УДК 004.93
DOI: 10.15587/1729-4061.2020.197319

Розробка генетичного методу для прогнозування показників здоров'я населення на основі нейромережевих моделей

Є. М. Федорченко, А.О. Олійник, О. О. Степаненко, Т.А. Зайко, С. К. Корнієнко, А. С. Харченко

Запропоновано генетичний метод для прогнозування показників здоров'я населення на основі нейромережевих моделей. Принципова відмінність запропонованого генетичного методу від існуючих аналогів полягає у використанні диплоїдного набору хромосом в особин популяції, яка еволюціонує. Така модифікація робить залежність фенотипу особини від генотипу менш детермінованою і, врешті, сприяє збереженню різноманітності генофонду популяції і варіабельності ознак фенотипу впродовж виконання алгоритму. Крім цього, запропоновано модифікацію генетичного оператору мутацій. На відміну від класичного методу, особини, які піддаються дії оператору мутації, обираються не випадковим чином, а у відповідності до їх мутаційної стійкості, що відповідає значенню функції пристосованості особини. Таким чином, мутують особини, що характеризуються гіршими значеннями цільової функції, а геном сильних особин залишається незмінним. У цьому випадку зменшується вірогідність втрати досягнутого впродовж еволюції екстремуму функції внаслідок дії оператора мутацій, а перехід до нового екстремуму здійснюється у випадку накопичення достатньої питомої ваги кращих ознак в популяції.

Порівняльний аналіз роботи моделей, синтезованих за допомогою розробленого генетичного методу, показав, що найкращі результати досягнуті у моделі на основі нейронної мережі довгої короткочасної пам'яті. Під час створення і навчання моделі на основі мережі довгої короткочасної пам'яті було досліджено можливість використання методу рою часток для оптимізації параметрів мережі. Результати експериментальних досліджень показали, що розроблена модель дає найменшу помилку передбачення кількості нових випадків туберкульозу – середня абсолютна помилка складає 6,139, що менше у порівнянні з моделями, побудованими за допомогою інших методів).

Практичне використання розроблених методів дасть можливість своєчасно коригувати плановані лікувально-діагностичні, профілактичні заходи, завчасно визначати необхідні ресурси для локалізації та ліквідації захворювань з метою збереження здоров'я населення.

Ключові слова: нейронні мережі, генетичний алгоритм, фенотип, модифікований генетичний оператор мутації, прогнозування показників здоров'я населення.

1. Вступ

Якість життя населення визначається різними показниками, зокрема показниками здоров'я, стан якого обумовлюється факторами навколишнього середо-

вища. Згідно з медичними дослідженнями, проведеними в останні роки [1], існує тісний зв'язок між техногенним забрудненням повітря на окремих територіях і підвищеною захворюваністю населення. За оцінкою Всесвітньої організації охорони здоров'я (ВООЗ), в даний час забруднення повітря є найбільшим фактором екологічного ризику для здоров'я [2]. Відповідно до цієї оцінки, близько 3,7 млн додаткових випадків смертності пов'язані з забрудненням атмосферного повітря, 4,3 млн – із забрудненням повітря всередині приміщень. Оскільки багато людей піддаються впливу як внутрішнього, так і зовнішнього забрудненого повітря, причини та кількість смертей від різних захворювань, викликаних різними джерелами, не можуть визначатися шляхом звичайного узагальнення даних. Найбільші проблеми зі здоров'ям, викликані прямим впливом забруднення повітря, пов'язані із захворюваннями кровообігу, респіраторними захворюваннями, раком, нервово-психічними розладами та деякими іншими [3, 4].

Отже, стан здоров'я та захворюваність населення регіону можна розглядати як похідні від навколишнього середовища.

Використання відомих методів статистики для прогнозування залежності показників здоров'я, а також математичних моделей, запропонованих у відомих літературних джерелах [1–14], характеризується певними обмеженнями та вимогами до цільових функцій. При використанні таких методів є неможливим підвищення точності прогнозу при зміні параметрів, наприклад, при прогнозуванні залежності показників здоров'я населення від обсягів викидів забруднюючих речовин у повітрі. Дані обмеження, при пошуку оптимальних рішень, не дозволяють підвищити точність прогнозу до необхідного значення, що обумовлює доцільність побудови моделей, які забезпечать більш високі показники точності прогнозу. Такими моделями можуть бути моделі на основі штучних нейронних мереж, які здатні до оброблення багатовимірних даних різних типів, а також характеризуються високими апроксимаційними та узагальнювальними властивостями.

Тому актуальним завданням є розроблення методів та моделей для прогнозування показників здоров'я населення на основі нейромережевих технологій.

2. Аналіз літературних даних та постановка проблеми

У роботі [5] наведено результати досліджень стосовно визначення математичної залежності показників здоров'я населення на основі нейромережевих моделей від обсягів викидів забруднюючих речовин. В запропонованій моделі незалежною змінною є обсяг викидів забруднюючих речовин, а залежною – показник захворюваності (1)

$$K_{morb} = f\left(x_{emiss}\right), \tag{1}$$

де K_{morb} – показник захворюваності, x_{emiss} – показники, що характеризують вплив обсягів викидів.

Виходячи з наведених даних і аналізу статистичних даних [2 – 5], можна дійти висновку, що шукана математична модель буде не детермінованою, а скоріш стохастичною.

У роботі [6] наведено результати досліджень, що окрім обсягів викидів забруднюючих речовин на рівень захворюваності, здійснює вплив множина інших факторів, точну кількість яких визначити досить проблематично. Якщо позначити ці фактори як x_1, x_2, …, x_n, то узагальнену модель залежності (1) можна представити у формі (2):

$$K_{morb} = f\left(x_{emiss}, x_1, x_2, ..., x_n\right). \quad (2)$$

Під час аналізу роботи [7] встановлено, що основним фактором впливу викидів на здоров'я людини є наявність в їх складі токсичних речовин. При вивченні впливу забруднення атмосферного повітря на здоров'я населення [5–7] визначено окрему групу хвороб. До цієї групи належать хронічні обструктивні захворювання легень, бронхів, бронхіальна астма, а також рак легенів, захворювання серцево-судинної і нервової системи.

Згідно з дослідженнями Центральної геофізичної обсерваторії [2], у 2015 році в атмосферне повітря в Україні було викинуто 4,5 млн. т шкідливих речовин, з яких 62 % – стаціонарними джерелами та 38 % – пересувними джерелами. Головними забруднювачами повітря є підприємства енергетики та металургії (55 % та 22 % всіх забруднень від стаціонарних джерел). Там фіксують підвищений вміст специфічних шкідливих речовин: формальдегіду, фенолу, фтористого водню, аміаку, особливо багато діоксиду азоту і оксиду вуглецю. Тому саме вплив цих токсичних речовин, викинутих від стаціонарних джерел, враховується в дослідженнях и було прийнято рішення про побудови моделі. Необхідно розробити модель для прогнозування рівня захворюваності на прикладі населення України, що захворіло внаслідок несприятливої екологічної ситуації у містах, залежно від видів та концентрації забруднюючих речовин. Практична цінність розробленої моделі полягає в тому, що вона може бути використана для прогнозування динаміки показників здоров'я у майбутньому для інших міст. Використання розробленої моделі дасть можливість своєчасно коригувати плановані лікувально-діагностичні, профілактичні заходи, завчасно визначати необхідні ресурси для локалізації та ліквідації захворювань з метою збереження здоров'я населення. Варто відзначити, що запропоновані у роботі методи синтезу моделей показників здоров'я населення можуть застосовуватися також для обробки даних з інших джерел та інших країн.

