УДК 004.89+004.032.26]:616.24-002.5(540) **DOI:** 10.31673/2412-9070.2025.021793
Д. В. НЕВІНСЬКИЙ[1], канд. техн. наук, доцент;
ORCID: 0000-0002-0962-072X
Д. І. МАРТЬЯНОВ[1], аспірант;
ORCID: 0009-0003-3919-4412
О. А. ГОСПОДАРСЬКИЙ[1], студент;
ORCID: 0009-0005-9088-3015
Я. І. ВИКЛЮК[1], доктор техн. наук, професор;
ORCID: 0000-0003-4766-4659
І. О. СЕМ'ЯНІВ[2], канд. мед. наук, доцент,
ORCID: 0000-0003-0340-0766
[1] Національний університет «Львівська політехніка», Львів
[2] Буковинський державний медичний університет, Чернівці

ВИЗНАЧЕННЯ ДЕТЕРМІНАНТІВ ТУБЕРКУЛЬОЗУ: АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА НЕЙРОННИХ МЕРЕЖ

Туберкульоз (ТБ) залишається однією з найсерйозніших інфекційних хвороб у світі, зокрема в Індії, де високий рівень захворюваності створює значні виклики для системи охорони здоров'я. Дослідження присвячене аналізу детермінантів поширення туберкульозу в Індії за допомогою методів машинного навчання (ML) та нейронних мереж (NN). Метою роботи є виявлення ключових факторів, що впливають на рівень захворюваності, та розробка точних прогнозних моделей для підтримки стратегій профілактики та лікування. На основі статистичних даних за 2019–2022 роки, що охоплюють демографічні характеристики, соціальні фактори та медичні показники, було проведено комплексний аналіз. Застосовано методи обробки даних, включаючи кореляційний аналіз, oversampling (SMOGN) для балансування вибірки, а також моделювання з використанням лінійних регресій (LM, Ridge, Lasso), алгоритмів ML (Decision Tree, K-Nearest Neighbors, Random Forest) та глибокої нейронної мережі. Результати показали, що лінійні моделі мають обмежену точність (R^2 Test до 0.600), тоді як Random Forest (R^2 Test = 0.832) та K-Nearest Neighbors (R^2 Test = 0.865) значно перевершують їх завдяки здатності враховувати нелінійні залежності. Найвищу точність продемонструвала нейронна мережа (R^2 Test = 0.822, RMSE Test = 0.433), що підкреслює її ефективність у виявленні складних взаємозв'язків. Ключовими факторами, що впливають на захворюваність, визначено чисельність населення (Population), гендерне співвідношення (Gender Ratio), кількість спеціалізованих центрів (Nodal_DR_TB_Centres_Per_Population) та міські характеристики (City_Encoded). Отримані результати підтверджують перспективність інтеграції ML та NN у медичні дослідження для прогнозування та контролю туберкульозу, що може сприяти розробці персоналізованих підходів до терапії та покращенню громадського здоров'я.

Ключові слова: машинне навчання; нейронні мережі; прогнозування захворюваності; oversampling; SMOGN; лінійна регресія; Random Forest; K-Nearest Neighbors; детермінанти; інтеграція.

Актуальність дослідження

Туберкульоз (ТБ) залишається однією з найбільш серйозних інфекційних хвороб у світі, незважаючи на значний прогрес у його лікуванні та профілактиці. За даними Всесвітньої організації охорони здоров'я (ВООЗ), Індія є однією з країн з найвищим рівнем захворюваності на ТБ, що становить значний виклик для системи охорони здоров'я країни.

Висока щільність населення, соціально-економічні фактори та нерівний доступ до медич-

них послуг створюють сприятливі умови для поширення інфекції. Особливо вразливими до туберкульозу залишаються малозабезпечені верстви населення, мігранти, люди з ослабленим імунітетом та особи, які страждають на супутні захворювання, зокрема ВІЛ/СНІД.

Окрім медичних аспектів, проблема туберкульозу має суттєві економічні та соціальні наслідки. Втрата працездатності, тривале лікування та необхідність ізоляції хворих створюють додаткове навантаження на економіку країни та родини пацієнтів.

Це дослідження спрямоване на аналіз поширення туберкульозу в Індії, виявлення основних тенденцій та оцінку факторів, що впливають на рівень захворюваності. Отримані результати можуть допомогти у розробці ефективних заходів боротьби з цією інфекцією та покращенні стратегій її профілактики. Метою цієї роботи є встановлення функціональних залежностей між основними факторами та поширенням туберкульозу на основі моделей машинного навчання та штучного інтелекту.

Огляд літературних джерел

Штучний інтелект (ШІ) та машинне навчання (ML) відіграють ключову роль у сучасній медицині, зокрема в діагностиці, лікуванні та прогнозуванні ефективності терапії. У роботах [1–3] розглядаються загальні питання застосування ML у сфері охорони здоров'я. Зокрема, досліджено науковометричний аналіз впровадження ML у системи охорони здоров'я, включаючи аналіз 145 публікацій, що висвітлюють епідеміологічний моніторинг та ранню діагностику хвороб. Окремо розглядаються питання інтеграції ШІ у медичні сервіси, що охоплюють автоматизацію взаємодії лікарів і пацієнтів, прогнозування захворювань та оптимізацію роботи лікарень. Також запропоновано методику побудови комплексної системи автоматизованої діагностики на основі ML, що дозволяє підвищити точність ухвалення медичних рішень та мінімізувати ризики лікарських помилок.

Важливе місце займають дослідження, присвячені використанню ML у діагностиці туберкульозу. У роботах [4–8] розглянуто можливості автоматичного виявлення хвороби за допомогою алгоритмів глибокого навчання на основі знімків КТ та рентгенограм, а також аналізу медичних записів. Окрему увагу приділено проблемам мультирезистентності туберкульозу та можливості застосування ШІ для прогнозування ефективності медикаментозного лікування. Запропоновано різні підходи до прогнозування успішності терапії пацієнтів, серед яких найбільш ефективним виявився алгоритм Decision Tree, що продемонстрував найвищу точність класифікації. Також досліджено нові методи автоматизованої діагностики, які поєднують глибоке навчання з оптимізаційними алгоритмами, що дозволяє значно підвищити якість діагностичних висновків. У деяких роботах розглядається інтеграція цифрової рентгенографії з технологіями ШІ для покращення раннього виявлення туберкульозу.

Окремий напрям досліджень пов'язаний із прогнозуванням ефективності лікування за допомогою ШІ. У роботах [9, 10] розглядаються ML-методи для фармакокінетичного моделювання та прогнозування дозування антибіотиків, а також аналізу прихильності пацієнтів до терапії. Запропоновані моделі продемонстрували високу точність у прогнозуванні ефективності лікування, що може суттєво покращити персоналізований підхід до терапії інфекційних захворювань.

