



В. В. Веселяк, Ю. І. Грицюк

Національний університет "Львівська політехніка", м. Львів, Україна

МЕТОДИ МАШИННОГО НАВЧАННЯ В ЕПІДЕМІОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

Досліджено особливості застосування методів машинного навчання для аналізу, передбачення та контролю епідеміологічного стану, виявлення тенденцій та моделей захворювань. Проведено дослідження та здійснено аналіз різноманітних методів машинного навчання для систематизації та порівняння епідеміологічних даних. Починаючи з класичних підходів, таких як лінійна регресія та дерева рішень, розглянуто їхні переваги та обмеження. Також розглянуто та запропоновано використання методів глибокого навчання, таких як згорткові нейронні мережі та рекурентні нейронні мережі, для ефективного аналізу складних взаємозв'язків у епідеміологічних даних. Детально досліджено та наведено переваги та обмеження кожного методу та їхні потенційні застосування для прогнозування епідеміологічної ситуації та виявлення ризикових чинників захворювань. Наведено порівняння точності різних методів на прикладі декількох вірусних захворювань. Проаналізовано наявні проблеми та обмеження у використанні машинного навчання у прогнозуванні інфекційних захворювань. Було проаналізовано також дослідження у разі поєднання декількох методів і їх результати. Розглянуто вимоги до вхідних даних і показники, які є важливими під час вибору вхідних даних для методів машинного навчання. Розглянуто останні високоякісні дослідження, де було застосовано методи машинного навчання до даних реального світу та оцінено ефективність цих моделей у прогнозуванні поширення та початку інфекційних захворювань. Розглянуто дослідження на вхідних даних різного характеру, отриманих із різних джерел та зібраних у різних регіонах світу щодо декількох інфекційних захворювань. Проаналізовано ефективність методів виявлення залежності між тенденціями захворювання та географічним розповсюдженням, а також враховано сезонність, особливості місцевості та інші відмінності. Наведено висновки щодо наявних можливостей та перспектив використання методів машинного навчання для аналізу та прогнозування епідеміологічних даних, а також розроблено список критеріїв, за якими можна оцінити доцільність використання методу машинного навчання для оцінювання та прогнозування епідеміологічного стану.

Ключові слова: модель машинного навчання; пандемія COVID-19; штучний інтелект; прогнозування.

Вступ / Introduction

Байдужість керівництва країни до розвитку нових інфекційних захворювань становлять зростаючу загрозу для її громадян. Адже багато епідемічних захворювань у світі (особливо тих, що передаються через посередників) відомі своєю високою чутливістю до довготермінових змін клімату та короткотермінових перепадів погоди. Застосування екологічних даних у дослідженні захворювань населення надає можливість виявлення векторно-екологічних взаємозв'язків і потенційного прогнозування ризику спалахів хвороб чи епідемій. Точні моделі прогнозування захворювань населення відчутно підвищують можливості запровадження заходів профілактики та контролю епідеміологічних ситуацій.

Епідеміологічні дані, які описують поширення захворювань серед населення країни, виконують ключову роль у прогнозуванні, управлінні та контролі епідемій. Зростаючий обсяг даних, їхнє різноманіття та складність точного їх отримання створюють потребу розроблення ефективних методів їх аналізу та оброблення. Сучасний стан методів машинного навчання відкриває нові перспективи для розуміння епідеміологічних проце-

сів і виявлення закономірностей у великих обсягах даних. Дослідники з різних країн розглядають не тільки класичні методи машинного навчання, такі як лінійна регресія, дерева рішень та часові ряди, а й новітні технології глибокого навчання, що набувають дедалі більшого значення в цій сфері.

Попередні дослідження показали потенціал методів машинного навчання у виявленні ризикових груп населення, прогнозуванні поширення хвороб та виявленні нових епідеміологічних трендів. Проте, їхнє успішне застосування потребує додаткового вивчення особливостей появи епідеміологічних даних, вибору відповідних алгоритмів, здатних навчатися на цих даних і узагальнюватися на небачені дані, потім виконувати завдання без явних настанов, а також здійснювати правильну інтерпретацію отриманих результатів. Висновки їхнього аналізу можуть виявитися корисними для дослідників, які працюють у сфері епідеміології та медичного аналізу даних, а також для фахівців, що цікавляться застосуванням методів машинного навчання для вирішення глобальних проблем здоров'я населення.

Об'єкт дослідження – машинне навчання для аналі-

Інформація про авторів:

Веселяк Валерій Валерійович, магістр, аспірант, кафедра програмного забезпечення. Email: valerii.v.veseliak@lpnu.ua

Грицюк Юрій Іванович, д-р техн. наук, професор, кафедра програмного забезпечення.

Email: yurii.i.hrytsiuk@lpnu.ua; <https://orcid.org/0000-0001-8183-3466>

Цитування за ДСТУ: Веселяк В. В., Грицюк Ю. І. Методи машинного навчання в епідеміологічних дослідженнях. Науковий вісник НЛТУ України. 2024, т. 34, № 4. С. 59–67.

Citation APA: Veseliak, V. V., & Hrytsiuk, Yu. I. (2024). Machine learning methods in epidemiological research. *Scientific Bulletin of UNFU*, 34(4), 59–67. <https://doi.org/10.36930/40340408>

зу епідеміологічних даних.

Предмет дослідження – методи побудови моделей, здатних навчатися на даних епідеміологічних досліджень, що дасть змогу отримувати точні прогнози епідеміологічної ситуації на наявних даних.

Мета роботи – проаналізувати методи машинного навчання, придатні для проведення епідеміологічних досліджень, що дасть змогу класифікувати набори даних на підставі розроблених моделей, а також передбачати майбутні стани епідеміологічних ситуацій з використанням цих моделей.

Для досягнення зазначеної мети визначено такі основні завдання дослідження:

1. Провести аналіз класичних методів машинного навчання, їхніх переваг та обмежень для аналізу епідеміологічних даних, що дасть змогу обрати придатну модель та її конфігурації для досягнення максимальної точності прогнозування епідеміологічної ситуації.
2. Розглянути новітні технології глибокого навчання та їхній потенціал у вирішенні епідеміологічних завдань, що дасть змогу використовувати нейронні мережі як більш точний метод оцінювання та прогнозування епідеміологічної ситуації.
3. Дослідити методи виявлення епідеміологічних груп ризику, прогнозування поширення хвороб та виявлення нових епідеміологічних трендів з використанням методів машинного навчання, що забезпечить швидке реагування та оповіщення у випадку епідемії та пандемії.
4. Проаналізувати особливості даних, які використовують в епідеміологічних дослідженнях, та розглянути методи їхнього оброблення та підготовки для подальшого аналізу, які допоможуть дійти висновків щодо поширення захворювань, виявити зв'язки між різними чинниками та ризиком виникнення захворювань, а також спрогнозувати можливий розвиток ситуації і розробити ефективні стратегії протидії епідеміям і пандеміям.
5. Визначити потенційні обмеження та виклики, пов'язані з застосуванням методів машинного навчання в епідеміологічних дослідженнях і запропонувати шляхи їх подолання, які уможливають краще використання цих методів для аналізу великих обсягів даних, виявлення складних зв'язків між різними чинниками та здійснити прогнозування розвитку епідемічних ситуацій з високою точністю.

