (SimpleImputer) та логарифмічним перетворенням цільової змінної (TransformedTargetRegressor з func = log1p, inverse_func = expm1). Логарифмічне перетворення зменшує вплив правосторонньої асиметрії розподілу KT10valueMax та покращує стійкість навчання.

Підбір гіперпараметрів виконано за допомогою бібліотеки Optuna (100 спроб, байєсівська оптимізація, напрямок – мінімізація RMSE) у поєднанні з модифікованим правилом одного стандартного відхилення (1-SE rule): серед усіх конфігурацій у межах

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

порогу RMSE_best + SE обирається та, що має найкращий сумарний ранг за критеріями mean_RMSE, std_RMSE та mean_R^2. Знайдені оптимальні значення гіперпараметрів наведено нижче:

max_iter = 494
max_depth = 5
min_samples_leaf = 13
learning_rate = 0.1412
l2_regularization = 0.213
max_leaf_nodes = 91

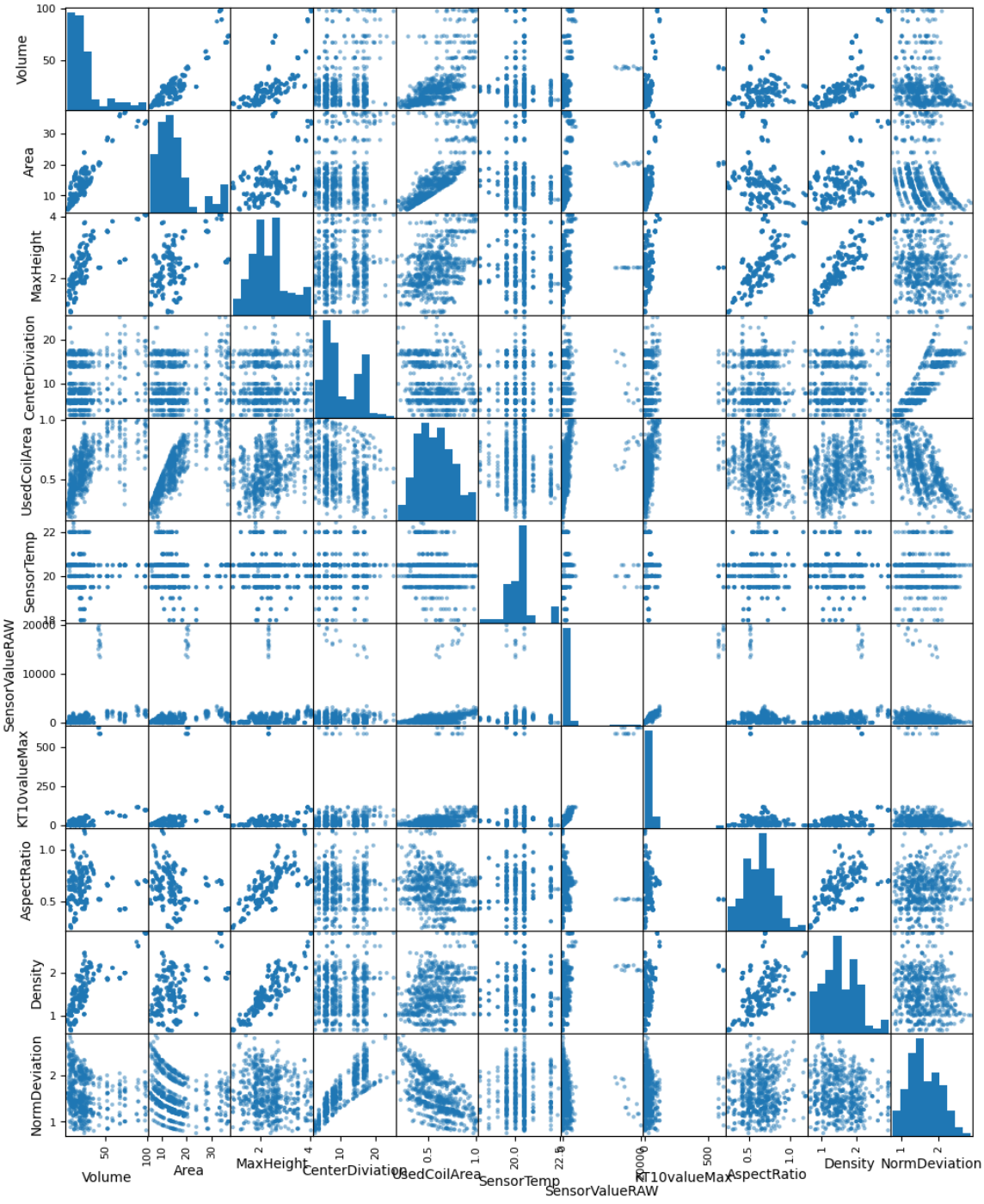


Рисунок 2 – Матриця кореляцій ознак із цільовою змінною KT10valueMax

Схема крос-валідації та загальна ефективність моделі. Для оцінювання моделі застосовано 10-fold GroupKFold крос-валідацію, де групою є мітка зразка (Label – тип руди). Така схема виключає витік інформації між спостереженнями одного типу матеріалу через межу навчальної та тестової підвибірок.

Результати крос-валідації подано в Таблиці 1.

Таблиця 1

Результати 10-fold GroupKFold крос-валідації (HistGradientBoostingRegressor)

Метрика	Значення по фолдах	Середнє
R^2	0,964; 0,841; 0,985; 0,992; 0,950; 0,948; 0,904; 0,858; 0,905; 0,973	0,932
RMSE	5,63; 9,43; 4,81; 16,14; 3,83; 4,56; 5,14; 4,98; 12,76; 27,72	9,50

Значення $R^2 > 0,93$ підтверджує високу пояснювальну здатність моделі. Суттєвий розкид RMSE між фолдами (від 3,8 до 27,7) відображає неоднорідність складності задачі для різних груп зразків, що є характерним для реальних промислових наборів даних.

Побудова кривої навчання. Криву навчання побудовано для 12 розмірів навчальної підвибірки від 20 % до 100 % загального обсягу з кроком, рівномірно розподіленим у діапазоні [0,2; 1,0]. Для кожного розміру застосовувалась та сама схема 10-fold GroupKFold, як метрику обрано RMSE на валідаційних фолдах. Обчислювались середнє та стандартне відхилення RMSE по фолдах.