В роботі [7] наведено результати досліджень, що характер і ступінь впливу токсичних речовин, їх здатність провокувати патологічні стани в організмі людини варіюють в залежності від комбінації метеорологічних і кліматичних факторів. Опади і високі температури, навпаки, сприяють інтенсивному розкладанню речовин. Більш висока температура біля поверхні землі в денний час змушує повітря підніматися вгору, що призводить до додаткової турбулентності. Як тільки повітря прогрівається до 10 градусів і вище, кількість шкідливих речовин починає накопичуватися в атмосфері. Вночі температура біля поверхні землі нижча, тому турбулентність зменшується. Це явище призводить до змен-

шення розсіювання відпрацьованих газів. Тому при побудові моделі буде враховуватися середня температура повітря та кількість опадів на місяць.

В роботі [8] наведено результати досліджень, що на показники захворюваності впливає якість медичного обслуговування населення. Тому у якості основних метрик, які доцільно враховувати при будуванні моделі залежності показників захворюваності, використаємо показник кількості лікарів (усіх спеціалізацій) у регіоні. А також застосуємо показник кількості лікарняних ліжок у стаціонарних відділеннях медичних закладів регіону, як кількісний показник обсягів медичного обслуговування.

В роботі [9] встановлено, що розподілення захворюваності в різних регіонах є статистичним, тому для моделювання такої залежності варто враховувати і кількість населення в регіоні.

Оскільки, згідно даних медичної статистики [3, 5], загальна захворюваність населення має різні показники у різних вікових групах (як правило, збільшується з віком). Виявлено тенденцію, що у людей похилого віку частіше виникають серцево-судині захворювання, туберкульоз, та рак (онкологічні захворювання) ніж у молодих. Тобто, високі показники захворюваності населення характерні для регіонів з високими частками людей похилого віку. Регіони з найбільшим показником середнього віку жителів потенційно є регіонами із несприятливими передумовами щодо захворюваності населення. Тому доцільно також враховувати середній вік населення у регіоні.

Таким чином, узагальнену модель залежності показників здоров'я від обсягів викидів з деяким припущенням можна привести до виду (3):

$$K_{morb} = f\left(x_{emiss}, x_{popul}, x_{temp}, x_{ra\,\mathrm{inf}\,all}, x_{docs}, x_{beds}\right), \tag{3}$$

де x_{popul} – показник, що характеризує вплив кількості населення, x_{temp} – середня температура повітря, $x_{rainfall}$ – кількість опадів, x_{docs} – показник, що характеризує вплив кількості лікарів, x_{beds} – показник, що характеризує вплив загальної кількості ліжок у стаціонарних відділеннях.

У статті [10] пропонується класичний регресійний аналіз для встановлення визначення математичної залежності показників здоров'я від обсягів викидів забруднюючих речовин. Показано, як класичний метод стохастичного прогнозування захворюваності досліджує взаємозв'язки показників між захворюваністю і факторами, що її зумовлюють, коли залежність між ними не є строго функціональної і спотворена впливом сторонніх факторів. Також показано, що при проведенні кореляційно-регресійного аналізу будуються різні кореляційні і регресійні моделі захворюваності. У цих моделях виділяють факторні і результативні показники (ознаки). Авторами роботи представлено регресійний аналіз, який показує вибір форми зв'язку і типу моделі для визначення розрахункових значень залежної змінної (результативної ознаки). В роботі розроблені неадаптивні регресивні моделі, які показують врахування всієї передісторії захворюваності на аналізованій території. Але для того, щоб їх побудувати, використовувалися всі наявні дані та спостереження останніх років, що володіють схо-

жими характеристиками. Так, якщо властивості процесу захворюваності змінилися, ймовірно, застарілі дані вже не допоможуть уточнити прогноз. Тому залишилося не вирішена проблема, пов’язана з тим що, неадаптивні моделі дозволяють отримати проекції захворюваності на довгостроковий термін. Такі моделі ігнорують локальні коливання епідемічних показників і погано підходять для короткострокового прогнозування. Варіантом подолання відповідних труднощів може бути розрахунок середньострокової оцінки захворюваності при досить великій ширині ковзного вікна. Отже, розроблювана модель повинна бути досить чутливої для того, щоб реагувати на поточні тенденції захворюваності для формування прогнозів на декілька тижнів вперед.

В роботі [11] наведено результати досліджень використання байєсівських мереж для прогнозування захворюваності. Показано, що байєсовські мережі є ефективним, компактним та інтуїтивно зрозумілим способом представлення знань, пов'язаних з невизначеністю. Представлено байєсівську мережу (БМ), як графічну модель, що відображає імовірнісні залежності множини змінних і дозволяє проводити імовірнісний висновок за допомогою цих змінних. Показано, що у медичній діагностиці найбільш ймовірний діагноз визначається як значення множини можливих діагнозів, що має максимум ймовірності наявності захворювання за умови конкретного набору даних. Ці дані включають в себе симптоми, результати тестів та інші ознаки. Побудова авторами БМ здійснюється як при великому, так і малому обсязі вихідних даних, однак алгоритми оцінки параметрів моделі складно обчислювальний. Тому авторами роботи було проаналізовано БМ на основі вузького ковзного вікна спостережень. В проаналізованій роботі залишилося не вирішена проблема, пов’язана з тим що, байєсівські мережі надають можливість лише короткострокового прогнозування захворюваності.

Робота [12] надає опис використанню штучних нейронних мереж (ШНМ) для встановлення залежності показників здоров’я населення хворобами від зовнішніх чинників. Авторами роботи показано, що ШНМ дозволяють моделювати різного роду залежності, в основі яких можуть бути лінійні моделі, узагальнено лінійні моделі і нелінійні моделі. Саме здатність ІНС до узагальнення і виділення прихованих залежностей між вхідними та вихідними даними лежить в основі отримання достовірних статистичних прогнозів. В роботі показано, що потенційна прогностична здатність нейронних мереж виявляється кращою за рахунок більш якісного поділу класів, обумовленого використанням гладких функцій трансформації. Функції забезпечують збереження інформації до етапу остаточного прийняття рішень.

Основним недоліком даної роботи являється те, що використання нейронних мереж потребує тривалих часових витрат на виконання процедури навчання, які часто не дозволяють застосовувати ШНМ в системах реального часу [13].Отже, проаналізувавши дану роботу можна дійти висновку, що ШНМ може бути досить ефективним математичним базисом для прогнозування залежності показників здоров’я населення від обсягів викидів забруднюючих речовин у повітрі.

Таким чином, на даний момент універсального способу прогнозування захворюваності не існує, внаслідок чого дослідники вимушені обирати прогнос-

тичні моделі, виходячи з порівняння результатів, отриманих за допомогою різних методів на основі емпіричних даних.

Проаналізувавши роботи [10–13] встановлено, що для вирішення поставленого завдання можуть ефективно використовуватися ШНМ, оскільки моделі на основі штучних нейронних мереж забезпечують можливість оброблення багатовимірних даних різних типів (тим самим реалізуючи функцію багатьох змінних), високу адаптивність до зовнішніх змін, забезпечують можливість синтезу моделей з високими апроксимаційними та узагальнювальними властивостями. Тому необхідно розробити метод побудови нейромережевих моделей на основі емпіричних даних, що дозволить синтезувати моделі залежності показників здоров’я від обсягів викидів забруднюючих речовин.

Використання традиційних методів статистики для прогнозування залежності показників здоров'я [14], а також математичних моделей, запропонованих в [14], характеризується певними обмеженнями та вимогами до цільових функцій. При використанні таких методів є неможливим підвищення точності прогнозу при зміні параметрів, наприклад, при прогнозуванні залежності показників здоров'я населення від обсягів викидів забруднюючих речовин у повітрі. Дані обмеження, при пошуку оптимальних рішень, не дозволяють підвищити точність прогнозу до необхідного значення. При застосуванні генетичних алгоритмів (ГА), заснованих на механізмах природного відбору і успадкування, дозволяє уникнути ряд обмежень, і тим самим підвищити точність прогнозу [15].

В ГА використовується еволюційний підхід [15], де пошук екстремуму цільової функції здійснюється одночасно за багатьма напрямками шляхом використання популяції можливих рішень. Перехід від однієї популяції до іншої дозволяє уникнути попадання в локальний оптимум, при цьому ГА характеризується поліноміальною складністю обчислень.