Аналіз останніх досліджень та публікацій. Відповідно до Глобального звіту про профілактику та контроль інфекцій від ВООЗ (Всесвітня організація охорони здоров'я) [38], наразі інфекційні захворювання становлять дедалі більшу частину глобальної проблеми охорони здоров'я: щороку мільйони смертей спричинені інфекційними захворюваннями. Для того, щоб розробити заходи щодо захисту від інфекційних захворювань, важливо визначити чинники поширення захворювання для застосування заходів контролю та профілактики [25]. Виявлення чинників поширення хвороби може дозволити робити прогнози, корисні для відповідних органів влади щодо прийняття відповідних рішень, наприклад, щодо закупівлі вакцин, кампаній з підвищення обізнаності громадськості та навчальних програм для медичних працівників [11, 29].

Алгоритми машинного навчання можуть сприяти контролю над інфекційними захворюваннями, допомагаючи як просторово, так і тимчасово прогнозувати еволюцію та поширення інфекційних захворювань [36]. Алгоритми машинного навчання вміють аналізувати об'ємні та складні набори даних, виявляючи закономірності та тенденції, які може бути складно виявити лю-

дям. Це робить їх вельми корисними для прогнозування інфекційних захворювань, які часто містять численні чинники, такі як демографічний склад населення, умови навколишнього середовища та індивідуальна поведінка. Останнім часом було проведено багато досліджень, де застосовували методи машинного навчання для прогнозування інфекційних захворювань, і отримані результати виявилися дуже перспективними. Однією з ключових проблем під час використання машинного навчання для прогнозування захворювань є необхідність наявності якісних і повних даних.

Системи нагляду за інфекційними захворюваннями часто збирають дані про різні чинники, такі як кількість зареєстрованих випадків, місця спалахів і демографічні дані інфікованих осіб. Однак, ці дані часто є неповними, упередженими або містять шум, що може вплинути на ефективність моделей машинного навчання. Окрім цього, багато інфекційних захворювань мають тривалий інкубаційний період, через що дані про минулі спалахи можуть не точно відображати поточні умови. Для подолання цих проблем дослідники використовують різноманітні алгоритми машинного навчання, в т.ч. дерева рішень, випадкові ліси, опорні векторні машини та глибокі нейронні мережі. Ці алгоритми застосовують до різноманітних джерел даних, таких як електронні медичні записи, геномні дані та публікації в соціальних мережах. Узагальнені результати досліджень показують, що алгоритми машинного навчання можуть ефективно прогнозувати поширення та початок інфекційних захворювань, причому їхню ефективність можна порівняти або вона навіть буде краща за традиційні статистичні методи. Отже, алгоритми машинного навчання мають великий потенціал у прогнозуванні початку інфекційних захворювань та їх поширення [28]. Наприклад, у деяких дослідженнях використовували машинне навчання для прогнозування кількості випадків конкретного захворювання в певному регіоні на підставі історичних даних і поточних умов [23].

Відповідно до результатів цього дослідження, реалізація прогнозу динаміки епідемічного захворювання щодо ризику інфікування можна досягнути за допомогою багатофакторної логістичної регресійної моделі на підставі модифікованого SEIR (англ. *Susceptible – Exposed – Infected – Recovered* – сприйнятливий – виявлені – заражені – видалені). Інші використовували машинне навчання для виявлення найбільш вірогідних джерел спалаху на підставі генетичного складу збудника та моделей інфекції [10, 27]. Аналізуючи результати, можна зрозуміти, що було досягнуто значних удосконалень для модельно-специфічних і модельно-агностичних підходів для інтерпретації моделі. Однак ці методи часто не можуть врахувати прихованих гетерогенних ефектів повного простору ознак, що може знизити точність моделі та ввести в оману дослідників залежно від вибору алгоритму. Ще інші використовували машинне навчання, щоб передбачити ймовірність зараження індивіда інфекційним захворюванням на підставі його особистих характеристик і поведінки [4].

Загалом використання машинного навчання для прогнозування інфекційних захворювань є перспективним напрямом досліджень із потенційним застосуванням у сфері охорони здоров'я, епідеміології та клінічної практики. Однак існують також значні проблеми та обмеження щодо використання машинного навчання в

цьому контексті, в т.ч. потребу у високоякісних даних, складність базових явищ, а також потенціал упередженості та надмірного оснащення [26].

Незважаючи на багатообіцяючі результати цих досліджень, існують також значні проблеми та обмеження щодо використання машинного навчання для прогнозування інфекційних захворювань. Однією з основних проблем є потреба у високоякісних даних, які часто важко отримати через такі проблеми, як відсутні значення, неповні записи та різні формати даних. Окрім цього, складність основних явищ, таких як передача інфекційних захворювань, може ускладнити розроблення точних моделей.

Матеріали та методи дослідження. У дослідженні було використано низку методів системного аналізу даних. А саме документальний аналіз з метою вивчення наукових досліджень та практичних розроблень, наукових публікацій та відгуків, які можуть сприяти отриманню повнішого уявлення про ризики та виклики, які виникають у процесі вибору методів машинного та глибокого навчання для аналізу епідеміологічних даних і їх впровадження.

Результати дослідження та їх обговорення / Research results and their discussion

Машинне навчання виявляє перспективи для розроблення моделей прогнозування та клінічних правил прогнозування, які можуть покращити надання медичної допомоги [25, 38]. Прихильники використання машинного навчання в охороні здоров'я стверджують, що ці системи покращують прогностичну точність, тому їх можна адаптувати для користі пацієнтів за рахунок більш точної діагностики, прогнозу та персоналізованої терапії [11, 29].

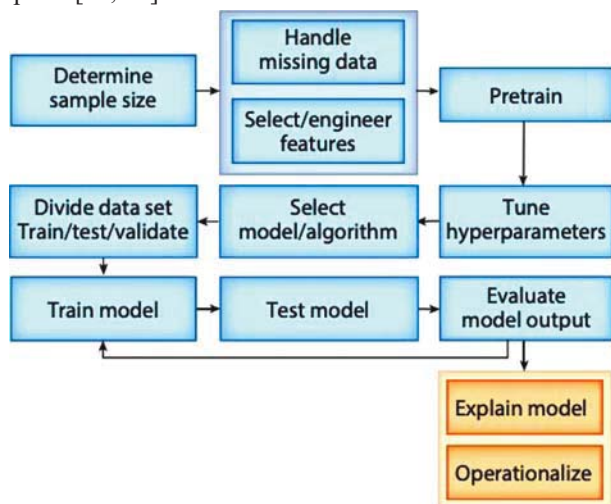


Рисунок. Схема розгортання і тренування моделей машинного навчання / Deployment and training of machine learning models [30]

Незважаючи на потенційні переваги машинного навчання, як можна побачити на рисунку, навчання та використання моделі машинного навчання є доволі складними в імплементації та інтерпретації та мають кілька потенційних варіантів неправильного використання [30]. По-перше, ці методи часто важко впроваджувати, вимагаючи багатоступеневих методів оброблення даних, а також створення тренувальних і тестових наборів. По-друге, ці методи часто важко інтерпретувати [23, 27]. Їх також називають як "чорна скринька", в якій

ці алгоритми можуть генерувати прогноз, але надають мало інформації щодо того, як було досягнуто результату [10]. Отже, дослідники не можуть знати, чи допоміг певний сигнал від важливої коваріанти у прогнозуванні, чи демографічні чинники призвели до прогнозування і відтворюють попередні упередження системи охорони здоров'я [4]. По-третє, оцінювання цих алгоритмів не має традиційного виводу "рівня значущості" для кожної коваріанти, що було загальним у більшості методів, які використовують в медичній літературі (наприклад, лінійна регресія, аналіз виживання, логістична регресія) [26, 28].