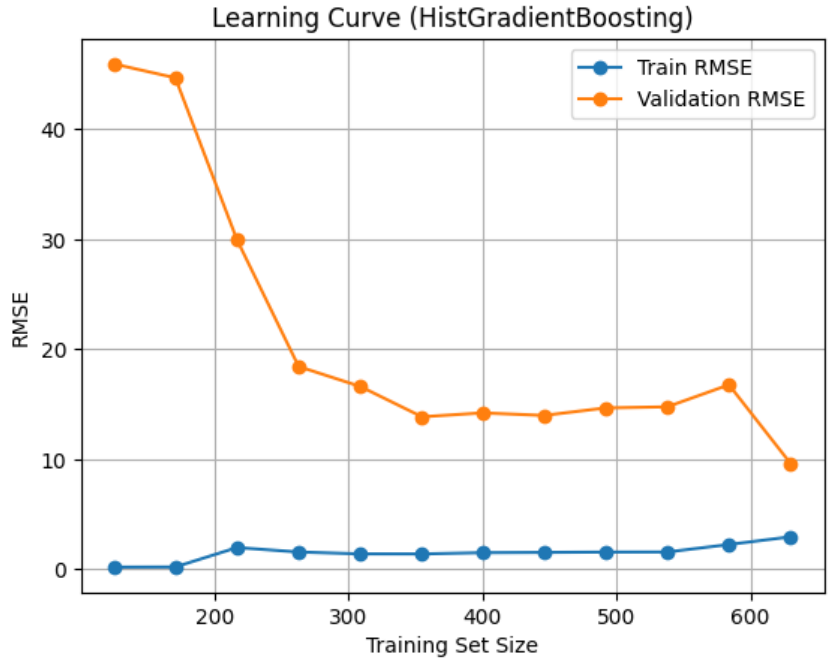


Рисунок 3 – Крива навчання моделі HistGradientBoosting

Спостерігається монотонне зниження RMSE зі збільшенням N із тенденцією до стабілізації у правій частині кривої – ознака входження в зону насичення. Навчальна RMSE суттєво нижча за валідаційну при малих N, що вказує на певний рівень перенавчання, який зменшується зі зростанням вибірки.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Параметрична степенева екстраполяція. Для апроксимації кривої навчання обрано степеневу функцію:

$$E(N) = a + b \cdot \left(\frac{N}{Nref}\right)^{-\gamma} \quad (1)$$

де E(N) – значення RMSE при обсязі вибірки N; γ – швидкість спадання похибки; b – масштаб початкового відхилення від асимптотичної межі; a – асимптотична RMSE, до якої прямує крива [1]; N_{ref} – медіана N (застосовується для нормування, що покращує числову стабільність підгонки).

Параметри підібрані методом нелінійної регресії (scipy.optimize.curve_fit) з ваговим урахуванням стандартного відхилення RMSE по фолдах (absolute_sigma=True). Результати підгонки наведено в Таблиці 2.

