DOI: 10.32347/2412-9933.2025.61.160-169
УДК 004.02, 004.67, 004.891.3, 616.24-002.5-02:316.342.6:316.62:314(477)

Господарчук Дмитро Віталійович
Студент кафедри диференціальних рівнянь та математичної статистики,
https://orcid.org/0009-0003-9425-4609
Львівський національний університеті імені Івана Франка, Львів
Невінський Денис Володимирович
Доцент кафедри електронних засобів інформаційно-комп'ютерних технологій,
https://orcid.org/0000-0002-0962-072X
Інститут телекомунікацій, радіоелектроніки та електронної техніки Національного університету «Львівська політехніка», Львів
Мартьянов Дмитро Ігорович
Аспірант кафедри систем штучного інтелекту,
https://orcid.org/0009-0003-3919-4412
Інститут комп'ютерних наук та інформаційних технологій Національного університету «Львівська політехніка», Львів
Виклюк Ярослав Ігорович
Професор кафедри систем штучного інтелекту,
https://orcid.org/0000-0003-4766-4659
Інститут комп'ютерних наук та інформаційних технологій Національного університету «Львівська політехніка», Львів
Сем'янів Ігор Олександрович
Доцент кафедри фтизіатрії та пульмонології,
https://orcid.org/0000-0003-0340-0766
Буковинський державний медичний університет, Чернівці

ОПТИМІЗАЦІЯ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ДЛЯ ОЦІНКИ РИЗИКУ ПОШИРЕННЯ ТУБЕРКУЛЬОЗУ

Анотація. *Туберкульоз (ТБ) залишається однією з найактуальніших проблем охорони здоров'я, особливо в країнах, що розвиваються. Високий рівень захворюваності та поширення мультирезистентних штамів Mycobacterium tuberculosis створюють значні виклики для сучасної медицини. Індія є однією з держав із найбільшим тягарем ТБ, тому оптимізація методів прогнозування поширення хвороби є надзвичайно важливою для ефективного впровадження заходів профілактики і лікування. Застосування методів машинного навчання (ML) дає можливість автоматизувати аналіз великих обсягів даних та виявляти ключові фактори ризику. Метою цього дослідження є розроблення ефективних моделей машинного навчання для оцінки ризику поширення ТБ в Індії на основі соціально-економічних, демографічних і медичних факторів. Для аналізу було використано набір даних, що містить 148 записів за період 2019–2022 рр., розбитих за штатами Індії. До основних змінних належить кількість виявлених випадків ТБ, показники успішності лікування, рівень смертності серед хворих, а також статус вживання тютюну й алкоголю серед пацієнтів. Дослідження включало попередню обробку даних, кореляційний аналіз та застосування методів машинного навчання. Було протестовано кілька моделей: лінійну регресію, регуляризовані моделі (Lasso та Ridge), метод опорних векторів (SVM), метод найближчих сусідів (KNN), випадковий ліс та дерево рішень. Аналіз засвідчив, що найкращу точність має модель SVM із оптимізованими параметрами, що продемонструвала найвищий коефіцієнт детермінації та найнижчу середньоквадратичну помилку. Порівняння інших моделей виявило значні переваги SVM над лінійною регресією та деревом рішень, які показали низьку узагальнюючу здатність. Визначення найбільш вагомих факторів у прогнозуванні поширення ТБ здійснено за допомогою методу Permutation Importance. Найбільший вплив мали такі фактори: географічне розташування (штат), кількість зареєстрованих випадків ТБ серед дітей, кількість жінок із ТБ, рівень смертності серед пацієнтів та інфраструктура для лікування лікарсько-стійкого ТБ. Виявлено, що соціальні фактори, такі як рівень споживання тютюну й алкоголю серед пацієнтів, також впливають на поширення хвороби, проте їхній внесок є менш значущим. Дослідження підтвердило*

ефективність застосування методів машинного навчання для прогнозування поширення туберкульозу. Оптимізована модель SVM забезпечила найкращі показники точності й узагальнюючої здатності. Аналіз вагомості факторів засвідчив, що найбільший вплив на поширення хвороби мають регіональні особливості, демографічні показники та рівень смертності. Отримані результати можуть бути використані для вдосконалення стратегій боротьби з ТБ, зокрема через цільове впровадження заходів у регіонах з високими ризиками. Використання ML-методів дає змогу покращити ефективність контролю над захворюванням, що є важливим кроком у глобальній боротьбі з туберкульозом.

Ключові слова: туберкульоз; машинне навчання; прогнозування; модель SVM; фактори впливу; штучний інтелект; регресійний аналіз

Актуальність дослідження

Туберкульоз (ТБ) є однією з важливих проблем сучасної медицини в країнах, що розиваються, і має прямий вплив на соціально-економічні показники розвитку суспільства. За даними ВООЗ, Індія входить до числа держав, на які припадає значна частина випадків ТБ у світі. Високий рівень захворюваності, разом із недостатньою діагностикою та лікуванням, створює серйозні проблеми для національної системи охорони здоров'я.

Вивчення поширення туберкульозу на основі статистичних даних дає змогу виявити закономірності, ключові фактори ризику та допомагає у прогнозуванні тенденцій захворюваності. Це дослідження особливо актуальне у 2025 р., враховуючи глобальне зростання резистентного туберкульозу, який потребує сучасних методів аналізу для ефективного втручання. Статистичний аналіз і прогнозування на основі наявних даних є важливим кроком для підвищення ефективності заходів профілактики та лікування ТБ.

Огляд літературних джерел

Штучний інтелект (ШІ) і машинне навчання (МН) на сьогодні стають важливими інструментами для вивчення й боротьби з інфекційними захворюваннями, такими як туберкульоз. Розглянемо основні дослідження, що висвітлюють використання МН у різних аспектах діагностики, прогнозування та лікування ТБ.