Методи машинного навчання. У роботі [24] було запропоновано методологію, повністю базовану на машинному навчанні, яка надає прогнози в реальному часі щодо активності грипу в США, збираючи дані з різних джерел, як клінічних, так і не клінічних. Створена система здатна прогнозувати наступні чотири тижні відповідно до звітів CDC (англ. *Centers for Disease Control and Prevention*, центри з контролю та профілактики захворювань у США) про ГРВІ. Автори роботи [5] провели аналіз та оцінку того, як модель машинного навчання LASSO (англ. *Least Absolute Shrinkage and Selection Operator*, метод оцінювання коефіцієнтів лінійної регресійної моделі) може допомогти у наданні корисних результатів прогнозування для введення різних патогенів у районах з різним кліматом. Було створено набір моделей LASSO для різних захворювань, різних країн та різних вікон прогнозування зі змінною складністю моделі.

Прогнозні моделі були майже точними у прогнозуванні спалаху, але були нечутливими під час захоплення розміру спалаху. У роботі [2] вивчили різні моделі, які використовують для прогнозування вибуху епідемії, спричиненого Гарячкою Західного Нілу. У дослідженні описано використання різних полігамних моделей в епідеміології для визначення поширення Гарячки Західного Нілу. Було визначено, що класифікатор випадкового лісу RFC (англ. *Random Forest Classifier*) показав кращі результати для прогнозування. Модель SIS (англ. *Susceptible Infected Susceptible*) стала найкращою моделлю поширення. Автори роботи [17] склали модель прогнозування для передбачення спалаху Африканської чуми свиней ASF (англ. *Advanced Streaming Format* – розширений формат потокового передавання) з використанням даних про спалахи ASF та метеорологічних даних. Для побудови моделі було використано алгоритми Random Forest. Модель досягла діапазону точності від 76,02 % до 84,64 %. Робота не містить всіх даних про місцевість, а використання інших технік добору ознак та технік глибокого навчання могло би забезпечити кращі результати. У роботі [16] показали, як моделі AR-MA (англ. *Autoregressive Moving-Average Model*, модель авторегресії – ковзного середнього) може бути використано для прогнозування поширення Хвороби Ньюкасла протягом одного місяця у провінції Китаю. Модель AR-MA з нульовим авторегресійним порядком та першим порядком ковзного середнього змогла забезпечити точні та надійні прогнози спалахів Ньюкасли. Параметри, необхідні для прогнозування, були оцінені за допомогою SPSS (англ. *Statistical Package for the Social Sciences* – статистичний пакет для соціальних наук). Щотижневі або щоденні дані могли би забезпечити кращі результати порівняно з щомісячними даними [24].

Системи охорони здоров'я по всьому світу перевантажені внаслідок швидкого зростання кількості пацієнтів із COVID-19. Для раннього втручання та оптимального розподілу ресурсів потрібна точна модель прогнозування появи пандемії. На відміну від роботи [22], де оцінювали кілька моделей окремо, у роботі [9] автори інтегрували чотири різні моделі машинного навчання в одну ансамблевую модель для прогнозування ризику смертності госпіталізованих пацієнтів з COVID-19.

Для розроблення ансамблевої моделі автори спочатку вибрали 14 важливих змінних (перебування у свідомості, стать, вік, наявність мокрот, азот креатиніну у крові, частота дихання, D-димер, кількість супутніх захворювань, кількість тромбоцитів, лихоманка, альбумін, насиченість киснем крові, лімфоцити та хронічна хвороба нирок) зі стартового набору 53-х змінних за допомогою методу ласо, ще одного методу машинного навчання, часто використовуваного для видалення неважливих змінних. З використанням цих 14 змінних автори розробили 6 моделей прогнозування методами машинного навчання (логістична регресія, метод опорних векторів, градієнтний бустінг, нейронна мережа, k -найближчий сусід та випадковий ліс), а потім інтегрували чотири найкращі прогнозные моделі (логістична регресія, метод опорних векторів, градієнтний бустінг та нейронна мережа) в одну. Ансамблева модель досягла площі під кривою (AUC) 0,96 і 0,92 для прогнозування смертності від COVID-19 у двох зовнішніх когортах. Автори дійшли висновку, що ця модель ефективна, вона дає змогу точно стратифікувати ризики у пацієнтів з COVID-19 у разі госпіталізації [22].

Методи глибокого навчання. У роботі [34] використовували стекову архітектуру для прогнозування кількості випадків малярії, оскільки традиційні моделі часових рядів неефективні для охоплення кожної окремої деталі в структурі даних. Вони використовували стекову архітектуру, яка може поєднувати різні алгоритми та моделі. Мережеві моделі ARIMA (англ. *Autoregressive Integrated Moving Average*, модель методу Бокса-Дженкінса), BP-ANN (англ. *Back Propagation Artificial Neural Network*, BP – модель штучної нейронної мережі), STL+ARIMA (англ. *Season (AI), Trend, And Residual*, LOESS – *Locally Estimated Scatterplot Smoothing*) та LSTM (англ. *Long short-term memory*, довгий ланцюг елементів короткотермінової пам'яті) були окремо реалізовані на малярії та метеорологічних даних. Ефективність порівнювали за допомогою таких показників оцінювання, як RMSE (англ. *The Root Mean Squared Error*, середньоквадратична помилка), MAD (англ. *Mean Absolute Deviation*, медіана абсолютних відхилень) і MASE (англ. *Mean Absolute Scaled Error*, середня абсолютна масштабована помилка). Значення RMSE, MAD і MASE зменшилися після використання стекової архітектури [34].