Застосування ГА вирішує проблему, використовуючи процес, подібний біологічному розвитку. Він працює як рекомбінація і мутація генетичних послідовностей. Рекомбінація і мутація – генетичні оператори, тобто вони керують генами (послідовністю кодів), що містять всю інформацію, необхідну для того, щоб створити функціональний організм з певними характеристиками (генотипом) [16].

У випадку генетичної оптимізації, використовуваної для вирішення завдань, пов'язаних з прогнозуванням, послідовність кодів зазвичай приймає форму ряду чисел. Як і в процесі біологічного відбору (де менш придатні члени популяції залишають менше потомства), менш придатні рішення видаляються. При цьому більш придатні рішення розмножуються, створюючи інше покоління рішень, яке може містити декілька кращих рішень, ніж попередні. Процес рекомбінації, випадкової мутації і відбору є надзвичайно дієвим механізмом вирішення даного завдання.

Основною метою роботи є дослідження можливості застосування ГА до вирішення завдання прогнозування показників здоров’я населення від обсягів викидів забруднюючих речовин у повітрі, при мінімальних часових витратах.

Як свідчить аналіз методів і засобів статистичного прогнозування [7–11] застосування ГА в даному аспекті не суперечить логіці і математичному базису, закладеним в цих методах. У зв'язку з цим доцільною є розробка моделі про-

гнозування залежності показників здоров'я населення від обсягів викидів забруднюючих речовин у повітрі із застосуванням ГА та модифікацією одного з операторів генетичного методу [16].

За останні роки запропоновано різні методи та програмні засоби [6–16], що використовують штучні нейронні мережі для прогнозування захворюваності. Проте відомі моделі часто не дозволяють забезпечити прийнятну достовірність результатів прогнозування. Така ситуація перш за все обумовлена тим, що зазвичай, архітектура моделі нейронної мережі, її топологія, значення параметрів обираються на основі експертної оцінки або емпірично. Такими параметрами можуть бути кількість вузлів шару, метод оптимізації мережі, розмір підвибірки, кількість епох навчання мережі та ін. Для підбору оптимальних значень цих параметрів можуть бути використані такі стохастичні методи, як метод рою часток і генетичні алгоритми. Об'єднання генетичних алгоритмів і нейронних мереж відомо в літературі під абревіатурою COGANN (Combinations of Genetic Algorithms and Neural Networks) [16]. Використання ГА для навчання нейронних мереж має такі переваги: генетичні алгоритми малочутливі до зростання розмірності вхідної множини даних, такі методи не вимагають диференційованості цільової функції, на кожній ітерації працюють з множиною рішень, що дозволяє їм більш детально досліджувати простір пошуку та виходити з областей локальних екстремумів. Тому для подолання зазначеної проблеми може бути розроблено генетичний алгоритм для підбору параметрів нейронної мережі.

3 Мета і задачі дослідження

Мета роботи – створення методу синтезу нейромережевих моделей на основі генетичного підходу для прогнозування показників здоров'я населення.

Для досягнення мети були поставлені такі завдання:

– розробити базис нейромережевої моделі залежності показників здоров'я від обсягів викидів забруднюючих речовин;

– розробити метод побудови нейромережевих моделей на основі довгої короткочасної пам'яті;

– виконати експериментальне дослідження запропонованого генетичного методу при синтезі нейромережевих моделей залежності показників здоров'я населення.

4. Розробка базису нейромережевої моделі залежності показників здоров'я від обсягів викидів забруднюючих речовин

Для побудови математичної моделі залежності показників здоров'я населення від обсягів викидів забруднюючих речовин у повітря використано штучні нейронні мережі на основі багатошарового персептрону [16]. У якості вхідних параметрів використовується показники обсягів викидів забруднюючих речовин в атмосферне повітря, кількості населення, середнього віку населення, середньої температури, кількості опадів, кількості лікарів у регіоні і кількості ліжок у стаціонарних відділеннях закладів охорони здоров'я у регіоні [1–3].

Вибір параметрів нейронної мережі, зокрема кількості нейронів прихованого шару, в більшості випадків є доволі складною задачею і виконується, як

правило, на підставі експертної оцінки. Однак існує декілька рекомендацій з цього приводу. Так, Hecht-Neilson [17] для обчислення верхньої мережі кількості прихованих елементів використовував теорему Колмогорова [17], що стверджує, що будь-яка функція n змінних може бути представлена як суперпозиція $2i+1$ одновимірних функцій. Ця мережа h дорівнює подвоєної кількості вхідних елементів і плюс одиниця (4):

$$h \le 2i + 1, \tag{4}$$

де i – кількості вхідних елементів.

Отже, модель залежності має вхідний шар мережі, що містить сім нейронів (по кількості вхідних параметрів). Розроблювану модель залежності (3) з використанням теореми Колмогорова представимо у вигляді [17]:

$$\mathrm{K}_{\mathrm{morb}} = \sum_{j=1}^{2n+1} p_i \left(\sum_{i=1}^{n} d_{ij}\left(x_i\right) \right), \tag{5}$$

де n – кількість вхідних параметрів, p_i та d_{ij} – безперервні функції, причому d_{ij} не залежать від K_{morb}. Ця формула показує реалізацію функцій багатьох змінних як операцію підсумовування і композицію функцій однієї змінної.

Звичайно, застосувати формулу (5) на практиці досить складно. Однак ця формула показує можливість реалізації складної залежності за допомогою відносно простої нейронної мережі, званої багатошаровим персептроном. Отже побудуємо тришаровий персептрон, що має вхідний шар, вихідний шар і прихований шар нейронів, який реалізує функцію активації. Така мережа реалізує наступне відображення [17]:

$$y = \sum_{i=1}^{h} \phi_i f\left(\omega_{i,0} x_1 + \omega_{i,1} x_1 + \omega_{i,2} x_2 + \ldots \omega_{i,n} x_n\right), \tag{6}$$

де φ_i – матриця ваг зв'язків між виходами нейронів прихованого шару і вихідним нейроном мережі, $\omega_{i,n}$ – матриця ваг зв'язків між вхідними нейронами і нейронами прихованого шару, які власне і реалізують функцію активації, f – функція активації нейрона прихованого шару.

Вхідний вектор мережі визначається як набір значень інцидентності, що надходять на вхідні нейрони за одну ітерацію навчання. Вихідний вектор мережі – це набір значень інцидентності на вихідних нейронах. Для розрахунку числа нейронів в прихованих шарах використано формулу оцінки числа семантичних ваг U_s для багатошарових персептронів з сигмоїдальними передавальними функціями [17]:

$$\frac{mN}{1+\log_2 N} \le U_s \le m\left(\frac{N}{m}+1\right)\left(n+m+1\right)+m, \tag{7}$$

де n – розмірність вхідного сигналу, m – розмірність вихідного сигналу, N – число елементів навчальної вибірки.

Число нейронів в прихованому шарі оцінюється за формулою:

$$U = \frac{U_s}{n+m}. \tag{8}$$

Отже, розроблювана модель залежності має прихований шар, що містить 12 нейронів (12<2*7), та вихідний шар, що містить один нейрон [18].

Одним з найважливіших аспектів нейронних мереж є функція активації (activation function), яка привносить в мережу нелінійність, роблячи їх універсальними апроксиматорами функцій [19].

Функція активації – це спосіб нормалізації вхідних даних. Тобто, якщо на вході велика кількість даних, пропустивши їх через функцію активації, отримаємо на виході дані в потрібному діапазоні. В розроблюваній мережі нейрони вхідного і прихованого шару у якості функції активації використовують ReLU [20] (Rectifier activation function). Перевагою використання функції активації ReLU є те, що вона позбавлена ресурсоємних операцій, відсутнє розростання або загасання градієнта та забезпечує швидке навчання.