ШІ також активно використовується для діагностики легеневих захворювань. У роботах [11, 12] досліджено застосування ML та DL для класифікації зображень легенів та ідентифікації таких захворювань, як хронічна обструктивна хвороба легень, астма та легеневий фіброз. Особливу увагу приділено використанню алгоритмів глибокого навчання, які забезпечують точнішу диференціацію між різними типами легеневих патологій, що є важливим для персоналізованого лікування.

У дослідженнях [13, 14] розглядаються перспективи використання ML у діагностиці інфекційних захворювань, зокрема технологій CRISPR для швидкої ідентифікації патогенів. Розглядається також роль ML у прогнозуванні спалахів інфекцій, розробці вакцин та визначенні молекулярних мішеней для лікування.

Аналіз літературних джерел показує, що використання методів штучного інтелекту та машинного навчання значно покращує діагностику, лікування та моніторинг туберкульозу. Алгоритми ML дозволяють не лише швидко та точно виявляти хворобу на основі медичних зображень, але й прогнозувати ефективність медикаментозного лікування, що особливо важливо для пацієнтів із мультирезистентними формами туберкульозу. Висока точність моделей, зокрема методів глибокого навчання та дерев рішень, демонструє їхню перспективність у клінічній практиці. Дослідження також підтверджують, що застосування ШІ у фармакокінетичному моделюванні та персоналізованій терапії може суттєво покращити якість лікування. Водночас залишається необхідність подальших досліджень щодо інтеграції ML у загальну систему охорони здоров'я, стандартизації алгоритмів та оцінки їхньої ефективності у різних клінічних умовах.

Методологія дослідження

У даному дослідженні використано комплексну методологію аналізу даних на основі підходів машинного навчання та штучного інтелекту. Дослідження складаються з наступних етапів:

1. **Збір даних.** Дані про кількість випадків туберкульозу та демографічні характеристики населення отримані з відкритих джерел та представлені у форматі таблиць. Інформація включає статистику по регіонах за кілька років.
2. **Обробка та очищення даних.** На цьому етапі перевірялась коректність даних, а також були згенеровані нові ознаки. Проведено попередню обробку, включаючи перевірку на відсутні значення, коригування форматів та усунення можливих аномалій.
3. **Обчислення кореляційних зв'язків.** Для оцінки залежності між чисельністю населення та рівнем захворюваності використано статистичні методи, зокрема кореляційний аналіз.
4. **Моделювання з використанням класичних лінійних моделей.** Були протестовані класичні лінійні моделі з використанням регуляризації. Обрано оптимальні характеристики за допомогою *GridSearchCV*.
5. **Визначення важливості факторів для лінійних моделей.** Виконано оцінку впливу різних ознак на результати прогнозування.
6. **Тестування моделей машинного навчання.** Протестовані такі моделі машинного навчання:
   - *Decision Tree;*
   - *K-Nearest Neighbors;*
   - *Random Forest.*
7. **Покращення точності моделей.** Для підвищення точності моделей використано методи *oversampling*. Також було побудовано власну нейронну мережу зворотного поширення помилки.
8. **Порівняльний аналіз точності моделей.** Проведено порівняння точності різних моделей машинного навчання та штучного інтелекту.
9. **Визначення ключових факторів для нейронних мереж.** Виконано аналіз впливу вхідних ознак на результати прогнозування нейронної мережі.

Запропонована методологія дозволяє отримати комплексне уявлення про досліджувану проблему та виявити ключові тенденції та залежності.

Результати та обговорення

Збір даних. Аналіз даних базувався на статистичних даних щодо поширення туберкульозу в Індії з 2019-го по 2022 рік. Дані охоплювали всі штати Індії та містили 28 полів, зокрема:

- *Year* – рік спостереження;
- *Active Case Finding TB cases diagnosed among tested* – кількість випадків туберкульозу, діагностованих серед протестованих;
- *MDR/RR TB DIAGNOSED MDR/RR patient diagnosed* – діагностовані випадки мультирезистентного туберкульозу;

- *Paediatric TB patients notified* – повідомлення про випадки туберкульозу серед дітей;
- *TB case notification total* – загальна кількість виявлених випадків туберкульозу;
- *TB Cases Notified Female* – кількість випадків серед жінок;
- *TB Cases Notified Male* – кількість випадків серед чоловіків;
- *TB patients with known Tobacco usage status* – інформація про пацієнтів із відомим статусом вживання тютюну;
- *TB-HIV co-infected patients Diagnosed* – кількість пацієнтів із ко-інфекцією ВІЛ/ТБ;
- *Treatment outcome of TB patients notified in (% Lost to follow up)* – відсоток пацієнтів, що припинили лікування;
- *Treatment outcome of TB patients notified in (Death Rate)* – рівень смертності;
- *Treatment outcome of TB patients notified in (Success Rate)* – рівень успішного лікування;
- *PMDT- Infrastructure No. of Nodal DR-TB centres* – кількість центрів лікування мультирезистентного туберкульозу;
- *TB- DM patients initiated on Anti-diabetic treatment* – кількість пацієнтів з ТБ та діабетом, що розпочали лікування;
- *TB-COVID 19 patients detected* – кількість пацієнтів з одночасною інфекцією ТБ та COVID-19;
- *Population* – загальна чисельність населення відповідного регіону.

Загальна кількість записів у наборі даних становила 144. У якості цільового значення використовувалось поле TB case notification total.

Обробка та очищення даних. На цьому етапі було виконано кілька важливих дій для забезпечення якості даних:

- Усунено неточності у даних за 2019 рік, де деякі показники були представлені у відносних, а не абсолютних значеннях.
- Згенеровано нові ознаки для покращення ефективності аналізу та моделювання:
  - *Gender Ratio* – співвідношення між кількістю жінок та чоловіків, яким було повідомлено про випадки туберкульозу.
  - *Detection Efficiency* – ефективність виявлення туберкульозу серед протестованих осіб.
  - *Treatment Success Rate Adjusted* – скоригований показник успіху лікування з урахуванням негативних результатів.
  - *Total Notified Cases by Gender* – загальна кількість повідомлених випадків туберкульозу серед чоловіків і жінок.
  - *Total Treatment Outcomes* – загальна кількість випадків із завершеним лікуванням.
  - *Tobacco_Alcohol Interaction* – взаємодія між використанням тютюну та алкоголю серед пацієнтів.
  - *HIV_ART Impact* – вплив антиретровірусної терапії (ART) на пацієнтів із ВІЛ та туберкульозом.
  - *City_Encoded* – закодована інформація про регіони або міста.
  - *Population* – чисельність населення відповідного регіону.
  - *Population_Category* – категоризація населення за розмірами регіонів.
  - *Nodal_DR_TB_Centres_Per_Population* – кількість центрів лікування DR-TB на одиницю населення.
  - *Year* – рік реєстрації випадків.
  - *MDR/RR TB DIAGNOSED* – кількість випадків мультирезистентного туберкульозу.
  - *TB patients with known Tobacco usage status* – пацієнти з відомим статусом вживання тютюну.
  - *TB patients with known Alcohol usage status* – пацієнти з відомим статусом вживання алкоголю.
  - *TB case notification total* – загальна кількість повідомлених випадків туберкульозу (цільова змінна).