Таблиця 2

Параметри степеневої апроксимації кривої навчання та їх 95 % довірчі інтервали

Параметр	Значення	95 % ДІ
a	10,795	[−13,58; 35,17]
b	3,073	[−26,43; 32,58]
γ	2,802	[−10,72; 16,32]
Зважений R^2	0,743	–

Прогнозована асимптотична RMSE (параметр a = 10,795) задає нижню межу досяжної якості за наявного рівня шуму та неповноти ознак. Широкі 95 % ДІ параметрів зумовлені відносно малим числом точок кривої навчання та значним розкидом RMSE між фолдами, що підкреслює необхідність урахування параметричної невизначеності при інтерпретації прогнозів.

Точкові та ймовірнісні оцінки за методом NLS. На основі підігнаної функції (1) отримано такі точкові оцінки мінімально достатнього обсягу вибірки:

– N для RMSE ≤ 12: 559 спостережень;

– N для малопомітного приросту якості (виграш від подвоєння вибірки < 0,2 RMSE): 1004 спостереження.

Монте-Карло симуляція (n–mc = 3000 зразків із коваріаційної матриці параметрів, 729 валідних після відкидання фізично нереалістичних) дала ймовірнісні оцінки:

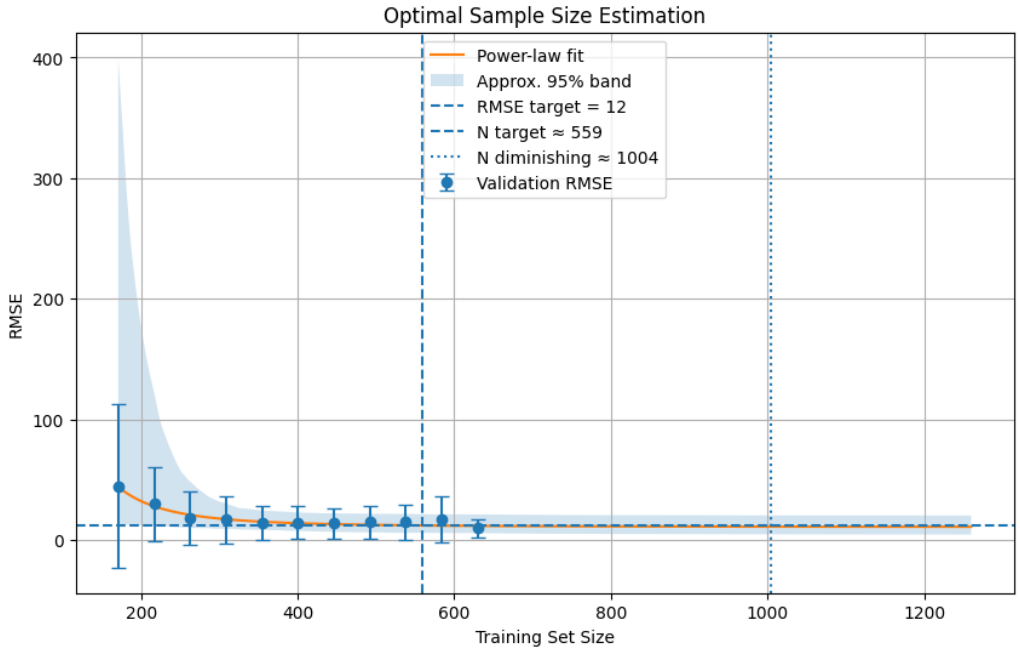
– P(RMSE ≤ 12) ≥ 50 %: N = 778;