Автори роботи [5] вперше використали методи глибокого навчання для прогнозування в реальному часі. Вони використовували RNN (англ. *Recurrent Neural Network*, рекурентна нейронна мережа), щоб зафіксувати довготермінову кореляцію в даних, і CNN (англ. *Convolutional Neural Network*, згорткові нейронні мережі), щоб об'єднати інформацію з багатьох різних джерел. Їхні методи забезпечили кращі результати, ніж звичайні базові авторегресійні моделі та моделі процесів Гаусса [5]. Тапак та ін. (2019) досліджували точність,

яку забезпечують часові ряди Random Forest, опорні векторні машини та моделі штучних нейронних мереж у моделюванні грипоподібних захворювань та виявленні спалахів захворювань. Для створення прогнозу використовували набір даних щотижневих спалахів грипоподібних захворювань в Ірані. Внаслідок було зроблено висновок, що модель часових рядів Random Forest перевершила інші три моделі в моделюванні даних.

Окрім цього, виявилося, що модель нейронної мережі була кращою у виявленні спалахів захворювань [39]. Соліман та ін. (2019) вивчали тенденції сезонного грипу в окрузі Даллас. Робота спрямована на оцінку спроможності прогнозування різних моделей глибокого навчання з використанням FNN (англ. *Feedforward Neural Network*, нейронна мережа прямого розповсюдження) та оцінку ефективності звичайних статистичних моделей. Ця робота містить розроблення ймовірнісної моделі прогнозування на підставі інформації про захворюваність на грип в країні шляхом об'єднання всіх моделей за допомогою BMA (англ. *Bayesian Model Averaging*, модель усереднення Байєса). Результати свідчать про те, що багатомодельний ансамбль прогнозів спалахів грипоподібних захворювань на підставі глибокого навчання і BMA дає майже однакову продуктивність, яка перевершує всі інші протестовані моделі [32].

У роботі [31] було розроблено моделі прогнозування на підставі даних про смертність та історичну захворюваність на інфекційні захворювання. Вони також інтегрували дані запитів пошукових систем і сезонну інформацію для кращих результатів. Для цілей прогнозування було створено різні моделі, які використовували для прогнозування захворюваності на десятьох інфекційних захворюваннях, у т.ч. лінійну модель, модель градієнтного бустінгу, аналіз часових рядів і моделі глибокого навчання (RNN). Було виявлено, що модель RNN перевершує всі інші моделі прогнозування [31]. Автори роботи [21] передбачили рівень ризику спалаху лихоманки Денге за допомогою ELM (англ. *Extreme Learning Machines*, машини екстремального навчання). Встановлено, що поширення лихоманки Денге пов'язане з кліматичними умовами регіону. Модель ELM використовує погодні дані як вхідні, а рівень ризику спалаху лихоманки – як цільовий. Модель ELM із 50 прихованими нейронами та бінарною сигмоїдною функцією активації працює найкраще порівняно з іншими використовуваними моделями ELM [19]. У роботі [20] автори використовували нелінійні методи, такі як ARX (англ. *Autoregressive Exogenous Input*, авторегресійна екзогенна модель), ARMAX (англ. *Autoregressive Moving Average with Exogenous Input*, авторегресійне плаваюче середнє з екзогенними вхідними даними), Deep Multilayer Perceptron і згорткову нейронну мережу (CNN), щоб передбачити захворюваність на грип на підставі даних соціальних мереж, таких як Twitter. Продукт твітів разом з даними CDC та твітів разом з даними Google також було внесено як нові функції. Глибока нейронна мережа, навчена новим доданим функціям, показала кращі результати порівняно з іншими методами. Модель зменшила середню помилку (MAE) аж до 25 % [20].

Тюрбе та ін. розробили модель машинного навчання для визначення того, чи показники результатів позитивні чи негативні на підставі фотографій швидких діагностичних тестів на ВІЛ (вірус імунodefіциту людини). Всього було використано 11 374 зображень, зроблених

за допомогою планшетів, які були позначені трьома експертами зі швидких діагностичних тестів, і потім використані як навчальний набір даних. Автори розробили чотири моделі, використовуючи цей набір даних; одна з них – це метод опорних векторів (SVM), а три – це різні моделі згорткових нейронних мереж. Одну з моделей згорткових нейронних мереж використано через її найкращу продуктивність з погляду чутливості та специфічності. Як пілотне випробування, ефективність моделі порівнювали з ефективністю п'ятихостових користувачів з різним рівнем підготовки (дві медсестри та три нових навчених робітники здоров'я спільноти). У візуальному тлумаченні швидкого діагностичного тестування рівні згоди остаточно користувачів були від 61 до 100 %. Модель машинного навчання продемонструвала кращу ефективність, ніж остаточно користувачі, за такими чотирма показниками: чутливість (95,6 % проти 97,8 %), специфічність (89 % проти 100 %), позитивне передбачене значення (88,7 % проти 100 %) та негативне передбачене значення (95,7 % проти 98 %). Автори дійшли висновку, що зображення швидкого діагностичного тесту, зроблені за допомогою мобільного пристрою, можуть стандартизувати тлумачення результатів тестів, зменшити помилки тлумачення та забезпечити платформу для навчання персоналу [33].

Методи часових рядів. Прогнозування часових рядів (англ. *Time Series Forecasting*) – це метод, за допомогою якого є можливість робити прогнози на підставі часово відзначених історичних даних. Історичні дані аналізують і отримані з них спостереження використовують для прийняття рішень. Два найбільш поширені методи часових рядів – це ARIMA та LSTM (глибоке навчання). Далі буде коротко охарактеризовано моделі ARIMA та LSTM.

Таблиця. Точність моделей передбачення грипу, малярії та гепатиту /
Accuracy of Prediction Models for Influenza, Malaria, and Hepatitis [3]

Influenza		Malaria		Hepatitis	
Model	Accuracy	Model	Accuracy	Model	Accuracy
Logistic regression	87.69	Logistic regression	89.23	Logistic regression	53.84
Naive Bayes	73.84	Naive Bayes	89.23	Naive Bayes	58.46
KNN	86.15	KNN	89.23	KNN	52.30
SVM	75.38	SVM	89.23	SVM	58.46
Decision Tree	90.01 (overfitting)	Decision Tree	84.61 (overfitting)	Decision Tree	53.84 (overfitting)
Bagging	89.23	Bagging	89.23	Bagging	53.84
Boosting D. Tree Clf.	90.76 (overfitting)	Boosting D. Tree Clf.	83.07 (overfitting)	Boosting D. Tree Clf.	52.30
Random Forest	89.23	Random Forest	84.61	Random Forest	55.76
Voting Clf.	87.69	Voting Clf.	89.23	Voting Clf.	53.84
Neural Network	84.6	Neural Network	83.7	Neural Network	55.4

Нарадимо, ARIMA (англ. Autoregressive Integrated Moving Average, іноді модель Бокса-Дженкінса) – інтегрована модель авторегресії (ковзного середнього) модель і методологія аналізу часових рядів. Є розширенням моделей ARMA для нестационарних тимчасових рядів, які можна зробити стаціонарними взяттям різниць деякого порядку від вихідного тимчасового ряду (так звані інтегровані або різностно-стаціонарні тимчасові ряди). Модель $ARIMA(p,d,q)$ означає, що різниці часового ряду порядку d підкоряються моделі $ARMA(p,q)$.