Обчислення кореляційних зв'язків

На основі отриманих факторів був проведений кореляційний аналіз, що дозволяє оцінити взаємозв'язок між різними змінними та визначити їх можливий вплив на загальні показники туберкульозу (ТБ) (таблиця 1).

Коефіцієнти кореляції між новими факторами та цільовим показником

Показник	Кореляція з TB case notification total
Gender Ratio	0.366
Detection Efficiency	-0.015
Treatment Success Rate Adjusted	0.082
Total Notified Cases by Gender	0.117
Total Treatment Outcomes	-0.006
Tobacco_Alcohol Interaction	0.052
HIV_ART Impact	0.123
City_Encoded	-0.152
Population	-0.106
Population_Category	-0.020
Nodal_DR_TB_Centres_Per_Population	0.038
Year	-0.136
MDR/RR TB Diagnosed	0.082
TB patients with known Tobacco usage status	0.095
TB patients with known Alcohol usage status	0.010

Аналіз показав, що найбільший позитивний взаємозв'язок спостерігається між загальною кількістю повідомлених випадків ТБ та показником *Gender Ratio* (0.365858). Це може свідчити про нерівномірний розподіл захворюваності серед різних гендерних груп. Вплив таких факторів, як *HIV_ART Impact* (0.123413) та *Total Notified Cases by Gender* (0.116796), також є помірно позитивним, що може свідчити про значущість ВІЛ-інфекції у загальній статистиці ТБ.

Значення кореляції для *Treatment Success Rate Adjusted* (0.081819) та *MDR/RR TB Diagnosed* (0.081947) свідчать про відносно слабкий, але позитивний взаємозв'язок із загальним числом випадків ТБ. Це може свідчити про те, що покращення успішності лікування не має значного впливу на загальну кількість випадків захворювання, що, ймовірно, пояснюється затримками у діагностиці та лікуванні.

Негативна кореляція спостерігається для таких змінних, як *City_Encoded* (-0.151714) та *Year* (-0.135872), що може вказувати на зміни у просторовому розподілі захворюваності та тенденції зменшення випадків ТБ у певних регіонах або періодах часу. Відносно слабка негативна кореляція з *Population* (-0.106322) може свідчити про те, що загальна чисельність населення не є визначальним фактором у розподілі випадків ТБ.

Наявність слабкої кореляції між *Tobacco_Alcohol Interaction* (0.051977), *TB patients with known Tobacco usage status* (0.095110) та *TB patients with known Alcohol usage status* (0.010252) вказує на потенційний вплив шкідливих звичок на захворюваність, хоча безпосередній зв'язок залишається незначним.

Отримані результати вказують на необхідність подальшого аналізу впливу соціальних та демографічних факторів на розповсюдження ТБ.

Моделювання з використанням класичних лінійних моделей

Для прогнозування залежної змінної було використано три класичні лінійні моделі: множинну лінійну регресію (LM), Lasso-регресію та Ridge-регресію. Основною метою моделювання було оцінити вплив різних факторів на рівень туберкульозу та порівняти ефективність мо-

делей у прогнозуванні. Множинна лінійна регресія (LM) є базовим підходом, який оцінює залежність між предикторами та цільовою змінною за допомогою методу найменших квадратів. Вона є чутливою до мультиколінеарності, тому для покращення стабільності моделі застосовувалися методи регуляризації.

Lasso-регресія (Least Absolute Shrinkage and Selection Operator) включає L1-регуляризацію, що дозволяє виконувати автоматичний відбір змінних, встановлюючи вагові коефіцієнти деяких предикторів рівними нулю. Це робить її корисною для розріджених моделей із великою кількістю ознак.

Ridge-регресія застосовує L2-регуляризацію, що штрафує великі значення коефіцієнтів, зменшуючи їхній вплив, але не обнуляючи. Це дозволяє зменшити проблему мультиколінеарності та покращити узагальнюючу здатність моделі.

Для налаштування моделей використовувався метод GridSearchCV, що дозволяє знаходити оптимальні параметри шляхом перебору можливих значень. Зокрема, оптимізація здійснювалася за параметром **α (альфа)**, який визначає ступінь регуляризації. Діапазон значень для підбору параметра був заданий наступним чином:

$$\alpha \in \{0.01, 5.0\}$$

Якість моделей оцінювалася за показниками RMSE (середньоквадратична помилка) на тренувальній та тестовій вибірках, а також коефіцієнтом детермінації R^2, що характеризує пояснювальну здатність моделі. Найкращі значення параметра регуляризації були отримані наступні: Lasso-регресія: $\alpha = 0.11$; Ridge-регресія: $\alpha = 4.91$.

Отримані результати представлені у таблиці 2.

Порівняння якості моделей

Модель	RMSE Train	RMSE Test	R^2 Train	R^2 Test
LM	0.775	0.726	0.377	0.500
Ridge	0.793	0.770	0.347	0.437
Lasso	0.902	0.960	0.155	0.126

Результати показали, що множинна лінійна регресія (LM) забезпечила найкращі показники на тестовій вибірці ($R^2 = 0.500$), що свідчить про її кращу прогнозтичну здатність у порівнянні з регуляризованими моделями. Ridge-регресія мала трохи гірший результат ($R^2 = 0.437$), але продемонструвала стабільність у прогнозах. Lasso-регресія, хоча і дозволяє автоматично відбирати важливі ознаки, показала найгірші результати ($R^2 = 0.126$), що вказує на можливу втрату важливої інформації через занадто жорстку регуляризацію.

Визначення важливості факторів для лінійних моделей

Проведений аналіз дозволив визначити ключові фактори, що впливають на поширення туберкульозу, шляхом оцінки вагових коефіцієнтів у моделях Lasso та Ridge-регресії. Завдяки використанню регуляризації було отримано стабільні оцінки важливості ознак, що дозволяє виділити найсуттєвіші предиктори для прогнозування рівня захворюваності.

Отримані результати показують, що найбільший позитивний вплив на прогнозовану змінну мають *"Total Notified Cases by Gender"* (0.6328 для Lasso, 0.4282 для Ridge), що підтверджує важливість врахування гендерного розподілу при аналізі епідеміологічних даних. Також значущими є *"Gender Ratio"* (0.3222 у Lasso, 0.3385 у Ridge) та *"TB patients with known Alcohol usage status"* (0.2700 у Lasso, 0.2047 у Ridge), що вказує на вплив демографічних факторів та соціальної поведінки населення.