Таким чином, перша модель буде складатися з зовнішнього шару (сім нейронів), одного прихованого повнозв'язаного шару (12 нейронів) та вихідного шару (один нейрон). Схема створеної мережі відображена на рис. 1.

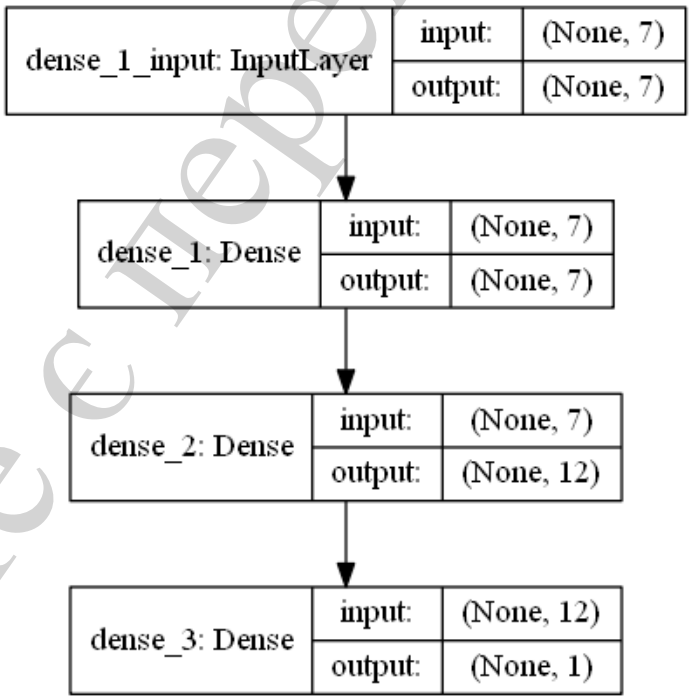

Рис. 1. Схема нейронної мережі з одним прихованим шаром

В ході роботи було створено декілька багатошарових нейромережевих моделей прямого поширення. Потім, у створену раніше модель було додано ще

один повнозв'язний прихований шар з 12 нейронів, який у якості функції активації також використовує ReLU. Схема створеної мережі відображена на рис. 2.

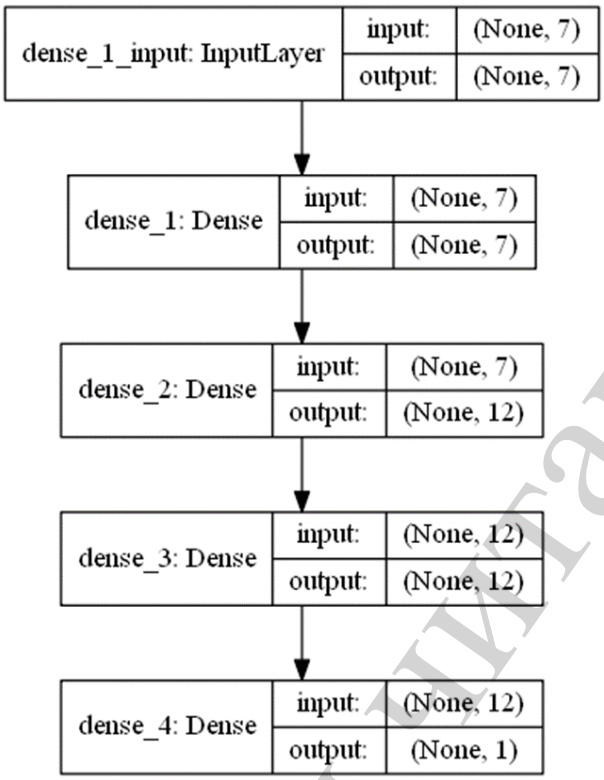

Рис. 2. Схема нейронної мережі з двома прихованими шарами

Отже, створено модель нейронної мережі з двома прихованими шарами. Модель складається з вхідного шару, що містить 7 нейронів (на кожний з них подається вхідний сигнал), 2 прихованих повнозв'язних шари (кожний з яких містить по 12 нейронів) і вихідного шару, що складається з одного нейрону (рис. 2).

Перенавчання є однією з суттєвих проблем, що ускладнюють практичне застосування нейронних мереж. Одним з прийомів запобігання ефекту перенавчання нейронної мережі є метод виключення (Dropout), що полягає у виключенні деяких нейронів мережі підчас процесу навчання [21].

Головна ідея Dropout – замість навчання однієї нейронної мережі навчити ансамбль декількох глибоких нейронних мереж (Deep Neural Network, DNN), а потім усереднити отримані результати.

Мережі для навчання отримуються за допомогою виключення з мережі (dropping out) нейронів з ймовірністю p, таким чином, ймовірність того, що нейрон залишиться в мережі, становить $q=1–p$. Виключення нейрона означає, що при будь-яких вхідних даних або параметрах він повертає 0.

Виключені нейрони не вносять свій внесок в процес навчання ні на одному з етапів алгоритму зворотного поширення помилки, тому виключення хоча б одного з нейронів рівносильно навчанню нової нейронної мережі.

В створену модель с двома прихованими шарами було додано виключення після першого прихованого шару (виключено 50% нейронів). Схема створеної мережі відображена на рис. 3.

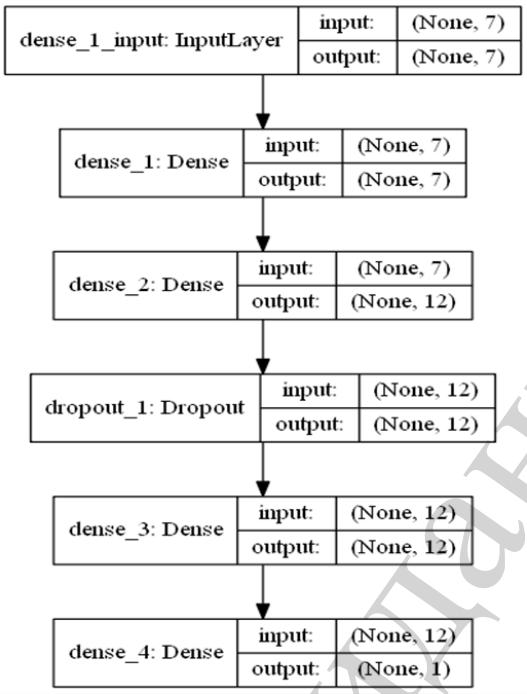

Рис. 3. Схема нейронної мережі з двома прихованими шарами і дропаутом

Таким чином, створено модель нейронної мережі з двома прихованими шарами і дропаутом. Вхідний шар мережі містить 7, повнозв'язний прихований шар (12 нейронів), дропаут (вимикає половину всіх нейронів шару), повнозв'язний прихований шар (12 нейронів і вихідний шар, що складається з одного нейрону. У якості функції активації використано ReLU. Первинна ініціалізації синаптичних ваг відповідає нормальному розподілу.

5. Розробка методу побудови нейромережевих моделей на основі довгої короткочасної пам'яті

Побудовані моделі (рис. 1–3) не вирішують проблему довготривалої залежності даних, що подаються на вхід. Оскільки представлені дані можна розглядати як часовий ряд, тому що значення розглянутих параметрів змінюються в часі. Для аналізу і прогнозування часових рядів можна використовувати моделі на основі нейронних мереж довгої короткочасної пам'яті (LSTM–long short-term memory) [21].

Розглянемо докладніше структуру LSTM-шару. Основним елементом такої мережі є запам'ятовуючий блок, який одночасно зі станом мережі h, обчислюється на кожному кроці, використовуючи поточне вхідне значення x^t і значення блоку на попередньому кроці r^{t-1}. Вхідний фільтр i^t визначає, наскільки значення блоку пам'яті на поточному кроці повинно впливати на результат. Значення фільтра варіюються від 0 (повністю ігнорувати вхідні значення) до 1, що забезпечується областю значень сигмоїдальною функцією:

$$i^t = \sigma\left(Y^i x^t + C^i h^{t-1}\right). \quad (9)$$

де C, Y – навчальні параметри нейронної мережі.