Аналіз сучасних досліджень свідчить про зростаючу роль методів машинного навчання та часових рядів у прогнозуванні епідемічних процесів. У дослідженні [1] проведено систематичний огляд літератури щодо застосування методів часових рядів для прогнозування епідемій. Автори виявили, що, незважаючи на широке використання цих методів у фінансах та метеорології, у сфері епідеміології їх застосування залишається обмеженим. Особливу увагу дослідники приділили впливу COVID-19 на використання таких методів у прогнозуванні інфекційних хвороб.

У роботі [2] досліджено ефективність узагальненої лінійної змішаної моделі (GLMM) та екстремальної нейронної мережі (ENN) у прогнозуванні ризику туберкульозу в провінції Західна Ява, Індонезія. Результати засвідчили, що такі фактори, як щільність населення та вікові групи, мають значний вплив на частоту випадків захворювання, а запропонований підхід може бути корисним для раннього виявлення ризиків.

Дослідження [3] було присвячене розробці моделі мультиплексних графових нейронних мереж (GNN) для прогнозування медичних наслідків. Запропонована методика уможливлює ефективно об'єднувати дані різних модальностей, таких як клінічні, зображувальні та геномні, що може значно покращити точність прогнозування.

У дослідженні [4] розглянуто можливості штучного інтелекту у прогнозуванні ефективності лікування туберкульозу. Було виявлено, що методи машинного навчання, зокрема нейронні мережі та випадковий ліс, демонструють високу точність у прогнозуванні тривалості лікування та ризику розвитку резистентності до ліків.

У роботі [5] представлено DeepDynaForecast – нову модель на основі графового глибокого навчання, яка використовує філогенетичні дерева для прогнозування динаміки поширення епідемій. Висновки авторів свідчать про значний потенціал цього підходу для раннього виявлення груп високого ризику.

Інше дослідження [6] зосереджене на застосуванні ансамблевого моделювання для прогнозування результатів лікування туберкульозу у дітей. Було показано, що запропонований підхід забезпечує високу точність передбачень порівняно з традиційними методами.

У роботі [7] представлено підхід до прогнозування перебігу туберкульозного менінгіту шляхом комбінованого аналізу зображень мозку та клінічних даних. Використання методів машинного навчання уможливило досягти високої точності в оцінці тяжкості захворювання та прогнозуванні його прогресії.

Дослідження [8] порівняло класичні статистичні моделі і методи глибокого навчання в прогнозуванні

коінфекції TB/HIV. Було встановлено, що такі моделі, як BiLSTM та CNN-LSTM, значно перевершують традиційні підходи щодо точності прогнозування.

У роботі [9] запропоновано нову методику прогнозування рівня захворюваності на туберкульоз у провінції Аньхой на основі машинного навчання. Використання алгоритмів випадкового лісу, LASSO та глибоких рекурентних нейронних мереж допомогло досягти високої точності прогнозування та виявлення ключових факторів впливу.

Дослідження [10] присвячене інтеграції машинного навчання в епідеміологічний аналіз прогнозування туберкульозу. Автори розглянули різні методи прогнозування захворюваності та виявили, що використання адаптивних моделей забезпечує точніші прогнози і покращує процес прийняття рішень у сфері громадського здоров'я.

У роботі [11] проаналізовано використання МН навчання для передбачення активності Mycobacterium tuberculosis in vitro. Дослідження охоплює широкий спектр моделей, включаючи байєсівські класифікатори та регресійні моделі, що дає змогу ефективно оцінювати потенційні лікарські засоби для боротьби з ТБ.

Отже, сучасні дослідження свідчать про ефективність використання методів машинного навчання, часових рядів та глибоких нейронних мереж у прогнозуванні інфекційних хвороб. Однак залишається необхідність у подальших дослідженнях для покращення точності моделей та їх адаптації до різних епідеміологічних сценаріїв.

Водночас дослідження, що фокусуються на локальних аспектах боротьби з ТБ у країнах із високим тягарем захворюваності, таких як Індія, залишаються відкритими й потребують подальшої уваги наукової спільноти.

Мета статті

Метою статті є аналіз впливу різних соціально-економічних, медичних та демографічних факторів на захворюваність на туберкульоз серед населення різних штатів Індії, з метою ідентифікації ключових чинників, які можуть сприяти формуванню більш ефективних стратегій контролю та профілактики хвороби.

Виклад основного матеріалу

Матеріали та методи

Опис набору даних

Набір даних для аналізу впливу різних соціально-економічних, медичних та демографічних факторів на захворюваність на ТБ складається зі згаданих полів та містить 148 записів. Дані отримані в період з 2019-го по 2022 р. з розбивкою по окремих штатах Індії. Цей набір даних включає інформацію про Виявлені випадки ТБ серед протестованих (активний пошук), Передбачувані випадки ТБ, протестовані серед обстежених (активний пошук), Діагностовані пацієнти з мультирезистентним/рифампіцин-резистентним ТБ, Зареєстровані випадки ТБ у дітей, Зареєстровані випадки ТБ серед жінок, Зареєстровані випадки ТБ серед чоловіків, Пацієнти з ТБ із відомим статусом вживання тютюну, Діагностовані пацієнти з коінфекцією ТБ-ВІЛ, Пацієнти з коінфекцією ТБ-ВІЛ, розпочаті на АРТ, Результат лікування пацієнтів із ТБ (% втрачено для подальшого спостереження), Результат лікування пацієнтів із ТБ (рівень смертності), Результат лікування пацієнтів із ТБ (рівень успішного лікування), Результат лікування пацієнтів із ТБ (рівень невдачі лікування), Результат лікування пацієнтів із коінфекцією ТБ-ВІЛ (рівень смертності), Результат лікування пацієнтів із коінфекцією ТБ-ВІЛ (рівень успішного лікування), Кількість центрів лікування лікарсько-стійкого ТБ, Пацієнти з ТБ і діабетом, розпочаті на протидіабетичне лікування, Результат лікування дітей із ТБ (рівень смертності), Результат лікування дітей із ТБ (рівень успішного лікування), Виявлені вагітні пацієнтки з ТБ, Діагностовані пацієнти з ТБ і діабетом серед протестованих, Пацієнти з ТБ із відомим статусом вживання алкоголю, Виявлені пацієнти з коінфекцією ТБ-COVID-19. В якості цільового поля слугувало Загальна кількість зареєстрованих випадків ТБ.