"Treatment Success Rate Adjusted" (0.1424 у Lasso, 0.1459 у Ridge) та *"HIV_ART Impact"* (0.1260 у Lasso, 0.1319 у Ridge) також мають помірний позитивний вплив, що може бути пов'язано з ефективністю лікувальних стратегій та супутніми захворюваннями, які впливають на стан пацієнтів з туберкульозом.

Значний негативний вплив спостерігається у змінної *"Population"* (-0.8813 у Lasso, -0.6212 у Ridge), що може вказувати на те, що в регіонах із вищою чисельністю населення загальна

кількість випадків туберкульозу може розподілятися більш рівномірно, зменшуючи відносну частку повідомлених випадків.

Деякі ознаки, такі як *"Tobacco_Alcohol Interaction"*, мали ваговий коефіцієнт рівний нулю у Lasso-регресії, що свідчить про їхню низьку інформативність та можливе усунення з моделі.

Вагові коефіцієнти ознак у моделях Lasso та Ridge

Feature	Lasso	Ridge
Gender Ratio	0.322	0.338
Detection Efficiency	-0.007	-0.006
Treatment Success Rate Adjusted	0.142	0.146
Total Notified Cases by Gender	0.633	0.428
Total Treatment Outcomes	-0.191	-0.202
Tobacco_Alcohol Interaction	0.000	-0.002
HIV_ART Impact	0.126	0.132
City_Encoded	-0.111	-0.113
Population	-0.881	-0.621
Population_Category	0.172	0.120
Nodal_DR_TB_Centres_Per_Population	-0.036	-0.049
Year	-0.130	-0.154
MDR/RR TB DIAGNOSED MDR/RR patient diagnosed	-0.066	-0.060
TB patients with known Tobacco usage status	0.143	0.192
TB patients with known Alcohol usage status	0.270	0.205

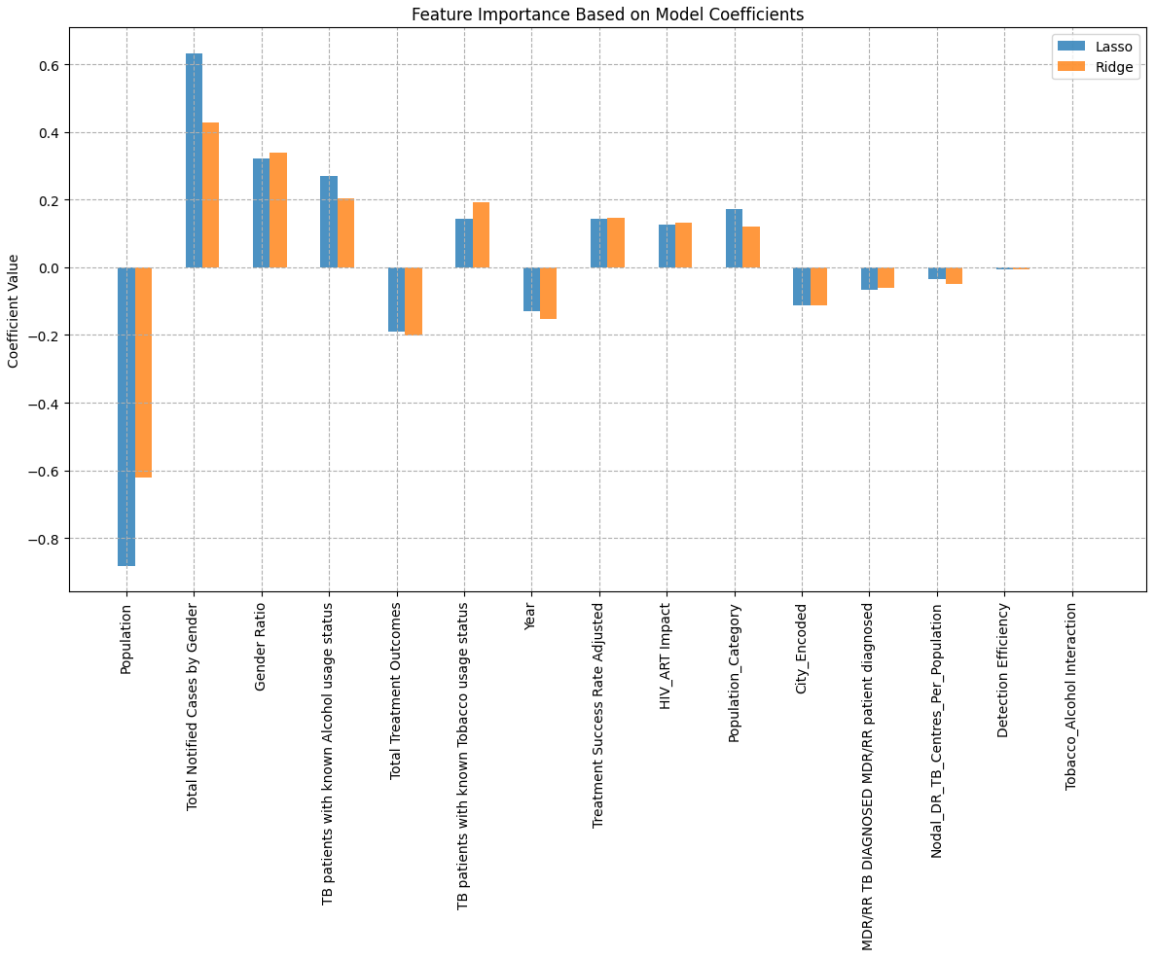

Рис. 1. Порівняльний аналіз важливості факторів для лінійних моделей

Тестування моделей машинного навчання

Для покращення якості прогнозування рівня захворюваності на туберкульоз було протестовано кілька моделей машинного навчання, зокрема *Decision Tree (дерево рішень), K-Nearest Neighbors (метод k найближчих сусідів)* та *Random Forest (випадковий ліс)*. Ці моделі мають різні підходи до побудови прогнозів та можуть по різному узагальнювати залежності у даних.

• *Decision Tree (Дерево рішень)* – алгоритм, що використовує деревоподібну структуру для ухвалення рішень. У процесі навчання модель поділяє простір ознак на області, в яких робить передбачення. Decision Tree добре адаптується до тренувальних даних, але має схильність до перенавчання (overfitting), якщо не застосовуються механізми обрізки (pruning).

• *K-Nearest Neighbors (Метод k найближчих сусідів, KNN)* – модель, яка базується на відстані між точками у багатовимірному просторі. Передбачення здійснюється на основі найближчих k-сусідів за допомогою голосування або середнього значення. Цей метод добре працює у випадках, коли дані мають локальні закономірності, але може бути чутливим до вибору k і масштабування ознак.

• *Random Forest (Випадковий ліс)* – ансамблевий метод, що об'єднує кілька дерев рішень для підвищення стабільності та точності. Використовує випадкову вибірку ознак та даних для навчання кожного дерева, що зменшує ймовірність перенавчання.

