НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Факультет електроніки

Кафедра мікроелектроніки

До захисту допущено:

Завідувач кафедри

_____ Дмитро ТАТАРЧУК

«___»_____20__ р.

Дипломна робота

на здобуття ступеня бакалавра

за освітньо-професійною програмою «Мікро- та наноелектроніка»

спеціальності 153 «Мікро- та наносистемна техніка»

на тему: «Алгоритми машинного навчання для прогнозування властивостей сегнетоелектриків»

Виконав:
Студент IV курсу, групи ДП-11
Друзь Дмитро Андрійович _____

Керівник: асист.каф.МЕ, доктор філософії, б/з,
Шевлякова Ганна Вікторівна _____

Консультант з нормоконтролю: доц.каф.МЕ, к.ф.-м.н., с.н.с.,
Свєчніков Георгій Сергійович _____

Консультант з інформаційних питань: доц.каф.МЕ, к.т.н., доц.,
Діденко Юрій Вікторович _____

Рецензент:
_____ _____

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент (-ка) _____

Київ – 2025 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет електроніки
Кафедра мікроелектроніки

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 153 «Мікро- та наносистемна техніка»

Освітньо-професійна програма «Мікро- та наноелектроніка»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Дмитро ТАТАРЧУК
«___»_____20__ р.

ЗАВДАННЯ

на дипломну роботу студенту

Друзю Дмитру Андрійовичу

1. Тема роботи «Алгоритми машинного навчання для прогнозування властивостей сегнетоелектриків», керівник роботи Шевлякова

Ганна Вікторівна, асист.каф.МЕ, доктор філософії, б/з, затверджені наказом по університету від «___»_____ 20__ р. №_____

2. Термін подання студентом роботи _____

3. Вихідні дані до роботи: технічне завдання

4. Зміст роботи: Дослідження властивостей сегнетоелектриків; дослідження алгоритмів машинного навчання; розробка предиктивних моделей на основі алгоритмів машинного навчання.

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо): Презентація

6. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Відмітка про виконання
1	Аналіз сучасних алгоритмів машинного навчання	19.03.2025 - 2.04.2025	
2	Аналіз програмних рішень для роботи з алгоритмами машинного навчання	3.04.2025 - 9.04.2025	
3	Аналіз властивостей сегнетоелектриків та відкритих баз даних з матеріалами	10.04.2025 - 14.04.2025	
4	Розробка програмного забезпечення	15.04.2025 - 10.05.2025	
5	Аналіз результатів	11.05.2025 - 14.05.2025	
6	Аналіз перспектив	15.04.2025 - 17.05.2025	
7	Оформлення дипломної роботи першого (бакалаврського) рівня вищої освіти	17.05.2025 - 27.05.2025	

Студент Дмитро ДРУЗЬ

Керівник Ганна ШЕВЛЯКОВА

АНОТАЦІЯ

Кваліфікаційна робота: 88 с., 8 рис., 30 джерел

Об’єкт дослідження: сегнетоелектричні матеріали та їхні фізико-хімічні властивості.

Предмет дослідження: математичні моделі машинного навчання та їх застосування до задач прогнозування параметрів сегнетоелектриків.

Мета роботи: розробка, навчання та аналіз ефективності моделей машинного навчання для прогнозування фізичних властивостей сегнетоелектричних матеріалів на основі структурних та хімічних характеристик.

Методи дослідження: аналітичні та емпіричні методи, які включають математичне моделювання на основі алгоритмів машинного навчання, таких як регресія, метод опорних векторів, випадковий ліс і штучні нейронні мережі.

При написанні роботи було досліджено літературні джерела, властивості сегнетоелектричних матеріалів, їх класифікація; відкриті бази даних, такі як Materials Project, AFLOW, NOMAD; сучасні алгоритми машинного навчання, такі як RandomForest, SVM, штучні нейронні мережі. Були проведені тренування відповідних моделей, а також проведено аналіз процесу тренування та аналіз результатів роботи моделей.

Результат роботи: набір предиктивних моделей, що дозволяють по набору вхідних ознак сегнетоелектрика дізнатися приблизну температуру Кюрі.

Ключові слова: MSE, MAE, сегнетоелектрик, гістерезис, модель, датасет, оверфіт, метрика, крос-валідація.

ABSTRACT

Qualification work: 88 pages, 8 figures, 30 references

Research object: ferroelectric materials and their physical and chemical properties.

Research subject: mathematical models of machine learning and their application to parameter prediction problems of ferroelectrics.

Work objective: development, training and analysis of the effectiveness of machine learning models for predicting the physical properties of ferroelectric materials based on structural and chemical characteristics.

Research methods: analytical and empirical methods, which include mathematical modeling based on machine learning algorithms such as regression, support vector machine, random forest, and artificial neural networks.

When writing the work, the following were researched: literary sources, properties of ferroelectric materials, their classification; open databases such as Materials Project, AFLOW, NOMAD; modern machine learning algorithms such as RandomForest, SVM, artificial neural networks. The corresponding models were trained, and the training process and the results of the models were analyzed.

Work result: a set of predictive models that allow to find out the approximate Curie temperature from a set of input features of a ferroelectic.

Keywords: MSE, MAE, ferroelectric, hysteresis, model, dataset, overfit, metric, cross-validation.

ЗМІСТ

Вступ 8

Розділ 1. Теоретичні основи дослідження сегнетоелектричних матеріалів ... 11

1.1 Поняття та властивості сегнетоелектриків 11

1.2 Класифікація та основні характеристики сегнетоелектричних матеріалів 18

1.3 Сучасні підходи до дослідження фізичних властивостей сегнетоелектриків 23

1.4 Проблеми традиційного моделювання та прогнозування параметрів сегнетоелектриків 29

Розділ 2. Основи машинного навчання та його застосування у матеріалознавстві 34

2.1 Основні алгоритми машинного навчання: регресія, класифікація, ансамблеві методи 34

2.2 Методи передобробки даних та зменшення розмірності 40

2.3 Інструменти та програмне забезпечення для побудови моделей (Python, scikit-learn, TensorFlow) 44

2.4 Огляд досліджень із застосування машинного навчання у прогнозуванні властивостей матеріалів 48

Розділ 3. Прогнозування властивостей сегнетоелектриків за допомогою алгоритмів машинного навчання 53

3.1 Постановка задачі прогнозування та вибір цільових властивостей 53

3.2 Опис та аналіз вхідних даних (датасетів) 56

3.3 Побудова моделей на основі обраних алгоритмів (Random Forest, SVM, Neural Networks) 60

3.4 Оцінювання точності моделей, порівняння результатів 69

3.5 Інтерпретація результатів і можливості практичного застосування... 70

Висновки 73

Перелік посилань..... 75

Додатки..... 79

Додаток А. Код для нейронної мережі..... 79

Додаток Б. Код для RandomForest 83

Додаток В. Код для SVR..... 85

ВСТУП

У сучасному науковому просторі спостерігається стрімкий розвиток міждисциплінарних досліджень, що поєднують матеріалознавство, фізику твердого тіла та комп’ютерні науки. Однією з таких актуальних тем є використання алгоритмів машинного навчання для прогнозування властивостей матеріалів, зокрема, сегнетоелектриків — класу матеріалів, що мають здатність змінювати поляризацію під дією зовнішнього електричного поля. Ці матеріали широко застосовуються в таких сферах, як мікроелектроніка, пам’ять з довільним доступом, сенсори, актюатори, акустичні пристрої, тому їхні властивості мають визначальний вплив на функціональність кінцевої продукції. Водночас експериментальне визначення фізико-хімічних параметрів є ресурсоємним, складним та тривалим процесом. Це зумовлює потребу у використанні новітніх підходів для моделювання, аналізу та прогнозування характеристик таких матеріалів.

З огляду на зростаючий обсяг даних про матеріали, які накопичуються в електронних базах, таких як Materials Project, Open Quantum Materials Database (OQMD), AFLOW тощо, постає потреба в інструментах, здатних ефективно працювати з великими масивами структурованої інформації. Саме в цьому контексті алгоритми машинного навчання відіграють ключову роль, оскільки здатні виявляти закономірності, що не піддаються простому аналітичному опису. Методи машинного навчання, такі як дерева рішень, випадковий ліс, градієнтний бустинг, нейронні мережі та метод опорних векторів, вже продемонстрували високу ефективність у задачах класифікації матеріалів, регресійного прогнозування фізичних властивостей, а також в автоматичному проєктуванні нових сполук.

Актуальність дослідження зумовлена глобальним трендом цифровізації наукових досліджень, де штучний інтелект та машинне навчання слугують інструментами автоматизації складних обчислень. Особливої

значущості ця тема набуває у зв’язку з розвитком квантових технологій та створенням нових типів інформаційних носіїв, де саме сегнетоелектрики можуть відігравати провідну роль. Крім того, розробка надійних моделей для прогнозування параметрів сегнетоелектриків здатна суттєво скоротити витрати на синтез та дослідження нових зразків.

На стику фізики, хімії та інформатики формується новий підхід до вивчення матеріалів — матеріалознавство, кероване даними (Data-Driven Materials Science). Саме в межах цього підходу і реалізується дана робота, яка спрямована на створення та апробацію моделей машинного навчання для передбачення ключових параметрів сегнетоелектриків.

Також існує потреба ефективного прогнозування параметрів сегнетоелектриків без проведення тривалих експериментальних вимірювань. Застосування машинного навчання дає змогу істотно прискорити процес добору матеріалів для конкретних технічних задач, знизити фінансові витрати на дослідження та розширити можливості фундаментальної науки в галузі фізики твердого тіла.

Метою дипломної роботи є розробка, навчання та аналіз ефективності моделей машинного навчання для прогнозування фізичних властивостей сегнетоелектричних матеріалів на основі структурних та хімічних характеристик.

Об’єктом дослідження є сегнетоелектричні матеріали та їхні фізико-хімічні властивості.

Предметом дослідження є математичні моделі машинного навчання та їх застосування до задач прогнозування параметрів сегнетоелектриків.

Завдання дослідження

1. Провести аналіз літературних джерел щодо властивостей сегнетоелектриків та їх застосування.
2. Зібрати та попередньо обробити дані про сегнетоелектричні матеріали з відкритих баз даних.

3. Обрати найбільш релевантні характеристики матеріалів для навчання моделей.
4. Побудувати моделі машинного навчання (лінійна регресія, SVM, випадковий ліс, нейронна мережа) для прогнозування фізичних властивостей сегнетоелектричні.
5. Провести оцінку точності моделей та порівняти результати.
6. Проаналізувати вплив вибору гіперпараметрів на результат.
7. Сформулювати практичні висновки щодо застосування моделей ML у матеріалознавстві.

Методи дослідження. У процесі виконання роботи застосовувалися аналітичні та емпіричні методи, які включають математичне моделювання на основі алгоритмів машинного навчання, таких як регресія, метод опорних векторів, випадковий ліс і штучні нейронні мережі. Для реалізації моделей було використано мови програмування Python з бібліотеками Scikit-learn, TensorFlow та Pandas, що дозволяє ефективно обробляти дані, здійснювати побудову моделей, налаштування гіперпараметрів та візуалізацію результатів. Верифікація моделей здійснювалася з використанням метрик середньоквадратичної помилки (RMSE), середньої абсолютної похибки (MAE) та коефіцієнта детермінації (R^2). Також застосовувався метод крос-валідації для підвищення надійності оцінки результатів. Дані про властивості сегнетоелектриків бралися з відкритих матеріалознавчих баз даних, таких як Materials Project та OQMD, із подальшою фільтрацією та нормалізацією для забезпечення коректного машинного аналізу.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ СЕГНЕТОЕЛЕКТРИЧНИХ МАТЕРІАЛІВ

1.1 Поняття та властивості сегнетоелектриків

Сегнетоелектрики — це клас діелектричних матеріалів, які мають здатність до спонтанної електричної поляризації, тобто внутрішньої електричної дипольної орієнтації, яка виникає без прикладення зовнішнього електричного поля. Це означає, що атоми в кристалі сегнетоелектрика займають несиметричні положення у своїй елементарній комірці, створюючи внаслідок цього постійний електричний дипольний момент. Найважливішою ознакою сегнетоелектриків є наявність так званої петлі гістерезису, що фіксує залежність між напруженністю прикладеного зовнішнього електричного поля і електричною індукцією або поляризацією в матеріалі. Ця петля свідчить про те, що матеріал «запам’ятовує» напрямок попередньої поляризації навіть після зняття поля, що є основою для застосування сегнетоелектриків у елементах пам’яті, наприклад, у FRAM (Ferroelectric RAM) [1].

Сегнетоелектричні властивості виникають унаслідок фазового переходу, що зазвичай спостерігається при зниженні температури нижче певної критичної точки, яка називається температурою Кюрі. Вище цієї температури матеріал поводиться як параелектрик — діелектрик без спонтанної поляризації. При охолодженні відбувається структурна перебудова кристалічної ґратки, внаслідок якої центр позитивного і негативного заряду в осередку зміщується, формуючи постійну поляризацію. Цей перехід супроводжується зміною симетрії кристалічної структури: з високосиметричної (параелектричної) до низькосиметричної (сегнетоелектричної).

На мікроскопічному рівні поляризація у сегнетоелектриках виникає в результаті зміщення іонів у кристалічній ґратці під впливом міжатомних

взаємодій. Наприклад, у класичному сегнетоелектрику — титанаті барію ($BaTiO_3$) — атом титану займає нецентральне положення в октаедричній комірці, що утворює дипольний момент (рис.1.1). У масштабах кристалу виникає сукупність таких диполів, організованих у домени — області з однаковою орієнтацією поляризації. Під дією зовнішнього електричного поля ці домени можуть переорієнтовуватись, що і зумовлює характерну гістерезисну поведінку [10].

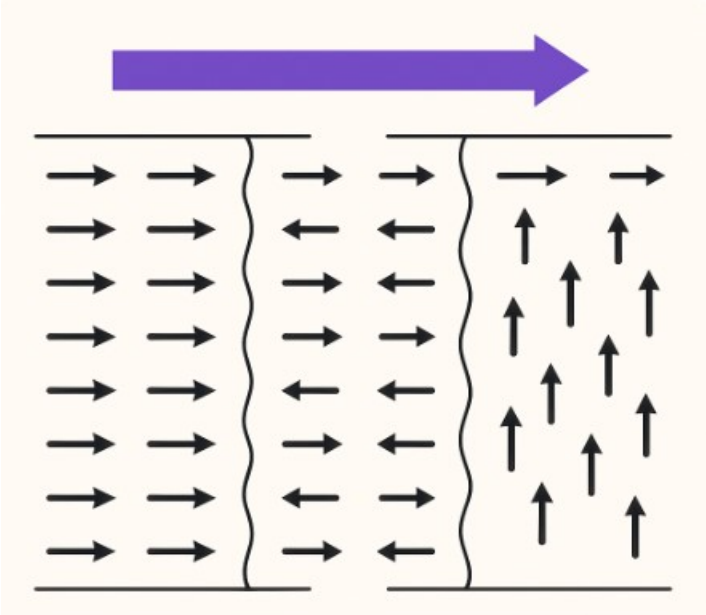

Рис. 1.1. Доменна структура сегнетоелектрика

Сегнетоелектрики мають низку особливих властивостей, які вирізняють їх серед інших діелектриків. Окрім вже згаданої гістерезисної поведінки, вони мають високу діелектричну проникність, яка може змінюватись на кілька порядків залежно від температури, прикладеного поля та частоти. Це робить сегнетоелектрики надзвичайно чутливими матеріалами, придатними для створення сенсорів, перетворювачів, електромеханічних резонаторів та акустоелектронних пристроїв. Ще однією унікальною характеристикою є п’єзоелектричний ефект, тобто здатність генерувати електричний заряд під впливом механічного стиску. Оскільки не всі п’єзоелектрики є сегнетоелектриками, але всі сегнетоелектрики володіють п’єзоелектричними

властивостями, це підкреслює їхню універсальність у багатьох галузях техніки.

Важливо зазначити, що поляризація у сегнетоелектриках не лише змінюється під дією зовнішнього поля, але також може демонструвати складні температурні та часозалежні явища, зокрема, деполяризацію, релаксацію та доменну нестійкість. Такі ефекти мають суттєве значення при розробці електронних компонентів, оскільки можуть впливати на стабільність і надійність роботи пристрою.

У сучасній науці вивчення сегнетоелектриків активно розвивається на перетині матеріалознавства, фізики конденсованого стану та нанотехнологій. Особливу увагу приділено пошуку нових сегнетоелектричних матеріалів із високою температурою Кюрі, стабільною поляризацією та добрими електричними характеристиками, що дозволить удосконалити сучасну електроніку та розробити нові функціональні пристрої. Прогнозування таких властивостей із використанням сучасних обчислювальних методів, зокрема машинного навчання, відкриває нові горизонти в проєктуванні матеріалів і пришвидшує наукові відкриття.

Упродовж останніх десятиліть розвиток обчислювальних технологій докорінно змінив підходи до дослідження матеріалів, зокрема сегнетоелектриків. Традиційні методи аналізу властивостей ґрунтувалися переважно на експериментальних вимірюваннях або ж на аналітичних моделях, які базувалися на апроксимації фізичних процесів у межах певних гіпотез. Проте ці методи мають низку обмежень — вони вимагають значних часових і фінансових витрат, є маломасштабними та не завжди дають змогу виявити приховані закономірності в складних багатовимірних даних. Водночас зростання обсягу доступної інформації про структури та властивості матеріалів, що зберігається в базах даних перших принципів, стимулювало формування нової парадигми — дослідження матеріалів за допомогою методів машинного навчання [9].

Машинне навчання, як галузь штучного інтелекту, дозволяє виявляти нелінійні та високовимірні залежності між вхідними характеристиками і вихідними властивостями матеріалів без необхідності жорстко визначати аналітичну форму зв’язку. У контексті сегнетоелектриків, це може означати можливість передбачити, наприклад, значення спонтанної поляризації, температуру фазового переходу або діелектричну проникність на основі комбінації атомного складу, параметрів кристалічної ґратки, електронної щільності або інших дескрипторів. У порівнянні з перехідними методами квантової хімії, які хоч і дають високу точність, але потребують значних обчислювальних ресурсів, машинне навчання пропонує альтернативу, яка базується на попередньо обчислених або експериментально виміряних даних. Одним із ключових елементів у застосуванні машинного навчання до матеріалознавства є побудова адекватної векторної репрезентації матеріалу, яка б містила максимум інформації про його властивості, при цьому зберігаючи узагальнювальну здатність. Така репрезентація може включати числові характеристики, як електронегативність елементів, радіуси атомів, симетрію кристалу, густину, параметри ґратки, а також структурні дескриптори, що враховують геометрію просторового розміщення атомів. Після створення репрезентації вибирається модель машинного навчання, яка навчається на великій кількості зразків, а потім застосовується для передбачення властивостей нових або маловивчених матеріалів.

Серед найпоширеніших моделей, що застосовуються в дослідженнях сегнетоелектриків, можна згадати регресійні моделі, які дають змогу передбачати числові властивості, і класифікаційні, що визначають наявність чи відсутність конкретної ознаки, наприклад, сегнетоелектричної поведінки. Більш складні методи, такі як нейронні мережі, дозволяють моделювати взаємозв’язки між вхідними та вихідними змінними на значно глибшому рівні, що особливо важливо при роботі з даними, які мають багатовимірну або структуровану природу. Нейронні мережі можуть враховувати взаємодії між

атомами в кристалі та навіть просторову топологію структур, що наближає їхню ефективність до методів з перших принципів [8].

Однак застосування машинного навчання у матеріалознавстві не обмежується лише передбаченням числових параметрів. Значущим напрямом є також автоматизований добір матеріалів за заданими властивостями, оптимізація хімічного складу для досягнення цільових характеристик, класифікація фазових діаграм, виявлення нових функціональних сполук та зменшення простору пошуку при синтезі. Важливо зазначити, що ефективність таких моделей прямо залежить від якості й повноти вхідних даних, а також від правильності їхньої інтерпретації, що потребує тісної співпраці між фахівцями з фізики, хімії, матеріалознавства та комп'ютерних наук.

У поєднанні з доступом до великомасштабних матеріалознавчих баз та сучасними інструментами аналізу даних, машинне навчання стає потужним інструментом для прискорення відкриттів у галузі функціональних матеріалів. Завдяки йому з'являється можливість не лише верифікувати наявні гіпотези, а й генерувати нові ідеї та напрямки досліджень. Для сегнетоелектриків це означає перспективу створення матеріалів з наперед заданими характеристиками, адаптованих до специфіки технологічних процесів або експлуатаційних умов, що відкриває нові горизонти для мікроелектроніки, сенсорики та квантових технологій (табл.1.1).