«Фільтр забування» (forget gate) дозволяє виключити при обчисленнях значення пам'яті попереднього кроку:

$$f^t = \sigma\left(Y^f x^t + C^f h^{t-1}\right). \quad (10)$$

На основі всіх даних, що надходять в момент часу t, обчислюється стан блоку пам'яті r^t на поточному кроці, використовуючи фільтри:

$$\tilde{r}^t = \tanh\left(Yx^t + Ch^{t-1}\right), \quad (11)$$

$$r^t = f^t \cdot r^{t-1} + i^t \cdot \tilde{r}^t. \quad (12)$$

Вихідний фільтр (output gate) аналогічний двом попереднім і має вигляд:

$$o^t = \sigma\left(Y^0 x^t + C^0 h^{t-1}\right). \quad (13)$$

Підсумкове значення LSTM-шару визначається вихідним фільтром (13) і нелінійної трансформацією над станом блоку пам'яті:

$$h^t = o^t \cdot \tanh\left(r^t\right). \quad (14)$$

Мережа отримує на вхід вісім параметрів – дані за попередній період: показник захворюваності, обсяг викидів забруднюючих речовин в атмосферне повітря від стаціонарних джерел, кількість населення, середній вік населення, середню температуру і середню кількість опадів, кількість лікарів в регіоні, кількість ліжок у стаціонарних відділеннях закладів охорони здоров'я у регіоні. Прихований LSTM шар складається з двадцяти нейронів, а вихідний шар з одного нейрону. У якості алгоритму оптимізації використано алгоритм adam [21]. Adam – це алгоритм оптимізації, який може використовуватися замість класичної процедури зниження випадкового градієнта, щоб оновити ітераційну вагу мережі, засновану на навчальних даних [22]. Алгоритм поєднує у собі переваги таких розширень класичного градієнтного спуску, як адаптивний градієнтний алгоритм (AdaGrad) і розповсюдження середнього квадрату (RMSProp).

Головна відмінність алгоритму полягає в середніх значень як градієнтів, так і других моментів градієнтів. Оновлення синаптичних ваг мережі при використанні алгоритму adam здійснюється таким чином:

$$m_\omega^{(t+1)} = \beta_1 m_\omega^{(t)} + \left(1-\beta_1\right)\nabla_\omega L^{(t)}, \quad (15)$$

$$v_{\omega}^{(t+1)} = \beta_2 v_{\omega}^{(t)} + \left(1-\beta_2\right)\left(\nabla_{\omega} L^{(t)}\right)^2, \quad (16)$$

$$\hat{m}_{\omega} = \frac{m_{\omega}^{(t+1)}}{1-\beta_1^t}, \quad (17)$$

$$\hat{v}_{\omega} = \frac{v_{\omega}^{(t+1)}}{1-\beta_2^t}, \quad (18)$$

$$\omega^{(t+1)} = \omega^{(t)} - \eta \frac{\hat{m}_{\omega}}{\sqrt{\hat{v}_{\omega} + \varepsilon}}, \quad (19)$$

де β_1, β_2 – гіперпараметри, що позначають експоненціальні темпи розпаду на момент оцінки; η – початковий рівень навчання; ε – це невелика константа, введена для чисельної стабільності; m_{ω} – експоненціальне рухоме середнє градієнта; v_{ω} – експоненціальне середнє квадрата градієнта; $\nabla_{\omega} L^{(t)}$ – значення градієнта за часом t; ω – вектор параметрів градієнтного спуска [23].

Зазвичай, архітектура моделі нейронної мережі, її топологія, значення макропараметрів обирається на основі експертної оцінки або емпірично. Для мереж такими параметрами можуть бути кількість вузлів шару довгої короткочасної пам'яті, оптимізатор, розмір підвибірки і кількість епох навчання.

Для вирішення даної проблеми було розроблено модифікацію генетичного алгоритму для оптимізації параметрів створених нейронних мереж.

Модель прогнозування будується на основі накопичених даних таких чинників: головних m_1, m_2 ,…, m_n (обсягів викидів) і допоміжних a_1, a_2, ..., a_n (кількість населення, опадів, лікарів, ліжок у стаціонарних відділеннях), де n – довжина актуальної частини ряду (кількість спостережень часового ряду), що становить 20–30 значень. Представимо ці дані як нечіткі часові ряди $F_1(t)$ і $F_2(t)$, де $F_1(t)$ відповідає головним, а $F_2(t)$ – допоміжним чинникам прогнозування [23]. Тоді залежність виду:

$$F_1(t) = \left(\left(F_1 - k\right), F_2\left(t-k\right)\right), ..., \left(\left(F_1\left(t-2\right)\right), F_2\left(t-2\right)\right), \left(\left(F_1\left(t-1\right)\right), F_2\left(t-1\right)\right), \quad (20)$$

називається факторною моделлю прогнозування k-го порядка на основі нечітких часових рядів [23].

Як випливає з аналізу джерел [16–23], для знаходження оптимального рішення за допомогою ГА вимагається "породити" близько двох – трьох мільйонів особин. Проте висока ресурсоємність визначення значення цільової функції для кожної особини може сильно збільшити час пошуку оптимуму.

Для вирішення цієї проблеми було прийнято рішення розробити модифікацію ГА, що дозволить істотно скоротити час оптимізації. У розробленій модифікації запропоновано змінені оператори схрещування, відбору і мутації, а та-

кож новий генетичний оператор селекції другого порядку за величиною вірогідності мутації.

Сутність запропонованої модифікації генетичного методу полягає в додаванні до каріотипу кожної особини ще однієї хромосоми з таким самим генним складом, тобто використовувати диплоїдний набір, що складається з двох гомологічних хромосом. Обидві хромосоми піддаються дії ти ж самих операторів з однаковими параметрами. Таким чином, при схрещуванні каріотип нащадка буде також складатися з двох гомологічних хромосом, як і у його батьків. Домінуючий ген в запропонованій модифікації обирається випадковим чином з двох алельних генів і використовується для обчислення значення функції пристосованості – фітнес функції, тобто, говорячи в термінах біології, визначає фенотип особини [23].

Позначимо особу як a_n^t, де n – означає номер особи, t – деякий момент часу еволюційного процесу. В якості вектору керуючих змінних приймемо $\overline{x}=(x_1, x_2, \ldots, x_m)$ – це найменша неподільна одиниця, що характеризує в математичній моделі (3) внутрішні параметри на кожному t-му кроці пошуку оптимального рішення.

Для опису особин введемо два типи варіабельних ознак, що відображають якісні і кількісні відмінності між особинами за ступенем їх вираженості. Якісні ознаки особин a_n^t, визначаються із узагальненої моделі (3) як $s(\overline{x})$, де кожній точці $\overline{x}$ відповідає a_n^t. В якості гена приймемо комбінацію $s_i(a_i)$, яка визначає фіксоване значення керуючої змінної x_i. Кожна особина характеризується m генами, а $s(\overline{x})=(s_1, s_2, \ldots, s_m)$ можна інтерпретувати хромосомою, що містить n зчеплених між собою генів, які слідують один за одним у строго визначеній послідовності. Хромосому особини a_n^t будемо позначати як x_n^t [23], тобто

$$x_n^t = x\left(a_n^t\right) = \left(x_1\left(a_n^t\right), x_2\left(a_n^t\right), \ldots, x_m\left(a_n^t\right)\right) = s\left(\overline{x}\right) = \left(s_1, s_2, \ldots, s_m\right). \quad (21)$$

Кількісні ознаки – ознаки, які виявляють мінливість, в зв'язку з чим ступінь їх вираженості можна охарактеризувати числом та розраховуються в роботі за формулою:

$$d\left(x_i^t, x_j^t\right) = \sum_{n=1}^{m} x_n\left(a_i^t\right) \cdot x_n\left(a_j^t\right), \quad (22)$$

де a_i^t a_j^t – особи, x_i^t x_j^t – гени, що нерівні за своїми значеннями, m – кількість позицій [23].