Отримані результати наведено у таблиці 4.

Порівняння якості моделей машинного навчання

Модель	RMSE Train	RMSE Test	R^2 Train	R^2 Test
Decision Tree	0.000	1.015	1.000	0.024
K-Nearest Neighbors	0.533	0.601	0.705	0.657
Random Forest	0.259	0.747	0.931	0.471

Результати показують, що Decision Tree повністю запам'ятало тренувальні дані (R^2 Train = 1.000), але погано узагальнює закономірності для тестової вибірки (R^2 Test = 0.024), що вказує на значне перенавчання.

K-Nearest Neighbors продемонстрував найкращі результати серед усіх моделей за тестовими метриками (RMSE Test = 0.601, R^2 Test = 0.657), що свідчить про хорошу узагальнюючу здатність цього алгоритму для прогнозування туберкульозу.

Random Forest також показав високу точність на тренувальній вибірці (R^2 Train = 0.931) і досить конкурентоспроможні результати на тестових даних (R^2 Test = 0.471). Це вказує на те, що ансамблеві методи можуть бути корисними для прогнозування рівня захворюваності.

Моделі LM, Ridge та Lasso продемонстрували слабші результати у порівнянні з KNN та Random Forest, що може бути пов'язано з лінійною природою цих методів та недостатньою здатністю до моделювання складних нелінійних залежностей.

Покращення точності моделей

Однією з проблем при побудові моделей регресії є дисбаланс у значеннях цільової змінної. У даному дослідженні для збільшення представленості рідкісних випадків застосовувався метод SMOGN (Synthetic Minority Over-sampling TEchnique for Regression), який є модифікацією класичного SMOTE (Synthetic Minority Over-sampling Technique) та використовується для регресійних задач. В якості гіперпараметру моделі, а саме кількості найближчих сусідів для синтетичного генератора, було обрано k=5. Метод SMOGN створює нові синтетичні точки, зберігаючи структуру вихідних даних та коригуючи дисбаланс у розподілі цільової змінної. Це дозволяє покращити роботу моделей регресії, які інакше можуть бути схильні до упередженості щодо більшості значень.

На рис. 2 представлено порівняння вихідного розподілу цільової змінної до та після застосування oversampling. Як видно з графіку, після балансування частка малопоширених значень значно зросла, що сприяє покращенню здатності моделей узагальнювати дані.

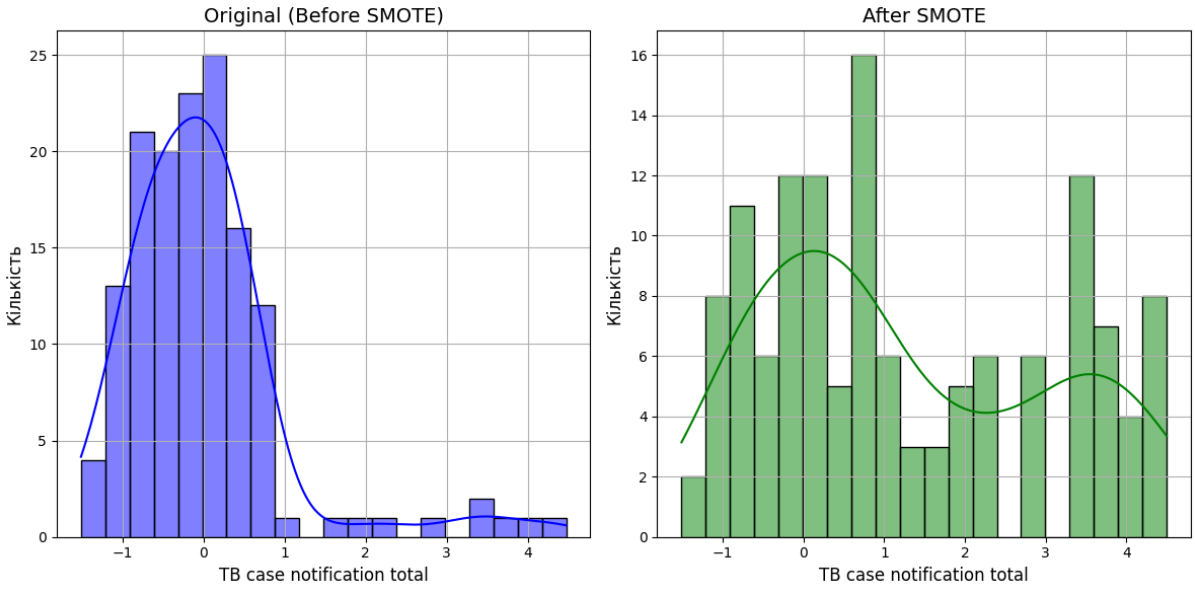

Рис. 2. Розподіл значень цільової комірки до і після застосування oversampling

Оцінка якості моделей після балансування вибірки

На основі оновленого набору даних були повторно натреновані всі моделі. Оцінка якості прогнозування після застосування oversampling представлена у таблиці 5.

Порівняння точності моделей після застосування oversampling

Модель	RMSE Train	RMSE Test	R^2 Train	R^2 Test
LM	0.920	1.056	0.733	0.600
Ridge	0.969	1.077	0.703	0.584
Lasso	1.084	1.063	0.629	0.595
Decision Tree	0.000	1.219	1.000	0.467
K-Nearest Neighbors	0.581	0.614	0.893	0.865
Random Forest	0.284	0.684	0.974	0.832

Порівняно з попередніми результатами, застосування oversampling та нейронних мереж дозволило суттєво покращити якість прогнозування.

1. Лінійні моделі (LM, Ridge, Lasso) показали незначне покращення точності. Значення R^2 Test зросли у всіх випадках (наприклад, для LM з 0.500 до 0.600), що свідчить про покращену узагальнюючу здатність моделей завдяки збалансованому розподілу цільової змінної.

2. Decision Tree залишається перенавченим, маючи R^2 Train = 1.000, але все ще демонструє погані результати на тестовій вибірці (R^2 Test = 0.467), що підтверджує його низьку стійкість до зміни вибірки.

3. Метод K-Nearest Neighbors (KNN) значно покращив результати. Значення R^2 Test зросло з 0.657 до 0.865, а RMSE Test зменшилося, що вказує на ефективніше узагальнення даних.

4. Random Forest також покращив якість прогнозування після балансування вибірки. Його R^2 Test підвищився з 0.471 до 0.832, що свідчить про стабільне узагальнення даних без значного перенавчання.

Покращення точності моделей

Для подальшого покращення прогнозування було створено глибоку нейронну мережу з трьома прихованими шарами по 128 нейронів кожен, що використовує функцію активації ReLU. Регуляризація L2 (λ=0.01) застосовувалася для запобігання перенавчанню. Також для

запобігання перенавчання використовувалися прошарки DropOut параметром активації нейронів 0.3.