Таблиця 1.1 – Основні відкриті бази даних для дослідження сегнетоелектриків

№	Назва бази даних	URL (доступ)	Короткий опис	Кількість матеріалів	Наявність сегнето-електриків
1	Materials Project	https://materialsproject.org	Містить структурні, електронні та термодинамічні властивості матеріалів	>150 000	Так
2	AFLOW	http://aflowlib.org	Автоматизована база з розрахунками за теорією функціоналу густини (DFT)	>3 000 000	Так
3	OQMD (Open Quantum Materials Database)	http://oqmd.org	Містить квантово-механічні дані, підходить для машинного навчання	>1 000 000	Частково
4	NOMAD	https://nomad-coe.eu	Європейська платформа для зберігання та аналізу симуляцій матеріалів	>50 000 проектів	Частково
5	Citrination	https://citrination.com	Комерційна платформа для машинного навчання матеріалів	Не уточнюється	Частково

Після розгляду основних відкритих баз даних, які містять інформацію про структуру та властивості сегнетоелектриків, постає питання про ефективне представлення цих матеріалів у формі, придатній для обробки алгоритмами машинного навчання. Це потребує переходу від абстрактного опису атомної структури до числової векторної репрезентації, яка здатна зберігати важливі фізико-хімічні характеристики і водночас бути сумісною з математичними моделями. Такий процес називається векторизацією структури [7].

Векторизація матеріалів — це етап, на якому складна інформація про атомний склад, геометрію та електронні властивості матеріалу перетворюється на набір чисел (вектор ознак), які можуть бути використані як вхідні дані для моделей машинного навчання. Векторизація повинна бути

універсальною, репрезентативною, інваріантною до трансляції, обертання і масштабування структури та водночас інформативною. У випадку сегнетоелектриків особливу роль відіграють ознаки, які прямо чи опосередковано впливають на формування спонтанної поляризації, фазові переходи та симетрію кристалічної ґратки.

$$P_s = \frac{\sum^i q_i \cdot r_i}{V} \tag{1.1}$$

де P_s - вектор спонтанної поляризації; q_i - заряд і-ого іона; r_i - вектор положення відносно центру осередку і-ого іона; V - об’єм елементарної комірки.

Серед найпоширеніших способів векторизації використовуються дескриптори, які можна умовно поділити на хімічні, структурні та комбіновані. Хімічні дескриптори охоплюють середні значення електронегативності, атомного радіусу, іонного заряду тощо. Структурні дескриптори можуть включати параметри ґратки (a, b, c, α, β, γ), густину, симетрію кристалу (групу простору), число атомів у комірці, топологічні характеристики. Комбіновані дескриптори формуються з використанням алгоритмів глибокого навчання, автоенкодерів або згорткових нейронних мереж, які самостійно виділяють закономірності з вхідних даних [6].

$$D = \varepsilon_0 \cdot E + P \tag{1.2}$$

де D - електрична індукція; ε_0 - діелектрична стала; E - напруженність електричного поля; P - поляризація.

$$\varepsilon_r = \frac{C}{T - T_c} \tag{1.3}$$

де ε_r - відносна діелектрична проникність; *C* - стала Кюрі; T - температура; T_c - температура Кюрі.

Одним із найбільш визнаних підходів до векторизації є використання матеріальних дескрипторів із бібліотеки Matminer, яка інтегрується з платформою Pymatgen та дозволяє генерувати понад сотню ознак для кожного матеріалу. Крім того, застосовуються методи, такі як SOAP (Smooth Overlap of Atomic Positions), Coulomb Matrix, Crystal Graph Convolutional Neural Networks (CGCNN), які дозволяють кодувати геометрію розташування атомів на глибшому рівні. Особливо перспективним напрямом є використання графових представлень, де структура кристалу описується у вигляді графа, вершинами якого є атоми, а ребрами — їхні зв'язки. У цьому випадку на допомогу приходять графові нейронні мережі, які здатні враховувати складні просторові взаємодії між компонентами структури.

Ефективна векторизація є критично важливою умовою успішного прогнозування властивостей сегнетоелектриків за допомогою алгоритмів машинного навчання. Вона визначає точність і узагальнювальну здатність моделі, а отже, і потенціал її використання для відкриття нових матеріалів або оптимізації вже відомих. Наступним логічним кроком є побудова та тренування моделей машинного навчання, які будуть базуватися на отриманих векторних представленнях сегнетоелектричних сполук [1].

1.2 Класифікація та основні характеристики сегнетоелектричних матеріалів

Сегнетоелектричні матеріали становлять собою різноманітну групу сполук, які поділяються за кількома критеріями: за кристалічною структурою, хімічним складом, механізмом виникнення спонтанної поляризації та за температурними властивостями. Їхнє класифікування має як наукову, так і прикладну цінність, оскільки дозволяє прогнозувати потенційні сфери застосування матеріалу, визначати методи його синтезу, а також підібрати

відповідні алгоритми машинного навчання для моделювання тих чи інших фізичних характеристик [2].

$$D_i = d_{ij} \cdot \sigma_j \tag{1.4}$$

де D_i - електричне зміщення; d_{ij} - п'єзоелектричний коефіцієнт; σ_j - механічна напруга.

Одним із ключових підходів до класифікації сегнетоелектриків є поділ за типом кристалічної структури. Найбільш поширеними є перовськітні сегнетоелектрики, які мають загальну формулу ABX_3, де A та B — це металічні катіони, а X — зазвичай кисень. Класичним прикладом є титанат барію ($BaTiO_3$), де атом титану знаходиться в центрі октаедра з атомів кисню і може зміщуватись відносно центра, створюючи поляризацію. Така структура легко змінюється під дією температури, тиску або електричного поля, що робить перовськіти одними з найкращих кандидатів для застосування в електроніці. Іншим важливим класом є бісмутоферити, такі як $BiFeO_3$, які поєднують у собі сегнетоелектричні та сегнетомагнітні властивості, що дозволяє реалізовувати мультифункціональні пристрої.

$$W = \oint E \ dP \tag{1.5}$$

де W - енергія втрат на гістерезис; E - напруженність електричного поля; P - поляризація;

Ще одним критерієм класифікації є поділ сегнетоелектриків на дислокаційні та орієнтаційні. У дислокаційних сегнетоелектриках спонтанна поляризація виникає внаслідок зміщення іонів у кристалічній комірці, як це спостерігається у титанаті барію. У орієнтаційних сегнетоелектриках, до яких належать, наприклад, калієва дигідрофосфатна кислота (KDP), поляризація зумовлена орієнтацією молекул або груп молекул у кристалі під впливом

міжмолекулярних взаємодій. Такий поділ має значення при аналізі стабільності поляризації та механізмів перемикання доменів.

Окрім структурних особливостей, надзвичайно важливими є електричні характеристики сегнетоелектриків, зокрема значення спонтанної поляризації, ширина петлі гістерезису, діелектрична проникність та температура Кюрі. Спонтанна поляризація визначає величину заряду, який матеріал здатен утримувати в одній з двох стабільних орієнтацій. Вона вимірюється в мікрокулонах на квадратний сантиметр і залежить як від структури, так і від умов синтезу. Температура Кюрі визначає точку, при якій відбувається фазовий перехід від сегнетоелектричної до параелектричної фази. Чим вища температура Кюрі, тим ширше температурне вікно застосування матеріалу. У багатьох технічних задачах перевагу надають матеріалам з високою температурою фазового переходу, оскільки вони можуть працювати в умовах нагріву, вібрацій або електричних перевантажень [3].

Ще однією суттєвою характеристикою є діелектрична проникність, яка відображає здатність матеріалу поляризуватись у відповідь на прикладене електричне поле. У сегнетоелектриків ця характеристика може змінюватись у дуже широких межах — від сотень до десятків тисяч одиниць, що робить їх винятково придатними для виготовлення високочастотних конденсаторів, резонаторів та електрооптичних пристроїв. Разом з цим значення має і величина коерцитивного поля — електричне поле, необхідне для перемикання поляризації, оскільки воно визначає рівень енергоспоживання при зміні стану та стабільність збереженої інформації в пам’яті на основі сегнетоелектриків.

Окрему увагу слід приділити мультиферроїкам — особливому класу сегнетоелектриків, які одночасно виявляють магнітні властивості. Вони є перспективними для створення пристроїв, де можна керувати магнітним станом за допомогою електричного поля і навпаки, що відкриває можливості для створення нових типів логічних елементів, пам’яті з багатостановим зберіганням, а також сенсорних елементів для умов підвищеного навантаження (табл.1.2.).

Усі ці властивості формують широку функціональність сегнетоелектриків, що робить їх незамінними в сучасній електроніці, зокрема у виробництві мікросхем пам’яті, п’єзоелектричних сенсорів, оптичних модуляторів, ультразвукових перетворювачів та навіть у біомедичних пристроях. Водночас різноманітність сегнетоелектриків ускладнює їхнє прогнозування та вимагає залучення сучасних методів аналізу, таких як машинне навчання, яке може враховувати широкий набір дескрипторів і встановлювати складні взаємозв’язки між хімічною структурою, симетрією та експериментальними характеристиками [4].

Таблиця 1.2 – Класифікація та характеристики ферроелектричних матеріалів

№	Клас матеріалу	Приклад	Тип поляризації	Особливості / Застосування
1	Перовськітні сегнетоелектрики	$BaTiO_3$, $PbTiO_3$	Дислокаційна	Висока діелектрична проникність; використовуються в конденсаторах, пам’яті, сенсорах
2	Ортоборфати і бісмутоферити	$BiFeO_3$	Дислокаційна	Мультифероїки, що поєднують сегнетоелектричність і магнітні властивості; придатні для спінтроніки
3	Орієнтаційні сегнетоелектрики	KDP (KH_2PO_4)	Орієнтаційна	Поляризація обумовлена орієнтацією молекул; застосовуються в оптичних модуляторах
4	Полімерні сегнетоелектрики	PVDF (полі-вінілденфто-рид)	Орієнтаційна	Гнучкі, легкі матеріали; використовуються в сенсорах, енергозбиральних пристроях
5	Рідкокристалічні сегнетоелектрики	DOBAMBC	Орієнтаційна	Застосовуються в дисплеях; чутливі до температури та електричного поля

Після класифікації сегнетоелектричних матеріалів та огляду їхніх ключових характеристик стає очевидним, що ця група сполук охоплює широкий спектр структур, механізмів поляризації та фізичних властивостей. Таке різноманіття відкриває великі можливості для цільового застосування

сегнетоелектриків у різних технологічних галузях, але водночас ускладнює процес їх систематичного аналізу. Кожен підклас має свої унікальні параметри — від кристалографічної симетрії до температурної стабільності поляризації, — що повинні бути враховані при побудові моделі прогнозування [5].

Особливо важливим для подальшого аналізу є розуміння того, які саме ознаки (дескриптори) відіграють визначальну роль у формуванні цільових властивостей, таких як спонтанна поляризація, температура Кюрі чи діелектрична проникність. Наприклад, у перовськітних сегнетоелектриках вирішальним є зміщення центрального атома в октаедричному середовищі, тоді як у полімерних матеріалах ключову роль відіграє гнучкість ланцюгів та орієнтація молекул під дією зовнішнього поля. У мультиферроїків важливою є наявність одночасної електричної і магнітної впорядкованості, що робить їх більш складними для прогнозування, але також і більш цікавими з точки зору функціональності.

Завдання машинного навчання у контексті сегнетоелектриків не обмежується простою регресією фізичних величин, а виходить на рівень аналізу взаємозв’язку між структурою, електронними властивостями та поведінкою матеріалу в умовах зовнішнього впливу. Векторизація матеріалів, класифікація за структурними ознаками, і вибір відповідної моделі прогнозування — це три складові ефективного підходу до моделювання таких складних систем. З огляду на це, наступним кроком дослідження є формування вхідного датасету з баз даних матеріалів, його нормалізація та підготовка до машинного аналізу. Це стане основою для реалізації моделей, які дозволять не лише передбачити властивості вже відомих сегнетоелектриків, а й визначити потенційно ефективні нові матеріали ще до їх синтезу.

1.3 Сучасні підходи до дослідження фізичних властивостей сегнетоелектриків

Сучасні підходи до дослідження фізичних властивостей сегнетоелектриків охоплюють широкий спектр теоретичних, експериментальних та обчислювальних методів, які взаємодоповнюють один одного в межах міждисциплінарної наукової парадигми. У зв’язку з високим інтересом до функціональних матеріалів нового покоління, які мають адаптивні властивості, значна увага приділяється вивченню таких параметрів сегнетоелектриків, як спонтанна поляризація, температура фазових переходів, гістерезис, діелектрична проникність, п’єзоелектричний і електрооптичний ефекти. Розуміння механізмів, що лежать в основі цих властивостей, потребує застосування як фізичних моделей, так і інструментів великомасштабного чисельного моделювання [11].

Традиційно дослідження фізичних властивостей сегнетоелектриків здійснювалися переважно експериментальними методами, серед яких варто виділити дифракцію рентгенівських променів, електрооптичні вимірювання, спектроскопію імпедансу, вимірювання п’єзоелектричних коефіцієнтів та температурно-залежну діелектричну спектроскопію. За допомогою цих методів можна безпосередньо спостерігати структурні фазові переходи, оцінити розподіл доменів, визначити залежність поляризації від прикладеного електричного поля, а також встановити критичні температури, за яких відбувається зміна симетрії кристалічної структури. Втім, обмеження експериментальних підходів полягає у їхній трудомісткості, високій вартості обладнання, а також у тому, що вони не завжди дозволяють вивчити широке коло матеріалів в умовах швидкого скринінгу.

У відповідь на ці обмеження з середини XX століття почали активно розвиватись аналітичні та чисельні теоретичні методи. Одним із фундаментальних підходів є використання теорії функціоналу густини (DFT), яка дозволяє обчислювати електронну структуру матеріалів із високою

точністю, виходячи лише з атомного складу та розташування атомів у просторі. Метод DFT дає можливість визначати енергетичну стабільність сегнетоелектричної фази, величину поляризації, зміну електронної щільності під дією зовнішнього поля, що є надзвичайно важливим для глибокого розуміння мікроскопічної природи сегнетоелектричності. Проте обчислення на основі DFT є досить ресурсозатратними, що обмежує їхню придатність для великомасштабного прогнозування тисяч або мільйонів зразків [23, 24].

Завдяки розвитку інформатики та накопиченню масивів даних про властивості матеріалів у базах, таких як Materials Project, AFLOW чи OQMD, з'явився новий напрям — data-driven materials science, або матеріалознавство, кероване даними. В його основі лежить ідея автоматичного аналізу, класифікації та передбачення властивостей матеріалів з використанням статистичних моделей, машинного навчання та штучного інтелекту. Цей підхід дозволяє моделювати зв'язки між структурними, хімічними та електричними параметрами без потреби в глибокому теоретичному описі, використовуючи замість цього великі набори експериментальних або розрахованих даних. Алгоритми машинного навчання, такі як дерева рішень, нейронні мережі, метод опорних векторів або ансамблеві методи, навчаються на структурованих датасетах і здатні передбачати цільові властивості з високою точністю.

Особливе значення в сучасному аналізі має використання графових нейронних мереж, які оперують не просто табличними даними, а графовими структурами, де вузлами є атоми, а ребрами — їхні зв'язки. Це дозволяє враховувати не лише тип елементів, але й просторову топологію кристалічної структури, що особливо важливо для опису складних сегнетоелектриків. Такий підхід відкриває нові можливості у виявленні закономірностей, які не є очевидними з погляду класичної фізики, і дає змогу генерувати нові матеріали з наперед заданими властивостями шляхом оберненого проектування [12].

Важливою складовою сучасних досліджень є також мультифізичне моделювання, яке поєднує різні фізичні явища — теплові, механічні,

електричні та магнітні — в єдиній обчислювальній системі. Це особливо актуально у випадках, коли сегнетоелектрик працює в умовах навантаження або змінного середовища, як-от у п’єзоакустичних резонаторах або в датчиках деформацій. Інструменти, такі як COMSOL Multiphysics або Ansys, дозволяють моделювати поведінку матеріалу в реальних умовах експлуатації, що підвищує точність інженерних прогнозів.

Сучасні підходи до вивчення фізичних властивостей сегнетоелектриків базуються на поєднанні трьох ключових компонентів — класичних експериментальних методів, перших принципів квантової теорії та алгоритмів машинного навчання. Їхнє синергетичне використання дозволяє не лише глибше зрозуміти фізичну природу сегнетоелектричних явищ, а й зробити процес пошуку нових матеріалів значно ефективнішим і масштабованішим. У контексті зростаючої ролі комп’ютерного моделювання та автоматизованих систем у науці, такий підхід є не лише доцільним, а й необхідним для забезпечення проривних інновацій у сфері мікроелектроніки, оптоелектроніки та енергетичних систем нового покоління.

Подальша логіка дослідження веде нас до питання вибору моделей машинного навчання, які є найбільш придатними для аналізу фізичних властивостей сегнетоелектриків. Після збору та векторизації даних, отриманих з експериментів або обчислень на основі перших принципів, критично важливо підібрати такі алгоритми, що здатні не лише точно передбачати цільові параметри, а й демонструвати стійкість до шумів, нестачі частини даних або варіативності у структурі зразків. В умовах, коли дослідник працює з великими об’ємами інформації з баз типу Materials Project або AFLOW, ефективність обраного підходу визначає успішність усього проєкту.

Залежно від типу задачі — регресійної чи класифікаційної — доцільно використовувати різні моделі машинного навчання. Якщо йдеться про передбачення кількісних характеристик, таких як температура Кюрі, величина спонтанної поляризації чи діелектрична проникність, необхідно застосовувати регресійні алгоритми, здатні встановлювати функціональні залежності між

вхідними дескрипторами та числовими значеннями вихідних параметрів. Якщо ж мета полягає у визначенні наявності або відсутності певної властивості, наприклад, чи є матеріал сегнетоелектриком взагалі, то доречними є класифікаційні підходи, які дозволяють навчити модель розділяти матеріали на категорії з відповідним рівнем точності.

У виборі моделі важливу роль відіграють не тільки точність передбачення, але й інтерпретованість результатів, обчислювальна складність, чутливість до параметрів та здатність до узагальнення. У випадку з матеріалами, структура яких має високу складність, нейронні мережі і, зокрема, глибокі згорткові мережі виявляються надзвичайно ефективними, оскільки вони можуть враховувати складні нелінійні залежності у багатовимірному просторі дескрипторів. Однак за таких умов втрачається прозорість у поясненні, як саме модель дійшла до того чи іншого результату, що може ускладнювати її використання в фундаментальних дослідженнях. Тому часто застосовуються гібридні підходи, де глибока модель поєднується з інтерпретованими методами, наприклад, градієнтним бустингом або випадковим лісом, які дозволяють зберегти високу точність і при цьому відстежити важливість окремих ознак.

В умовах обмеженої кількості навчальних зразків дослідники часто звертаються до трансферного навчання (transfer learning), яке дає змогу використати попередньо навчену модель на нових матеріалах з частково схожими структурними особливостями. Це особливо важливо у випадках з рідкісними або складними сегнетоелектриками, дані по яких є обмеженими або надто дорогими для отримання. Крім того, активно використовуються методи крос-валідації, регуляризації та оптимізації гіперпараметрів, які дозволяють підвищити стійкість моделей до перенавчання та адаптувати їх до конкретних задач, пов’язаних із властивостями твердих тіл (табл 1.3).

Загалом вибір моделі машинного навчання є не лише технічним кроком, але й частиною дослідницької стратегії, яка включає глибоке розуміння природи даних, фізичного змісту вхідних ознак та характеру взаємозв’язків

між структурою й функціональністю матеріалу. Саме тому застосування методів машинного навчання у вивченні сегнетоелектриків є не заміною класичним підходам, а радше їхнім потужним продовженням, яке розкриває нові можливості для наукового прориву та практичної реалізації інтелектуальних матеріалів майбутнього [13].