Хоча деякі дослідники неохоче уникали використання ARIMA для аналізу епідеміологічних даних COVID-19, Алабдулраззак та ін. [20] продемонстрували, що техніка ARIMA показала точний та валідний прогноз, особливо найкраща модель підгонки ARIMA для прог-

Методологію для прогнозування майбутніх станів епідеміологічних ситуації засобами часових рядів описано в таких дослідженнях [3]:

1. Історичні дані захворюваності на грип, малярію та гепатит було зібрано з Центру контролю та профілактики захворювань за період п'яти років (2017-2021) для шта-ту Нью-Йорк, США.
2. Зібрані дані потім було попередньо оброблено та нормалізовано, оскільки дані часових рядів мають вхідні значення з різними масштабами.
3. Дані потім використовували для навчання моделей часових рядів LSTM.
4. Модель з найнижчим показником RMSE та, яка могла відображати сезонність та варіації в даних, було використано для остаточного прогнозу (таблиця).

На підставі попередніх досліджень було встановлено, що моделі LSTM виявилися кращими за традиційні моделі, такі як ARIMA, і краще захоплювали сезонні тенденції в даних. Тому моделі LSTM використовували для прогнозування часових рядів. Було розроблено три моделі LSTM, одну для кожного захворювання. Модель, розроблена для грипу та гепатиту, містить два шари LSTM, кожний з 64 та 32 одиницями відповідно. Вихід шару LSTM під'єднано до густого шару з 32 нейронами. Вихід з цього шару під'єднано до остаточного вихідного шару. Середньоквадратичну помилку було використано як функцію втрат, а для оптимізації ваг мережі використовували оптимізатор adam – один із найефективніших алгоритмів оптимізації у навчанні нейронних мереж, який ідеї RMSProp та оптимізатора імпульсу. Модель, розроблена для малярії, містить два шари LSTM, кожний з 128 та 64 одиницями відповідно. Вихід з LSTM шарів під'єднано до густого шару з 64 нейронами. Вихід цього шару під'єднано до остаточного вихідного шару.

нозу підтверджених випадків одужання від COVID-19. Незважаючи на багато динамічних особливостей, що ґрунтують на новизні вірусу та характеру хвороби, фактичні значення для більшості періодів були в межах прогнозу моделі на рівні довіри 95 %. Кореляція Пірсона показала високі кореляції між точками прогнозу та фактичними записаними даними ($r = 0,996$) [20]. Це підтверджує, чому ARIMA є однією з найбільш точних моделей з задовільними результатами та мінімальною похибкою.

Обговорення результатів дослідження. У дослідженні [40] автори представили новий підхід до прогнозування епідемій, який об'єднує глибоке навчання з класичними моделями епідеміологічної динаміки. Цей гібридний підхід ставить за мету покращення точності прогнозів епідемій шляхом використання властивостей

обох методів і врахування взаємозв'язків у даних. Основними компонентами цього дослідження є використання великих наборів даних і алгоритмів глибокого навчання для аналізу тенденцій епідемій, а також моделювання динаміки епідемій за допомогою компартментальних моделей, зокрема моделі SEIR. Результати дослідження показали значне покращення точності прогнозів порівняно з традиційними методами, що свідчить про потенції гібридного підходу у керуванні та прогнозуванні епідемій. Важливим висновком є те, що цей підхід може бути застосовано в різних сферах охорони здоров'я, він має перспективи для подальшого розвитку.

У роботі [8] розглянуто прогнозування епідемічного процесу COVID-19 за допомогою методу Random Forest. Автори досліджують цю модель для прогнозування динаміки захворюваності в різних країнах, таких як Німеччина, Японія, Південна Корея та Україна. Метою роботи є оцінювання точності прогнозування на різні терміни (3, 7, 10, 14, 21 та 30 днів) та аналіз адекватності моделей. Результати моделювання показали, що цей метод може бути ефективно використаною для передбачення епідемічних процесів. Для аналізу було використано дані нових випадків захворювання на COVID-19, що дало змогу провести порівняльний аналіз між країнами. Дослідження показало, що прогнозування за допомогою методу Random Forest забезпечує високу точність завдяки здатності враховувати різноманітні чинники, які впливають на поширення вірусу. Враховуючи специфічні епідеміологічні дані та різні заходи контролю, прийняті в окремих країнах, модель допомагає визначити найбільш ефективні стратегії боротьби з пандемією на підставі достовірних прогнозів. Це дослідження підтвердило, що методи машинного навчання, зокрема Random Forest, може бути ефективно використано для моделювання та прогнозування епідемічних процесів, що сприяє розробленню науково обґрунтованих стратегій боротьби з пандемією.

Публікація [7] є актуальним дослідженням аналізу та прогнозування даних про COVID-19 за допомогою моделей машинного навчання. У цьому дослідженні автори використали різні підходи машинного навчання для аналізу великого обсягу даних про поширення COVID-19, в т.ч. класифікацію, кластеризацію та прогнозування. У цьому дослідженні було проаналізовано доступні дані щодо поширення COVID-19 та використано різноманітні моделі машинного навчання для прогнозування спалаху хвороби. Автори ретельно дослідили перспективи застосування таких моделей, зокрема алгоритмів класифікації, регресії та кластеризації, для аналізу динаміки поширення захворювання та визначення ключових чинників, що впливають на його поширення. Додатково, автори розглянули можливості моделей машинного навчання у визначенні ефективності заходів контролю та прогнозуванні подальшого розвитку ситуації. Внаслідок дослідження встановлено, що моделі машинного навчання, такі як алгоритми класифікації, регресії та кластеризації, показали значну ефективність у прогнозуванні поширення COVID-19. За допомогою цих моделей вдалося визначити ключові чинники, що впливають на поширення захворювання, такі як густина населення, рівень мобільності, наявність протиепідемічних заходів та інші. Дослідження також виявило, що з використанням цих моделей можна прогнозувати ефективність заходів контролю та розвиток ситуації в майбутньому.

Автори роботи [35] вирішують критичну проблему прогнозування тенденцій розповсюдження COVID-19 за допомогою комбінації логістичних моделей і методів машинного навчання. Автори використали дані про COVID-19 з університету Джона Гопкінса, що охоплюють період з 22 січня 2020 р. по 16 червня 2020 р. Окрім цього, дані про спалах SARS у Китаї в 2003 р. використовували для навчання логістичної моделі. Логістична модель, відома своєю простотою та ефективністю підгонки даних часових рядів, було використано для прогнозування верхніх меж випадків COVID-19. Потім цю модель було інтегровано з моделлю FbProphet, інструментом прогнозування часових рядів, який було використано для уточнення прогнозів, отриманих з логістичної моделі.

Комбінована модель передбачила, що глобальний спалах COVID-19 досягне піку наприкінці жовтня 2020 р. з приблизно 14,12 мільйона загальних випадків. Точність моделі перевірено на підставі відомих тенденцій даних, що демонструє її стійкість у прогнозуванні кривих епідемії для різних країн, в т.ч. Бразилію, Індію, Перу та Індонезію.