Методологія дослідження

Аналіз даних здійснено за допомогою методів МН для визначення ключових факторів, що впливають на поширення захворювання, та побудови прогнозних моделей. Методика дослідження включає кілька основних етапів: попередню обробку даних, аналіз взаємозв'язків між змінними, побудову й оцінку моделей машинного навчання.

1. *Попередня обробка даних.* На початковому етапі було здійснено завантаження та огляд набору даних. Проведено перевірку наявності пропущених значень, які за необхідності видалено або заповнено відповідними статистичними методами. Було виключено нерелевантні змінні, зокрема текстові атрибути, що не мали значення для подальшого аналізу. Для забезпечення коректної роботи алгоритмів МН виконано стандартизацію числових змінних.

2. *Аналіз взаємозв'язку даних.* Для визначення взаємозв'язків між змінними застосовано кореляційний аналіз із використанням теплової карти.

3. *Лінійна регресія.* Для побудови базової моделі прогнозування використано метод лінійної регресії. Модель навчено на основі навчальної вибірки і перевірено її узагальнюючу здатність тестовому наборі. Проведено порівняння точності з додатковим

перетворенням вихідних факторів методом головних компонент (PCA). Оцінку якості моделі здійснено за метриками середньоквадратичної помилки (RMSE) та коефіцієнта детермінації (R^2).

4. *Регуляризація.* Для запобігання перенавчанню та покращення узагальнюючої здатності моделей застосовано регуляризаційні методи Lasso та Ridge регресії. Оптимізацію гіперпараметра alpha для кожного методу виконано за допомогою GridSearchCV для досягнення найкращої точності прогнозування.

5. *Побудова та порівняння моделей ML.* З метою порівняння ефективності різних методів МН реалізовано такі моделі:

– Дерево рішень;
– Метод найближчих сусідів (KNN);
– Метод опорних векторів (SVM);
– Випадковий ліс.

Для кожної моделі виконано навчання і тестування, а також оцінено точність прогнозування за метриками RMSE та R^2.

6. *Порівняльний аналіз.* Підсумковий етап передбачав порівняльний аналіз отриманих моделей. Побудовано таблицю з результатами оцінки точності (RMSE, R^2) для всіх розглянутих моделей та візуалізовано їхнє порівняння у вигляді стовпчастого графіка. На основі отриманих даних зроблено висновки щодо ключових факторів, які найбільше впливають на прогноз поширення ТБ в Індії залежно від обраної моделі машинного навчання.

Результати дослідження

1. *Попередня обробка даних.* На першому етапі було виявлено, що деякі записи в набори даних мають неузгоджений формат даних. Зокрема, у наступних полях дані за 2019 р. представлені в процентах, а після 2019 р. – в абсолютних значеннях:

– результати лікування хворих на т ТБ, зареєстрованих у (% втрачено для подальшого спостереження);

– результати лікування хворих на ТБ, зареєстрованих у (рівень смертності);

– результати лікування хворих на ТБ, повідомлені в (success rate полях);

– результати лікування хворих на ТБ, зареєстрованих у (рівень невдач лікування);

– результати лікування пацієнтів з ТБ та ВІЛ, зареєстрованих у (рівень смертності);

– результати лікування пацієнтів з ТБ та ВІЛ, зареєстрованих у (показник успіху).

Для виправлення цієї невідповідності всі абсолютні значення було конвертовано у процентні.

Для забезпечення коректності аналізу було проведено відповідне завантаження даних з офіційних джерел та зведення всіх значень до єдиного типу. Після цього всі числові дані було стандартизовано для подальшої обробки:

$$\bar{x} = \frac{x - \mu}{\sigma},$$

де $\bar{x}$ – стандартизоване значення; x – вихідне значення; μ – середнє значення, ознаки; σ – стандартне відхилення ознаки.

2. *Аналіз взаємозв'язку даних*

На основі отриманого набору даних було побудовано теплову карту коефіцієнтів кореляції між факторами моделі (рис. 1).