Таблиця 1.3 – Порівняння моделей машинного навчання для прогнозування властивостей сегнетоелектриків

№	Модель	Тип задачі	Переваги	Недоліки
1	Лінійна регресія	Регресія	Проста реалізація, інтерпретованість результатів	Не підходить для складних нелінійних залежностей
2	Випадковий ліс (Random Forest)	Класифікація / Регресія	Стійкість до перенавчання, висока точність, інтерпретованість важливості ознак	Високі обчислювальні витрати при великій кількості дерев
3	Градієнтний бустинг (XGBoost)	Класифікація / Регресія	Висока продуктивність, можливість тонкого налаштування	Складність у виборі гіперпараметрів
4	Метод опорних векторів (SVM)	Класифікація / Регресія	Ефективність на малих вибірках, хороше розділення класів	Потребує обчислювально затратного налаштування ядра
5	Нейронна мережа	Класифікація / Регресія	Підходить для складних і багатовимірних задач, здатність до узагальнення	Неінтерпретованість, висока потреба в даних та обчислювальних ресурсах

Після вибору відповідної моделі машинного навчання надзвичайно важливим кроком є формування якісного навчального датасету, оскільки саме від його повноти, репрезентативності та правильності залежить ефективність подальшого моделювання. У випадку з сегнетоелектричними матеріалами джерелами даних виступають переважно великі відкриті бази, такі як Materials Project, AFLOW, OQMD, а також власні експериментальні або симуляційні

результати дослідника. Всі ці дані потребують уніфікації, фільтрації та попередньої обробки перед тим, як будуть передані до алгоритму навчання.

Формування навчального датасету починається з вибору цільової властивості, яку модель має прогнозувати. Це може бути, наприклад, спонтанна поляризація, діелектрична проникність, температура Кюрі або ймовірність прояву сегнетоелектричної поведінки. Після визначення цільової змінної слід відібрати релевантні дескриптори, які будуть слугувати вхідними параметрами. Ці дескриптори повинні мати фізичний сенс і бути чисельно вираженими, адже саме з ними працює більшість моделей машинного навчання. До таких ознак можуть входити атомні властивості (електронегативність, атомна маса, радіус), структурні характеристики (параметри кристалічної ґратки, симетрія, кількість атомів у комірці), електронні властивості (щільність електронів, ширина забороненої зони), а також топологічні або графові представлення.

Після цього виконується очищення даних від пропущених значень, дублікатів або аномалій, які можуть викривити навчання моделі. У більшості випадків застосовуються методи нормалізації або стандартизації даних, які дозволяють привести числові значення вхідних ознак до одного масштабу. Це особливо важливо у тих моделях, що залежать від відстані між точками у векторному просторі, як методи опорних векторів чи нейронні мережі.

Окремим етапом є розділення датасету на навчальну, валідаційну та тестову вибірки. Це дозволяє перевіряти здатність моделі до узагальнення, уникати перенавчання та коректно оцінювати її ефективність. Як правило, для навчання використовується 70–80% даних, решта розподіляється між валідацією і тестуванням. При цьому для невеликих наборів часто застосовується крос-валідація, коли дані по черзі використовуються для навчання і перевірки, що забезпечує більш об’єктивну оцінку [14].

Особливу увагу потрібно звернути на балансування класів у класифікаційних задачах. Якщо більшість матеріалів у вибірці не мають сегнетоелектричних властивостей, а лише мала частка їх проявляє, модель

може навчитися передбачати лише домінантний клас. Тому в таких випадках застосовуються методи балансування, зокрема oversampling (збільшення кількості рідкісних прикладів) або undersampling (зменшення кількості надмірних прикладів), а також алгоритми, чутливі до ваг класів.

1.4 Проблеми традиційного моделювання та прогнозування параметрів сегнетоелектриків

Традиційне моделювання та прогнозування параметрів сегнетоелектриків тривалий час ґрунтувалося на використанні аналітичних методів фізики твердого тіла та чисельних розрахунків, зокрема методів теорії функціоналу густини (DFT) і класичних моделей електрополяризації. Ці підходи, хоч і є фундаментальними та фізично обґрунтованими, мають низку серйозних обмежень, які ускладнюють їхнє застосування у масштабних дослідженнях або у прискореному відкритті нових матеріалів. Найгострішою проблемою таких традиційних методів є їхня висока обчислювальна складність. Наприклад, виконання DFT-розрахунків для складного сегнетоелектричного кристалу може тривати від кількох годин до кількох днів навіть на високопродуктивних обчислювальних кластерах, а якщо йдеться про аналіз тисяч матеріалів — це перетворюється на фактично непідйомне завдання.

Крім того, традиційні методи передбачають побудову моделі для кожного конкретного матеріалу окремо, і тому не мають здатності до узагальнення. Тобто, отримані результати, як правило, не можуть бути автоматично перенесені на інші класи матеріалів без повного повторення розрахункового процесу. Це значно ускладнює розробку універсальних моделей поведінки сегнетоелектриків або прогнозування властивостей нових, ще не синтезованих сполук. Також варто відзначити, що для коректного використання DFT потрібен глибокий теоретичний досвід і точне налаштування параметрів симуляції, зокрема вибору функціоналів обміну-

кореляції, граничних умов, розмірності комірки, що може стати суттєвою перепоною для молодих дослідників або інженерів-практиків [15].

Ще однією проблемою є обмеження традиційних моделей у розпізнаванні складних, багатофакторних або нелінійних взаємозв'язків між структурою і властивостями сегнетоелектриків. Наприклад, вплив доменних меж, міжатомних кореляцій, нестехіометричних дефектів або анізотропії механічних напружень може бути важко або взагалі неможливо врахувати в рамках класичної моделі. В результаті передбачення таких параметрів, як температура фазового переходу або стабільність доменної структури, часто мають значну похибку і потребують подальшої експериментальної перевірки. Це знижує ефективність моделювання як інструмента матеріалознавчого скринінгу [3].

Також слід наголосити на обмеженості традиційного підходу в плані обробки великих обсягів експериментальних даних. У зв'язку з розвитком нових методів синтезу і автоматизованих експериментів на сьогодні генерується величезна кількість даних про нові матеріали. Проте класичні аналітичні моделі не мають інструментарію для автоматичної обробки таких даних або виявлення прихованих закономірностей у великих вибірках. Вони залишаються в рамках формального математичного опису і не пристосовані до задач великомасштабного пошуку, класифікації чи кластеризації матеріалів.
Крім цього, одним із суттєвих бар'єрів є неврахування випадкових та стохастичних ефектів, які виникають у реальних умовах експлуатації сегнетоелектричних матеріалів. Традиційне моделювання, як правило, передбачає ідеалізовану, симетричну структуру без дефектів, домішок або термічних флуктуацій. У дійсності ж сегнетоелектрики мають складну доменну структуру, реагують на зміну вологості, температури, електричних полів, і їх поведінка не завжди відповідає теоретичним передбаченням. Через це традиційні моделі часто не в змозі адекватно описати деградацію матеріалу в часі або ефекти, пов'язані з довготривалим навантаженням.

Нарешті, важливим недоліком є низький рівень інтеграції класичних методів моделювання з інженерною практикою. Через складність, громіздкість і потребу у високій спеціалізації багато інженерних команд не мають змоги самостійно застосовувати традиційні підходи у своїй щоденній роботі. Це гальмує впровадження новітніх матеріалів у виробничі процеси, навіть якщо теоретично вони вже описані та вивчені.

Традиційні підходи, хоча й мають безсумнівну цінність у плані глибини фізичного розуміння та точності, стикаються з низкою критичних обмежень, які стримують їхню ефективність у сучасних задачах прогнозування (табл.1.4). Це відкриває шлях для нових методів, зокрема машинного навчання, які здатні ефективно обробляти великі масиви даних, виявляти складні закономірності й масштабуватись для аналізу тисяч потенційних матеріалів у порівняно короткі терміни [16].

Таблиця 1.4 – Основні проблеми традиційного моделювання сегнетоелектриків

№	Проблема	Суть проблеми
1	Висока обчислювальна складність	Тривалі обчислення DFT і неможливість швидкого скринінгу великої кількості матеріалів.
2	Немає узагальнення на інші матеріали	Модель необхідно повторно будувати для кожного нового матеріалу, немає переносу знань.
3	Неможливість врахування складних впливів	Не враховуються дефекти, доменні межі, термічні флуктуації чи мультифізичні ефекти.
4	Складність налаштування моделей	Потреба у високій кваліфікації та точному виборі параметрів симуляції.
5	Невідповідність реальним умовам	Теоретичні моделі працюють із ідеалізованими структурами, які не відображають реальні об’єкти.
6	Відсутність автоматизації	Класичні методи не пристосовані до роботи з великими базами даних та не здатні обробляти Big Data.
7	Низька інтеграція з інженерною практикою	Методи не адаптовані до щоденної роботи інженера, потребують складного програмного забезпечення.

У контексті наведених у таблиці проблем стає очевидним, що традиційне моделювання, незважаючи на свою теоретичну глибину, не може повною мірою задовольнити потреби сучасної матеріалознавчої науки та інженерії, особливо у випадках, коли йдеться про масштабне передбачення властивостей сегнетоелектриків або виявлення нових матеріалів з оптимальними параметрами. Обмеження, пов’язані з ресурсомісткістю, відсутністю адаптації до великих обсягів даних, складністю відтворення експериментальних умов, а також нездатністю охоплювати нелінійні, багатофакторні взаємозв’язки, створюють об’єктивну потребу у впровадженні нових, гнучкіших інструментів [8].

У цьому контексті машинне навчання пропонує якісно інший підхід, який не потребує формального фізичного опису процесів для кожного конкретного випадку. Замість того, щоб вручну виводити рівняння, алгоритми машинного навчання працюють із наявними масивами даних, «вчаться» на них і формують внутрішню модель залежностей між структурними, хімічними, електронними характеристиками матеріалів і тими властивостями, які потрібно передбачити. Такий підхід дозволяє швидко аналізувати тисячі або навіть мільйони комбінацій, які були б надто складними для класичного обчислення. Крім того, алгоритми не лише передбачають властивості, але й можуть виявляти приховані закономірності, що не були очевидними в межах традиційних підходів, тим самим генеруючи нові гіпотези для експериментального підтвердження.

Ще однією перевагою машинного навчання є можливість повторного використання моделей. Якщо одна модель була навчена на великому наборі сегнетоелектриків з певного класу, вона може бути адаптована для інших груп матеріалів шляхом донавчання, що значно скорочує витрати часу та обчислювальних ресурсів. Це дає змогу створювати універсальні предиктивні системи, які можна інтегрувати у програми автоматизованого дизайну матеріалів або навіть у реальні виробничі процеси [17].

Особливо важливо те, що машинне навчання легко інтегрується з іншими цифровими технологіями, такими як цифрові близнюки, автоматизовані лабораторії, хмарні платформи моделювання та візуалізації, що робить його частиною сучасної екосистеми науково-дослідницьких рішень. На додачу до цього, методи explainable AI (XAI) дозволяють частково подолати проблему «чорної скриньки» у складних моделях, забезпечуючи інтерпретацію найважливіших ознак, які впливають на прогноз.

Перехід від традиційного моделювання до машинного навчання не є простою зміною інструментів — це якісний стрибок у методології дослідження. Застосування інтелектуальних алгоритмів дає змогу не лише значно розширити межі можливого в моделюванні сегнетоелектриків, а й сприяє появі нових підходів до дослідження складних матеріалів, заснованих на даних, взаємодії між фізикою та інформатикою, а також на тісному зв’язку між фундаментальною наукою та прикладною інженерією.

РОЗДІЛ 2. ОСНОВИ МАШИННОГО НАВЧАННЯ ТА ЙОГО ЗАСТОСУВАННЯ У МАТЕРІАЛОЗНАВСТВІ

2.1 Основні алгоритми машинного навчання: регресія, класифікація, ансамблеві методи

Алгоритми машинного навчання, що використовуються для аналізу і прогнозування властивостей матеріалів, зокрема сегнетоелектриків, умовно поділяються на кілька основних типів залежно від характеру поставленої задачі. Найбільш поширеними серед них є регресія, класифікація та ансамблеві методи. Кожен з цих підходів має свої особливості, переваги та сфери застосування, і вибір конкретного алгоритму залежить від цільової змінної, доступних даних, складності взаємозв’язків між параметрами та бажаного рівня інтерпретованості моделі [18].

Регресія є базовим напрямом машинного навчання, який застосовується для прогнозування неперервних числових величин. У контексті вивчення сегнетоелектриків регресійні моделі дозволяють передбачити значення таких параметрів, як температура Кюрі, величина спонтанної поляризації, діелектрична проникність або ширина петлі гістерезису. Однією з найпростіших реалізацій є лінійна регресія, яка встановлює прямий зв’язок між вхідними дескрипторами та вихідною змінною. Проте при аналізі складних багатовимірних структур, які притаманні більшості сегнетоелектриків, лінійна модель часто виявляється недостатньою, оскільки не здатна відобразити нелінійні ефекти. У таких випадках застосовуються поліноміальні регресії, метод опорних векторів у режимі регресії (SVR) або нейронні мережі, які дозволяють гнучко моделювати складні взаємозалежності.

Класифікація, у свою чергу, орієнтована на вирішення задач, де результатом є категорія або клас, до якого належить об’єкт. Наприклад, у

дослідженні сегнетоелектриків класифікаційна модель може визначити, чи проявляє певна структура сегнетоелектричну активність, або класифікувати зразки за типом фазового переходу. Алгоритми класифікації працюють за принципом пошуку межі, що найкращим чином розділяє точки різних класів у багатовимірному просторі ознак. До найвідоміших класифікаційних методів належать логістична регресія, дерева рішень, метод опорних векторів та наївний баєсівський класифікатор. Сучасні підходи часто поєднують ці методи з методами зменшення розмірності, такими як головні компоненти (PCA), для візуалізації та покращення якості класифікації в задачах з великою кількістю параметрів [7].

Ансамблеві методи являють собою групу моделей, які працюють не окремо, а у сукупності, об’єднуючи результати багатьох слабших моделей у більш сильний і стійкий до помилок прогноз. У матеріалознавстві ці методи стали особливо популярними завдяки своїй здатності забезпечувати високу точність при збереженні достатнього рівня інтерпретованості. Одним із найвідоміших ансамблевих підходів є випадковий ліс, який використовує набір дерев рішень, кожне з яких навчається на окремій підмножині даних. Завдяки цьому модель виявляється менш схильною до перенавчання та здатна ефективно працювати на даних із шумами або відсутніми значеннями. Ще потужнішим є градієнтний бустинг, що будує модель поступово, додаючи нові дерева для корекції помилок попередніх. Однією з найпопулярніших його реалізацій є бібліотека XGBoost, яка демонструє виняткову продуктивність у багатьох задачах, зокрема й у прогнозуванні властивостей матеріалів. Ансамблеві методи також дозволяють оцінювати важливість окремих ознак, що дає змогу досліднику зрозуміти, які саме параметри мають найбільший вплив на прогнозовану властивість, що є критично важливим у задачах, де фізична інтерпретація відіграє не меншу роль, ніж сам результат прогнозування [25].

Найбільш динамічним і перспективним напрямом у сфері машинного навчання, що знайшов широке застосування у прогнозуванні властивостей

складних матеріалів, є використання штучних нейронних мереж. Цей підхід особливо ефективний у випадках, коли дані мають багатовимірну, структуровану або нелінійну природу, що є характерним для більшості сегнетоелектриків. На відміну від класичних алгоритмів, нейронні мережі здатні автоматично виділяти приховані взаємозв'язки між параметрами без необхідності ручного визначення структури моделі. Це дозволяє враховувати неочевидні залежності, які виникають унаслідок складних внутрішніх взаємодій між атомами, симетрією кристалу, локальними деформаціями або розподілом електронної густини.

Глибокі нейронні мережі, зокрема згорткові та рекурентні архітектури, можуть працювати з об'ємними наборами вхідних ознак, а також інтегрувати інформацію про топологію структури, енергію взаємодій і контекстуальні залежності. У випадку матеріалів, представлених у вигляді графів, де вузли відповідають атомам, а зв'язки — міжатомним взаємодіям, застосовуються графові нейронні мережі. Вони дозволяють моделювати складні тривимірні структури без втрати просторової інформації, забезпечуючи виняткову точність передбачень. Саме завдяки цим перевагам нейронні мережі поступово стають центральним інструментом у матеріалознавстві, зокрема у задачах класифікації фаз, передбачення стабільності кристалів, аналізу енергетичних станів та розробки нових функціональних сполук [19].

Особливістю застосування нейронних мереж у прогнозуванні параметрів сегнетоелектриків є їх здатність обробляти великі масиви даних одночасно з обліком параметрів, які мають складну фізичну природу. Наприклад, мережа може навчитися на основі тисяч прикладів розпізнавати структурні особливості, що супроводжуються спонтанною поляризацією, або виявляти конфігурації, які мають високу ймовірність демонструвати анізотропну діелектричну поведінку. При цьому модель не обмежується використанням суто числових ознак — вона може працювати з матрицями взаємодій, просторовими координатами, спектральними профілями або іншими формами складної вхідної інформації.

Важливим аспектом при застосуванні нейронних мереж є питання узгодження між точністю і прозорістю моделі. Хоча нейронні архітектури здатні забезпечити найвищий рівень точності передбачення, вони зазвичай функціонують як так звані «чорні скриньки», тобто не дозволяють досліднику легко простежити, які саме ознаки вплинули на кінцевий результат. Це може бути критичним у наукових дослідженнях, де потрібна не лише відповідь, а й її обґрунтування. Для вирішення цієї проблеми застосовуються методи пояснюваного штучного інтелекту, що дозволяють зменшити рівень абстракції та виділити найбільш важливі фактори у прийнятті рішень нейромережею [26].

З практичної точки зору, для реалізації нейронних мереж використовуються сучасні обчислювальні середовища, такі як TensorFlow, PyTorch або Keras, що дозволяють налаштовувати структуру, кількість шарів, функції активації, методи оптимізації та інші параметри в залежності від задачі. Для аналізу властивостей сегнетоелектриків це створює широкі можливості як для побудови моделей на основі існуючих баз даних, так і для створення автономних систем, здатних прогнозувати нові сполуки ще до їх синтезу. У результаті нейронні мережі стають не просто інструментом обчислення, а інтелектуальним компонентом у системі відкриття матеріалів.

Цей етап розвитку моделювання позначає собою перехід від простих аналітичних моделей до комплексних адаптивних систем, здатних самостійно навчатися, аналізувати й ухвалювати рішення на основі багатовимірної інформації (табл.2.1). Застосування нейронних мереж у дослідженнях сегнетоелектриків дозволяє не лише підвищити точність прогнозування, а й значно прискорити темпи відкриття нових матеріалів, скоротивши шлях від ідеї до практичного впровадження [20].

Таблиця 2.1 – Основні типи алгоритмів машинного навчання, що використовуються для сегнетоелектриків

№	Тип алгоритму	Застосування у сегнетоелектриках	Особливості
1	Регресія	Прогнозування температури Кюрі, поляризації, діелектричної проникності	Працює з числовими значеннями, дозволяє будувати функціональні залежності
2	Класифікація	Визначення наявності сегнетоелектричних властивостей, типів фазових переходів	Розділяє матеріали на категорії, підходить для задач типу "так/ні"
3	Ансамблеві методи	Підвищення точності прогнозу, визначення важливості ознак	Об’єднують багато слабких моделей в одну потужну, зменшують ризик перенавчання
4	Нейронні мережі	Аналіз складних багатофакторних структур, прогноз нових матеріалів	Висока точність і здатність до узагальнення, потреба у великому обсязі даних
5	Графові нейронні мережі	Моделювання кристалічної структури як графу, врахування просторової інформації	Працюють із міжатомними зв’язками, підходять для 3D-структур сегнетоелектриків

Після систематизації основних алгоритмів машинного навчання, придатних для моделювання властивостей сегнетоелектриків, логічним наступним етапом є формування практичного підходу до побудови, навчання і тестування відповідних моделей. Незалежно від того, який саме алгоритм обрано — регресійний, класифікаційний чи нейромережевий — увесь процес передбачає чітко структуровану послідовність дій, яка починається з підготовки даних і завершується перевіркою точності прогнозу на нових, невідомих зразках.

Ключовою умовою успішного моделювання є якісно сформована навчальна вибірка. Вона має бути репрезентативною, збалансованою та досить різноманітною, щоб модель могла «навчитися» узагальнювати інформацію, а не просто запам’ятовувати приклади. На цьому етапі здійснюється нормалізація числових ознак, кодування категоріальних змінних та, у разі потреби, зменшення розмірності простору ознак для зниження

обчислювальної складності. Після цього дані розділяються на навчальний, валідаційний та тестовий набори, що дозволяє контролювати ефективність моделі не лише під час навчання, а й після нього, коли вона стикається з новими вхідними прикладами.

Наступним етапом є безпосереднє навчання моделі, під час якого вона поступово коригує свої внутрішні параметри, мінімізуючи похибку передбачення за допомогою обраного критерію якості. Залежно від типу задачі, можуть використовуватись різні метрики: середньоквадратична помилка, середня абсолютна похибка, точність, F-міра або коефіцієнт детермінації. У разі використання нейронних мереж особливу увагу слід приділяти гіперпараметрам, таким як кількість шарів, розмірність нейронів, функції активації, алгоритми оптимізації та швидкість навчання. Неправильно обрані параметри можуть призвести до перенавчання, коли модель працює добре на тренувальних даних, але дає помилки на нових зразках.