Дослідження виявило важливість інтеграції моделей логістичного зростання з передовими методами машинного навчання для підвищення точності прогнозів епідемії. Це також засвідчує варіабельність прогресування епідемії в різних країнах через різні заходи контролю та відповіді громадської охорони здоров'я. Результати роботи показали, що поєднуючи переваги логістичних моделей і машинного навчання є ефективним способом для прогнозування траєкторії пандемії COVID-19.

У публікації [6] автори дослідили застосування інтерпретованої моделі машинного навчання для аналізу епідемій. Вони запропонували підхід, який поєднує порівняльне навчання та модель NARMAX для визначення динаміки поширення епідемії. Дослідження спрямоване на розроблення методології, яка дає змогу не тільки передбачати розвиток епідемій, але й розуміти чинники, що впливають на їх поширення. автори розробили та застосували інтерпретовану модель машинного навчання, яка базується на порівняльному навчанні та моделі NARMAX (англ. *Nonlinear Autoregressive Moving Average with eXogenous inputs*), для аналізу динаміки епідемій. Результати дослідження показали, що цю модель можна успішно використовувати для прогнозування та аналізу поширення епідемій, а також для виявлення ключових чинників, що впливають на їх розвиток.

Зокрема, автори з'ясували, що модель здатна точно передбачати тенденції поширення епідемій на підставі наявних даних. Окрім цього, вони виявили важливі параметри та залежності між різними чинниками, такими як погода, густина населення, міра карантину тощо, які мають великий вплив на динаміку епідемічного процесу.

У дослідженні [12] розглянуто застосування моделей машинного навчання для передбачення поширення COVID-19 та розроблення стратегій управління кризовою ситуацією. Дослідження охоплює аналіз доступних даних про хворобу, використання різних моделей машинного навчання, таких як модель SEIR (англ. *Susceptible-Exposed-Infectious-Recovered*) та регресійні моделі, для прогнозування та оцінювання ризику. Результати показали, що моделі машинного навчання можуть ефективно прогнозувати поширення COVID-19 та допомагати управлінню кризовою ситуацією, зокрема, визначити

оптимальні стратегії мітингу та інших обмежувальних заходів, а також визначати потребу в ресурсах для лікування та протиепідемічних заходах. Автори використовували різні моделі машинного навчання для аналізу динаміки поширення COVID-19 та прогнозування його подальшого розвитку. Зокрема, було використано модель SEIR (англ. *Susceptible-Exposed-Infectious-Recovered*), що дає змогу моделювати динаміку епідемічного процесу, та регресійні моделі для прогнозування кількості випадків захворювання на підставі наявних даних. Результати дослідження показали, що моделі машинного навчання можуть успішно передбачати динаміку поширення COVID-19 з високою точністю. Аналіз даних також допомагає визначити необхідність ресурсів для медичного обслуговування та лікування хворих, а також оцінити ефективність введених протиепідемічних заходів. Це дає змогу урядовим структурам краще реагувати на кризові ситуації та забезпечувати населенню необхідну підтримку та захист.

У роботі [15] автори пропонують інтегровану архітектуру системи Інтернету речей (IoT), що базується на моделях машинного навчання, для прогнозування росту та тенденцій епідемічних вибухів COVID-19 з метою виявлення високоризикових місць. У дослідженні використовували різні методи машинного навчання, такі як методи класифікації, регресії та кластеризації, для аналізу великих обсягів даних, отриманих з сенсорів і пристроїв IoT. Зокрема, використовували алгоритми нейронних мереж, дерев рішень та ансамблів моделей для прогнозування та класифікації даних про поширення хвороби.

Результати дослідження показали, що розроблена архітектура може успішно прогнозувати рівень ризику та тенденції епідемічних вибухів COVID-19 на підставі зібраних даних. Застосування цієї системи дає змогу ефективно ідентифікувати високоризикові місця та вживати належних заходів для запобігання поширенню хвороби. Отже, результати цього дослідження свідчать про великий потенціал інтегрованої архітектури машинного навчання та IoT для управління епідемічними вибухами та забезпечення публічного здоров'я.

У публікації [32] здійснено порівняльне оцінювання моделей часових рядів для прогнозування випадків захворювання на грип. Дослідження базується на аналізі даних про грипоподібні захворювання, отриманих з лікувальних закладів в Ірані. Автори розглядали різні моделі часових рядів для прогнозування спалахів грипу, такі як авторегресійна інтегрована модель скінчених різниць (ARIMA), адаптивні нейронні мережі (ANN) та авторегресійна модель ковзного середнього (ARMA). У дослідженні використовували дані з лікувальних закладів Ірану для побудови та порівняння різних моделей у прогнозуванні інцидентів грипу. Результати дослідження вказують на те, що модель ARIMA демонструє кращу точність у прогнозуванні виникнення грипу порівняно з іншими моделями. Використання цієї моделі може допомогти у вчасному виявленні та прогнозуванні інцидентів грипу, що є важливим для планування і впровадження ефективних заходів з громадського здоров'я.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – отримала подальший розвиток методика ана-

лізу методів машинного навчання, придатних для проведення епідеміологічних досліджень, що дасть змогу класифікувати набори даних на підставі розроблених моделей, а також передбачати майбутні результати з використанням цих моделей.

Практична значущість результатів дослідження – розроблену методику аналізу методів машинного навчання, придатну для проведення епідеміологічних досліджень на відповідних наборах даних, можна використати в медичних закладах для прогнозування та розвитку епідеміологічних ситуацій, що стане важливим і необхідним підґрунтям для прийняття управлінських рішень щодо запобігання такої надзвичайної ситуації як епідемія, або виходу з неї.

Висновки / Conclusions

Проаналізовано методи машинного навчання та методи нейронних мереж, придатні для проведення епідеміологічних досліджень, що дасть змогу класифікувати набори даних на підставі розроблених моделей, а також уможливить передбачати майбутні результати на підставі цих моделей. За результатами проведеного дослідження можна дійти таких основних висновків.

1. Проаналізовано ключові особливості використання методів машинного навчання у контексті аналізу епідеміологічних даних, охоплюючи широкий спектр підходів від класичних до глибокого навчання. Вказано на особливості використання цих методів для виявлення епідеміологічних груп ризику, прогнозування поширення хвороб, виявлення нових епідеміологічних трендів і формування передбачень для контролю та оцінювання епідеміологічної ситуації.
2. Виявлено, що успішне впровадження методів машинного навчання для аналізу епідеміологічних даних потребує не тільки технічної експертизи та навичок, але й глибокого розуміння унікальних особливостей цієї області та адаптації організаційних процесів. Ключовим є створення сприятливого середовища для взаємодії між фахівцями у галузі епідеміології та науковцями з інформаційних технологій, а також постійне вдосконалення методів аналізу, що враховують оновлення у наборі даних і технологічні зміни.
3. Представлено основні ризики, пов'язані з використанням методів машинного навчання для аналізу епідеміологічних даних, у т.ч. технічні труднощі та неоднозначність у тлумаченні результатів. Запропоновано стратегії вибору відповідного методу машинного навчання для вирішення завдань аналізу та прогнозування епідеміологічної ситуації.
4. Встановлено, що для досягнення якісних результатів у разі застосування методів машинного навчання дуже важливою є якість вихідного набору даних. Проаналізовано вимоги до вихідних даних і ключових показників під час їх добору для методів машинного навчання.
5. З'ясовано основні переваги та ефективність впровадження методів машинного навчання для аналізу епідеміологічних даних, що можуть привести до значного покращення прогнозування та контролю за епідеміями, сприяючи швидкому виявленню потенційних загроз для громадського здоров'я та розроблення ефективних стратегій запобігання їх поширенню.