– P(виграш від подвоєння < 0,2) ≥ 50 %: N = 1197.

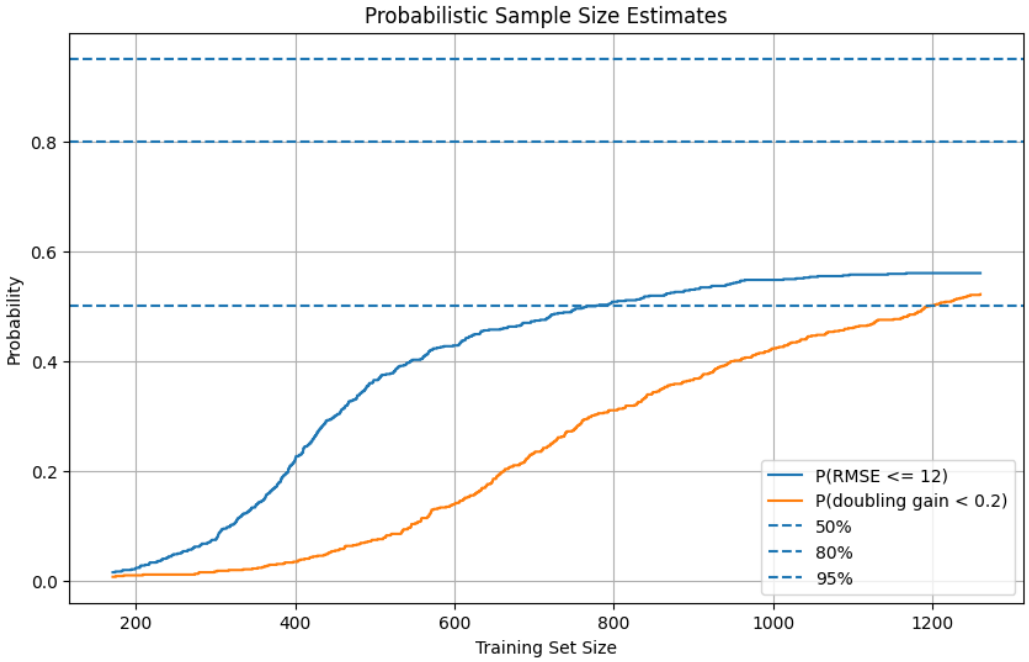
GP-based learning-type curve. Як більш просунутий статистичний підхід застосовано GP-based learning-type curve [6]. Залишки NLS-підгонки апроксимовано гауссівським процесом із ядром:

$$k(x, x') = C^2 \cdot RBF(\ell) + WhiteKernel(\sigma^2) \quad (2)$$

де вхід x = log(N). Підібрані параметри ядра: $C^2 = 0{,}0188^2$; $\ell = 0{,}955$; $\sigma^2 = 3{,}73 \times 10^{-4}$. Мале значення рівня шуму WhiteKernel підтверджує, що залишки NLS є структурованими, а не випадковими, – тобто детермінований скелет (1) вловлює основну тенденцію, а GP описує лише незначне систематичне відхилення.