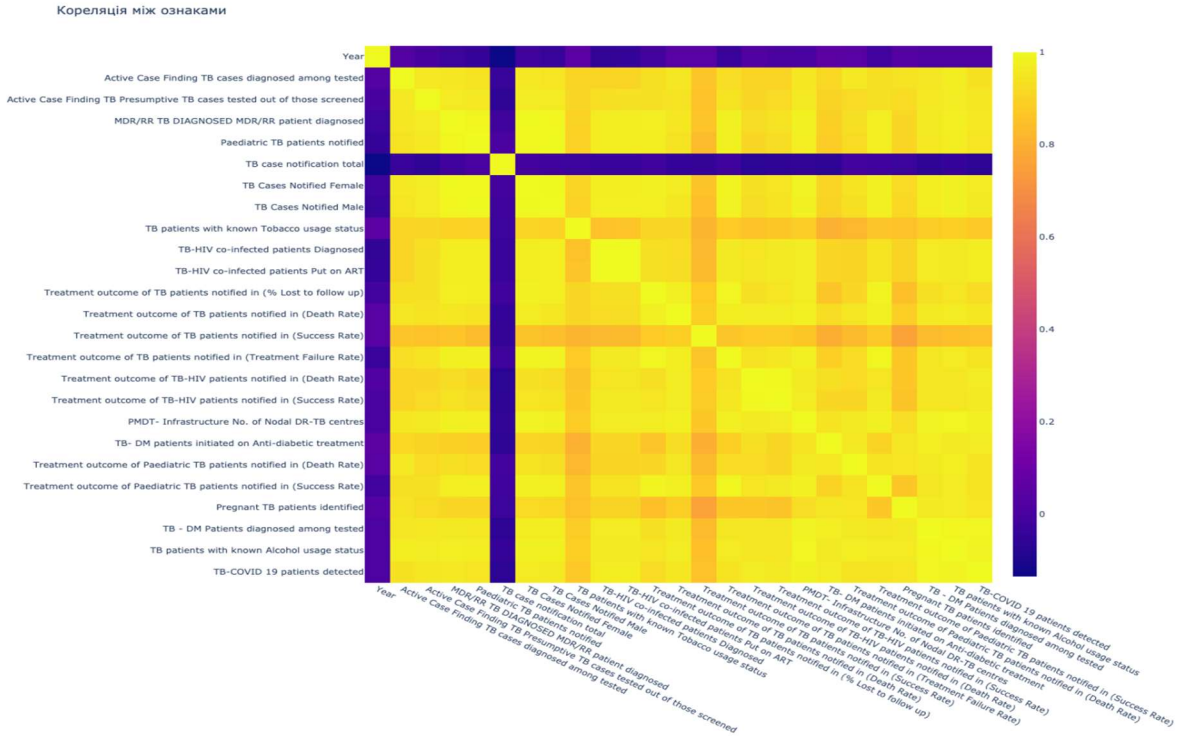

Рисунок 1 – Теплова карта кореляції між факторами набору даних

На основі побудованої теплової карти було виявлено, що між усіма досліджуваними змінними існує тісний взаємозв'язок. Це свідчить про високу залежність факторів один від одного, що може ускладнювати побудову прогнозних моделей. Однак аналіз кореляційних коефіцієнтів засвідчив, що жоден із досліджуваних факторів не має чіткої лінійної залежності з цільовою змінною *TB case notification total*. Це означає, що використання традиційних методів лінійного регресійного аналізу може бути недостатньо ефективним для прогнозування. Також з розрахунку видно, що колонка *рік* не корелює з жодним із вихідних болів. Це означає що в Індії відсутня часова динаміка в досліджуваних факторах.

З огляду на це, у подальших дослідженнях ознаки були перетворені методом головних компонент (PCA), який є більш стійким до мультиколінеарності та уможливлює усунути проблему взаємної кореляції даних. Цей підхід допомагає краще моделювати нелінійні взаємозв'язки між факторами та покращити точність прогнозування.

3. *Лінійна регресія*

Для побудови базової моделі прогнозування було використано метод лінійної регресії. Основна мета цього етапу полягала у визначенні здатності лінійної моделі точно прогнозувати цільову змінну *TB case notification total* та оцінити вплив використання методу головних компонент (PCA) на результати прогнозування. На першому етапі дані було розділено у співвідношенні 70:30 на навчальну та тестову вибірки. Перед розбиттям дані перемішувалися, щоб уникнути впливу впорядкованості записів. Далі було побудовано дві моделі лінійної регресії:

1. Лінійна регресія без попередніх перетворень – використовувались усі фактори у вихідному вигляді.

2. Лінійна регресія з використанням PCA – попередньо виконано зниження розмірності даних до двох, що дає змогу усунути кореляцію між ознаками.

Оцінку якості моделей здійснено за метриками середньоквадратичної помилки (RMSE) та коефіцієнта детермінації (R^2) на навчальній та тестовій вибірках. Результати наведено в табл. 1.

Таблиця 1 – ***Оцінка точності моделей***

Модель	RMSE train	RMSE test	R^2 train	R^2 test
Лінійна регресія	367	29000	0.96	-2.28
Лінійна регресія (PCA)	9406	8749	0.02	0.01

Додатково, на рис. 2 представлено графічне порівняння прогнозів, отриманих за допомогою обох моделей для навчальної та тестової вибірок.

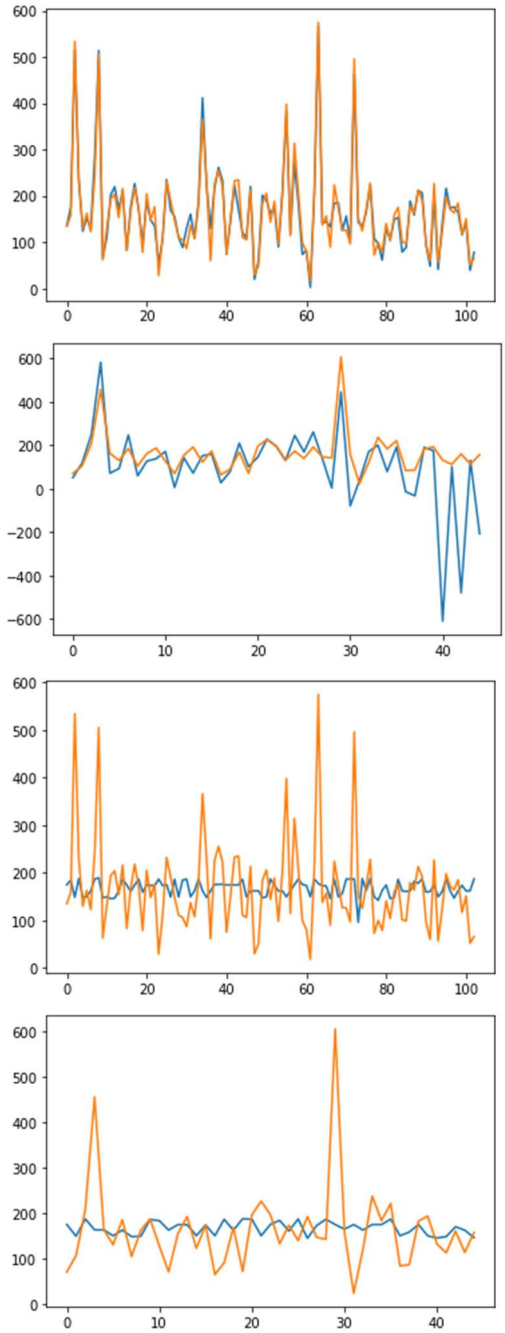

Рисунок 2 – Порівняльний прогноз лінійними моделями на навчальних та тестових наборах даних

Аналіз отриманих результатів засвідчує, що використання методу PCA суттєво погіршує точність прогнозування. Це пояснюється тим, що PCA, зменшуючи кількість змінних, втрачає значну частину інформації, необхідної для ефективного навчання моделі. Водночас звичайна лінійна регресія демонструє високу точність на навчальних даних, але значно гірші результати на тестових, що свідчить про перенавчання (overfitting) моделі.