References

1. Ajagbe, S. A., & Adigun, M. O. (2024). Deep learning techniques for detection and prediction of pandemic diseases: A systematic literature review. *Multimedia Tools and Applications*, 83, 5893–5927. <https://doi.org/10.1007/s11042-023-15805-z>
2. Ajith, A., Manoj, K., Kiran, H., Pillai, P. J., & Nair, J. J. (2020). A study on prediction and spreading of epidemic diseases. 2020

- International Conference on Communication and Signal Processing (ICCSP)* (pp. 1265–1268). <https://doi.org/10.1109/ICCSP48568.2020.9182147>
3. Al-Abdulrazzaq, D. (2019). Use of non-invasive parameters and machine-learning algorithms for predicting future risk of type 2 diabetes: A retrospective cohort study of health data from Kuwait. *Frontiers in Endocrinology*. <https://doi.org/10.3389/fendo.2019.00624>
4. Atkinson, A., Ellenberger, B., Piezzi, V., Kaspar, T., Salazar-Vizcaya, L., Endrich, O., Leichtle, A. B., & Marschall, J. (2022). Extending outbreak investigation with machine learning and graph theory: Benefits of new tools with application to a nosocomial outbreak of a multidrug-resistant organism. *Infection Control & Hospital Epidemiology*, 16, 1–7. Advance online publication. <https://doi.org/10.1017/ice.2022.66>
5. Chen, Y., Chu, C. W., Chen, M. I. C., & Cook, A. R. (2018). The utility of LASSO-based models for real-time forecasts of endemic infectious diseases: A cross-country comparison. *Journal of Biomedical Informatics*, 81, 16–30. <https://doi.org/10.1016/j.jbi.2018.02.014>
6. Cheng, S., Zhou, J., Luo, X., Zhou, T., & Jiang, B. (2022). The application of interpretable machine learning model based on comparative learning and NARMAX in epidemic research. In 2022 *IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)* (pp. 1071–1076). Dalian, China. IEEE. <https://doi.org/10.1109/ICA-ICA54878.2022.9844583>
7. Chrin, R., & Wang, S. (2022). Analysis and Prediction of COVID-19 Data using Machine Learning Models. In Proceedings of the 2021 *10th International Conference on Computing and Pattern Recognition (ICCPR 21)* (pp. 296–301). Association for Computing Machinery. <https://doi.org/10.1145/3497623.3497671>
8. Chumachenko, D., Menailov, I., Bazilevych, K., & Krivtsov, S. (2021). Forecasting of COVID-19 Epidemic Process by Random Forest Method. In 2021 *IEEE 8th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T)* (pp. 491–494). Kharkiv, Ukraine. IEEE. <https://doi.org/10.1109/PICST54195.2021.9772149>
9. Gao, Y., Cai, G. Y., & Fang, W., et al. (2020). Machine learning based early warning system enable accurate mortality risk prediction for COVID-19. *Nature Communications*, 11, Article ID 5033. <https://doi.org/10.1038/s41467-020-18684-2>
10. Ghannam, R. B., & Techtman, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal*, 19, 1092–1107. <https://doi.org/10.1016/j.csbj.2021.01.028>
11. Gianfredi, V., Santangelo, O. E., & Provenzano, S. (2021). Correlation between flu and Wikipedias pages visualization. *Acta Biomedica*, 92, Article ID e2021056. <https://doi.org/10.23750/abm.v92i1.9790>
12. Gupta, R., Pandey, G., Chaudhary, P., & Pal, S. K. (2020). Machine Learning Models for Government to Predict COVID-19 Outbreak. Digital Government: Research and Practice, 1(4), Article ID 26, 6 p. <https://doi.org/10.1145/3411761>
13. Huang, A. A., & Huang, S. Y. (2023). Technical report: Machine-learning pipeline for medical research and quality-improvement initiatives. *Cureus*, 15(10), Article ID e46549. <https://doi.org/10.7759/cureus.46549>
14. Kłosowski, P. (2020). Artificial intelligence application to improve the performance of distance learning servers during the coronavirus pandemic threat period. 2020 *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Poznan, Poland, pp. 110–115. <https://doi.org/10.23919/SPA50552.2020.9241301>
15. Kumar, C. R. J., Arunsi, M. B., & Majid, M. A. (2023). A Machine Learning-driven IoT Architecture for Predicting the Growth and Trend of Covid-19 Epidemic Outbreaks to Identify High-risk Locations. In 2023 *20th Learning and Technology Conference (L&T)* (pp. 120–127). Jeddah, Saudi Arabia. <https://doi.org/10.1109/LT58159.2023.10092331>
16. Li, F., & Luan, P. (2011). ARMA model for predicting the number of new outbreaks of Newcastle disease during the month. In Proceedings of the 2011 *IEEE International Conference on Computer Science and Automation Engineering* (pp. 660–663). <https://doi.org/10.1109/CSAE.2011.5952933>
17. Liang, R., Lu, Y., Qu, X., Su, Q., Li, C., Xia, S., Liu, Y., Zhang, Q., Cao, X., Chen, Q., & Niu, B. (2019). Prediction for global African swine fever outbreaks based on a combination of random forest algorithms and meteorological data. *Transboundary and Emerging Diseases*. <https://doi.org/10.1111/tbed.13424>
18. Liu, M., Liu, Y., & Liu, J. (2023). Epidemiology-aware deep learning for infectious disease dynamics prediction. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM 23) (pp. 4084–4088). Association for Computing Machinery. <https://doi.org/10.1145/3583780.3615139>
19. Mitra, A., Soman, B., & Singh, G. (2019). Integrating multiple data sources and learning models to predict infectious diseases in China. *AMIA Joint Summits on Translational Science Proceedings*, 2019, 680–685. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6568090/>
20. Molaei, S., Khansari, M., Veisi, H., et al. (2019). Predicting the spread of influenza epidemics by analyzing Twitter messages. *Health Technology*, 9, 517–532. <https://doi.org/10.1007/s12553-019-00309-4>
21. Najjar, A. M., Irawan, M. I., & Adzkiya, D. (2018). Extreme learning machine method for dengue hemorrhagic fever outbreak risk level prediction. 2018 *International Conference on Smart Computing and Electronic Enterprise (ICSCEE)* (pp. 1–5). <https://doi.org/10.1109/ICSCEE.2018.8538409>
22. Ono, S., & Goto, T. (n.d.). Introduction to supervised machine learning in clinical epidemiology. Department of Eat-loss Medicine, Graduate School of Medicine, The University of Tokyo. Department of Clinical Epidemiology and Health Economics, The University of Tokyo. TXP Medical Co. Ltd. <https://doi.org/10.37737/acc.22009>
23. Palaniappan, S., V., R., David, B., & Pathur Nisha, S. (2022). Prediction of epidemic disease dynamics on the infection risk using machine learning algorithms. *SN Computer Science*, 3, 47 p. <https://doi.org/10.1007/s42979-021-00902-3>
24. Pramod, A., Abhishek, J. S., & Suganthi, K. (2023). Epidemic outbreak prediction using machine learning models. School of Electronics Engineering, Vellore Institute of Technology Chennai Campus. <https://doi.org/10.48550/arXiv.2310.19760>
25. Provenzano, S., Santangelo, O. E., Giordano, D., Alagna, E., Piazza, D., Genovese, D., Calamusa, G., & Firenze, A. (2019). Predicting disease outbreaks: Evaluating measles infection with Wikipedia Trends. *Recenti Progressi in Medicina*, 110, 292–296. <https://doi.org/10.1701/3182.31610>
26. Roth, J. A., Battegay, M., Juchler, F., Vogt, J. E., & Widmer, A. F. (2018). Introduction to machine learning in digital healthcare epidemiology. *Infection Control & Hospital Epidemiology*, 39, 1457–1462. <https://doi.org/10.1017/ice.2018.265>
27. Roy, S., Biswas, P., & Ghosh, P. (2021). Spatiotemporal tracing of pandemic spread from infection data. *Scientific Reports*, 11, Article ID 17689. <https://doi.org/10.1038/s41598-021-97207-5>
28. Santangelo, O. E., Gentile, V., Pizzo, S., Giordano, D., & Cedrone, F. (2023). Machine learning and prediction of infectious diseases: A systematic review. *Machine Learning and Knowledge Extraction*, 5, 175–198. <https://doi.org/10.3390/make5010013>
29. Santangelo, O., Provenzano, S., Piazza, D., Giordano, D., Calamusa, G., & Firenze, A. (2019). Digital epidemiology: Assessment of measles infection through Google Trends mechanism in Italy. *Ann Ig*, 31, 385–391. <https://doi.org/10.7416/ai.2019.2300>
30. Santillana, M., et al. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Computational Biology*, 11(10), Article ID e1004513. <https://doi.org/10.1371/journal.pcbi.1004513>
31. Soliman, M., Lyubchich, V., & Gel, Y. R. (2019). Complementing the power of deep learning with statistical model fusion: Pro-