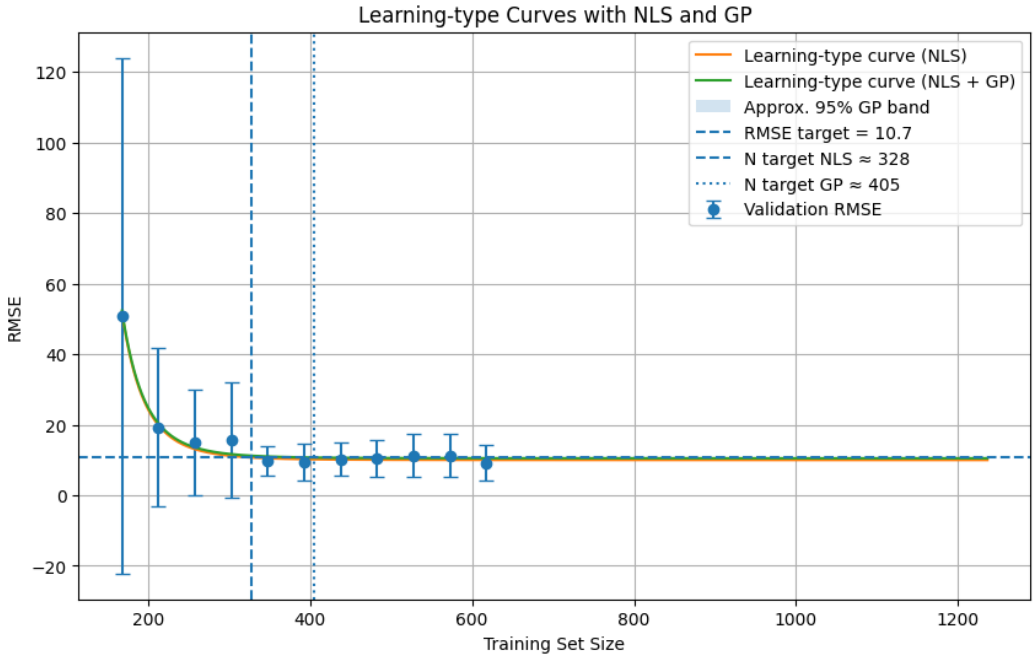
а



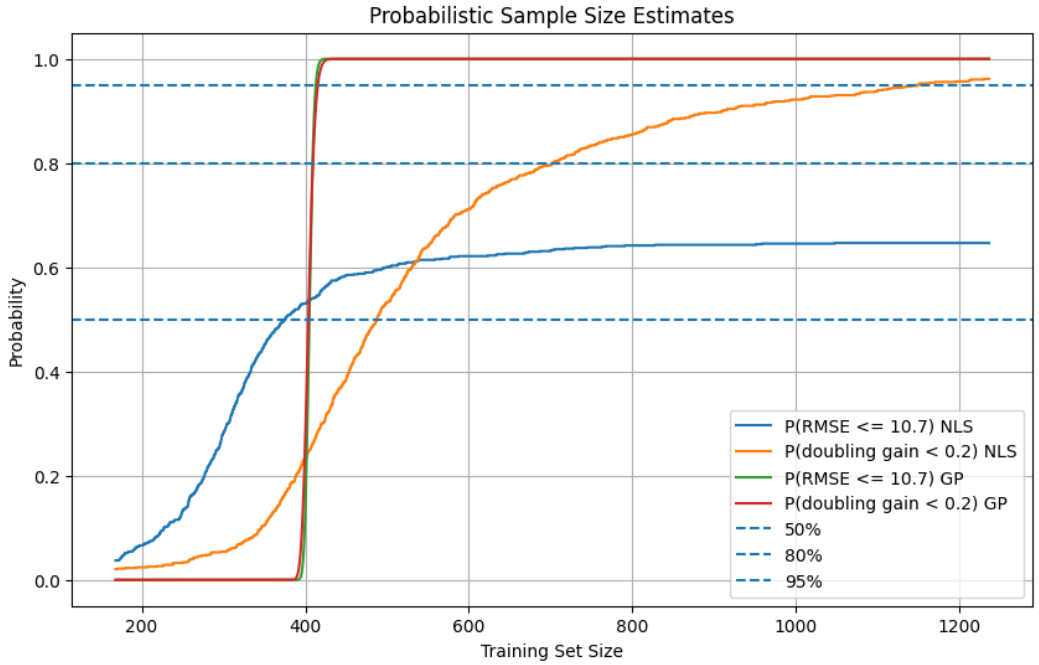
б

Рисунок 4 – Степенева екстраполяція кривої навчання з 95 ДІ (Монте-Карло)

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)



а



б

Рисунок 5 – NLS та NLS+GP криві навчання з довірчими смугами

Порівняння точкових та ймовірнісних оцінок за методами NLS та NLS+GP наведено у Таблиці 3.

Порівняння NLS та NLS+GP виявляє закономірну відмінність: GP-based підхід дає більш консервативну, але статистично обґрунтовану оцінку необхідного N для досягнення RMSE ≤ 12 (756 проти 559). Оцінки порогу малопомітного приросту якості збігаються (1004). Максимальний розмір навчальної вибірки в експерименті становив 630 спостережень (≈ 90 % від загального обсягу датасету 699 після виділення

валідаційного фолду). Ця величина перевищує точкову оцінку NLS (559), проте залишається нижчою за ймовірнісні пороги NLS+GP: для надійності 80 % необхідно 782 спостереження, для 95 % – 810. З огляду на широкі довірчі інтервали параметрів степеневої функції, рекомендується орієнтуватися на оцінки NLS+GP та розширити загальний датасет до рівня, що забезпечить навчальну вибірку обсягом 780–810 зразків.

Таблиця 3

Порівняння оцінок мінімально достатнього обсягу навчальної вибірки

Критерій	NLS	NLS+GP
N для RMSE ≤ 12 (точкова оцінка)	559	756
N для P(RMSE ≤ 12) ≥ 50 %	778	756
N для P(RMSE ≤ 12) ≥ 80 %	–	782