4. *Регуляризація*

Для запобігання перенавчанню та покращення узагальнюючої здатності моделей у цьому дослідженні застосовано регуляризаційні методи Lasso та Ridge регресії. Регуляризація уможливлює зменшити складність моделі, накладаючи штраф на великі значення коефіцієнтів, що допомагає боротися з перенавчанням.

На початковому етапі для базових моделей параметр *alpha* було встановлено на рівні 0.9. Далі, для оптимізації цього гіперпараметра, застосовано метод GridSearchCV, у якому досліджувалося 100 значень *alpha* у діапазоні від 0 до 1 за логарифмічним розподілом.

Результати обчислень представлено в табл. 2.

Таблиця 2 – ***Оцінка точності моделей після регуляризації***

Модель	RMSE train	RMSE test	R^2 train	R^2 test
Ridge регресія	2070	5716	0.78	0.35
Lasso регресія	2722	4136	0.71	0.53
Оптимізована Lasso регресія	2889	4229	0.70	0.52
Оптимізована Ridge регресія	2162	5519	0.77	0.37

Як видно з табл. 2, використання регуляризації суттєво покращило узагальнюючу здатність моделей порівняно зі звичайною лінійною регресією.

- Ridge регресія значно зменшила помилку на тестових даних порівняно зі звичайною лінійною регресією, покращивши R^2 з -2.2812 до 0.3532.
- Lasso регресія показала ще кращий результат, отримавши $R^2 = 0.5320$ на тестових даних, що є найкращим показником серед розглянутих моделей.
- Оптимізовані варіанти моделей Ridge та Lasso дали незначне покращення порівняно з базовими регуляризованими моделями, що свідчить про стабільність вибраного підходу.

Отже, регуляризація дала змогу значно покращити узагальнюючу здатність лінійних моделей, а метод Lasso показав найкращий баланс між точністю на навчальних і тестових даних.

5. *Побудова та порівняння моделей ML*

З метою порівняння ефективності різних методів МН було протестовано вищезазначені моделі з такими значеннями гіперпараметрів:

- **Дерево рішень та випадковий ліс** реалізовано зі стандартними гіперпараметрами. Це означає, що модель дерева рішень розгалужувалась до досягнення повного узгодження з навчальними даними, а випадковий ліс використовував ансамбль дерев для підвищення стійкості до шуму в даних.
- Метод опорних векторів (SVM) було налаштовано з радіальним базисним ядром (RBF kernel) та параметром C = 10000. Радіальне базисне ядро дає змогу моделі захоплювати нелінійні залежності у даних, а високий параметр C змушує модель надавати більше значення правильному класифікуванню кожного прикладу, що може приводити до більш точних, але потенційно менш узагальнюючих моделей.
- Метод найближчих сусідів (KNN) використовував k = 5. Це означає, що для прогнозування значення для нового зразка враховувалися п'ять найближчих сусідів у просторі ознак.

Результати навчання та тестування точності наведено в табл. 3.

Таблиця 3 – ***Оцінка точності моделей***

Модель	RMSE train	RMSE test	R^2 train	R^2 test
Дерево рішень	0.00	4235	1.00	0.52
Випадковий ліс	897	2846	0.90	0.67
SVM	134	1868	0.98	0.78
KNN	1477	2354	0.84	0.73

Як видно з таблиці, найкращими моделями за точністю прогнозування є метод опорних векторів (SVM) та метод найближчих сусідів (KNN).

- SVM показав найнижче значення RMSE на тестових даних (1868.04) та найвище значення R^2 (0.7886), що свідчить про високу точність узагальнення.
- KNN також показав хороші результати, маючи RMSE = 2354.43 та $R^2 = 0.7336$.
- Випадковий ліс також продемонстрував високу ефективність, отримавши RMSE = 2846.79 та $R^2 = 0.6779$.
- Дерево рішень, хоча і навчилося ідеально на тренувальній вибірці (RMSE = 0), на тестовій вибірці показало помітно гіршу узагальнюючу здатність ($R^2 = 0.5207$), що свідчить про перенавчання.

Отже, найкращою моделлю у цьому дослідженні виявився метод опорних векторів (SVM), оскільки він показав найкращі результати як на навчальних, так і на тестових даних.

6. *Порівняльний аналіз*

Оскільки метод опорних векторів (SVM) показав найкращу узагальнюючу здатність серед базових моделей, його вибрано для подальшого уточнення гіперпараметра C.

Для цього застосовано RandomizedSearchCV – метод випадкового пошуку найкращого значення гіперпараметра. На відміну від GridSearchCV, де перевіряються всі можливі комбінації, Randomized-SearchCV випадково вибирає кілька значень, що значно зменшує час обчислень при збереженні високої ймовірності знайти оптимальний параметр.

Гіперпараметр C досліджувався в діапазоні від 0 до 10 000 за логарифмічним розподілом. Значення C контролює баланс між складністю моделі та її узагальнюючою здатністю. Високі значення змушують модель приділяти більше уваги правильній класифікації кожного окремого зразка, що може призводити до перенавчання. В результаті пошуку гіперпараметрів найкращу точність показала модель зі значенням C = 4982.