Для навчання мережі та уникнення перенавчання використовувалася рання зупинка навчання на основі критерію мінімізації помилки на валідаційній вибірці (val_loss). Модель зупиняла навчання, якщо помилка не покращувалася протягом 100 епох, при цьому зберігалася найкраща конфігурація вагів:

early_stopping = EarlyStopping(monitor='val_loss', patience=100, restore_best_weights=True, verbose=2)

Результати роботи нейронної мережі порівняно з іншими моделями представлені у таблиці 6.

Точність прогнозування для нейронної мережі

Модель	RMSE Train	RMSE Test	R^2 Train	R^2 Test
LM	0.920	1.056	0.733	0.600
Ridge	0.969	1.077	0.703	0.584
Lasso	1.084	1.063	0.629	0.595
Decision Tree	0.000	1.219	1.000	0.467
K-Nearest Neighbors	0.581	0.614	0.893	0.865
Random Forest	0.284	0.684	0.974	0.832
Neural Network	0.249	0.433	0.936	0.822

На рис. 3 представлено графіки динаміки навчання нейронної мережі, що демонструють зниження помилки на тренувальній та валідаційній вибірках. На рис. 4 наведено порівняльний аналіз прогнозу нейронної мережі з фактичними значеннями.

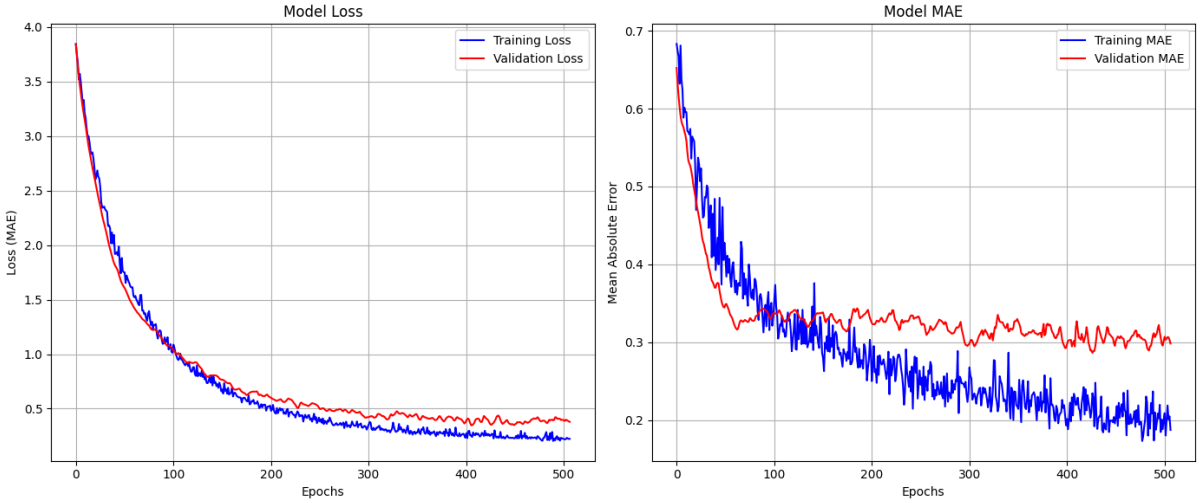

Рис. 3. Динаміка функції втрат та середньоквадратичної помилки у процесі навчання нейронної мережі

Як видно з таблиці та рисунків, використання глибокої нейронної мережі суттєво покращило точність прогнозування у порівнянні з іншими моделями. RMSE Test знизився до 0.433, що є кращим результатом серед усіх моделей. Значення R^2 Test = 0.822 вказує на високу пояснювальну здатність моделі.

Порівняно з попередніми моделями:

• Нейронна мережа перевершила всі лінійні моделі (LM, Ridge, Lasso), які мали значно вищі RMSE та нижчий R^2 на тестовій вибірці.

• Random Forest та KNN показали конкурентоспроможні результати, але поступилися нейронній мережі за точністю.

• Decision Tree продовжує демонструвати перенавчання, маючи R^2 Train = 1.000, але поганий результат на тестових даних (R^2 Test = 0.467).

Таким чином, нейронна мережа продемонструвала найкращу загальну точність, що свідчить про її здатність виявляти складні взаємозв'язки між факторами захворюваності.