N для P(RMSE ≤ 12) ≥ 95 %	–	810
N для малопомітного приросту (≥ 50 %)	1197	1004
N для малопомітного приросту (≥ 80 %)	–	1160
Асимптотична RMSE (параметр a)	10,795	10,795

Таким чином, у даній роботі крива навчання розглядається як базовий емпіричний інструмент, параметрична екстраполяція – як основний практичний спосіб кількісного оцінювання достатнього обсягу вибірки, а GP-based learning-type curve – як більш просунутий статистичний варіант, що додатково характеризує невизначеність прогнозу. Така ієрархія підходів дозволяє, з одного боку, спиратися на фактично спостережувану поведінку моделі на реальних даних, а з іншого – оцінювати її ймовірну поведінку в режимах, які ще не були безпосередньо досягнуті в експерименті.

Висновки. Проведене дослідження підтвердило, що задача визначення мінімально достатнього обсягу навчальної вибірки для системи сенсорного сортування руд є розв'язуваною емпіричними методами на основі аналізу кривих навчання та їх екстраполяції.

Модель HistGradientBoostingRegressor продемонструвала високу якість на реальних сенсорних даних ($R^2 = 0{,}93$), що підтверджує придатність обраного класу моделей до поставленої задачі. Побудована крива навчання демонструє монотонне спадання похибки зі збільшенням N з тенденцією до стабілізації у правій частині — що свідчить про входження в зону насичення та зменшення граничного ефекту від додаткових спостережень.