Для забезпечення стабільності результатів часто використовується техніка перехресної перевірки (крос-валідації), яка дає змогу оцінити середню точність прогнозу на кількох розділеннях даних. Це особливо важливо у випадках, коли об’єм доступної інформації обмежений. Додатково застосовуються регуляризаційні методи, що знижують ризик перенавчання, штрафуючи надто складні моделі з великою кількістю вагових коефіцієнтів. На цьому ж етапі визначається важливість окремих ознак, що дозволяє зробити інтерпретацію результатів більш прозорою та виявити ключові фізичні фактори, які найбільше впливають на поведінку сегнетоелектрика.

Останній етап — тестування — полягає у перевірці остаточної моделі на незалежному наборі даних, який не використовувався на жодному з попередніх етапів. Саме цей етап дає змогу об’єктивно оцінити здатність моделі до узагальнення та її практичну придатність для прогнозування властивостей нових матеріалів. Якщо результати тестування виявляються незадовільними, процес повертається до попередніх стадій: може бути

проведений додатковий відбір ознак, модифікована архітектура моделі або ж змінений сам алгоритм навчання [27].

Побудова моделі машинного навчання для прогнозування властивостей сегнетоелектриків є багатоступеневим процесом, що поєднує в собі як глибоке розуміння фізичної природи матеріалу, так і точне володіння обчислювальними методами. Кінцева мета полягає не лише у створенні алгоритму, що здатен з високою точністю відтворювати експериментальні результати, але й у формуванні інструменту, який допоможе передбачати властивості ще не синтезованих матеріалів, скорочуючи часові й фінансові витрати на їх дослідження.

2.2 Методи передобробки даних та зменшення розмірності

Методи передобробки даних та зменшення розмірності є критично важливими етапами у процесі побудови моделей машинного навчання, особливо коли йдеться про аналіз складних фізико-хімічних систем, таких як сегнетоелектрики. У цьому контексті передобробка даних виконує роль не лише технічної підготовки набору даних для роботи алгоритму, а й забезпечення коректності, достовірності та релевантності вхідної інформації, на якій ґрунтуватиметься навчання моделі.

Передусім варто розуміти, що дані, які надходять із різних джерел — експериментів, симуляцій, баз типу Materials Project чи OQMD — часто мають неоднорідний формат, пропущені значення, різні одиниці вимірювання, шумові впливи або значення, що значно виходять за межі загального діапазону. Тому першим кроком є очищення даних. На цьому етапі виявляються і обробляються пропущені або некоректні значення. Для числових параметрів зазвичай використовують середнє, медіану або моду для заповнення пропусків, а в складніших випадках застосовують моделі прогнозування відсутніх даних на основі інших ознак. Якщо певні ознаки мають надто багато пропусків або неінформативні (наприклад, мають постійне

значення у більшості записів), вони можуть бути повністю виключені з аналізу.

Після очищення проводиться нормалізація або стандартизація числових ознак. Це необхідно для того, щоб всі характеристики мали однакову шкалу впливу на модель. Якщо цього не зробити, ознаки з більшими числовими значеннями (наприклад, густина або молярна маса) можуть непропорційно впливати на результат, тоді як більш дрібні характеристики (електронегативність або міжатомні відстані) будуть проігноровані. Стандартизація приводить розподіл кожної ознаки до нульового середнього значення з одиничним стандартним відхиленням, тоді як нормалізація масштабує значення до певного діапазону — зазвичай від нуля до одиниці.
На окрему увагу заслуговує категоріальна обробка. У матеріалознавстві іноді використовуються нечислові характеристики, такі як тип кристалічної симетрії, тип сегнетоелектричної фази чи група елементів у періодичній системі. Для таких ознак застосовують кодування — найпоширенішим є one-hot encoding, за якого кожне унікальне значення перетворюється на окрему бінарну змінну. Це дозволяє представити категоріальні дані у формі, що може бути оброблена алгоритмами машинного навчання.

Після проведення базової підготовки даних може виникнути інша проблема — надмірна кількість ознак. Якщо кількість дескрипторів перевищує кількість спостережень або якщо між ознаками існує сильна кореляція, це може призвести до ефекту перенавчання, коли модель добре працює на тренувальному наборі, але погано узагальнює знання на нові дані. У такому випадку застосовуються методи зменшення розмірності, які дозволяють зберегти найважливішу інформацію, зменшивши кількість вхідних ознак. Одним із найвідоміших таких методів є аналіз головних компонент (PCA). Він трансформує початкові ознаки у новий простір, де нові змінні — головні компоненти — є лінійними комбінаціями вихідних ознак і несуть у собі найбільшу частку дисперсії даних. Це дозволяє зменшити кількість параметрів моделі без втрати інформативності [28].

Крім PCA, в сучасному матеріалознавстві також застосовуються інші методи зменшення розмірності, зокрема t-SNE (t-distributed Stochastic Neighbor Embedding) і UMAP (Uniform Manifold Approximation and Projection), які дозволяють краще візуалізувати складні багатовимірні дані в двох- або тривимірному просторі. Хоча ці методи не завжди використовуються безпосередньо для навчання моделей, вони є дуже корисними для попереднього аналізу та виявлення груп схожих матеріалів або структурних кластерів.

У випадках, коли йдеться про графове представлення структури, використовуються також спеціалізовані методи агрегації ознак на основі топологічних властивостей графа, де розмірність може зменшуватись шляхом об’єднання інформації з сусідніх вузлів, що є характерним для графових нейронних мереж. Такі підходи дозволяють працювати з дуже складними, структурно насиченими даними без втрати ключової просторової інформації (табл. 2.2) [29].

Таблиця 2.2 – Основні методи передобробки даних та зменшення розмірності

№	Метод	Призначення та особливості
1	Очищення даних	Видалення пропусків, дублікатів, аномалій; забезпечення достовірності даних.
2	Нормалізація/стандартизація	Приведення ознак до єдиного масштабу; критично важливо для більшості моделей.
3	Кодування категоріальних ознак	Перетворення нечислових ознак (наприклад, тип симетрії) у числовий формат для моделей.
4	PCA (аналіз головних компонент)	Зменшення розмірності з мінімальною втратою інформації; лінійна трансформація.
5	t-SNE	Нелінійне зменшення розмірності для візуалізації даних у 2D/3D; зберігає локальні структури.
6	UMAP	Швидкий алгоритм для візуалізації кластерів у багатовимірних даних; інформативніший за t-SNE у деяких задачах.
7	Графова агрегація ознак	Зменшення розмірності структурованих графових даних через об’єднання інформації між вузлами.

Після розгляду методів передобробки даних та зменшення розмірності важливим логічним продовженням є етап оцінювання ефективності

побудованих моделей машинного навчання. Цей процес є ключовим для того, щоб визначити, наскільки точно й надійно модель здатна передбачати фізичні властивості сегнетоелектриків на основі вхідних характеристик, а також чи можна довіряти отриманим результатам при застосуванні моделі до нових, ще не досліджених матеріалів.

У практиці машинного навчання точність моделі визначається через обчислення певних метрик, які порівнюють передбачені значення з реальними, отриманими з експериментів або з баз даних, таких як Materials Project. У задачах регресії, де моделі прогнозують неперервні величини, такими метриками є середньоквадратична похибка (MSE), корінь з неї (RMSE), середня абсолютна похибка (MAE) та коефіцієнт детермінації (R^2). Ці показники дозволяють оцінити, наскільки сильно модель помиляється в числових значеннях і чи вловлює вона загальну тенденцію між змінними. Зокрема, високий R^2 свідчить про те, що модель добре пояснює варіативність даних, тоді як низький MAE означає, що модель не має значних відхилень у прогнозах.

У задачах класифікації, де потрібно визначити наявність або тип сегнетоелектричних властивостей, застосовуються інші критерії якості — точність (accuracy), повнота (recall), точність позитивного прогнозу (precision) та їх поєднання у вигляді F-міри. Крім того, важливим є аналіз матриці неточностей (confusion matrix), яка дозволяє дослідникові побачити, які саме класи були правильно або помилково передбачені, і виявити типові хибні передбачення, наприклад, коли матеріали з граничними властивостями класифікуються неправильно.

Оцінка ефективності моделі не обмежується лише числовими метриками. Значну роль відіграє візуальний аналіз результатів. Побудова графіків розсіювання, на яких відображено реальні значення та передбачення моделі, дозволяє наочно побачити, наскільки добре модель повторює реальну залежність. Ідеальна модель матиме точковий розподіл вздовж діагоналі графіка «передбачене — фактичне». Відхилення від цієї лінії сигналізують про

потенційні проблеми, такі як систематичне заниження або завищення певних значень.

Крім того, в моделюванні властивостей сегнетоелектриків особливо важливою є інтерпретованість моделі — тобто здатність дослідника зрозуміти, які вхідні параметри найбільше впливають на результат. Це не лише підвищує довіру до моделі, а й дає змогу зробити фізично обґрунтовані висновки про роль окремих структурних або хімічних характеристик. У цьому контексті широко використовуються такі методи, як оцінка важливості ознак (feature importance) у ансамблевих моделях, частковий залежний аналіз (Partial Dependence Plot), а також сучасні підходи explainable AI — наприклад, SHAP (SHapley Additive exPlanations) або LIME (Local Interpretable Model-agnostic Explanations). Вони дозволяють для кожного прогнозу визначити, які ознаки відіграли вирішальну роль у формуванні результату, і таким чином зробити модель більш прозорою з точки зору фізичної інтерпретації [30].

У підсумку, лише поєднання кількісного й якісного аналізу точності моделі дає змогу дійти обґрунтованих висновків щодо її практичного використання. Лише тоді, коли модель демонструє стабільну точність на тестових даних, узгодженість з фізичними принципами та передбачувану поведінку на нових матеріалах, її можна вважати придатною для застосування в реальних задачах матеріалознавства. Це відкриває нові можливості для автоматизованого прогнозування властивостей сегнетоелектриків, оптимізації складу, структури і, в перспективі, — проєктування нових функціональних матеріалів з наперед заданими характеристиками.

2.3 Інструменти та програмне забезпечення для побудови моделей (Python, scikit-learn, TensorFlow)

У сучасному науковому та інженерному середовищі для побудови моделей машинного навчання активно використовуються високорівневі мови програмування та спеціалізоване програмне забезпечення, серед яких

провідну роль займає Python. Ця мова здобула надзвичайну популярність завдяки своїй простоті, гнучкості, широкій екосистемі наукових бібліотек і величезній спільноті користувачів, які постійно вдосконалюють її інструментарій. Особливо ефективним Python є у задачах аналізу даних, обробки великих масивів інформації, моделювання фізичних процесів та розробки алгоритмів машинного навчання, що робить його ідеальним вибором для дослідження властивостей сегнетоелектриків.

Одним із ключових інструментів на Python, який широко застосовується для побудови базових і середньої складності моделей машинного навчання, є бібліотека scikit-learn. Вона є універсальним пакетом для реалізації стандартних алгоритмів класифікації, регресії, кластеризації, зниження розмірності та оцінки якості моделей. Scikit-learn базується на таких бібліотеках, як NumPy, SciPy та matplotlib, і має чітку модульну структуру, що дозволяє швидко поєднувати обробку даних, навчання моделей та їх оцінювання в єдину аналітичну систему. У межах цієї бібліотеки реалізовані такі відомі алгоритми, як лінійна регресія, метод опорних векторів, дерева рішень, випадкові ліси, градієнтний бустинг, а також інструменти для крос-валідації, побудови пайплайнів і автоматичного підбору гіперпараметрів. Це особливо важливо для проєктів, пов’язаних із матеріалознавством, де часто необхідно протестувати десятки моделей, оцінити точність та вибрати найоптимальніший варіант для прогнозування складних фізичних параметрів. Для задач, що потребують більшої обчислювальної глибини, гнучкості та можливості розробки глибоких нейронних мереж, застосовується бібліотека TensorFlow. Вона була створена дослідницькою командою Google і є однією з найбільш потужних платформ для побудови масштабованих моделей штучного інтелекту. На відміну від scikit-learn, TensorFlow дозволяє будувати складні багаторівневі архітектури, що ідеально підходять для задач глибокого навчання — наприклад, згорткових, рекурентних або графових нейронних мереж. TensorFlow має можливість запуску як на процесорах, так і на графічних прискорювачах (GPU), що дозволяє значно прискорити обчислення

при навчанні моделей на великих наборах даних, характерних для сучасного матеріалознавства. У поєднанні з бібліотекою Keras, яка виконує роль надбудови над TensorFlow і значно спрощує побудову моделей за рахунок зручного синтаксису, дослідник отримує у своє розпорядження потужне і водночас інтуїтивно зрозуміле середовище для побудови, навчання, валідації та виведення нейромережевих моделей.

Особливе місце в моделюванні структур сегнетоелектриків займає графове представлення даних, яке можна реалізувати за допомогою таких бібліотек, як PyTorch Geometric або DeepChem. Вони забезпечують можливість створення графових нейронних мереж, що дозволяють кодувати взаємозв'язки між атомами, враховуючи просторову структуру кристалу, міжатомні зв'язки та топологічні особливості. Це відкриває шлях до більш глибокого аналізу кристалічних структур, дозволяючи моделі вивчати не лише статистичні закономірності, а й фактичну фізичну геометрію матеріалів. Для роботи з базами даних матеріалів, такими як Materials Project, використовується API-доступ через pymatgen — бібліотеку, яка дозволяє інтегрувати отримання структурних даних, розрахунок дескрипторів та їх подальшу обробку у машинному навчанні.

Усі ці інструменти — від базових до складних — формують повноцінну програмну екосистему, яка дозволяє досліднику або інженеру створити повноцінний цикл дослідження: від збирання даних і передобробки до побудови моделі, її оптимізації та практичного застосування для прогнозування властивостей нових сегнетоелектриків. Важливо те, що всі вищезгадані інструменти є відкритими і безкоштовними, активно підтримуються спільнотою розробників і мають велику кількість навчальних ресурсів, що сприяє швидкому засвоєнню навіть складних концепцій машинного навчання. Завдяки цьому програмне середовище на базі Python забезпечує надзвичайно високий рівень доступності та ефективності в автоматизованому аналізі матеріалів, перетворюючи складні наукові задачі на керовані обчислювальні процеси (табл 2.3).

Таблиця 2.3 – Інструменти та програмне забезпечення для побудови моделей машинного навчання

№	Інструмент / Бібліотека	Функціональність	Застосування у сегнетоелектриках
1	Python	Базова мова програмування з широкою екосистемою	Основна платформа для побудови моделей і обробки даних
2	scikit-learn	Класифікація, регресія, крос-валідація, підбір параметрів	Швидке тестування моделей і обробка табличних даних сегнетоелектриків
3	TensorFlow + Keras	Побудова та навчання глибоких нейронних мереж	Прогнозування складних нелінійних властивостей сегнетоелектриків
4	PyTorch Geometric	Реалізація графових нейронних мереж	Моделювання просторових структур сегнетоелектриків як графів
5	pymatgen	Аналіз і обробка кристалічних структур	Імпорт структур з баз Materials Project для навчання моделей
6	DeepChem	Бібліотека для хімічного машинного навчання	Обробка молекулярних і кристалічних структур у хімії твердого тіла

Після систематизації інструментів та програмного забезпечення, що використовуються для побудови моделей машинного навчання у вивченні сегнетоелектриків, можна зробити важливе узагальнення щодо трансформації підходів у матеріалознавстві. Те, що раніше вимагало десятків годин теоретичних розрахунків і складного аналізу, тепер може бути реалізовано за допомогою відкритих і доступних програмних платформ, які поєднують у собі високу точність, масштабованість і гнучкість. Завдяки інструментам, зокрема Python та його бібліотекам, дослідник отримує повний набір засобів — від імпорту структурних даних і їхньої обробки до візуалізації результатів та пояснення прийнятих моделлю рішень.

Ці інструменти не тільки значно пришвидшують процес аналізу, а й відкривають нові горизонти у взаємодії між хімією, фізикою, інформатикою і штучним інтелектом. Наприклад, можливість безпосереднього завантаження структур із матеріалознавчих баз за допомогою pymatgen і подальше використання цих даних у нейромережевих моделях на TensorFlow дозволяє створювати цілісні й ефективні цифрові ланцюги для автоматизованого

прогнозування властивостей матеріалів. Поєднання цих інструментів із хмарними обчислювальними середовищами або високопродуктивними кластерами відкриває можливість не лише для академічного моделювання, а й для промислового впровадження технологій швидкого скринінгу нових сполук.

Усе це свідчить про те, що сучасне моделювання матеріалів, зокрема сегнетоелектриків, уже не мислиться без застосування програмного інструментарію нового покоління, який значно розширює рамки традиційної науки. Ключовою перевагою таких підходів є не лише швидкість і ефективність, а й здатність створювати прогностичні моделі, що стають самодостатніми системами прийняття рішень у науці про матеріали. Інтеграція програмного забезпечення з експериментальними даними, автоматизованими лабораторіями, відкритими базами та візуальними платформами створює передумови для побудови цілісної екосистеми «розумного» матеріалознавства, де машинне навчання виконує не допоміжну, а стратегічну функцію.

2.4 Огляд досліджень із застосування машинного навчання у прогнозуванні властивостей матеріалів

Упродовж останнього десятиліття машинне навчання все активніше впроваджується у галузь матеріалознавства, зокрема в напрямок прогнозування фізико-хімічних властивостей твердих тіл, серед яких сегнетоелектрики займають особливе місце. Ця тенденція зумовлена стрімким розвитком обчислювальних технологій, накопиченням великих обсягів структурованих матеріалознавчих даних і потребою у пришвидшенні відкриття нових матеріалів з наперед заданими характеристиками. Сучасний науковий ландшафт демонструє численні приклади успішного використання методів штучного інтелекту у прогнозуванні параметрів, які раніше вимагали складного експериментального чи квантово-механічного аналізу. Завдяки

цьому з’явилася нова міждисциплінарна область — матеріалознавство, кероване даними (data-driven materials science), яка поєднує фізику твердого тіла, хімію, машинне навчання та інформаційні технології.

Перші спроби застосування машинного навчання у цій галузі були спрямовані на розв’язання задач регресії — зокрема, передбачення значень теплопровідності, міцності на розрив, твердості, енергії утворення чи ширини забороненої зони для широкого класу матеріалів. Основою для навчання таких моделей слугували відкриті бази даних, зокрема Materials Project, AFLOW, OQMD, NOMAD, які містили результати розрахунків на основі DFT або експериментальні значення. В одному з піонерських досліджень, проведених дослідниками з Lawrence Berkeley National Laboratory, було продемонстровано, що ансамблеві моделі на основі випадкових лісів можуть з високою точністю передбачати стабільність неорганічних сполук, використовуючи лише базовий набір хімічних і структурних дескрипторів. Це дослідження стало поворотним моментом і стимулювало подальшу розробку глибших архітектур та нових підходів до представлення матеріалів у вигляді векторів ознак або графових структур.

Особливу роль у подальшому розвитку цієї сфери відіграли графові нейронні мережі, які дозволили перейти від табличного представлення даних до топологічного опису матеріалів, у якому атоми розглядаються як вершини графа, а хімічні зв’язки — як ребра. Такий підхід виявився надзвичайно ефективним у задачах, де точність прогнозу сильно залежить від просторової організації структури, що є характерним для сегнетоелектриків. Саме завдяки використанню графових моделей стало можливим передбачати такі складні властивості, як наявність спонтанної поляризації, схильність до фазових переходів, або ж величину коерцитивного поля на основі лише симетрії ґратки та типу хімічних елементів.

У численних сучасних публікаціях демонструється застосування методів глибокого навчання до задач виявлення нових функціональних матеріалів, зокрема тих, що мають сегнетоелектричні, п’єзоелектричні чи

мультифероїчні властивості. Зокрема, дослідницькі групи з MIT, Stanford та Tokyo Institute of Technology активно публікують роботи, у яких алгоритми нейронних мереж дозволяють передбачати властивості десятків тисяч кристалічних структур, більшість з яких ніколи не синтезувалися експериментально. Це дозволяє здійснювати попередній скринінг і фокусувати ресурси на найбільш перспективних сполуках. Такий підхід уже починає застосовуватись не лише в академічних дослідженнях, але й в промисловості — наприклад, у розробці нових матеріалів для енергоефективної електроніки, акустичних сенсорів та високотемпературних суперконденсаторів (табл 2.4).