Фінальним етапом дослідження стало порівняння точності моделей та визначення ключових факторів, що впливають на поширення ТБ в Індії. Для цього було побудовано стовпчасту діаграму (рис. 3), яка відображає результати оцінки кожної моделі та зведені всі результати в підсумкову табл. 4.

Після оптимізації параметра C модель SVM продемонструвала найкращий результат серед усіх розглянутих методів:

- Найнижче значення RMSE на тестових даних (1401.26), що вказує на мінімальну середню помилку прогнозу.
- Найвище значення R^2 на тестових даних (0.8415), що підтверджує високу узагальнюючу здатність моделі.

Для порівняння:

- Базова версія SVM мала RMSE test = 1868.04 та R^2 test = 0.7886, що свідчить про суттєве покращення після підбору параметра C.
- Метод KNN (другий за точністю) мав RMSE test = 2354.43 та R^2 test = 0.7336, що є гіршим результатом порівняно з оптимізованим SVM.
- Випадковий ліс мав RMSE test = 2846.79 та R^2 test = 0.6779, що робить його третім за ефективністю методом.

Отже, оптимізована модель SVM з параметром C, знайденим через RandomizedSearchCV, показала найкращу здатність до узагальнення і була обрана як фінальна модель для аналізу ключових факторів, що впливають на поширення ТБ.

Таблиця 4 – ***Підсумкова таблиця точності моделей***

Модель	RMSE train	RMSE test	R^2 train	R^2 test
Лінійна регресія	367	29000	0.96	-2.28
Лінійна регресія (PCA)	9406	8749	0.02	0.01
Ridge регресія	2070	5716	0.78	0.35
Lasso регресія	2722	4136	0.71	0.53
Оптимізована Lasso регресія	2889	4229	0.70	0.52
Оптимізована Ridge регресія	2162	5519	0.77	0.37
Дерево рішень	0.00	4235	1.00	0.52
Випадковий ліс	897	2846	0.90	0.67
SVM	134	1868	0.98	0.78
KNN	1477	2354	0.84	0.73
Оптимізований SVM	307	**1401**	0.96	**0.84**

Рисунок 3 – Порівняльна діаграма точності моделей

Визначення найбільш вагомих факторів

Для визначення найважливіших факторів у моделі опорних векторів (SVM) було використано метод Permutation importance. Цей метод оцінює вплив кожної ознаки на точність моделі шляхом випадкової перестановки її значень і повторного обчислення метрики якості. Перевагою цього методу є те, що він не залежить від конкретної моделі і дає змогу оцінити вплив факторів навіть для складних алгоритмів, таких як SVM, які не мають вбудованих механізмів визначення важливості ознак. На відміну від аналізу коефіцієнтів у лінійних моделях, permutation_importance враховує нелінійні взаємозв'язки між ознаками. Він дає більш інтуїтивно зрозумілу оцінку важливості, оскільки безпосередньо показує, наскільки знижується точність після порушення структури даних.

Аналіз важливості ознак засвідчив, що найбільший вплив на прогноз поширення ТБ мають такі фактори:

1. Місто (City) – 1.3970.
2. Кількість зареєстрованих педіатричних пацієнтів з ТБ (Paediatric TB patients notified) – 0.7639.
3. Кількість зареєстрованих жінок з ТБ (TB Cases Notified Female) – 0.5395.
4. Рівень смертності серед зареєстрованих пацієнтів з ТБ (Treatment outcome – Death Rate) – 0.1806.
5. Рік (Year) – 0.1100.

Ці фактори мають найбільший внесок у прогнозування поширення хвороби, що вказує на необхідність їх особливого врахування під час планування заходів з боротьби з ТБ.

Менш значущі, але все ще важливі фактори:

- Інфраструктура для лікування лікарсько-стійкого ТБ (PMDT-Infrastructure) – 0.0883.
- Кількість діагностованих пацієнтів з мультирезистентним ТБ (MDR/RR TB diagnosed) – 0.0867.
- Статус споживання тютюну серед хворих на ТБ (TB patients with known Tobacco usage status) – 0.0713.
- Рівень успішного лікування серед дітей з ТБ (Paediatric TB success rate) – 0.0433.

Висновки

Проведено порівняльний аналіз різних методів машинного навчання для прогнозування поширення туберкульозу в Індії. Використано такі моделі: лінійна регресія, регуляризовані моделі Lasso та Ridge, метод опорних векторів (SVM), випадковий ліс, дерево рішень та метод найближчих сусідів (KNN). Виявлено, що найкращою моделлю є оптимізований SVM, який забезпечив найвищу точність прогнозування ($R^2 = 0.8415$, RMSE = 1401.26), що уможливлює ефективніше оцінювати фактори ризику поширення ТБ.

Проаналізовано вплив попередньої обробки даних та вибору методу на результати прогнозування. Використано метод головних компонент (PCA), кореляційний аналіз та стандартизацію даних. Виявлено, що застосування PCA не покращує, а навпаки, знижує точність прогнозування, оскільки втрачається частина значущої інформації. Водночас регуляризація моделей значно підвищує їхню узагальнюючу здатність, що підтверджено кращими показниками RMSE та R^2 у Lasso-регресії порівняно з базовими моделями.

Найбільший вплив на прогнозування мають географічне розташування (місто), гендерний розподіл хворих, рівень смертності та рівень дитячого ТБ. З медичної точки зору, географічне розташування зажди має значення на поширення ТБ, оскільки, доступність до діагностики і надання якісної допомоги, в різних регіонах значно відрізняється. Рівень смертності в регіоні вказує також на несвоєчасність або взагалі відсутність лікування. Щодо гендерного розподілу, чоловіки завжди складали групу ризику, оскільки вони значно частіше хворіють. Рівень дитячого ж ТБ високий у тих регіонах, які найменше приділяють уваги профілактиці у вигляді вакцинації дітей та недостатній діагностиці активних форм туберкульозу у дитячому віці.