Параметрична екстраполяція степеневою функцією дозволила перейти від якісного аналізу кривої до кількісних оцінок достатнього обсягу вибірки. За результатами GP-based learning-type curve, для досягнення RMSE ≤ 12 з надійністю 95 % необхідно 810 спостережень у навчальній вибірці. Максимальний розмір навчальної вибірки в експерименті становив 630 спостережень — близько 90 % від загального обсягу датасету після виділення валідаційного фолду, — що є дещо меншим за рекомендований поріг. Поріг малопомітного приросту якості оцінено на рівні N ≈ 1004, після якого подальше розширення вибірки не забезпечує статистично значущого покращення якості моделі.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Таким чином, для досягнення цільового рівня похибки з достатньою статистичною надійністю рекомендується розширити датасет до рівня, що забезпечить навчальну вибірку обсягом 780–810 зразків.

ЛІТЕРАТУРА

1. Кісельов Б. Г., Сенько А. О. Вплив адитивних стохастичних збурень на нижню межу узагальнювальної похибки моделей регресії в сенсорних системах. Комп'ютерні інтелектуальні системи та мережі : матеріали XIX Всеукраїнської науково-практичної WEB-конференції аспірантів, студентів та молодих вчених, 25–27 березня 2026 р. Кривий Ріг : Криворізький національний університет, 2026. С. 156–159.

2. Viering T., Loog M. The Shape of Learning Curves: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. Vol. 45, no. 12. P. 15050–15067. DOI: 10.1109/TPAMI.2021.3085003.

3. Figueroa R. L., Zeng-Treitler Q., Kandula S., Ngo L. H. Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making. 2012. Vol. 12. Article 8. DOI: 10.1186/1472-6947-12-8.

4. Beleites C., Neugebauer U., Bocklitz T., Krafft C., Popp J. Sample size planning for classification models. Analytica Chimica Acta. 2013. Vol. 760. P. 25–33. DOI: 10.1016/j.aca.2012.11.007.

5. Vabalas A., Gowen E., Poliakoff E., Casson A. J. Machine learning algorithm validation with a limited sample size. PLoS ONE. 2019. Vol. 14, no. 11. Article e0224365. DOI: 10.1371/journal.pone.0224365.

6. Snell K. I. E., Archer L., Ensor J. et al. Sample size requirements for training clinical prediction models using participant-level meta-analysis. Statistics in Medicine. 2024. Vol. 43, no. 15. P. 2945–2975. DOI: 10.1002/sim.10121.

7. Kaplan J., McCandlish S., Henighan T. et al. Scaling Laws for Neural Language Models. arXiv. 2020. arXiv:2001.08361. DOI: 10.48550/arXiv.2001.08361.

8. Zöller M. A., Huber M. F. Benchmark and Survey of Automated Machine Learning Frameworks. Journal of Artificial Intelligence Research. 2021. Vol. 70. P. 409–472. DOI: 10.1613/jair.1.11854.

9. Ke G., Meng Q., Finley T. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 3146–3154.

10. Domingos P. A few useful things to know about machine learning. Communications of the ACM. 2012. Vol. 55, no. 10. P. 78–87. DOI: 10.1145/2347736.2347755.

11. Cortes C., Jackel L. D., Solla S. A., Vapnik V., Denker J. S. Learning curves: Asymptotic values and rate of convergence. In: Advances in Neural Information Processing Systems. 1994. Vol. 6. P. 327–334.

12. Mukherjee S., Tamayo P., Rogers S. et al. Estimating dataset size requirements for classifying DNA microarray data. Journal of Computational Biology. 2003. Vol. 10, no. 2. P. 119–142. DOI: 10.1089/106652703321825928.

REFERENCES

1. Kiselov, B. H., & Senko, A. O. (2026). *Vplyv adytyvnykh stokhastychnykh zburen' na nyzhniu mezhu uzahalniuval'noi pokhybky modelei rehresii v sensornykh systemakh* [Influence of additive stochastic perturbations on the lower bound of regression model generalisation error in sensor systems]. In *Proceedings of the XIX All-Ukrainian Scientific-Practical WEB Conference* (pp. 156–159). KNU. [in Ukrainian]

2. Viering, T., & Loog, M. (2023). The shape of learning curves: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(12), 15050–15067. https://doi.org/10.1109/TPAMI.2021.3085003

3. Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making, 12*, Article 8. https://doi.org/10.1186/1472-6947-12-8

4. Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta, 760*, 25–33. https://doi.org/10.1016/j.aca.2012.11.007

5. Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE, 14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

6. Snell, K. I. E., Archer, L., Ensor, J., Maier, A., Debray, T. P. A., Burdett, S., Riley, R. D., & Ensor, J. (2024). Sample size requirements for training clinical prediction models using participant-level meta-analysis. *Statistics in Medicine, 43*(15), 2945–2975. https://doi.org/10.1002/sim.10121

7. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models* [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2001.08361

8. Zöller, M. A., & Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research, 70*, 409–472. https://doi.org/10.1613/jair.1.11854

9. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.

10. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

11. Cortes, C., Jackel, L. D., Solla, S. A., Vapnik, V., & Denker, J. S. (1994). Learning curves: Asymptotic values and rate of convergence. *Advances in Neural Information Processing Systems, 6*, 327–334.

12. Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., & Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology, 10*(2), 119–142. https://doi.org/10.1089/106652703321825928

Received 17.04.2026.
Accepted 22.04.2026.
Published 30.04.2026

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Empirical determination of the minimum sufficient training sample size for machine learning models at a given error level

This article addresses the problem of empirically determining the minimum sufficient training sample size for machine learning regression models in ore sensor sorting systems. The relevance of this topic stems from the significant costs associated with creating representative datasets in the mining industry. The problem lies in the need to transition from a qualitative analysis of the learning curve to a quantitative assessment of the sufficient sample size for a given error level. The aim of the study is to evaluate a hierarchy of approaches: learning curve – parametric power-law extrapolation – GP-based learning-type curve. Methods: 10-fold GroupKFold cross-validation, Bayesian hyperparameter tuning (Optuna), nonlinear regression, Gaussian processes. Results: $R^2 = 0.93$ on test folds; the minimum sufficient sample size for RMSE ≤ 12 was estimated in the range of 559–810 observations. Key conclusion: the proposed method allows for a well-founded determination of the threshold beyond which further expansion of the sample ceases to yield practical benefits.

Keywords: machine learning; learning curve; minimum sample size; power-law approximation; Gaussian process; ore sorting; cross-validation; HistGradientBoosting; Neural Scaling Laws; extrapolation.

Кісельов Богдан Геннадійович – аспірант кафедри автоматизації, комп'ютерних наук і технологій, Криворізький національний університет, Кривий Ріг, Україна.
ORCID: http://orcid.org/0009-0007-9338-1031
Сенько Антон Олександрович – канд. техн. наук, кафедра комп'ютерних систем та мереж, Криворізький національний університет, Кривий Ріг, Україна.
ORCID: http://orcid.org/0000-0002-4104-8372
Купін Андрій Іванович – д.т.н., професор, кафедра комп'ютерних систем та мереж, Криворізький національний університет, Кривий Ріг, Україна.
ORCID: http://orcid.org/0000-0001-7569-1721
Балик Дмитро Костянтинович – інженер-програміст, ТОВ «НВП Гамаюн».
ORCID: http://orcid.org/0009-0000-4768-8576

Bohdan Kiselov – PhD Student, Department of Automation, Computer Sciences and Technology, Kryvyi Rih National University, Ukraine.
ORCID: http://orcid.org/0009-0007-9338-1031
Anton Senko – Candidate of Technical Sciences, Department of Computer Systems and Networks, Kryvyi Rih National University, Kryvyi Rih, Ukraine.
ORCID: http://orcid.org/0000-0002-4104-8372
Andrii Kupin – Doctor of Technical Sciences, Professor, Department of Computer Systems and Networks, Kryvyi Rih National University, Kryvyi Rih, Ukraine.
ORCID: http://orcid.org/0000-0001-7569-1721
Dmytro Balyk – "SPE Hamaiun" LLC.
ORCID: http://orcid.org/0009-0000-4768-8576