- babilistic forecasting of influenza in Dallas County, Texas, USA. *Epidemics*, 28, Article ID 100345. <https://doi.org/10.1016/j.epidem.2019.05.004>
32. Tapak, L., et al. (2019). Comparative evaluation of time series models for predicting influenza outbreaks: Application of influenza-like illness data from sentinel sites of healthcare centers in Iran. *BMC Research Notes*, 12, Article ID 353. <https://doi.org/10.1186/s13104-019-4393-y>
 33. Turbe, V., Herbst, C., Mngomezulu, T., Meshkinfamfard, S., Dlamini, N., Mhlongo, T., Smit, T., Cherepanova, V., Shimada, K., Budd, J., Arsenov, N., Gray, S., Pillay, D., Herbst, A., Shahmanesh, M., & McKendry, R. (2021). Deep learning of HIV field-based rapid tests. *Nature Medicine*, 27, 1–6. <https://doi.org/10.1038/s41591-021-01384-9>
 34. Wang, M., et al. (2019). A novel model for malaria prediction based on ensemble algorithms. *PLoS ONE*, 14(12), Article ID e0226910. <https://doi.org/10.1371/journal.pone.0226910>
 35. Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning techniques. *Chaos, Solitons & Fractals*, 139, Article ID 110058. <https://doi.org/10.1016/j.chaos.2020.110058>
 36. Wiemken, T. L., & Kelley, R. R. (2019). Machine learning in epidemiology and health outcomes research. *Annual Review of Public Health*, 40, 21–36. <https://doi.org/10.1146/annurev-publhealth-040119-094437>
 37. World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva, Switzerland: World Health Organization, Licence: CC BY-NC-SA 3.0 IGO. URL: <https://www.who.int/publications/i/item/9789240029200>
 38. World Health Organization. (2022). *Global Report on Infection Prevention and Control*. Geneva, Switzerland: World Health Organization, Licence: CC BY-NC-SA 3.0 IGO. URL: <https://iris.who.int/bitstream/handle/10665/354489/9789240051164-eng.pdf?sequence=1>
 39. Wu, Y., Yang, Y., Nishiura, H., & Saitoh, M. (2018). Deep learning for epidemiological predictions. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR 18)* (pp. 1085–1088). Association for Computing Machinery. <https://doi.org/10.1145/3209978.3210077>
 40. Xu, B., Chen, X., A, Y., & Chen, Z. (2024). Epidemic Prediction Algorithm Based on Deep Learning and Epidemic Dynamics. In *Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application (ICMLCA 23)* (pp. 1021–1026). Association for Computing Machinery. <https://doi.org/10.1145/3650215.3650396>

V. V. Veseliak, Yu. I. Hrytsiuk

Lviv Polytechnic National University, Lviv, Ukraine

MACHINE LEARNING METHODS IN EPIDEMIOLOGICAL RESEARCH

This study investigates the application of machine learning methods for analyzing, predicting, and controlling epidemiological conditions, as well as identifying disease trends and patterns. The research encompasses a comprehensive analysis of various machine learning techniques for epidemiological data analysis. Starting with classical approaches such as linear regression and decision trees, their advantages and limitations are discussed. Additionally, the study explores the use of deep learning methods, such as convolutional neural networks and recurrent neural networks, for effectively analyzing complex relationships in epidemiological data. Each method's benefits and constraints, along with their potential applications for predicting epidemiological situations and identifying disease risk factors, are detailed. A comparison of the accuracy of different methods is provided using examples of several viral diseases. The challenges and limitations in utilizing machine learning for infectious disease prediction are analyzed. The study also examines the outcomes of combining multiple methods and their results. The requirements for input data and important indicators for selecting input data for machine learning methods are considered. Recent high-quality studies that applied machine learning methods to real-world data and evaluated the effectiveness of these models in predicting the spread and onset of infectious diseases are reviewed. The research examines data from various sources and regions around the world for several infectious diseases. The effectiveness of methods in detecting dependencies between disease trends and geographic distribution, as well as accounting for seasonality, local characteristics, and other differences, is analyzed. Conclusions are drawn regarding the current capabilities and future prospects of using machine learning methods for analyzing and predicting epidemiological data. A list of criteria is developed to assess the suitability of machine learning methods for evaluating and forecasting epidemiological conditions.

Keywords: machine learning model; COVID-19 pandemic; artificial intelligence; forecasting.