На першому етапі відбувається ініціалізація популяції. Генний склад кожної з двох гомологічних (H,H') хромосом особин обирається випадковим чином. Для визначення фенотипу особини з кожної алелі обирається ген, що буде визначатися як домінантний і буде визначати фенотип особини, тобто приймає участь в обчисленні функції пристосованості особини. Визначення фенотипу особини можна представити у вигляді формули [21–23]:

$$F_j = \sum_{i=1}^{m} rand\left[H_j g_i ; H_j^{'} g_i \right], \tag{23}$$

де F_j – фенотип j-ої особини, m – кількість генів у хромосомах та $H_j g_i$ – i-ий ген в парі гомологічних хромосом j-ої особини.

Таким чином, фактично визначаються аргументи фітнес-функції особини. Після обчислення функцій пристосованості і відбору особин популяції проводиться схрещування. Генотип особини-нащадка має таку ж структуру, як і генотип батьків, тобто складається з двох гомологічних хромосом. До нащадків застосовується оператор мутацій. При цьому, мутувати може будь-яка алель пари гомологічної хромосом, але в кожній алелі мутує лише один ген [24].

Далі еволюцію популяції P^t будемо представляти як чергування поколінь, в процесі якого особини змінюють свої варіабельні ознаки:

$$\eta_{сер}(t) = \frac{1}{t} \sum_{n=1}^{m} \eta\left(a_n^t\right), \tag{24}$$

де сукупність з m генотипів всіх особин $(a_1^t, a_2^t, ..., a_m^t)$, що утворюють популяцію P^t і хромосомний набір $(x_1^t, x_2^t, \ldots, x_m^t)$, який повністю містить в собі генетичну інформацію про популяції P^t в цілому.

Процедура вибору «найкращого» рішення з популяції P^t враховує не тільки значення функції пристосованості F_j, але і структуру хромосом x_i^t, отже її можна представити у вигляді [25]:

$$d\left(a^t, a_i^t\right) = \min_{l=1,m} d\left(x\left(a^t\right), x\left(a_l^t\right)\right) \tag{25}$$

при умові, що $\eta(a_1^t) < \eta(a^t)$, де a_i – «найкраща» особина в популяції P^t, a_i^t – особина, що виключається з популяції P^t, $d(x(a^t), x(a_1^t))$ – міра «близькості» генотипів особин.

Далі, як і в класичному методі, цикл повторюється до настання умов закінчення виконання оптимізації.

Підсумовуючи, можна сказати, що запропонований метод відрізняється від класичного генетичного методу використанням не однієї хромосоми, а пари гомологічних хромосом, і додаванням етапу визначення тих генів алелі, які будуть приймати участь у визначенні значення функції пристосованості особини. Результатом такої модифікації є підтримання досить високої варіабельності ознак (генів) в популяції (генофонду популяції) під час еволюції, яка, в той же час, може мати невеликий вплив на фенотип особин.

Зазначену модифікацію методу було використано для оптимізації LSTM [26] нейронної мережі: кількості вузлів мережі, функції оптимізації при навчанні, розміру підвибірки та кількості епох навчання.

Іншою запропонованою модифікацією генетичного методу є модифікація оператора мутацій. На відміну від класичного застосування цього оператору, коли мутації піддаються усі особини генерації з певною імовірністю, пропону-

ється ввести поняття мутаційної стійкості особини і здійснюється згідно з таким розподілом:

$$x_i^1 = \begin{cases} x_i^{'}, \ P\left(x_i^{'}\right) = \dfrac{\eta\left(x^{'}\right)}{\eta\left(x^{'}\right) + \eta\left(x^{''}\right)}; \\ x_i^{''}, P\left(x_i^{''}\right) = 1 - P\left(x^{'}\right), \end{cases} \tag{26}$$

де x_i^1 – нащадок, $\eta(x')$, $\eta(x'')$ – значення функції пристосованості, що оцінюють, відповідно, батьківські кодування x' і x'' [27–29].

Обчислене значення функції пристосованості особини (фітнес функції) може бути інтерпретоване як значення мутаційної стійкості особини. Отже, пропонується на кожній ітерації методу після обчислення функції пристосованості проводити ранжування особин отриманої генерації за значенням мутаційної стійкості. На відміну від класичного оператора, на початку вказується не вірогідність мутації, а частка особин, які піддаються дії оператора (25).

$$\mathrm{K}_{\mathrm{mut}} = H_{gen} \cdot R_{mut}, \tag{27}$$

де K_{mut} – кількість особин, що піддаються дії мутації, H_{gen} – кількість особин отриманої генерації, R_{mut} – частка особин генерації, що піддаються дії мутації [29–31].

Фактично, пропонується застосовувати оператор тільки до особин з найнижчим значенням функції пристосованості. У цьому випадку при потраплянні популяції в область локального екстремуму функції, застосований оператор мутацій має забезпечити вихід з такої області. При чому він не змінює отримані найкращі особини на момент застосування значень, а тільки за рахунок більш слабких особин проводить пошук. Визначена частка особин, що піддаються дії оператора, повинна бути достатньою для забезпечення створення потенціалу для подальшої еволюції усієї популяції.

Такі мутації повинні бути більш «м'якими» у сенсі збереження знайдених на попередніх ітераціях алгоритму найкращих значень і мають нівелювати небезпеку втрати екстремуму функції при їх застосуванні не зупиняючи пошук нових кращих значень.

Таким чином, розроблено модифікований генетичний метод для параметричного синтезу моделі на основі нейронної мережі довгої короткочасної пам'яті, яка використовує модифікацію оператора мутації. Модифікований оператор мутації дозволяє проводити пошук оптимальних значень, виключаючи втрату надбаних під час пошуку кращих рішень.

6. Експериментальне дослідження модифікованого генетичного методу при синтезі моделей залежності показників здоров'я населення

Для розробки і тестування моделі залежності показників здоров'я від обсягів викидів забруднюючих речовин використано статистичну інформацію про обсяги викидів забруднюючих речовин та діоксиду вуглецю в атмосферне пові-

тря від стаціонарних джерел забруднення. Також використовувалася інформація про рівень захворюваності по таким показникам, як кількість випадків захворювань системи кровообігу (зареєстровані в амбулаторних установах), кількість нових випадків туберкульозу і кількість зареєстрованих випадків раку. Враховуючи той факт, що наведені дані виражені в абсолютних значеннях, доцільно зробити поправку на кількість населення у регіоні. Тому використано також дані про кількість населення у регіонах по роках [2].

В розроблених моделях (рис. 1 – 3) застосовуються статистичні дані про середню температури в регіоні і рівень опадів, кількість лікарів у регіоні, кількість ліжок у стаціонарних відділеннях.

Для вирішення задачі обрано програмне середовище на основі мови програмування Python, для прискорення виконання обчислень використано GPU NVIDIA GeForce GTS 450 з підтримкою архітектури CUDA [41]. Для зручної роботи з масивами даних і формування датасетів використано бібліотеку NumPy – пакет Python для наукових обчислень. Для будування моделей нейронної мережі та роботи з ними обрано бібліотеку Keras [42] і бібліотеку Theano [43].

У якості оцінки прогнозних моделей використано mae (Mean Absolute Error) – середню абсолютну помилку [15]. До початку створення та тестування моделей було проведено первинну обробку даних. Враховуючи різну розмірність даних, було проведено стандартизацію вхідних даних. Дані були перетворені таким чином, щоб їх середнє значення дорівнювало 0, а дисперсія 1. В ході роботи було створено і досліджено декілька моделей на основі штучних нейронних мереж. Результат навчання і роботи першої створеної мережі відображено на рис. 4.