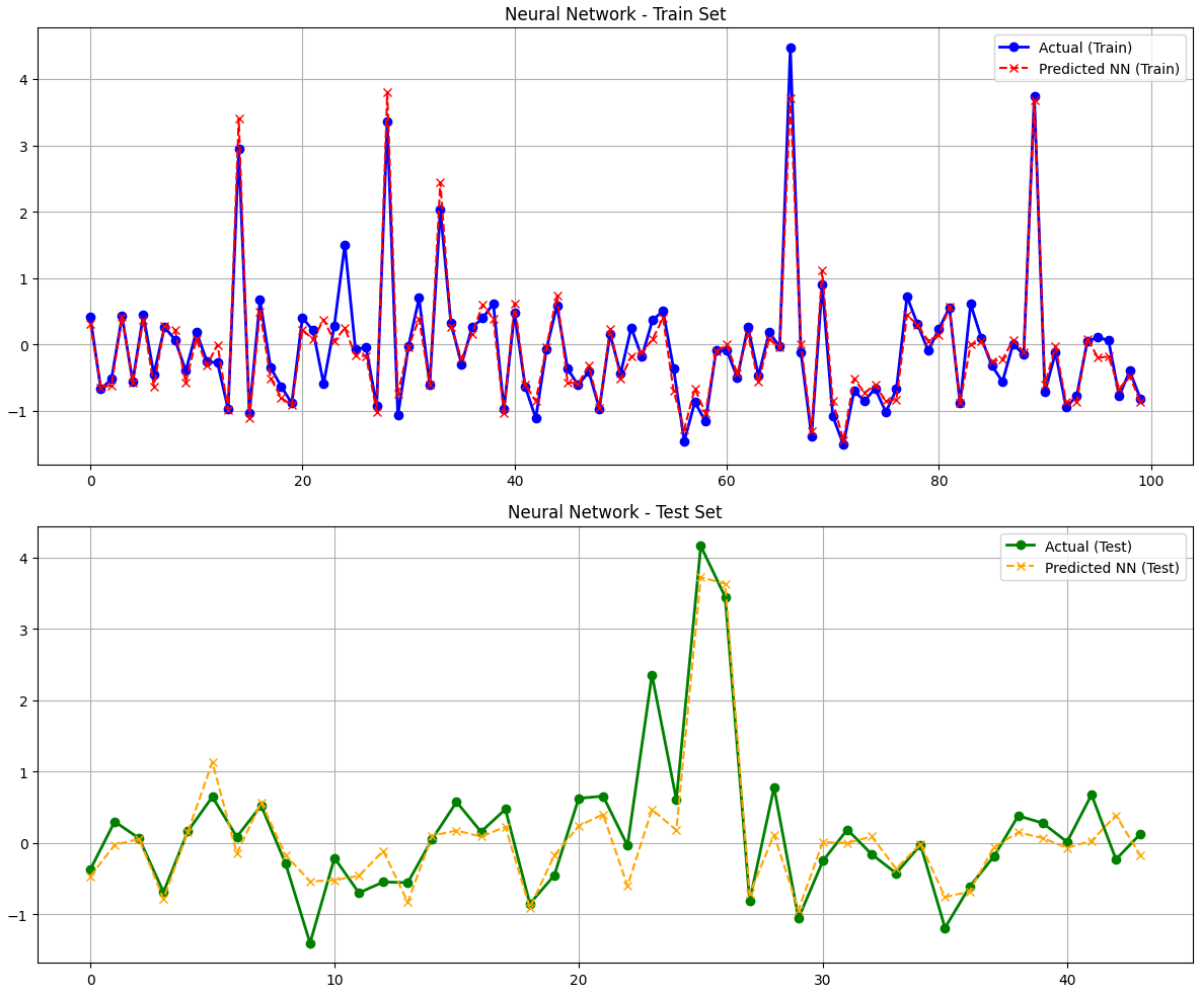

Рис. 4. Порівняльний аналіз прогнозу динаміки поширення туберкульозу за допомогою нейронної мережі

Визначення ключових факторів для нейронних мереж

Для оцінки впливу вхідних ознак на результати прогнозування нейронної мережі було використано метод permutation_importance – це метод, який вимірює вплив кожної змінної на модель шляхом випадкового перемішування її значень і повторного розрахунку метрик точності. Основні етапи методу:

1. Визначається базова точність моделі (наприклад, R^2 або RMSE).
2. По черзі перемішуються значення окремих предикторів, а модель знову прогнозує результати.
3. Визначається зміна точності прогнозування після перемішування. Чим більше зменшується точність, тим важливішим є цей фактор.

Цей підхід є зручним, оскільки дозволяє оцінити реальний вплив змінних без необхідності повторного навчання моделі, що значно пришвидшує процес аналізу.

Отримані результати представлені у таблиці 7.

Важливість ознак для нейронної мережі (Permutation Importance)

Ознака	Важливість
Population	0.623
Gender Ratio	0.602

Nodal_DR_TB_Centres_Per_Population	0.454
City_Encoded	0.370
TB patients with known Alcohol usage status	0.146
Year	0.078
HIV_ART Impact	0.038
Treatment Success Rate Adjusted	0.031
Total Notified Cases by Gender	0.017
Population_Category	0.010
MDR/RR TB DIAGNOSED MDR/RR patient diagnosed	0.003
Total Treatment Outcomes	0.002
Detection Efficiency	-0.000
TB patients with known Tobacco usage status	-0.004
Tobacco_Alcohol Interaction	-0.011

Як видно з таблиці, найбільш значущими факторами для прогнозування туберкульозу у нейронній мережі є Population (0.623) та Gender Ratio (0.602). Це узгоджується з результатами лінійних моделей, де Population також мала сильний негативний вплив, а Gender Ratio відігравало важливу роль у розподілі захворюваності. Nodal_DR_TB_Centres_Per_Population (0.454) та City_Encoded (0.370) мають високий вплив у нейронній мережі, що вказує на значущість інфраструктури центрів діагностики ТБ та міського середовища у прогнозуванні захворюваності. У лінійних моделях ці фактори мали менший вплив, що свідчить про здатність нейронної мережі краще виявляти складні закономірності. Соціальні фактори, такі як TB patients with known Alcohol usage status (0.146), мають дещо більший вплив у нейронній мережі порівняно з лінійними моделями, що може свідчити про взаємозв'язок шкідливих звичок із розповсюдженням хвороби. Фактор Year (0.078) має помірний вплив, що свідчить про часові тенденції у динаміці захворюваності. Total Notified Cases by Gender (0.017) та HIV_ART Impact (0.038) виявилися менш значущими у нейронній мережі, хоча у лінійних моделях вони мали вищі коефіцієнти. Це може свідчити про те, що нейронна мережа виявляє більш складні взаємозв'язки, а не лише прямий лінійний вплив цих змінних. Мінімальний або від'ємний вплив спостерігається у таких змінних, як Detection Efficiency (-0.000), TB patients with known Tobacco usage status (-0.004) та Tobacco_Alcohol Interaction (-0.011), що узгоджується з результатами лінійних моделей, де ці фактори також мали слабкий вплив або обнулялися у Lasso-регресії.

Порівняння з лінійними моделями

У лінійних моделях Population мала негативний вплив, а у нейронній мережі ця змінна є найважливішим предиктором. Це підтверджує, що нейронні мережі здатні знаходити більш складні залежності, які можуть не бути очевидними при лінійному аналізі.

У лінійних моделях Gender Ratio теж мав високу важливість, що підтверджує сталість цього фактору у різних підходах.

Лінійні моделі значну увагу приділяли Total Notified Cases by Gender, тоді як у нейронній мережі ця змінна відіграє меншу роль. Це може бути наслідком того, що нейромережа може використовувати інші фактори для побудови прогнозу.

Міські та регіональні фактори (City_Encoded та Nodal_DR_TB_Centres_Per_Population) мали нижчий вплив у лінійних моделях, але суттєво підвищили свою значущість у нейронній

мережі. Це може свідчити про те, що у лінійних моделях такі фактори погано описують залежності, а нейронна мережа знаходить більш складні зв'язки між ними.

Висновки

Регресійні моделі (Linear Regression, Ridge, Lasso) показали обмежену точність прогнозування через неможливість врахування складних нелінійних взаємозв'язків між факторами. Регуляризація в Lasso та Ridge допомогла дещо покращити стабільність моделей, але їхня узагальнююча здатність залишалася нижчою порівняно з моделями машинного навчання.

Ансамблеві моделі, такі як Random Forest та K-Nearest Neighbors (KNN), суттєво перевершили лінійні методи, забезпечивши R^2 Test ≈ 0.832 для Random Forest та 0.865 для KNN. Це свідчить про їхню здатність виявляти нелінійні закономірності та взаємозв'язки між факторами ризику.

Використання методу SMOGN для балансування вибірки дозволило покращити якість прогнозування, особливо для моделей, що працюють із нерівномірними розподілами даних. Це допомогло підвищити точність тестових прогнозів, зокрема для KNN та нейронних мереж.

Найкращі результати були отримані за допомогою глибокої нейронної мережі, яка досягла R^2 Test = 0.822 при RMSE Test = 0.433. Використання регуляризації L2 та dropout дозволило уникнути перенавчання, а застосування ранньої зупинки навчання допомогло зберегти оптимальні ваги.