Отримані результати можуть бути використані для оптимізації заходів боротьби з ТБ та розроблення цільових програм медичного втручання. Розуміючи фактори, які мають найбільший вплив на поширеність туберкульозної інфекції, можна створювати стратегію, яка є ефективною не глобально, а враховувати слабкість медичної системи в конкретному регіоні.

Список літератури

1. Batoure Bamana A., Shafiee Kamalabad M., Oberski D. L. A systematic literature review of time series methods applied to epidemic prediction *Informatics in Medicine Unlocked*, 50, art. no. 101571, 2024 DOI: 10.1016/j.imu.2024.101571.

2. Arisanti R., Pontoh R. S., Winarni S., Nurhasanah Y., Pertiwi A. P., Aini S. D. N. Integrating Generalized Linear Mixed Models with Extreme Neural Network: Enhancing Pulmonary Tuberculosis Risk Modeling in West Java, Indonesia *Communications in Mathematical Biology and Neuroscience*, 2024, art. no. 85, 2024 DOI: 10.28919/cmbn/8748.

3. D'Souza N. S., Wang H., Giovannini A., Foncubierta-Rodriguez A., Beck K. L., Boyko O., Syeda-Mahmood T. F. Fusing modalities by multiplexed graph neural networks for outcome prediction from medical data and beyond *Medical Image Analysis*, 93, art. no. 103064, 2024 DOI: 10.1016/j.media.2023.103064.

4. Zhang F., Zhang F., Li L., Pang Y. Clinical utilization of artificial intelligence in predicting therapeutic efficacy in pulmonary tuberculosis *Journal of Infection and Public Health*, 17 (4), pp. 632-641, 2024 DOI: 10.1016/j.jiph.2024.02.012.

5. Sun C., Fang R., Salemi M., Prosperi M., Magalis B. R. Deep Dyna Forecast: Phylogenetic-informed graph deep learning for epidemic transmission dynamic prediction *PLoS Computational Biology*, 20 (4), art. no. e1011351, 2024 DOI: 10.1371/journal.pcbi.1011351.

6. Yilmaz Y. Stacked ensemble modeling for improved tuberculosis treatment outcome prediction in pediatric cases *Concurrency and Computation: Practice and Experience*, 36 (13), art. no. e8089, 2024 DOI: 10.1002/cpe.8089.

7. Canas L. S., Dong T. H. K., Beasley D., Donovan J., Cleary J. O., et al. Computer-aided prognosis of tuberculous meningitis combining imaging and non-imaging data *Scientific Reports*, 14 (1), art. no. 17581, 2024 DOI: 10.1038/s41598-024-68308-8.

8. Abade A., Porto L. F., Scholze A. R., Kuntath D., Barros N. D. S., et al. A comparative analysis of classical and machine learning methods for forecasting TB/HIV co-infection *Scientific Reports*, 14 (1), art. no. 18991, 2024 DOI: 10.1038/s41598-024-69580-4.

9. Zhang Y., Ma H., Wang H., Xia Q., Wu S., et al. Forecasting the trend of tuberculosis incidence in Anhui Province based on machine learning optimization algorithm, 2013–2023 *BMC Pulmonary Medicine*, 24 (1), art. no. 536, 2024 DOI: 10.1186/s12890-024-03296-z.

10. Hamna Mariyam K B, Anuwat Jirawattanapanit, Sayooj Aby Jose, Karuna Mathew. A comprehensive study on tuberculosis prediction models: Integrating machine learning into epidemiological analysis *Journal of Theoretical Biology*, 597, art. no. 111988, 2025 DOI: 10.1016/j.jtbi.2024.111988.

11. Lane T. R., Urbina F., Rank L., Gerlach J., Riabova O., et al. Machine Learning Models for Mycobacterium tuberculosis in Vitro Activity: Prediction and Target Visualization *Molecular Pharmaceutics*, 19 (2), pp. 674–689, 2022 DOI: 10.1021/acs.molpharmaceut.1c00791.

Стаття надійшла до редколегії 24.02.2025

Hospodarchuk Dmytro
Student of the Department of Differential Equations and Mathematical Statistics,
https://orcid.org/ 0009-0003-9425-4609
Ivan Franko National University of Lviv, Lviv
Nevinskyi Denys
Associate Professor of the Department of Electronic Means of Information and Computer Technologies,
https://orcid.org/0000-0002-0962-072X
Institute of Telecommunications, Radio Electronics, and Electronic Engineering, Lviv Polytechnic National University, Lviv
Martjanov Dmytro
Postgraduate Student of the Department of Artificial Intelligence,
https://orcid.org/0009-0003-3919-4412
Institute of Computer Science and Information Technology, Lviv Polytechnic National University, Lviv
Vyklyuk Yaroslav
Professor of the Department of Artificial Intelligence,
https://orcid.org/0000-0003-4766-4659
Institute of Computer Science and Information Technology, Lviv Polytechnic National University, Lviv
Semianiv Ihor
Associate Professor of the Department of Phthisiology and Pulmonology,
https://orcid.org/0000-0003-0340-0766
Bukovinian State Medical University, Chernivtsi

OPTIMIZATION OF MACHINE LEARNING MODELS FOR ASSESSING THE RISK OF TUBERCULOSIS SPREAD

Abstract. *Tuberculosis (TB) remains one of the most pressing public health issues, especially in developing countries. The high incidence rate and the spread of multidrug-resistant strains of "Mycobacterium tuberculosis" pose significant challenges to modern medicine. India is one of the countries with the highest TB burden, making the optimization of disease spread prediction methods crucial for the effective implementation of prevention and treatment measures. The application of machine learning (ML) methods enables the automation of large-scale data analysis and the identification of key risk factors. This study aims to develop effective machine learning models for assessing the risk of TB spread in India based on socio-economic, demographic, and medical factors. A dataset containing 148 records from the period 2019–2022, categorized by Indian states, was used for analysis. Key variables included the number of detected TB cases, treatment success rates, mortality rates among patients, and the tobacco and alcohol consumption status of patients. The study involved data preprocessing, correlation analysis, and the application of machine learning methods. Several models were tested: linear regression, regularized models (Lasso and Ridge), support vector machine*