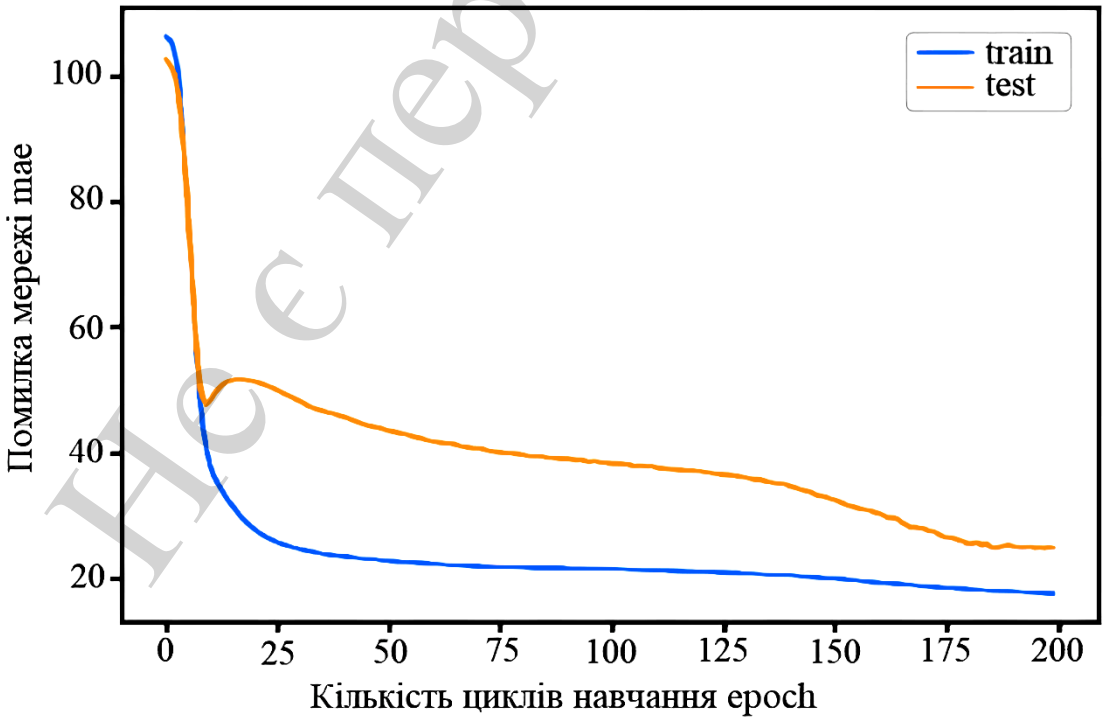

Рис. 4. Значення метрики мережі (mae) для показника «Кількість нових випадків туберкульозу»

З рис. 1 видно, що спостерігається поступове зниження значень помилки впродовж тренування мережі. Вочевидь, в районі 10–15 епохи тренування досягає локального екстремуму. Про локальність мінімуму помилки може свідчити подальше поступове зменшення помилки мережі. Отже, доцільним в даному випадку є подальше тренування моделі.

Для покращення метрики нейронних мереж, їх збіжності, витрат на навчання, тощо існує декілька підходів пов'язаних з підбором оптимальної топології мережі та методів навчання. Так, у створену раніше модель було додано ще один повнозв'язний прихований шар з 12 нейронів.

Друга модель складається з вхідного шару, що містить 7 нейронів, 2 прихованих повнозв'язних шарів (кожний з яких містить по 12 нейронів) і вихідного шару, що складається з одного нейрону. Таким чином, мережа має 264505 параметрів (синаптичних ваг), що можуть бути треновані. Навчання мережі з двома прихованими шарами проводилося впродовж 100 епох (розмір підвибірки 75) і сплітом валідаційної вибірки, що дорівнював 0,1. Результати навчання мережі наведені на рис. 5.

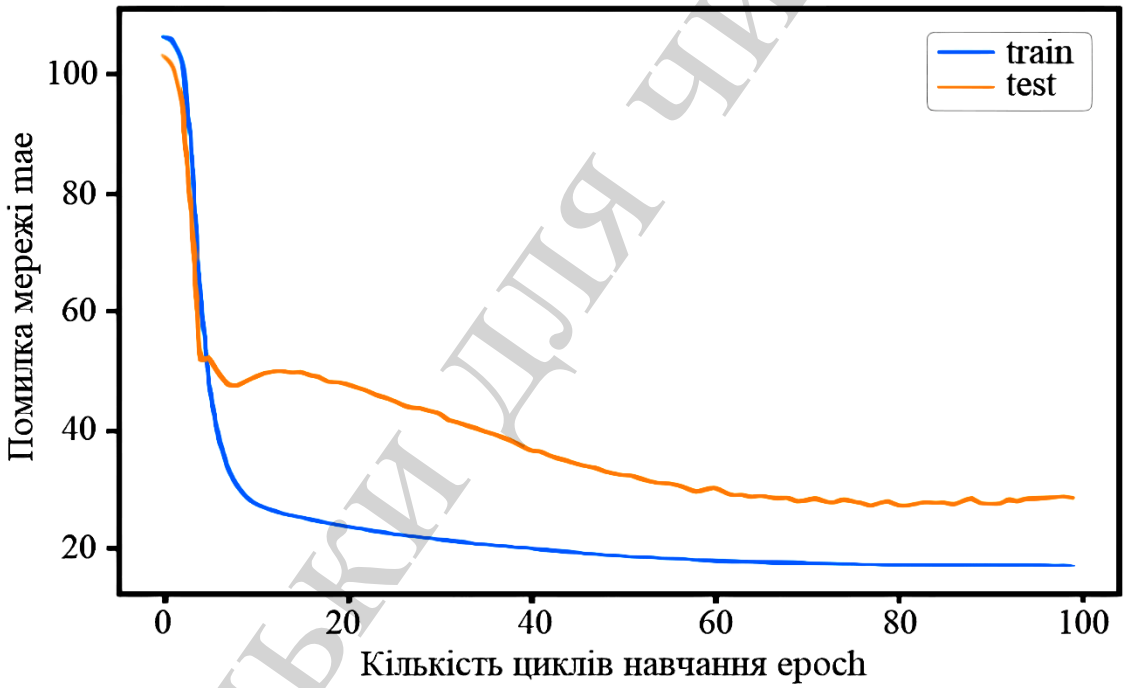

Рис. 5. Значення метрики мережі (mae) для показника «Кількість нових випадків туберкульозу»

В даному випадку спостерігається схожа з попередньою моделлю картина – поступове зниження значень помилки впродовж тренування мережі. Наприкінці тренування спостерігається збіг мережі. Враховуючи попередній досвід тренування MLP мережі, тренування відбувалося в продовж 100 епох. Як і в попередньому випадку, для показника «Число нових випадків туберкульозу» спостерігається досягнення локального мінімуму помилки.

Одним з прийомів запобігання ефекту перенавчання нейронної мережі є метод виключення (Dropout) [31], що полягає в виключенні деяких нейронів мережі підчас процесу навчання.

Третя модель складається з вхідного шару, що містить 7 нейронів, 2 прихованих повнозв'язних шарів (кожний з яких містить по 12 нейронів), дропаута і вихідного шару, що складається з одного нейрону. Таким чином, мережа має 266273 параметри (синаптичних ваг), що можуть бути треновані. Результати навчання мережі відображені на рис. 6.