Аналіз показав, що найбільший вплив на рівень захворюваності на туберкульоз мають Population (чисельність населення), Gender Ratio (гендерне співвідношення), Nodal_DR_TB_Centres_Per_Population (кількість спеціалізованих центрів на душу населення) та City_Encoded (міські фактори). Ці змінні залишалися важливими незалежно від вибору моделі, що підтверджує їхню ключову роль у прогнозуванні.

Лінійні моделі надавали більше значення Total Notified Cases by Gender та HIV_ART Impact, тоді як у нейронній мережі ключовими факторами стали Population, Gender Ratio, міські характеристики та кількість спеціалізованих центрів. Це демонструє, що глибокі моделі можуть краще виявляти складні зв'язки у даних.

Дослідження підтвердило, що нейронні мережі та ансамблеві моделі (Random Forest, KNN) є найбільш ефективними для прогнозування рівня захворюваності на туберкульоз. Використання oversampling та глибокого навчання дозволило значно підвищити якість прогнозування, що може бути корисним для подальших медичних досліджень та розробки стратегій боротьби з поширенням туберкульозу.

Список літератури

1. Kumari, S., & Bhatia, M. (2022). Machine Learning Techniques For Public Health System: A Scientometric Review. ICCSEA 2022. DOI: 10.1109/ICCSEA54677.2022.9936149

2. Abraham, S., Mamatha, G., Airbail, H., et al. (2022). Diagnosing Patient Health Conditions and Improving the Patient Experience: An Application of AI and ML. Healthcare and Knowledge Management for Society 5.0. DOI: 10.1201/9781003168638-2

3. Varela-Rey, I., Bandín-Vilar, E., Toja-Camba, F.J., et al. (2024). Artificial Intelligence and Machine Learning Applications to Pharmacokinetic Modeling and Dose Prediction of Antibiotics: A Scoping Review. Antibiotics, 13(12), 1203. DOI: 10.3390/antibiotics13121203

4. Siddiqui, A.K., & Garg, V.K. (2021). Diagnosis of Pulmonary Tuberculosis through Intelligent Techniques: A Review. ICCS 2021, 189-193. DOI: 10.1109/ICCS54944.2021.00045

5. Hassan, Y.M., Mohamed, A.S., Hassan, Y.M., & El-Sayed, W.M. (2025). Recent developments and future directions in point-of-care next-generation CRISPR-based rapid diagnosis. Clinical and Experimental Medicine, 25(1), 33. DOI: 10.1007/s10238-024-01540-8

6. Yadav, P., Rastogi, V., Yadav, A., & Parashar, P. (2024). Artificial Intelligence: A promising tool in diagnosis of respiratory diseases. Intelligent Pharmacy, 2(6), 784-791. DOI: 10.1016/j.ipha.2024.05.002

7. Qureshi, H., Shah, Z., Raja, M.A.Z., et al. (2024). Machine learning investigation of tuberculosis with medicine immunity impact. Diagnostic Microbiology and Infectious Disease, 110(3), 116472. DOI: 10.1016/j.diagmicrobio.2024.116472

8. Fayaz, S.A., Babu, L., Paridayal, L., et al. (2024). Machine learning algorithms to predict treatment success for patients with pulmonary tuberculosis. PLoS ONE, 19(10), e0309151. DOI: 10.1371/journal.pone.0309151

9. Rabie, A.H., & Saleh, A.I. (2024). Diseases diagnosis based on artificial intelligence and ensemble classification. Artificial Intelligence in Medicine, 148, 102753. DOI: 10.1016/j.artmed.2023.102753

10. Al Meslamani, A.Z., Sobrino, I., & de la Fuente, J. (2024). Machine learning in infectious diseases: potential applications and limitations. Annals of Medicine, 56(1), 2362869. DOI: 10.1080/07853890.2024.2362869

11. Escorcia-Gutierrez, J., Soto-Diaz, R., Madera, N., et al. (2023). Computer-Aided Diagnosis for Tuberculosis Classification with Water Strider Optimization Algorithm. Computer Systems Science and Engineering, 46(2), 1337-1353. DOI: 10.32604/csse.2023.035253

12. Dolma, K.G., Paul, A.K., Rahmatullah, M., et al. (2023). AI and TB: A New Insight in Digital Chest Radiography. Lecture Notes in Computational Vision and Biomechanics, 37, 439-450. DOI: 10.1007/978-981-19-0151-5_37

13. Kulkarni, M., Golechha, S., Raj, R., et al. (2022). Predicting treatment adherence of tuberculosis patients at scale. Proceedings of Machine Learning Research, 193, 35-61.

14. Sharma, M., & Singh, P. (2021). Use of Artificial Intelligence in Research and Clinical Decision Making for Combating Mycobacterial Diseases. Artificial Intelligence and Machine Learning in Healthcare, 183-215. DOI: 10.1007/978-981-16-0811-7_9

D. Nevinskyi, D. Martjanov, O. Hospodarskyy, Y. Vyklyuk, I. Semianiv

DETERMINING THE FACTORS OF TUBERCULOSIS: ANALYSIS OF MACHINE LEARNING AND NEURAL NETWORK METHODS

Tuberculosis (TB) remains one of the most serious infectious diseases globally, particularly in India, where its high incidence poses significant challenges to the healthcare system. This study focuses on analyzing the determinants of TB prevalence in India using machine learning (ML) and neural network (NN) methods. The objective is to identify key factors influencing TB incidence and develop accurate predictive models to support prevention and treatment strategies. Based on statistical data from 2019–2022, encompassing demographic characteristics, social factors, and medical indicators, a comprehensive analysis was conducted. Data processing techniques, including correlation analysis, oversampling (SMOGN) for sample balancing, and modeling with linear regressions (LM, Ridge, Lasso), ML algorithms (Decision Tree, K-Nearest Neighbors, Random Forest), and a deep neural network were employed. Results revealed that linear models exhibited limited accuracy (R^2 Test up to 0.600), while Random Forest (R^2 Test = 0.832) and K-Nearest Neighbors (R^2 Test = 0.865) significantly outperformed them due to their ability to capture nonlinear relationships.

The highest accuracy was achieved by the neural network (R^2 Test = 0.822, RMSE Test = 0.433), highlighting its effectiveness in detecting complex interdependencies. Key factors influencing TB incidence included population size (Population), gender ratio (Gender Ratio), the number of specialized centers (Nodal_DR_TB_Centres_Per_Population), and urban characteristics (City_Encoded). These findings underscore the potential of integrating ML and NN into medical research for TB forecasting and control, offering valuable insights for developing personalized therapeutic approaches and improving public health outcomes.

Keywords: machine learning; neural networks; disease prediction; oversampling; SMOGN; linear regression; Random Forest; K-Nearest Neighbors; determinants; integration.