(SVM), k-nearest neighbors (KNN), random forest, and decision tree. The analysis showed that the best accuracy was achieved by the SVM model with optimized parameters, demonstrating the highest coefficient of determination and the lowest root mean square error. The comparison of other models revealed significant advantages of SVM over linear regression and decision trees, which exhibited low generalization capability. The most influential factors in predicting TB spread were determined using the Permutation Importance method. The most significant factors included geographic location (state), the number of registered TB cases among children, the number of women with TB, the mortality rate among patients, and the infrastructure available for treating drug-resistant TB. It was also found that social factors, such as tobacco and alcohol consumption among patients, influence the disease spread, although their contribution is less significant. The study confirmed the effectiveness of applying machine learning methods to predict tuberculosis spread. The optimized SVM model provided the best accuracy and generalization capability. Factor importance analysis revealed that regional characteristics, demographic indicators, and mortality rates have the greatest impact on disease spread. The obtained results can be used to improve TB control strategies, particularly through targeted interventions in high-risk regions. The use of ML methods enhances disease control efficiency, which is an essential step in the global fight against tuberculosis.

Keywords: tuberculosis; machine learning; prediction; SVM model; influencing factors; artificial intelligence; regression analysis

References

1. Batoure Bamana, A., Shafiee Kamalabad, M., & Oberski, D. L. (2024). A systematic literature review of time series methods applied to epidemic prediction. *Informatics in Medicine Unlocked,* 50, 101571. https://doi.org/10.1016/j.imu.2024.101571.

2. Arisanti, R., Pontoh, R. S., Winarni, S., Nurhasanah, Y., Pertiwi, A. P., & Aini, S. D. N. (2024). Integrating generalized linear mixed models with extreme neural network: Enhancing pulmonary tuberculosis risk modeling in West Java, Indonesia. *Communications in Mathematical Biology and Neuroscience, 2024*, 85. https://doi.org/10.28919/cmbn/8748.

3. D'Souza, N. S., Wang, H., Giovannini, A., Foncubierta-Rodriguez, A., Beck, K. L., Boyko, O., & Syeda-Mahmood, T. F. (2024). Fusing modalities by multiplexed graph neural networks for outcome prediction from medical data and beyond. *Medical Image Analysis, 93*, 103064. https://doi.org/10.1016/j.media.2023.103064.

4. Zhang, F., Zhang, F., Li, L., & Pang, Y. (2024). Clinical utilization of artificial intelligence in predicting therapeutic efficacy in pulmonary tuberculosis. *Journal of Infection and Public Health,* 17 (4), 632–641. https://doi.org/10.1016/j.jiph.2024.02.012.

5. Sun, C., Fang, R., Salemi, M., Prosperi, M., & Magalis, B. R. (2024). DeepDynaForecast: Phylogenetic-informed graph deep learning for epidemic transmission dynamic prediction. *PLoS Computational Biology,* 20 (4), e1011351. https://doi.org/10.1371/journal.pcbi.1011351.

6. Yilmaz, Y. (2024). Stacked ensemble modeling for improved tuberculosis treatment outcome prediction in pediatric cases. *Concurrency and Computation: Practice and Experience, 36*(13), e8089. https://doi.org/10.1002/cpe.8089.

7. Canas, L. S., Dong, T. H. K., Beasley, D., Donovan, J., Cleary, J. O., et al. (2024). Computer-aided prognosis of tuberculous meningitis combining imaging and non-imaging data. *Scientific Reports,* 14 (1), 17581. https://doi.org/10.1038/s41598-024-68308-8.

8. Abade, A., Porto, L. F., Scholze, A. R., Kuntath, D., Barros, N. D. S., et al. (2024). A comparative analysis of classical and machine learning methods for forecasting TB/HIV co-infection. *Scientific Reports,* 14 (1), 18991. https://doi.org/10.1038/s41598-024-69580-4.

9. Zhang, Y., Ma, H., Wang, H., Xia, Q., Wu, S., et al. (2024). Forecasting the trend of tuberculosis incidence in Anhui Province based on machine learning optimization algorithm, 2013–2023. *BMC Pulmonary Medicine,* 24 (1), 536. https://doi.org/10.1186/s12890-024-03296-z.

10. Hamna Mariyam K B, Anuwat Jirawattanapanit, Sayooj Aby Jose, Karuna Mathew. A comprehensive study on tuberculosis prediction models: Integrating machine learning into epidemiological analysis *Journal of Theoretical Biology*, 597, art. no. 111988, 2025 DOI: 10.1016/j.jtbi.2024.111988.

11. Lane, T. R., Urbina, F., Rank, L., Gerlach, J., Riabova, O., et al. (2022). Machine learning models for *Mycobacterium tuberculosis* in vitro activity: Prediction and target visualization. *Molecular Pharmaceutics,* 19 (2), 674–689. https://doi.org/10.1021/acs.molpharmaceut.1c00791.

Посилання на публікацію

APA Hospodarchuk D., Nevinskyi D., Martjanov D., Vyklyuk Ya., Semianiv I. (2025). Optimization of machine learning models for assessing the risk of tuberculosis spread. *Management of Development of Complex Systems*, 61, 160–169, dx.doi.org\10.32347/2412-9933.2025.61.160-169.

ДСТУ Господарчук Д. В., Невінський Д. В., Мартьянов Д. І., Виклюк Я. І., Сем'янів І. О. Оптимізація моделей машинного навчання для оцінки ризику поширення туберкульозу. *Управління розвитком складних систем.* Київ, 2025. № 61. С. 160 – 169, dx.doi.org\10.32347/2412-9933.2025.61.160-169.