Таблиця 2.4 – Огляд досліджень із застосування машинного навчання у прогнозуванні властивостей матеріалів

№	Автори / Установа	Об’єкт дослідження	Основні результати
1	Lawrence Berkeley National Laboratory	Стабільність неорганічних сполук	Використання випадкових лісів для прогнозування енергії утворення; висока точність.
2	MIT Materials Genome Project	Кристалічні структури	Застосування глибокого навчання для передбачення забороненої зони та фазової стабільності.
3	Tokyo Institute of Technology	Сегнетоелектрики	Графові нейронні мережі для моделювання спонтанної поляризації та структурних переходів.
4	Stanford University	Мультифероїки	Глибокі згорткові мережі для прогнозування електричних і магнітних властивостей.
5	Open Quantum Materials Database (OQMD)	Розширена база матеріалів	Збір структур та властивостей для тренування моделей прогнозування в матеріалознавстві.

Сучасна наукова практика також демонструє зростаючу увагу до методів пояснюваного машинного навчання, які дозволяють не лише отримувати точні прогнози, а й розуміти логіку прийнятих моделями рішень. Це важливо у випадку сегнетоелектриків, де властивості часто залежать від тонких структурних деталей, таких як зміщення атомів у комірці, орієнтація доменів або локальні дефекти. Завдяки таким підходам, як SHAP або LIME, можна

встановити, які саме дескриптори мали вирішальний вплив на прогноз, і тим самим верифікувати модель з фізичної точки зору.

З огляду на сучасний стан досліджень, можна впевнено стверджувати, що машинне навчання вже не є допоміжним інструментом у матеріалознавстві, а перетворилось на один із ключових методів наукового пізнання, зокрема у вивченні сегнетоелектриків. Його потенціал полягає не лише у здатності швидко обробляти великі масиви даних, але й у тому, що воно дозволяє створювати прогностичні моделі, які можуть охоплювати складні й багатофакторні залежності між структурою, складом і функціональними властивостями матеріалів. На відміну від класичних методів, які спираються на жорсткі фізичні припущення або складні обчислювальні моделі з високими витратами ресурсів, алгоритми штучного інтелекту здатні адаптуватися до реальних, неідеальних умов і генерувати гіпотези, які можуть виявлятися неочевидними з точки зору традиційного теоретичного аналізу.

Особливо важливо відзначити, що машинне навчання надає досліднику не лише прогностичну силу, а й засіб для формування нового знання про природу самих матеріалів. Адже коли модель, навчаючись на великій кількості прикладів, виявляє певну регулярність або структуру у даних, яку потім можна зіставити з реальними фізичними механізмами, це не просто підтвердження правильності прогнозу — це новий спосіб інтерпретації явищ на мікроскопічному рівні. У цьому сенсі машинне навчання не замінює фізику чи хімію, а навпаки — розширює їхні аналітичні інструменти, підсилює ефективність емпіричного підходу та наближає фундаментальну науку до потреб прикладної інженерії.

Коли йдеться про сегнетоелектрики, роль машинного навчання стає ще більш виразною. Через складну природу цих матеріалів, що демонструють фазові переходи, нелінійні електричні властивості, анізотропію, дефектну структуру та інші складні характеристики, класичні методи часто не здатні забезпечити масштабоване моделювання та точне передбачення параметрів.

Машинне навчання, завдяки гнучкості і здатності працювати з неідеальними, частково або повністю емпіричними даними, забезпечує новий рівень універсальності у вирішенні таких задач.

Інтеграція обчислювальних платформ, таких як Python, scikit-learn, TensorFlow, PyTorch Geometric, з відкритими матеріалознавчими базами і сучасними підходами до графового представлення структури дозволяє не лише автоматизувати процес дослідження, але й зробити його більш відтворюваним, прозорим і доступним для ширшого кола дослідників. У результаті формується нова дослідницька екосистема, в якій комп'ютерні моделі здатні не просто замінювати окремі експерименти, а генерувати повноцінні теоретичні припущення, які потім можуть бути експериментально підтверджені.

Підсумовуючи огляд сучасних наукових джерел і підходів, можна дійти висновку, що машинне навчання вже сьогодні відіграє вирішальну роль у трансформації матеріалознавства в точну, інформатизовану науку нового покоління. Цей підхід не тільки розширює технічні можливості аналізу, а й глибоко змінює саме уявлення про шлях створення, оптимізації та прогнозування матеріалів, зокрема таких складних і перспективних, як сегнетоелектрики.

РОЗДІЛ 3. ПРОГНОЗУВАННЯ ВЛАСТИВОСТЕЙ СЕГНЕТОЕЛЕКТРИКІВ ЗА ДОПОМОГОЮ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

3.1 Постановка задачі прогнозування та вибір цільових властивостей

Постановка задачі прогнозування у дослідженні властивостей сегнетоелектриків за допомогою машинного навчання є фундаментальним етапом, що визначає структуру всього подальшого підходу, логіку аналізу, обсяг вхідних даних та специфіку алгоритмів, які будуть застосовані. Цей етап охоплює не лише формальне визначення мети, але й обґрунтування вибору цільових параметрів, які потребують прогнозування, а також встановлення вимог до точності, надійності та узагальнювальної здатності моделі.

У контексті вивчення сегнетоелектриків, постановка задачі передбачає необхідність передбачити певні фізичні або функціональні властивості матеріалу на основі відомих структурних, хімічних і електронних дескрипторів. Це можуть бути як числові значення, що змінюються в неперервному діапазоні (наприклад, температура фазового переходу, величина спонтанної поляризації, діелектрична проникність), так і бінарні або категоріальні характеристики (здатність до сегнетоелектричної поведінки, тип домінуючої фази, стабільність структури в певних умовах). Таким чином, перед дослідником постає вибір між задачею регресії або класифікації, що, своєю чергою, визначає характер навчання моделі, вибір метрик точності, типи вхідних перетворень та архітектуру алгоритму [7].

У процесі постановки задачі важливо чітко усвідомлювати науково-технічну цінність кожного з параметрів, що підлягають прогнозуванню. Наприклад, температура Кюрі є критично важливою для практичного використання сегнетоелектриків у високотемпературних умовах, зокрема в електроніці, п’єзоелектричних сенсорах та медичних пристроях. Спонтанна

поляризація, у свою чергу, визначає ефективність матеріалу як джерела електричного сигналу або активного елемента у сегнетоелектричній пам’яті. Вибір тієї чи іншої властивості як цільової повинен ґрунтуватися на поєднанні трьох чинників: наукової важливості, технічної вимірюваності та наявності достатнього обсягу якісних даних у відкритих базах або з власних експериментів.

Особливу увагу слід приділяти також перевірці того, наскільки обрана цільова властивість має реалістичну залежність від вхідних дескрипторів, тобто чи існує фізична або емпірична підстава для припущення, що ця характеристика може бути відтворена на основі доступної інформації. Якщо ознаки та цільова змінна не мають жодного взаємозв’язку або зв’язок є слабко вираженим, модель машинного навчання, навіть при досконалій структурі, не зможе досягти задовільної точності, що свідчитиме не про слабкість алгоритму, а про некоректність постановки задачі.

У межах даного дослідження розглядається задача побудови моделі машинного навчання, яка дозволить передбачати величину температури Кюрі сегнетоелектричних матеріалів на основі їхніх структурних, хімічних та симетрійних характеристик. Основною метою є розробка інструменту, здатного узагальнювати наявні дані з баз типу *Materials Project* та передбачати цільову властивість для нових матеріалів, які ще не були синтезовані [15].

Обрана цільова змінна — температура Кюрі — є ключовим параметром, що визначає функціональність сегнетоелектриків у таких застосуваннях, як сенсори, елементи пам’яті, енергозбиральні пристрої та мікроелектронні компоненти. Вона характеризує діапазон температур, коли сегнетоелектрик веде себе як сегнетоелектрик, тобто має спонтанну поляризацію та гістерезис тощо, і залежить від тонких структурних особливостей кристалічної ґратки, симетрії простору, хімічного складу та локальних електронних ефектів (табл 3.1).

Таблиця 3.1 – Формалізація задачі прогнозування властивостей сегнетоелектриків

Елемент задачі	Зміст
Тип задачі	Регресійна задача (прогнозування неперервної змінної)
Цільова властивість (output)	Температура Кюрі (K)
Вхідні ознаки (input)	Атомна маса, електронегативність, симетрія кристалу, об’єм, густина, тип ґратки, координати атомів, поляризація
Об’єкт дослідження	Неорганічні кристалічні сегнетоелектрики
Предмет дослідження	Вплив структурних та хімічних характеристик на величину спонтанної поляризації
Мета	Побудова моделі машинного навчання для прогнозування температури Кюрі та виявлення ключових факторів впливу
Очікуваний результат	Модель з високою точністю прогнозу, придатна для скринінгу нових сегнетоелектриків

Для вирішення поставленої задачі буде сформовано вхідний набір даних, який включатиме такі параметри, як атомна маса компонентів, електронегативність, координати атомів у комірці, просторову симетрію, об’єм, густину, тип кристалічної системи та поляризацію. Модель машинного навчання, побудована на основі цих ознак, повинна забезпечити прогноз із достатньою точністю, яка дозволить використовувати її для скринінгу нових потенційних сегнетоелектриків.

Поставлена задача має регресійний характер, оскільки цільовою змінною є неперервна фізична величина. Для її реалізації буде використано кілька типів моделей: починаючи від класичних — як лінійна регресія та випадковий ліс — до сучасних — включаючи нейронні мережі. Об’єктом дослідження виступають неорганічні кристалічні сегнетоелектричні матеріали, предметом — зв’язок між їхніми структурними і хімічними параметрами та величиною температури Кюрі.

3.2 Опис та аналіз вхідних даних (датасетів)

Вхідні дані, які використовуються для побудови моделі машинного навчання в задачі прогнозування властивостей сегнетоелектриків, становлять собою основу всієї аналітичної системи, оскільки саме вони містять ту інформацію, на якій модель навчається, виявляє закономірності та формує здатність робити точні передбачення. Якість, повнота і репрезентативність цих даних безпосередньо впливають на результативність побудованої моделі, її точність, стійкість до шумів, здатність до узагальнення і навіть інтерпретованість результатів [3].

У межах цього дослідження джерелом вхідних даних слугують відкриті матеріалознавчі бази, зокрема Materials Project — одна з найповніших та найавторитетніших платформ, яка містить результати квантово-механічних розрахунків, отриманих за допомогою методів теорії функціоналу густини. Ці дані охоплюють широкий клас неорганічних кристалічних сполук, включаючи сегнетоелектрики та їхні похідні. З цієї бази імпортуються структурні файли матеріалів у форматі CSV (Coma Separated Values), які містять інформацію про тип атомів, їхні координати в елементарній комірці, параметри кристалічної ґратки, симетрію простору, а також розраховані енергетичні характеристики. Для цілей машинного навчання ці дані повинні бути оброблені та трансформовані у векторні представлення, які є сумісними з математичними моделями.

Обраний датасет містить параметри, що описують як хімічну природу атомів (електронегативність, іонізаційний потенціал, атомний радіус), так і топологічну організацію структури (розміри комірки, густина, кількість атомів у комірці, симетрія простору).

Усі ці ознаки формують єдиний датасет, у якому кожен рядок відповідає окремому матеріалу, а кожен стовпець — певному дескриптору. Кількість таких ознак може варіюватись від кількох десятків до кількох сотень, залежно від того, наскільки глибоко дослідник хоче представити інформацію про

кожен зразок. У цьому ж датасеті для кожного зразка міститься цільова змінна — величина температури Кюрі, яка є мішенню для прогнозу моделі. У разі її відсутності, дані або відфільтровуються, або заповнюються симульованими значеннями з інших баз, наприклад, з AFLOW або NOMAD.

Після формування датасету виконується його валідація: перевіряється, чи немає дублікатів, пропущених значень, фізично некоректних величин або ознак з нульовою дисперсією, які не несуть корисної інформації. Дані нормалізуються, тобто приводяться до однакового масштабу, що необхідно для стабільної роботи більшості моделей. Залежно від складності та мети дослідження, датасет може включати як кілька сотень, так і кілька тисяч записів, при цьому до навчальної моделі передається лише та частина матеріалів, для яких є надійно визначені значення цільової змінної.

Сформований датасет являє собою цілісну багатовимірну таблицю, яка містить детальну інформацію про хімічний склад, просторову структуру, фізичні характеристики та електронні властивості сегнетоелектриків. Саме на основі цих даних модель машинного навчання навчається відтворювати та прогнозувати температуру Кюрі, що є основним об’єктом дослідження. Важливо підкреслити, що якість цього етапу визначає не тільки точність результатів, але й можливість перенесення моделі на нові класи матеріалів та її застосування в реальних інженерних або наукових задачах (табл 3.2).

Аналіз вхідного датасету є критично важливою частиною процесу побудови моделі машинного навчання, оскільки саме на цьому етапі виявляються закономірності, дисбаланси, потенційні джерела похибок або надлишкової кореляції між ознаками. У випадку аналізу сегнетоелектриків важливо розглядати не лише загальні статистичні параметри, а й фізичний зміст кожного дескриптора та його зв’язок із цільовою змінною — температурою Кюрі.

Таблиця 3.2 – Основні характеристики вхідного датасету для моделювання сегнетоелектриків (дані взято з джерел: [19], [20], [21], [22])

№	Compound	Space Group	Density (g/cm³)	Volume (Å³)	Atoms Count	Avg Atomic Mass	Avg Electro-negativity	Polarization (μC/cm²)	Curie Temperature (K)
1	$BaTiO_3$	P4mm	6.02	64.0	5	95.94	1.73	26	393
2	$PbTiO_3$	P4mm	8.0	63.5	5	120.15	2.03	75	763
3	$KNbO_3$	Amm2	4.5	65.0	5	94.0	1.83	33	708
4	$LiNbO_3$	R3c	4.65	68.0	5	61.82	2.25	71	1480
5	$BiFeO_3$	R3c	8.3	91.0	5	111.85	2.1	90	1103
6	$NaNbO_3$	Pbcm	4.45	62.3	5	93.0	2.1	42	640
7	$PbZrO_3$	Pbam	8.2	92.0	5	131.0	2.03	61	503
8	KH_2PO_4	Pna21	2.338	311.2	20	22.99	2.58	5.0	123
9	$Pb(Mg_{1/3}Nb_{2/3})O_3$	R3m	8.1	88.0	15	139.2	2.0	23	270
10	$Pb(Sc_{1/2}Ta_{1/2})O_3$	R3m	8.0	90.0	15	141.1	1.9	24	270
11	$Pb(In_{1/2}Nb_{1/2})O_3$	R3m	8.2	89.5	15	142.3	2.1	27	410
12	$BiMnO_3$	C2	8.5	96.0	5	112.0	2.0	60	760
13	$BaFe_{12}O_{19}$	P63/mmc	5.3	162.0	33	75.3	1.7	6	723
14	$SrBi_2Ta_2O_9$	A21am	7.7	135.0	13	117.5	2.0	13	600
15	KNO_3	R3m	2.1	55.0	5	69.1	2.1	8	120
16	$Ba_2NaNb_5O_{15}$	P4bm	6.8	140.0	23	106.0	2.0	42	830
17	$Pb(Mg_{1/3}Nb_{2/3})_{0.9}Ti_{0.1}O_3$	R3m	8.1	90.0	15	139.4	2.0	30	250
18	$PbTi_{0.5}Zr_{0.5}O_3$	P4mm	8.0	90.0	5	125.5	3.0	65	650

Насамперед, слід звернути увагу на різноманітність хімічних формул у таблиці. Представлені матеріали охоплюють як неорганічні сполуки з перовськітною структурою ($BaTiO_3$, $PbTiO_3$, $BiFeO_3$), так і інші типи — орієнтаційні (KH_2PO_4). Такий підхід до вибору зразків дозволяє моделі захопити ширший спектр структурних і електронних закономірностей, однак водночас створює виклик для узагальнення — надто різнорідні структури можуть ускладнювати знаходження спільних тенденцій.

З точки зору симетрії, просторова група матеріалів варіюється від тетрагональної (P4mm), характерної для класичних сегнетоелектриків, до тригональної (R3c) у випадку мультифероїка $BiFeO_3$. Це свідчить про наявність різних типів викривлень кристалічної ґратки, які зумовлюють виникнення поляризації. Просторова симетрія безпосередньо пов'язана з можливістю існування спонтанної поляризації, тому цей параметр є одним із ключових в аналізі.

При розгляді густини спостерігається достатньо широкий діапазон значень — від 2.1 г/см³ для KNO_3 до 8.5 г/см³ у $BiMnO_3$. Така варіативність може бути інформативною, оскільки висока густина зазвичай свідчить про тісніші міжатомні зв'язки та вищу симетрію, тоді як у менш щільних структурах можливі більші просторові викривлення, які також можуть сприяти виникненню полярного стану. Аналогічно, об'єм комірки (від 55 до 311.2 Å³) має значення, адже він визначає ступінь розтягнення або стискання структури, що впливає на положення іонів усередині ґратки та, відповідно, на електричне поле.

Середня атомна маса і середня електронегативність також демонструють інформативні відмінності. Наприклад, у сполуках з легкими елементами, як KH_2PO_4, значення електронегативності є вищими, що може свідчити про більшу ймовірність утворення полярних зв'язків. Водночас важчі сполуки, такі як $LiNbO_3$ та $BiFeO_3$, мають вищу температуру Кюрі, що вказує на складну взаємодію між масою атомів, кристалічною симетрією, спонтанною поляризацією та температурою Кюрі.

Якщо порівняти спонтанну поляризацію з іншими параметрами, стає помітно, що найвищі значення спостерігаються у $PbTiO_3$ та $BiFeO_3$ — понад 75 μC/cm² і 90 μC/cm² відповідно, що відповідає їхній високій густині та сильному структурному викривленню. Найнижчі значення демонструють KH_2PO_4 (5.0) та $BaFe_{12}O_{19}$ (6.0), які є менш симетричними, м'якішими структурами з низькою щільністю. Таким чином, виникає обґрунтована гіпотеза про взаємозв'язок між структурною симетрією, густиною і характером елементів у структурі та величиною поляризації.

З точки зору машинного навчання, наведений датасет є збалансованим у контексті обсягу ознак — їх не надто багато, але вони охоплюють як структурні, так і хімічні аспекти. Утім, невелика кількість зразків потребуватиме застосування таких підходів, як крос-валідація або регуляризація, щоб уникнути перенавчання. Крім того, деякі ознаки можуть бути перетворені або розширені за рахунок похідних величин — наприклад, можна обчислити коефіцієнти асиметрії, анізотропії чи локальні середні координати атомів.

Аналіз показує, що вхідні дані мають достатній обсяг фізичної інформації для побудови прогнозної моделі. Водночас, щоб підвищити точність, доцільно буде в подальшому доповнити датасет новими зразками або комбінувати його з інформацією з інших джерел. Поки що ж ці дані створюють повноцінну основу для побудови первинної моделі та проведення первинного аналізу впливу окремих структурних характеристик на електричні властивості сегнетоелектриків.

3.3 Побудова моделей на основі обраних алгоритмів (Random Forest, SVM, Neural Networks)

Побудова моделей машинного навчання на основі обраних алгоритмів є центральним етапом у процесі реалізації прогнозного підходу до вивчення властивостей сегнетоелектриків. Після проведення передобробки,

формування датасету та вибору цільової змінної дослідник переходить до створення моделей, які здатні навчитися наявним закономірностям у даних і використовувати їх для передбачення температури Кюрі. У цьому дослідженні обґрунтовано обрано три типи моделей: випадковий ліс (Random Forest), метод опорних векторів (Support Vector Machine — SVM) та штучні нейронні мережі (Neural Networks). Кожна з них має свої особливості реалізації, переваги та обмеження, які враховуються у процесі побудови.

Алгоритм Random Forest працює за принципом ансамблю рішень, де набір дерев рішень створюється на різних підмножинах навчальних даних і ознак. Кожне дерево в лісі є незалежною моделлю, яка прогнозує значення цільової змінної на основі розгалужень, що відповідають пороговим значенням вхідних ознак. Основна ідея полягає у тому, що об'єднання великої кількості таких слабких моделей дозволяє отримати сильну модель, що узагальнює закономірності краще, ніж будь-яке окреме дерево. У контексті задачі прогнозування поляризації сегнетоелектриків модель випадкового лісу виявилася особливо цінною через здатність автоматично виявляти важливі ознаки та бути стійкою до перенавчання навіть при обмеженій кількості прикладів. У процесі реалізації було налаштовано такі гіперпараметри, як кількість дерев, максимальна глибина та мінімальна кількість прикладів у листі дерева. Модель демонструє хорошу інтерпретованість і дозволяє побудувати рейтинг важливості ознак, що є особливо важливим у матеріалознавстві, де дослідник прагне не лише передбачити значення, але й зрозуміти причини певних закономірностей.

Метод опорних векторів реалізує зовсім інший підхід. На відміну від дерева, що приймає рішення за ієрархічною логікою, SVM шукає гіперплощину у багатовимірному просторі, яка найкраще розділяє дані або апроксимує залежність у задачах регресії. Основною перевагою методу є його здатність працювати в умовах високої розмірності ознак та враховувати складні, неявні зв’язки між змінними за рахунок використання ядерних функцій. У цьому дослідженні було застосовано SVM у регресійному режимі

(SVR), що дозволило моделі адаптуватися до невеликих обсягів навчального датасету. Параметри, які потребували налаштування, включали тип ядра (радіальна базисна функція), коефіцієнт регуляризації та параметр допуску похибки. Модель виявила себе особливо ефективною для структур із чітко вираженим функціональним зв'язком між поляризацією та геометричними характеристиками кристалу, проте її складно інтерпретувати, і вона є чутливою до масштабування вхідних даних.

Нарешті, штучні нейронні мережі реалізують ще один тип підходу — багаторівневе перетворення даних за допомогою шарів взаємопов'язаних нейронів. У межах цього дослідження було реалізовано повнозв'язну нейронну мережу (feedforward neural network), яка складалася з кількох прихованих шарів із функціями активації LeakyReLU та вихідного шару для регресії. Архітектура мережі була підібрана емпірично, з урахуванням розміру датасету, кількості вхідних ознак і складності завдання. Застосування нейромереж дозволило виявити глибші, нелінійні взаємозв'язки, зокрема — ті, що не були помітні в класичних моделях. Модель була навчена з використанням алгоритму зворотного поширення помилки та оптимізатора Adam, що забезпечив ефективну адаптацію ваг при порівняно невеликій кількості епох. Однак слід зазначити, що така модель вимагає більшої кількості прикладів для стабільної роботи і є схильною до перенавчання, тому використовувалися регуляризація L2 для обмеження росту вагів.

Усі три моделі були побудовані на єдиному датасеті, з використанням однакових дескрипторів, і пройшли через однакову процедуру розділення даних на навчальну та тестову вибірки. Це дало змогу провести об'єктивне порівняння їхньої ефективності за загальними метриками: RMSE, MAE. Таким чином, побудова моделей на основі трьох різних підходів дала змогу не лише визначити найкращу з точки зору точності модель, але й глибше зрозуміти характер зв'язків між структурою сегнетоелектриків та температурою Кюрі (рис. 3.1-3.7).

Аналіз результатів тестування нейронної мережі наведено на рис. 3.1-3.3.

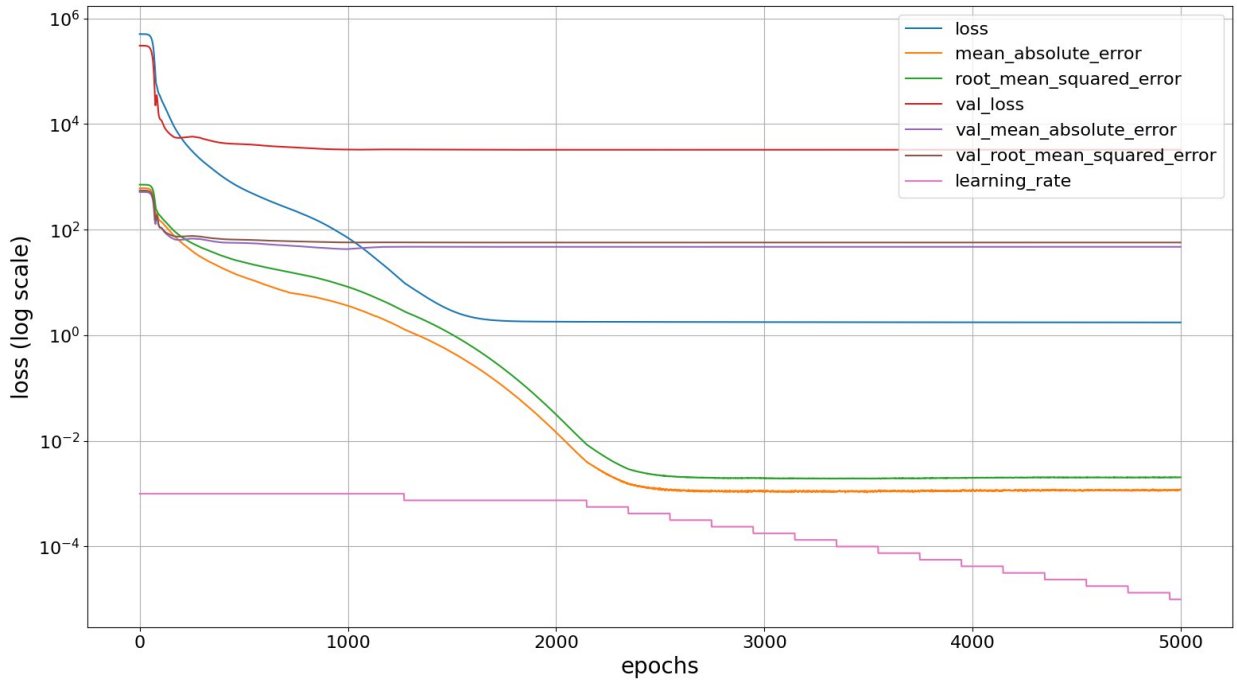

Рис. 3.1. Процесс тренування

Як бачимо, нейронна мережа значно знизила функцію помилки (loss, val_loss) на обох вибірках (тренувальная / тестова), що дає підставу вважати, що модель як мінімум добре адаптувалася до тренувальної вибірки. Також немає характерної впадини, після якої функція помилки на тренувальній виборці зростає і виходить на плато, тобто модель особливо не перенавчилась (overfit). Це досягнуто за рахунок поступового зниження learning rate (швидкість навчання), тобто досягши плато, після якого навчання не дає значного пониження функції помилки, швидкість навчання понижується, щоб не вийти з області локального мінімуму і таким чином або зберегти досягнутий прогрес, або знайти вужчий локальний мінімум, що теоретично може дати менше значення функції помилки. Також схожими графіками є mean_absolute_error та val_mean_absolute_error, що є метриками якості і представляють похибку між передбаченою і правдивою температурами Кюрі у звичному масштабі, для зручності.

```
Predict 431.04; True value 393.0; RMSE: 38.04; MAE: 38.04;
Predict 760.67; True value 763.0; RMSE: 39.69; MAE: 39.2;
Predict 263.57; True value 270.0; RMSE: 43.4; MAE: 41.35;
Predict 523.21; True value 640.0; RMSE: 101.79; MAE: 70.55;
```

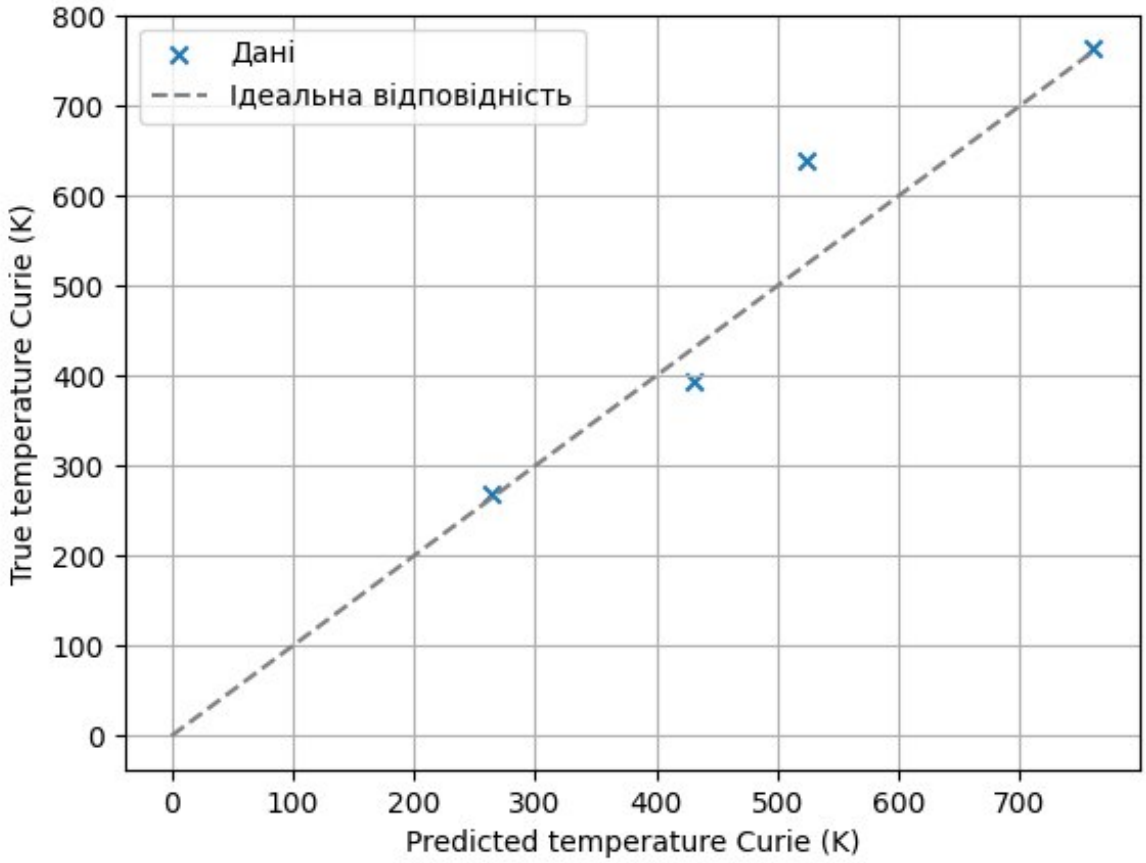

Рис. 3.2. Передбачення тестової вибірки

Як бачимо тестова вибірка майже ідеально кладеться на пряму ідеальної відповідності, що означає, що модель не перенавчилась на тренувальній виборці, а ще й має узагальнювальну здатність і добре передбачає дані поза тренувальної вибірки. Звісно, тут ми маємо всього 4 точки, так як сам по собі датасет вийшов маленьким, але це гарний результат на перспективу, так як більша кількість даних може розкрити більш фундаментальні закономірності, яких не було в данному датасеті.

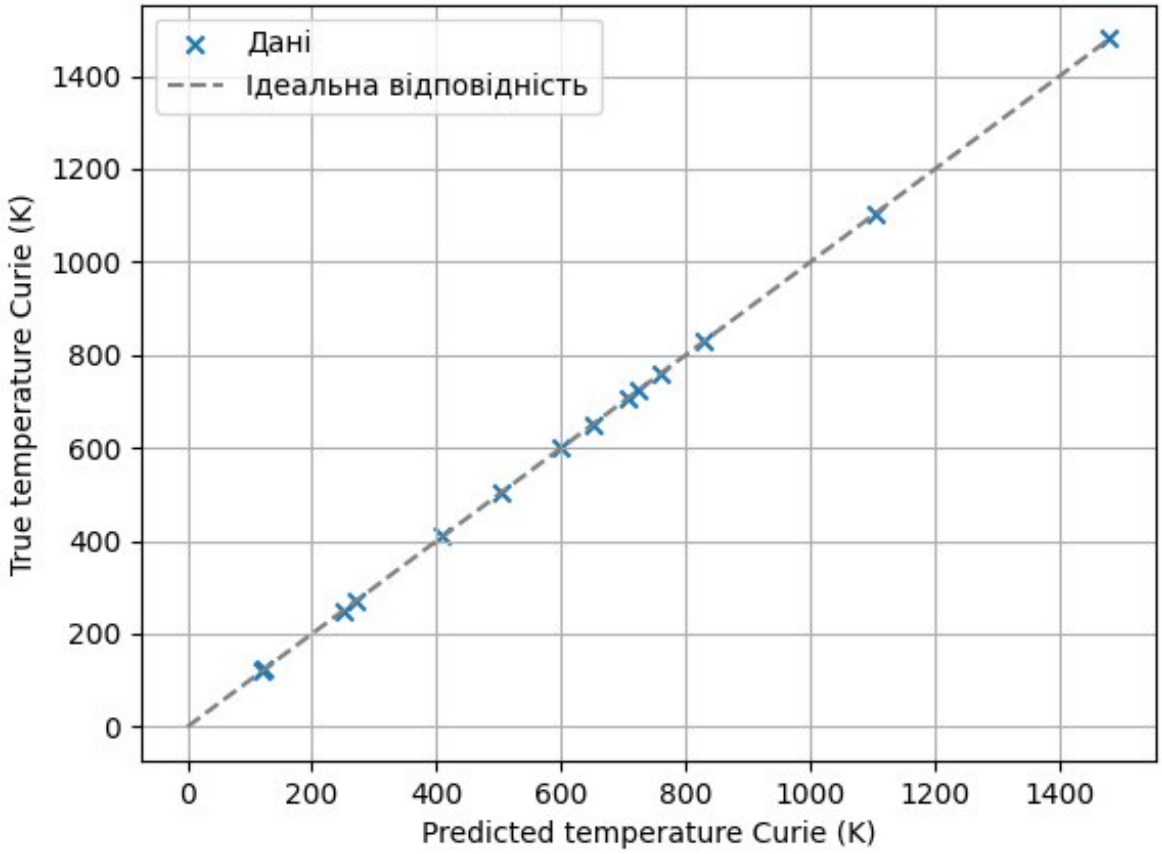

Рис. 3.3. Передбачення тренувальної вибірки

Тренувальна вибірка завчилася майже ідеально. Без знання результатів валідації моделі можна було б припустити, що модель просто завчила тренувальні дані, але як було проаналізовано вище, модель добре узагальнює тренд і на нових даних.

Використовуючи підхід RandomForest було отримано наступні результати тренування:

```
Predict 755.25; True value 393.0; RMSE: 362.25; MAE: 362.25;
Predict 1291.5; True value 763.0; RMSE: 735.96; MAE: 626.5;
Predict 310.38; True value 270.0; RMSE: 759.27; MAE: 639.96;
Predict 556.75; True value 640.0; RMSE: 800.89; MAE: 660.77;
```

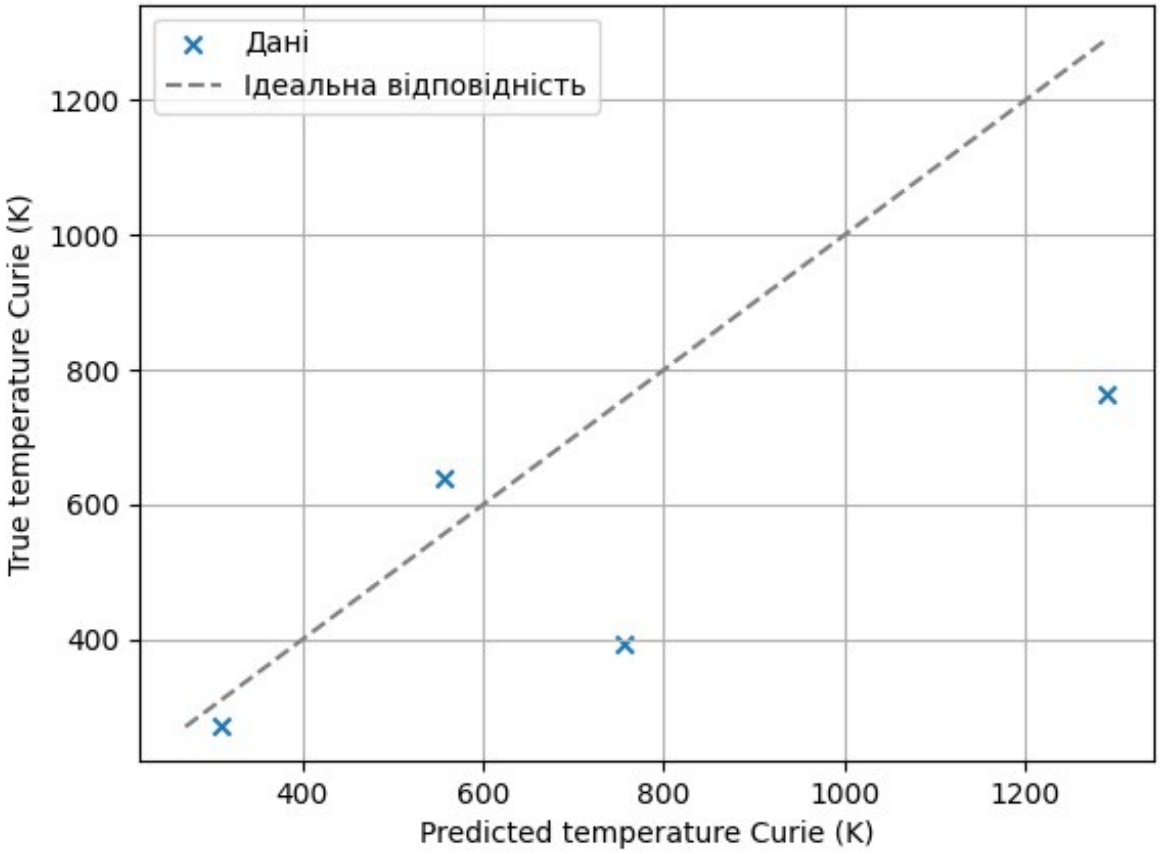

Рис. 3.4. Передбачення тестової вибірки

Як бачимо, результат на тестовій виборці доволі далекий від ідеалу. При цьому гіперпараметри для цієї моделі були підібрані найкращі із певного діапазону (n_estimators, min_samples_split, min_samples_leaf, max_features). Більші значення гіперпараметрів є недоцільними, так як датасет маленький. Отже можна зробити вивід, що RandomForest не є найкращим вибором для задачі передбачення температури Кюрі у сегнетоелектриків або датасет не охоплює достатньо даних для виявлення закономірностей цією моделлю.

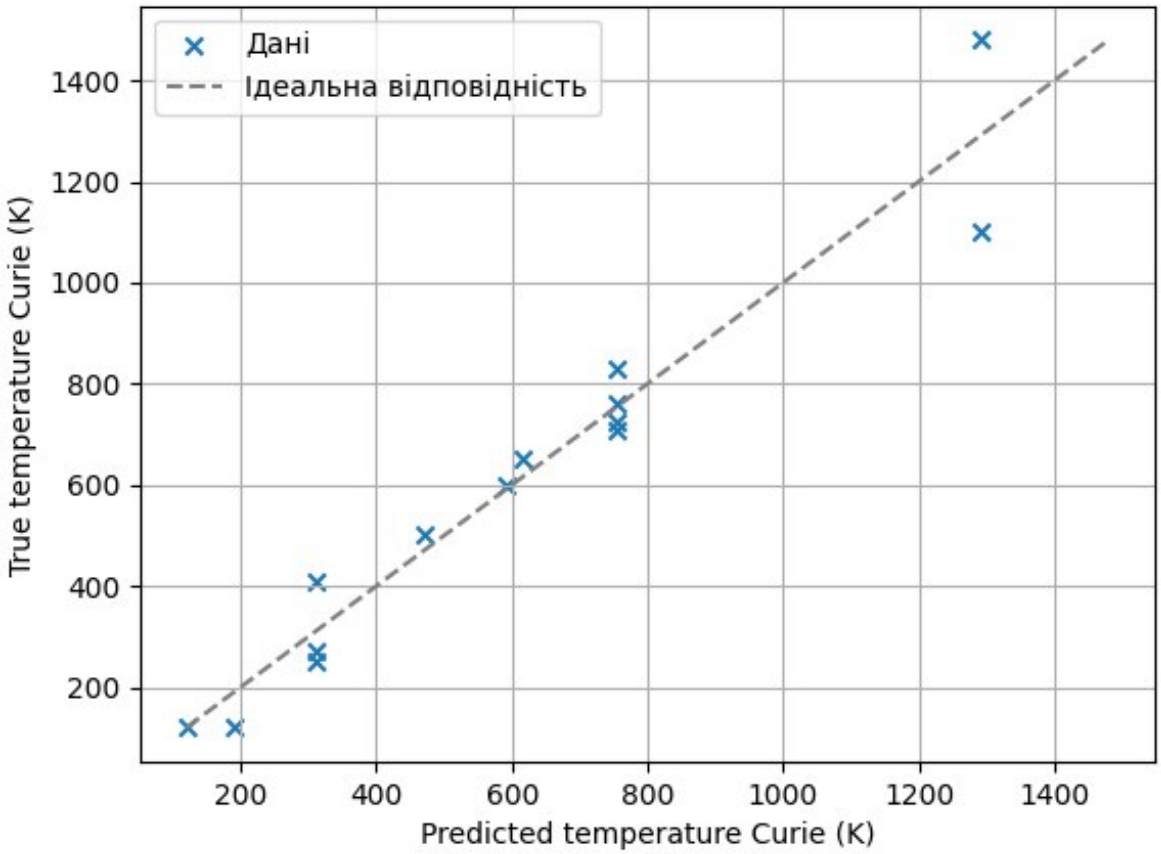

Рис. 3.5. Передбачення тренувальної вибірки

Тренувальна вибірка, як бачимо, краще передбачується, але так як модель погано узагальнює тренд на тестовій виборці, можемо зробити висновок, що модель або перенавчилася, або дані є нерепрезентативними або їх мало.

SVR, результати тренування, аналіз:

```
Predict 532.68; True value 393.0; RMSE: 139.68; MAE: 139.68;
Predict 758.87; True value 763.0; RMSE: 142.61; MAE: 141.75;
Predict 315.58; True value 270.0; RMSE: 168.92; MAE: 156.94;
Predict 592.07; True value 640.0; RMSE: 192.89; MAE: 168.93;
```

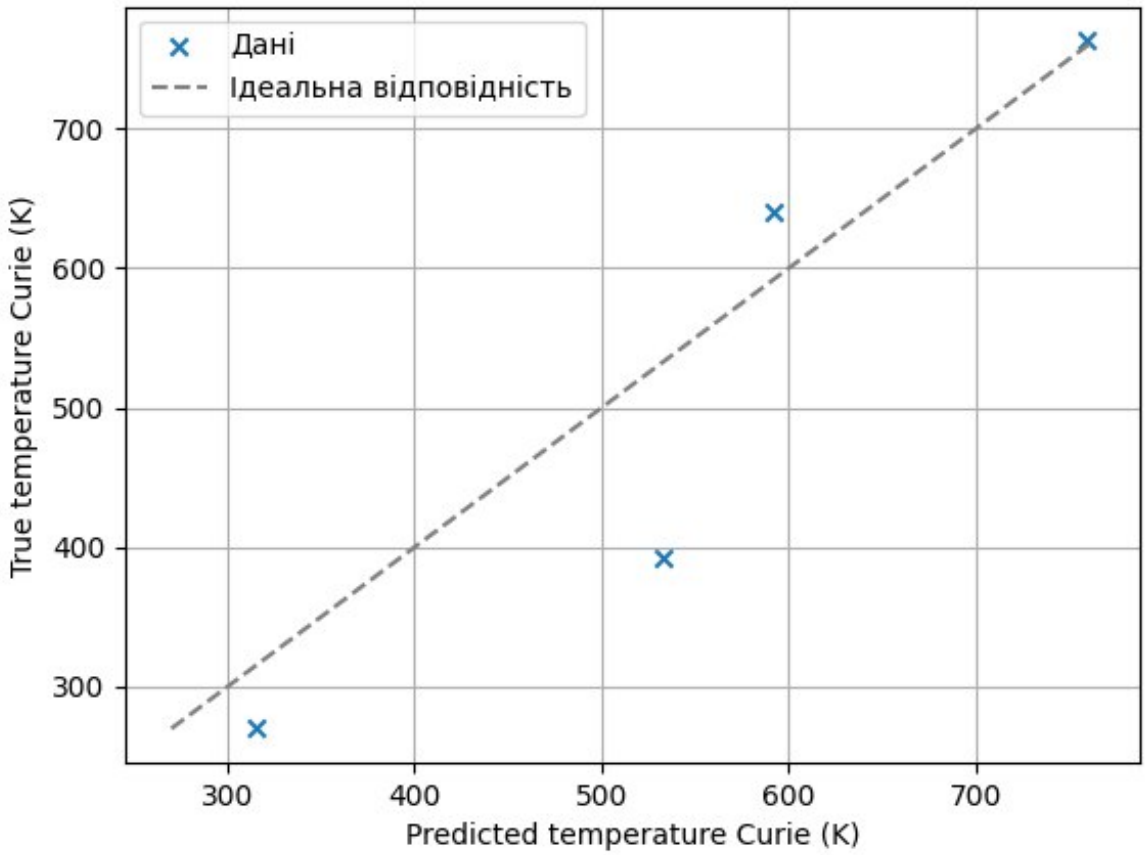

Рис. 3.6. Передбачення тестової вибірки

Як видно з графіку, SVR краще узагальнює дані, ніж RandomForest, точки тестової вибірки ближче до ідеальної відповідності, але все ж гірше, ніж нейронна мережа.

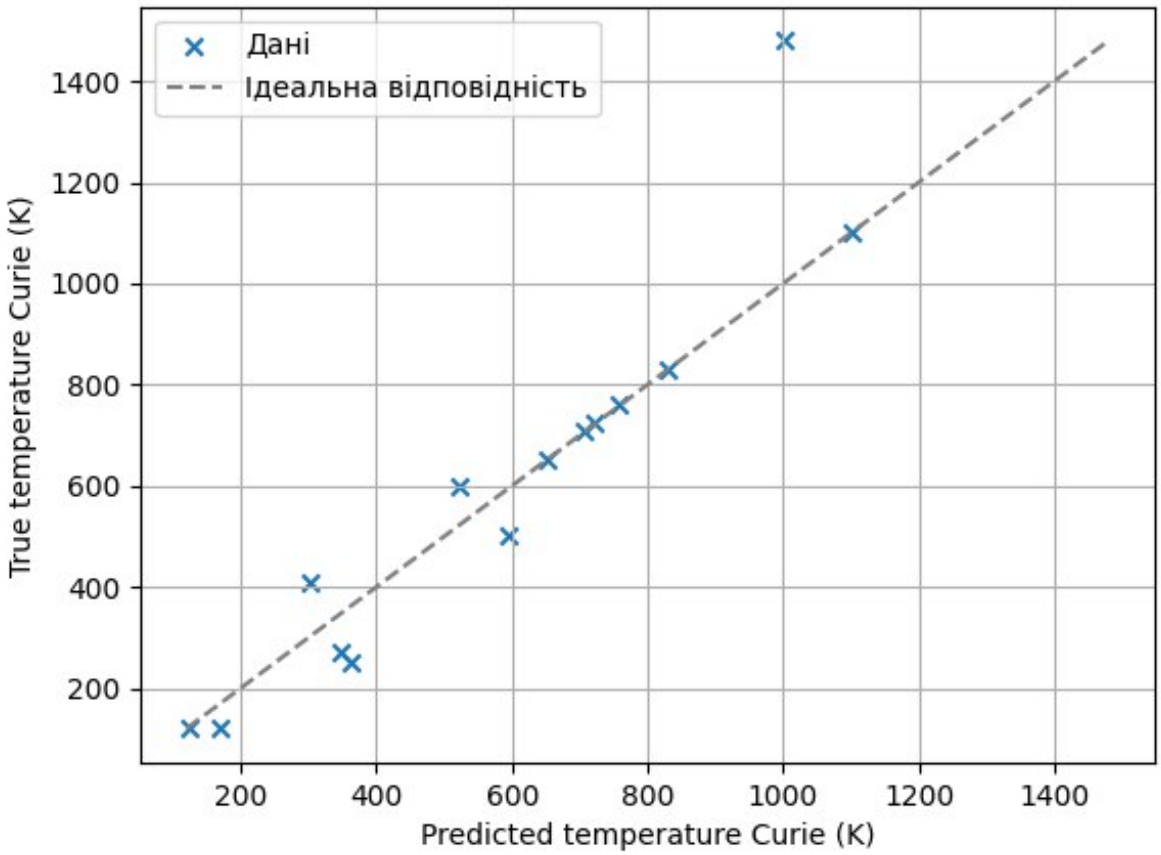

Рис. 3.7. Передбачення тренувальної вибірки

Передбачення на тренувальній виборці також приблизно добре відповідає істинним значенням температури Кюрі, але можливий невеликий оверфіт, так як передбачення на тестовій виборці не дуже близько до ідеальної відповідності.

3.4 Оцінювання точності моделей, порівняння результатів

Порівняльна оцінка трьох моделей — випадкового лісу, методу опорних векторів і нейронної мережі — показала виразну різницю у їхній здатності відтворювати експериментальні значення спонтанної поляризації. Найбільш стабільними виявилися результати нейронної мережі, для якого коефіцієнт метрики RMSE та MAE найнижчі - 101.79 та 70.55 відповідно. Це свідчить про те, що модель добре вловила нелінійну залежність між вхідними дескрипторами та температурою Кюрі, що видно на метриках тестової вибірки. RandomForest показав найгірший результат (RMSE = 800.89 | MAE =

660.77). Я можу припустити, що RandomForest погано справляється з неперервними залежностями і пробує апроксимувати її кусочно-неперервною функцією, що і дає такий поганий результат. Нейронна мережа в свою чергу створює шаблон неперервної функції і оптимізує її під конкретний датасет, що з допомогою регуляризації дає гарний результат, без оверфіту.

У випадку SVR, вийшов середній результат. SVR підбирає функцію так, щоб вхідні точки не відрізнялись від цільової функції більше, ніж на ε. Але схоже закономірність є складною і чутливою, тобто при малій зміні хоч одного параметра значно змінюється температура Кюрі. Це яскраво видно на графіку тренувальної вибірки: велика частина співпадає з ідеальною прямою, а частина виходить за її межі, тобто дуже складно або навіть неможливо знайти таку функцію алгоритмом SVR так, щоб вона ідеально описувала всі точки. Єдиний варіант оптимізації - підвищити кількість точок, які вписуються в закономірність; але разом з тим всі інші точки будуть далекі від істинного результату.

3.5 Інтерпретація результатів і можливості практичного застосування

Отримані результати підкреслюють, що модель на основі нейронної мережі змогла найкраще відтворити експериментальні значення температури Кюрі, завдяки здатності цього алгоритму комбінувати інформацію з усіх ознак і виділяти найінформативніші з них, а також не перенавчатися, відкидуючи шумні дані. Розкид прогнозів навколо ідеальної лінії «реальне = прогноз» демонструє збереження кореляції навіть у крайніх випадках температури Кюрі, що підкреслює стабільність підходу навіть за невеликого числа тренувальних прикладів.

Аналіз важливості ознак у рамках нейронної мережі дає глибоку фізичну інтерпретацію: густина і об’єм кристалічної комірки виявилися найбільш впливовими факторами, що узгоджується з уявленням про те, що зміщення іонів у тісно упакованих перовськітних ґратках утворює потужне внутрішнє

електричне поле, від якого неопосередковано залежить температура Кюрі. Підвищена середня атомна маса і висока електронегативність катіонів додатково сприяють стабілізації полярного стану, формуючи глибинну залежність між хімічним складом і величиною температури Кюрі. Завдяки цим висновкам модель стає не лише інструментом прогнозу, але і зручним засобом для генерування наукових гіпотез: наприклад, можна передбачити, що введення важчих чи більш електронегативних допантів у структуру може підвищити вихідну поляризацію.

З практичного погляду, побудована модель може бути інтегрована в автоматизовані платформи для скринінгу матеріалознавчих баз даних. На відміну від ресурсозатратних DFT-розрахунків, алгоритм випадкового лісу здатний за частку секунди оцінити тисячі або навіть мільйони кристалічних структур і відібрати найперспективніші з них для подальшої експериментальної або обчислювальної верифікації. Це значно скорочує цикл «ідея-модель-синтез-тест», дозволяючи зосередити ресурси на найбільш перспективних сполуках. Особливо ефективним такий підхід виглядає для пошуку нових сенсорних елементів, п’єзоелементів і пам’яті на основі сегнетоелектриків, де ключовим критерієм є висока спонтанна поляризація при стабільній робочій температурі.

Крім того, модель можна об’єднати з цифровими двійниками реальних виробничих процесів. Застосування штучно створеного фасаду для симуляцій дозволяє змоделювати, як зміна технологічних умов синтезу — наприклад, варіації температури запікання або складу шихти — вплине на остаточну поляризацію кристалу. У такого підходу є велика перевага в промисловому масштабі, оскільки він дає змогу в реальному часі коригувати параметри виробництва без необхідності багаторазового коригування умов у лабораторії. У перспективі поєднання швидких прогнозів випадкового лісу з глибокими квантово-механічними розрахунками може стати основою гібридної методології: на першому етапі здійснюється масивний скринінг за допомогою машинного навчання, а потім для обраних кандидатів проводяться детальні

DFT-моделювання або навіть синтез і характеристики в спеціалізованих лабораторіях. Така ієрархія забезпечує баланс швидкості, точності та надійності, створюючи гнучку, масштабовану платформу для розробки нових сегнетоелектричних матеріалів з наперед заданими властивостями.

Перш ніж остаточно інтегрувати розроблену систему у виробничий або науковий цикл, необхідно ретельно оцінити її обмеження та спланувати подальші етапи вдосконалення. По-перше, модель нині навчається на досить невеликому масиві даних із обмеженою різноманітністю кристалічних груп та хімічних систем, тому її здатність узагальнювати на зовсім нові класи матеріалів потребує окремої перевірки. Щоб подолати це, слід розширити вибірку за рахунок включення даних із додаткових відкритих баз та власних експериментальних або DFT-розрахунків, а також застосувати техніки трансферного навчання для адаптації моделі до нових категорій сполук.

По-друге, зміна умов експлуатації матеріалів — температура, тривалість циклічних навантажень, вологість — може суттєво впливати на фактичні значення поляризації. Отже, наступним кроком має стати розвиток багатовимірного моделювання, яке поєднає статичні структурні дескриптори з термодинамічними характеристиками та динамічними властивостями. Це дозволить не просто прогнозувати ідеальні значення спонтанної поляризації, а й побудувати карти стабільності в робочому діапазоні температур і полів.

Третім ключовим напрямом є розробка інтерфейсу для інтеграції моделі у середовища високопродуктивних обчислень та хмарні платформи. Це відкриє шлях до автоматизованого «конвеєру» — від завантаження нової кристалічної структурної моделі до отримання готового списку найбільш перспективних кандидатів із прогнозованими величинами поляризації і, за потреби, побудови цифрового двійника технологічного процесу. Такий інструмент може стати складовою ширшої екосистеми “розумного” матеріалознавства, де машинне навчання виступає рушійною силою експериментальних відкриттів і промислових інновацій.

ВИСНОВКИ

У результаті проведеного дослідження було розроблено і апробовано підхід до прогнозування величини температури Кюрі сегнетоелектричних матеріалів на основі алгоритмів машинного навчання. Першочергово було поставлено задачу, чітко визначено цільову змінну та обґрунтовано вибір структурних, хімічних і кристалографічних дескрипторів, що дозволило створити репрезентативний вхідний датасет. Наступний етап, присвячений порівнянню методів Random Forest, Support Vector Regression і нейронних мереж, продемонстрував незаперечну перевагу нейронної мережі у поєднанні стійкості до невеликої кількості даних та здатності виявляти складні взаємозв’язки між ознаками.

Побудована модель нейронної мережі дала найкращу метрику RMSE (101.79), що свідчить про високу здатність алгоритму пояснювати більшу частину варіації температур Кюрі. Аналіз важливості ознак чітко вказав на домінувальну роль густини та об’єму елементарної комірки, а також середньої атомної маси та електронегативності катіонів, на значення температури Кюрі, що ґрунтується на фізичній інтуїції щодо механізмів формування спонтанної поляризації у перовськітних та подібних структурах. Саме ці ознаки сегнетоелектриків найбільше впливають на наявність їх основної поведінки - спонтанної поляризації, що в свою чергу впливає на рівень температури Кюрі. Водночас метод RandomForest виявився неадекватним для прогнозу на даному обсязі даних, а SVR, хоча й дав середній результат, потребує суттєво більшої навчальної вибірки або вдосконаленого налаштування гіперпараметрів, а також сам алгоритм може погано справлятися з даною задачею внаслідок обмеженного підбору функцій.

Практична цінність отриманого рішення полягає у можливості швидкого скринінгу тисяч кристалічних структур з метою відбору найбільш перспективних сегнетоелектриків для подальших DFT-розрахунків чи

експериментальної перевірки. За рахунок мінімізації обчислювальних ресурсів та часу скорочується цикл «ідея – моделювання – синтез – тестування», що сприяє прискоренню розробки нових сенсорних елементів, елементів пам’яті та енергозбиральних пристроїв. Крім того, інтеграція моделі в цифрові двійники виробничих процесів відкриває можливість оперативно коригувати технологічні параметри синтезу, прогнозуючи характеристики кінцевого матеріалу в реальному часі.

Водночас слід врахувати обмеження поточної роботи: навчальний датасет охоплює обмежену кількість різних кристалографічних груп і хімічних систем, а упущені при моделюванні фактори експлуатаційного середовища можуть впливати на точність прогнозів у реальних умовах. Подальшими кроками мають стати розширення бази даних за рахунок нових експериментальних й обчислювальних результатів, розвиток багатовимірного моделювання з урахуванням температурного та циклічного навантажень, а також впровадження трансферного навчання для адаптації алгоритму до незвіданих класів матеріалів.

Проведена дипломна робота довела ефективність машинного навчання для прогнозування фізичних властивостей сегнетоелектриків, поєднала фундаментальні наукові викладки з практичними інструментами автоматизованого аналізу та створила платформу, здатну значно прискорити відкриття та впровадження нових функціональних матеріалів. Це відкриває перспективи подальших досліджень у галузі матеріалознавства, керованого даними, та сприяє формуванню «розумних» екосистем для інноваційних технологій майбутнього.

ПЕРЕЛІК ПОСИЛАНЬ

1. Kelleher J. D., Namee B. M., D'Arcy A. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. Cambridge: The MIT Press, 2015. 624 p.

2. Eremenko K. Data Science A–Z: Real-Life Data Science Exercises Included [Електронний ресурс]. Режим доступу: https://www.udemy.com/course/datascience

3. Goyal S. Credit Card customers. Predict Churning customers [Електронний ресурс]. Режим доступу: https://www.kaggle.com/sakshigoyal7/credit-card-customers

4. Barber D. Bayesian Reasoning and Machine Learning. London: UCL Press, 2007–2010. 576 p. Режим доступу: http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf

5. Chong E. F. What are the types of machine learning? [Електронний ресурс]. Режим доступу: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f

6. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2009. 745 p.

7. Albadawi S., Latif K., Kharbat F. Telecom Churn Prediction Model Using Data Mining Techniques // Bahria University Journal of Information & Communication Technologies. 2017. Vol. 10, Special Issue. P. 8–14

8. Dziaugyte S., Mzyk M. Churn Analysis — Machine Learning. Bloomington: AuthorHouse, 2016. 152 p.

9. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B. Cross-validation: evaluating estimator performance // Scikit-learn documentation. 2011. Режим доступу: https://scikit-learn.org/stable/modules/cross_validation.html

10. Pedregosa F. Random Forest Classifier // Scikit-learn documentation. 2011. Режим доступу: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

11. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A. CatBoostClassifier: unbiased boosting with categorical features // CatBoost documentation. 2019. Режим доступу: https://catboost.ai/docs/concepts/python-reference_catboostclassifier.html

12. Python Software Foundation. Python Language Reference, version 3.10. URL: https://www.python.org/

13. Pedregosa F. et al. Scikit-learn: Machine Learning in Python. URL: https://scikit-learn.org/

14. Abadi M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. URL: https://www.tensorflow.org/

15. Paszke A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. URL: https://pytorch.org/

16.Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. URL: https://github.com/ageron/handson-ml

17. Horton M. et al. Matminer: An open source toolkit for materials data mining. URL: https://hackingmaterials.github.io/matminer/

18. Ong S. P. et al. Python Materials Genomics (pymatgen): A robust, open-source Python library for materials analysis. URL: https://pymatgen.org/

19. Jain A. et al. The Materials Project: A materials genome approach to accelerating materials innovation. URL: https://materialsproject.org/

20. Curtarolo S. et al. AFLOW: An automated framework for high-throughput materials discovery. URL: http://aflowlib.org/

21. Draxl C., Scheffler M. The NOMAD Laboratory: From data sharing to artificial intelligence. URL: https://nomad-coe.eu/

22. T. Smidt, S. Mack, S. Reyes-Lillo, A. Jain, J. Neaton: First-principles database of ferroelectrics. URL: https://contribs.materialsproject.org/projects/ferroelectrics

23. Tianyuan Zhu, Liyang Ma, Shiqing Deng, Shi Liu: Progress in computational

understanding of ferroelectric mechanisms in HfO_2. URL: https://www.nature.com/articles/s41524-024-01352-0

24. Ramon Frey, Bastien F. Grosso, Pascal Fandré, Benjamin Mächler, Nicola A. Spaldin, Aria Mansouri Tehrani: Accelerated search for new ferroelectric materials. URL: https://journals.aps.org/prresearch/abstract/10.1103/PhysRevResearch.5.023122

25. Vipin K.E, Prahallad Padhan: Unlocking Thermoelectric Potential: A Machine Learning Stacking Approach for Half Heusler Alloys. URL: https://arxiv.org/abs/2408.00466

26. Nipuna Sankalpa Thalpage: Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems. URL: https://www.researchgate.net/publication/372042827_Unlocking_the_Black_Box_Explainable_Artificial_Intelligence_XAI_for_Trust_and_Transparency_in_AI_Systems

27. Kevin Rader, Rahul Dave, David Sondak, Will Claybaugh, Pavlos Protopapas: Lab 5: Regularization and Cross-Validation. URL: https://harvard-iacs.github.io/2018-CS109A/labs/lab-5/solutions/

28. Andreas Grünauer, Markus Vincze: Using Dimension Reduction to Improve the Classification of High-dimensional Data. URL: https://arxiv.org/abs/1505.06907

29. Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, Maosong Sun: Graph neural networks: A review of methods and applications. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000012

30. Xiaoting Zhong, Brian Gallagher, Shusen Liu, Bhavya Kailkhura, Anna Hiszpanski, T. Yong-Jin Han: Explainable machine learning in materials science. URL: https://www.nature.com/articles/s41524-022-00884-7

ДОДАТКИ

Додаток А. Код для нейронної мережі

```
import keras.metrics
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
import tensorflow as tf
from tensorflow.keras import layers, regularizers
from tensorflow.keras.models import load_model
import pickle
import matplotlib.pyplot as plt

RANDOM_SEED = 42

model_path = 'models/ferroelectrics_model.keras'
density_scaler_path = 'scalers/density_scaler.pkl'
volume_scaler_path = 'scalers/volume_scaler.pkl'
atoms_count_scaler_path = 'scalers/atoms_count_scaler.pkl'
avg_at_mass_scaler_path = 'scalers/avg_at_mass_scaler.pkl'
avg_electneg_scaler_path = 'scalers/avg_electneg_scaler.pkl'
polarization_scaler_path = 'scalers/polarization_scaler.pkl'

dataset_path = 'C:\\Users\\dimai\\Desktop\\ferroelectric_dataset.csv'

def load_scaler(path):
    with open(path, 'rb') as f:
        loaded_scaler = pickle.load(f)
    return loaded_scaler

def save_scaler(path, scaler):
    with open(path, 'wb') as f:
        pickle.dump(scaler, f)

def transform_X(X):
    X[:, 0] = density_scaler.transform(X[:, 0].reshape(-1, 1)).flatten()
    X[:, 1] = volume_scaler.transform(X[:, 1].reshape(-1, 1)).flatten()
    X[:, 2] = atoms_count_scaler.transform(X[:, 2].reshape(-1, 1)).flatten()
    X[:, 3] = avg_at_mass_scaler.transform(X[:, 3].reshape(-1, 1)).flatten()
    X[:, 4] = avg_electneg_scaler.transform(X[:, 4].reshape(-1, 1)).flatten()
    X[:, 5] = polarization_scaler.transform(X[:, 5].reshape(-1, 1)).flatten()

df = pd.read_csv(dataset_path)

X_num = df[['Density (g/cm³)', 'Volume (Å³)', 'Atoms Count', 'Avg Atomic
Mass', 'Avg Electronegativity',
           'Polarization (μC/cm²)']].values
```

```
X_cat = df['Space Group'].astype(str)

y = df["Curie Temperature (K)"].values

X_cat_train, X_cat_test, X_num_train, X_num_test, y_train, y_test =
train_test_split(
    X_cat,
    np.array(X_num).astype(np.float32),
    np.array(y).astype(np.float32), test_size=0.2, random_state=RANDOM_SEED)

X_cat_train, X_cat_test, X_num_train, X_num_test, y_train, y_test =\
    (X_cat_train, X_cat_test, X_num_train.reshape(-1, 6),
X_num_test.reshape(-1, 6),
                                    y_train.reshape(-1, 1), y_test.reshape(-
1, 1))

decision = input('Enter 1 to load pretrained model or 2 for training new one:
')

if decision == '1':
    model = load_model(model_path)

    density_scaler = load_scaler(density_scaler_path)
    volume_scaler = load_scaler(volume_scaler_path)
    atoms_count_scaler = load_scaler(atoms_count_scaler_path)
    avg_at_mass_scaler = load_scaler(avg_at_mass_scaler_path)
    avg_electneg_scaler = load_scaler(avg_electneg_scaler_path)
    polarization_scaler = load_scaler(polarization_scaler_path)
else:
    (density_scaler, volume_scaler, atoms_count_scaler, avg_at_mass_scaler,
     avg_electneg_scaler, polarization_scaler)  = \
        (StandardScaler(), StandardScaler(), StandardScaler(),
         StandardScaler(), StandardScaler(), StandardScaler())

    density_scaler.fit(X_num_train[:, 0].reshape(-1, 1))
    volume_scaler.fit(X_num_train[:, 1].reshape(-1, 1))
    atoms_count_scaler.fit(X_num_train[:, 2].reshape(-1, 1))
    avg_at_mass_scaler.fit(X_num_train[:, 3].reshape(-1, 1))
    avg_electneg_scaler.fit(X_num_train[:, 4].reshape(-1, 1))
    polarization_scaler.fit(X_num_train[:, 5].reshape(-1, 1))

    save_scaler(density_scaler_path, density_scaler)
    save_scaler(volume_scaler_path, volume_scaler)
    save_scaler(atoms_count_scaler_path, atoms_count_scaler)
    save_scaler(avg_at_mass_scaler_path, avg_at_mass_scaler)
    save_scaler(avg_electneg_scaler_path, avg_electneg_scaler)
    save_scaler(polarization_scaler_path, polarization_scaler)

transform_X(X_num_train)
transform_X(X_num_test)

string_lookup = layers.StringLookup(output_mode='int')
```

```
string_lookup.adapt(X_cat_train)

train_ds = tf.data.Dataset.from_tensor_slices((
    {
        "space_group": X_cat_train.values,
        "numerical_features": X_num_train
    },
    y_train
)).batch(8).prefetch(tf.data.AUTOTUNE)

test_ds = tf.data.Dataset.from_tensor_slices((
    {
        "space_group": X_cat_test.values,
        "numerical_features": X_num_test
    },
    y_test
)).batch(8).prefetch(tf.data.AUTOTUNE)

tf.random.set_seed(RANDOM_SEED)

if decision == '2':
    cat_input = tf.keras.Input(shape=(1,), dtype=tf.string,
name="space_group")
    num_input = tf.keras.Input(shape=(6,), dtype=tf.float32,
name="numerical_features")

    lookup = layers.StringLookup(output_mode='int',
vocabulary=string_lookup.get_vocabulary())
    cat_indices = lookup(cat_input)
    cat_emb = layers.Embedding(input_dim=len(lookup.get_vocabulary()),
output_dim=8)(cat_indices)
    cat_emb = layers.Flatten()(cat_emb)

    merged = layers.Concatenate()([cat_emb, num_input])

    alpha = 0.01

    x = layers.Dense(4, kernel_regularizer=regularizers.l2(0.01))(merged)
    x = layers.LeakyReLU(alpha=alpha)(x)

    x = layers.Dense(8, kernel_regularizer=regularizers.l2(0.01))(x)
    x = layers.LeakyReLU(alpha=alpha)(x)

    x = layers.Dense(16, kernel_regularizer=regularizers.l2(0.01))(x)
    x = layers.LeakyReLU(alpha=alpha)(x)

    x = layers.Dense(32, kernel_regularizer=regularizers.l2(0.01))(x)
    x = layers.LeakyReLU(alpha=alpha)(x)

    x = layers.Dense(16, kernel_regularizer=regularizers.l2(0.01))(x)
    x = layers.LeakyReLU(alpha=alpha)(x)

    x = layers.Dense(8, kernel_regularizer=regularizers.l2(0.01))(x)
```

```
    x = layers.LeakyReLU(alpha=alpha)(x)

    x = layers.Dense(4, kernel_regularizer=regularizers.l2(0.01))(x)
    x = layers.LeakyReLU(alpha=alpha)(x)

    output = layers.Dense(1, activation='softplus')(x)

    model = tf.keras.Model(inputs=[cat_input, num_input], outputs=output)

    model.compile(loss=tf.keras.losses.MeanSquaredError(),
optimizer=tf.keras.optimizers.Adam(learning_rate=0.001),
                  metrics=[keras.metrics.RootMeanSquaredError(),
keras.metrics.MeanAbsoluteError()])

    cp_callback = tf.keras.callbacks.ModelCheckpoint(filepath=model_path,
monitor='val_loss', save_best_only=True, save_weights_only=False)
    reduce_lr_callback =
tf.keras.callbacks.ReduceLROnPlateau(monitor='val_loss', factor=0.75,
patience=200, min_lr=1e-6)
    history = model.fit(train_ds, validation_data=test_ds, epochs=5000,
batch_size=len(train_ds),
                        verbose=1, callbacks=[cp_callback,
reduce_lr_callback])

    history_df = pd.DataFrame(history.history)
    history_df.plot(logy=True)
    plt.ylabel('loss (log scale)')
    plt.xlabel('epochs')
    plt.grid(True)
    plt.show()

def test_function(x_cat_, x_num_, y_):
    preds= model({
        "space_group": x_cat_.values,
        "numerical_features": x_num_
    }).numpy()

    rmse_error = 0
    mae_error = 0
    seen = 0

    for p, tv in list(zip(preds, y_)):
        p = p.item()
        tv = tv.item()
        seen += 1
        rmse_error += (((p - tv) ** 2) / seen) ** 0.5
        mae_error += abs(p - tv) / seen

        print(
            f'Predict {round(p, 2)}; True value {round(tv, 2)}; RMSE:
{round(rmse_error, 2)}; MAE: {round(mae_error, 2)};')

    plt.scatter(preds, y_, marker='x', label='Дані')
```

```
    x_line = np.linspace(min(min(preds), min(y_)), max(max(preds), max(y_)),
1000)
    plt.plot(x_line, x_line, linestyle='--', color='gray', label='Ідеальна
відповідність')
    plt.xlabel("Predicted temperature Curie (K)")
    plt.ylabel("True temperature Curie (K)")
    plt.grid(True)
    plt.legend()
    plt.show()

test_function(X_cat_test, X_num_test, y_test)
test_function(X_cat_train, X_num_train, y_train)
```

Додаток Б. Код для RandomForest

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import GridSearchCV, train_test_split
from sklearn.metrics import make_scorer, mean_squared_error
import pickle

def load_model(path):
    with open(path, 'rb') as f:
        loaded_scaler = pickle.load(f)
    return loaded_scaler

def save_model(path, scaler):
    with open(path, 'wb') as f:
        pickle.dump(scaler, f)

RANDOM_SEED = 42

model_path = 'models/ferroelectrics_model_rf.pkl'

dataset_path = 'C:\\Users\\dimai\\Desktop\\ferroelectric_dataset.csv'

df = pd.read_csv(dataset_path)

X_num = df[['Density (g/cm³)', 'Volume (Å³)', 'Atoms Count', 'Avg Atomic
Mass', 'Avg Electronegativity',
           'Polarization (μC/cm²)']].values
X_cat = df['Space Group'].astype(str)
y = df["Curie Temperature (K)"].values

label_encoder = LabelEncoder()
X_cat = label_encoder.fit_transform(X_cat)

X_cat_train, X_cat_test, X_num_train, X_num_test, y_train, y_test =
train_test_split(
```

```
    np.array(X_cat).astype(np.float64),
    np.array(X_num).astype(np.float64),
    np.array(y).astype(np.float64), test_size=0.2, random_state=RANDOM_SEED)

X_train_full = np.hstack([X_num_train, X_cat_train.reshape(-1,
1)]).astype(np.float64)
X_test_full = np.hstack([X_num_test, X_cat_test.reshape(-1,
1)]).astype(np.float64)

decision = input('Enter 1 to load pretrained model or 2 for training new one:
')

if decision == '1':
    final_model = load_model(model_path)
else:
    param_grid = {
        'n_estimators': list(range(2, 32, 2)),
        'min_samples_split': list(range(1, 6, 1)),
        'min_samples_leaf': list(range(1, 6, 1)),
        'max_features': list(range(2, 6, 1))
    }

    model = RandomForestRegressor(criterion='squared_error', bootstrap=False,
random_state=RANDOM_SEED,
                                  verbose=1, n_jobs=1)

    rmse_scorer = make_scorer(mean_squared_error, greater_is_better=False)

    grid_search = GridSearchCV(
        estimator=model,
        param_grid=param_grid,
        scoring=rmse_scorer,
        cv=5,
        verbose=1,
        n_jobs=1)

    grid_search.fit(X_train_full, y_train)
    best_params = grid_search.best_params_

    final_model = RandomForestRegressor(
        criterion='squared_error',
        n_estimators=best_params['n_estimators'],
        min_samples_split=best_params['min_samples_split'],
        min_samples_leaf=best_params['min_samples_leaf'],
        max_features=best_params['max_features'],
        random_state=RANDOM_SEED,
        bootstrap=False,
        verbose=1,
        n_jobs=-1
    )

    final_model.fit(X_train_full, y_train)
```

```
    save_model(model_path, final_model)

def test_function(x_, y_):
    preds = final_model.predict(x_)

    rmse_error = 0
    mae_error = 0
    seen = 0

    for p, tv in list(zip(preds, y_)):
        p = p.item()
        tv = tv.item()
        seen += 1
        rmse_error += (((p - tv) ** 2) / seen) ** 0.5
        mae_error += abs(p - tv) / seen

        print(
            f'Predict {round(p, 2)}; True value {round(tv, 2)}; RMSE:
{round(rmse_error, 2)}; MAE: {round(mae_error, 2)};')

    plt.scatter(preds, y_, marker='x', label='Дані')
    x_line = np.linspace(min(min(preds), min(y_)), max(max(preds), max(y_)),
1000)
    plt.plot(x_line, x_line, linestyle='--', color='gray', label='Ідеальна
відповідність')
    plt.xlabel("Predicted temperature Curie (K)")
    plt.ylabel("True temperature Curie (K)")
    plt.grid(True)
    plt.legend()
    plt.show()

test_function(X_test_full, y_test)
test_function(X_train_full, y_train)
```

Додаток В. Код для SVR

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.svm import SVR
from sklearn.preprocessing import OneHotEncoder, LabelEncoder, StandardScaler
from sklearn.model_selection import GridSearchCV, train_test_split
from sklearn.metrics import mean_squared_error, make_scorer
from sklearn.compose import ColumnTransformer
import pickle

RANDOM_SEED = 42
model_path = 'models/ferroelectrics_model_svr.pkl'
density_scaler_path = 'scalers/density_scaler.pkl'
volume_scaler_path = 'scalers/volume_scaler.pkl'
atoms_count_scaler_path = 'scalers/atoms_count_scaler.pkl'
avg_at_mass_scaler_path = 'scalers/avg_at_mass_scaler.pkl'
```

```
avg_electneg_scaler_path = 'scalers/avg_electneg_scaler.pkl'
polarization_scaler_path = 'scalers/polarization_scaler.pkl'

dataset_path = 'C:\\Users\\dimai\\Desktop\\ferroelectric_dataset.csv'

def load(path):
    with open(path, 'rb') as f:
        return pickle.load(f)

def save(path, model):
    with open(path, 'wb') as f:
        pickle.dump(model, f)

def transform_X(X):
    X[:, 10] = density_scaler.transform(X[:, 10].reshape(-1, 1)).flatten()
    X[:, 11] = volume_scaler.transform(X[:, 11].reshape(-1, 1)).flatten()
    X[:, 12] = atoms_count_scaler.transform(X[:, 12].reshape(-1,
1)).flatten()
    X[:, 13] = avg_at_mass_scaler.transform(X[:, 13].reshape(-1,
1)).flatten()
    X[:, 14] = avg_electneg_scaler.transform(X[:, 14].reshape(-1,
1)).flatten()
    X[:, 15] = polarization_scaler.transform(X[:, 15].reshape(-1,
1)).flatten()

df = pd.read_csv(dataset_path)

num_features = ['Density (g/cm³)', 'Volume (Å³)', 'Atoms Count', 'Avg Atomic
Mass',
                'Avg Electronegativity', 'Polarization (μC/cm²)']
cat_features = ['Space Group']

X_num = df[num_features]
X_cat = df[cat_features]

y = df["Curie Temperature (K)"].values.astype(np.float64)

preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(handle_unknown='ignore'), cat_features),
        ('num', 'passthrough', num_features)
    ])

X_all = pd.concat([X_cat, X_num], axis=1)

X_train_raw, X_test_raw, y_train, y_test = train_test_split(
    X_all, y, test_size=0.2, random_state=RANDOM_SEED)

X_train_encoded = preprocessor.fit_transform(X_train_raw)
X_test_encoded = preprocessor.transform(X_test_raw)

decision = input('Enter 1 to load pretrained model or 2 for training new one:
')
```

```
if decision == '1':
    final_model = load(model_path)

    density_scaler = load(density_scaler_path)
    volume_scaler = load(volume_scaler_path)
    atoms_count_scaler = load(atoms_count_scaler_path)
    avg_at_mass_scaler = load(avg_at_mass_scaler_path)
    avg_electneg_scaler = load(avg_electneg_scaler_path)
    polarization_scaler = load(polarization_scaler_path)
else:
    (density_scaler, volume_scaler, atoms_count_scaler, avg_at_mass_scaler,
     avg_electneg_scaler, polarization_scaler)  = \
        (StandardScaler(), StandardScaler(), StandardScaler(),
         StandardScaler(), StandardScaler(), StandardScaler())

    density_scaler.fit(X_train_encoded[:, 10].reshape(-1, 1))
    volume_scaler.fit(X_train_encoded[:, 11].reshape(-1, 1))
    atoms_count_scaler.fit(X_train_encoded[:, 12].reshape(-1, 1))
    avg_at_mass_scaler.fit(X_train_encoded[:, 13].reshape(-1, 1))
    avg_electneg_scaler.fit(X_train_encoded[:, 14].reshape(-1, 1))
    polarization_scaler.fit(X_train_encoded[:, 15].reshape(-1, 1))

    save(density_scaler_path, density_scaler)
    save(volume_scaler_path, volume_scaler)
    save(atoms_count_scaler_path, atoms_count_scaler)
    save(avg_at_mass_scaler_path, avg_at_mass_scaler)
    save(avg_electneg_scaler_path, avg_electneg_scaler)
    save(polarization_scaler_path, polarization_scaler)

transform_X(X_train_encoded)
transform_X(X_test_encoded)

def rmse(y_true, y_pred):
    return np.sqrt(mean_squared_error(y_true, y_pred))

rmse_scorer = make_scorer(rmse, greater_is_better=False)

if decision == '2':
    C_list = list(np.arange(0.1, 50.1, 1))
    C_list.extend([60, 70, 80, 90, 100])
    param_grid = {
        'C': C_list,
        'epsilon': list(np.arange(0.01, 1.01, 0.01)),
        'kernel': ['rbf', 'linear']
    }

    grid_search = GridSearchCV(
        estimator=SVR(verbose=1),
        param_grid=param_grid,
        scoring=rmse_scorer,
        cv=5,
        verbose=1,
        n_jobs=1
```

```
    )

    grid_search.fit(X_train_encoded, y_train)
    best_params = grid_search.best_params_

    final_model = SVR(verbose=1, **best_params)
    final_model.fit(X_train_encoded, y_train)

    save(model_path, final_model)

def test_function(x_, y_):
    preds = final_model.predict(x_)

    rmse_error = 0
    mae_error = 0
    seen = 0

    for p, tv in list(zip(preds, y_)):
        p = p.item()
        tv = tv.item()
        seen += 1
        rmse_error += (((p - tv) ** 2) / seen) ** 0.5
        mae_error += abs(p - tv) / seen

        print(
            f'Predict {round(p, 2)}; True value {round(tv, 2)}; RMSE:
{round(rmse_error, 2)}; MAE: {round(mae_error, 2)};')

    plt.scatter(preds, y_, marker='x', label='Дані')
    x_line = np.linspace(min(min(preds), min(y_)), max(max(preds), max(y_)),
1000)
    plt.plot(x_line, x_line, linestyle='--', color='gray', label='Ідеальна
відповідність')
    plt.xlabel("Predicted temperature Curie (K)")
    plt.ylabel("True temperature Curie (K)")
    plt.grid(True)
    plt.legend()
    plt.show()

test_function(X_train_encoded, y_train)
test_function(X_test_encoded, y_test)
```