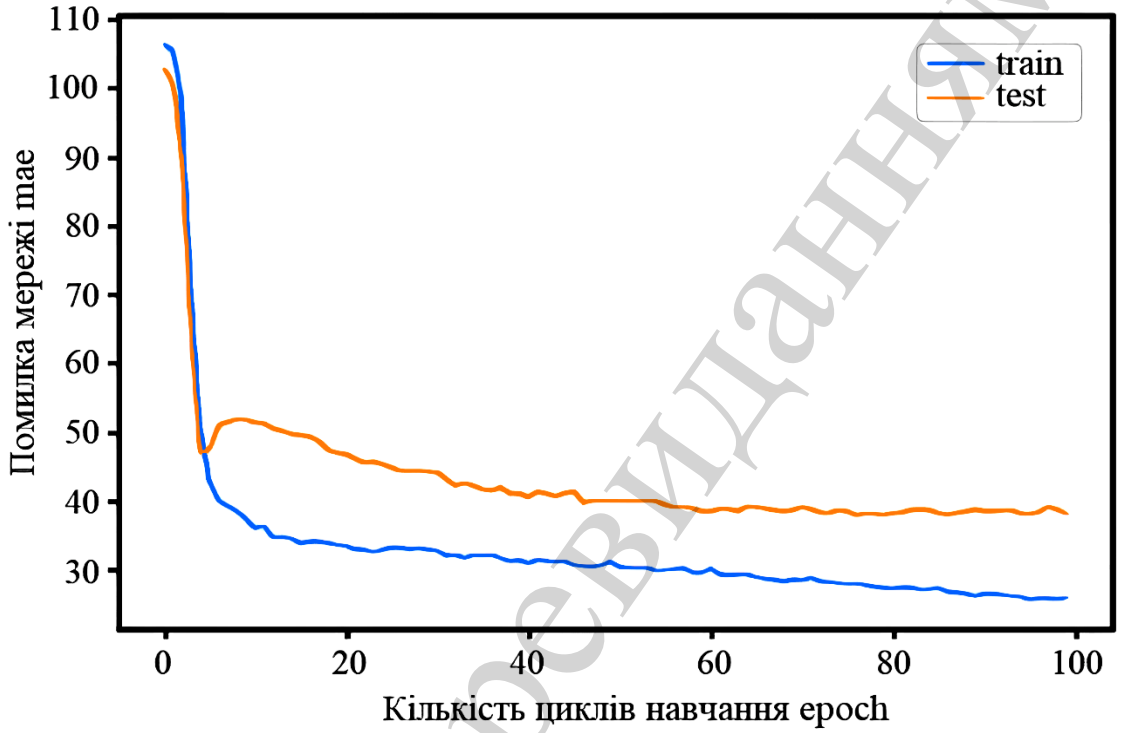

Рис. 6. Значення метрики мережі (mae) для показника «Кількість нових випадків туберкульозу»

Як і в попередніх випадку, під час тренування спостерігається збіг мережі, але дещо раніше – приблизно під час 60–70 епохи навчання. В усіх трьох випадках для показника «Число нових випадків туберкульозу» спостерігається знаходження локального мінімуму помилки мережі, що вочевидь є особливістю моделі на основі багатошарового перцептрону для даного показника захворюваності. Крім цього, завдяки використанню дропауту, крива тренування втрачає плавність і перетворюється на ломану.

Представлені дані можна розглядати як часовий рядок, тобто значення розглянутих параметрів змінюються в часі. Для аналізу і прогнозування часових рядів можна використовувати моделі на основі нейронних мереж довгої короткочасної пам'яті [16].

Мережа з використанням LSTM шару отримує на вхід вісім параметрів. Прихований LSTM шар складається з двадцяти нейронів, а вихідний шар з одного нейрону. Результати випробувань моделі наведені на рис. 7.

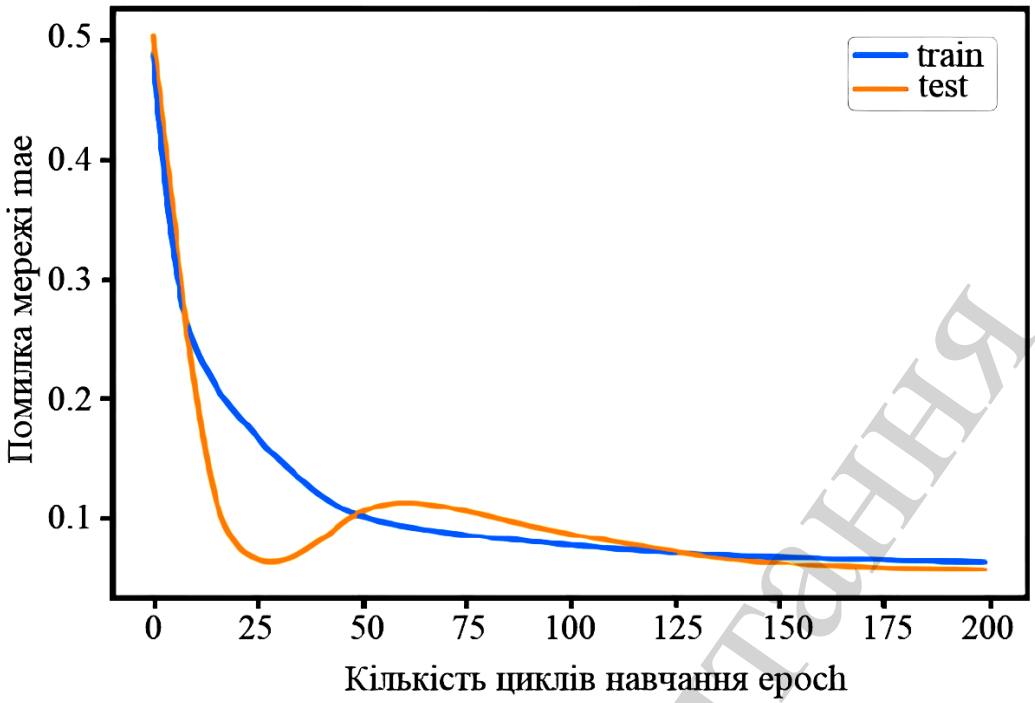

Рис. 7. Значення метрики LSTMмережі (mae) підчас навчання для показника «Кількість випадків захворювань системи туберкульозу»

На рис. 7 відображено зміну значення помилки LSTM мережі (mae) підчас навчання для показника «Кількість випадків захворювань системи туберкульозу». Для даного показника спостерігається досягнення локального мінімуму с подальшим виходом з нього. Під кінець тренування спостерігається відсутність подальшого зменшення значення помилки моделі, тому можна вважати, що під час тренування досягнуто глобального мінімуму помилки, а мереже вважається тренованою. У табл. 1 відображено порівняльні результати значень середньої абсолютної помилки (MAE), отримані під час випробування різних типів моделей (логістична регресія, багатошарові нейромережеві моделі та ін.), створених в ході дослідження.

Таким чином, з табл. 1 видно, що модель на основі штучної нейронної мережі короткої довгочасної пам’яті з 50 вузлами ДКЧП шару дає найменшу помилку у порівнянні з наведеними методами. А саме помилка передбачення числа нових випадків туберкульозу (MAE становить 6.139) та числа захворювань системи кровообігу (MAE становить 441.889), що є прийнятним показником. А для передбачення числа всіх зареєстрованих випадків раку найменша середня абсолютна помилка (становить 156.387) відповідає випадковий ліс.

У ході роботи для оптимізації мережі довгої короткочасної пам’яті було використано метод рою часток. Результати виконання алгоритму (визначення найменшого значення помилки мережі на кожній ітерації алгоритму) відображено на рис. 9.

Відзначимо, що побудована мережа довгої короткочасної пам’яті на тестовій вибірці дозволила отримати значення помилки RMSE на рівні 127.087, що є прийнятним показником для розв’язуваної практичної задачі.

Таблиця 1

Значення середньої абсолютної помилки під час випробувань моделей, створених в ході дослідження проблеми

Прогнозований параметр	Тип прогнозуючої моделі								
	Логістична регресія	Метод опорних векторів	Метод найменших квадратів	Випадковий ліс	Метод найближчого сусіда	Багатошаровий перцептрон з одним прихованим шаром (128 нейронів)	Багатошаровий перцептрон з двома прихованими шарами (128, 1024 і 128 нейронів)	Багатошаровий перцептрон з двома прихованими шарами і дропаутами (128, дропаут (0,5), 1024, дропаут (0,5), 128 нейронів)	Мережа довгої короткочасної пам'яті з 50 вузлами ДКЧП шару
Число нових випадків туберкульозу	29.764	40.271	22.957	7.671	8.236	23.445	21.675	22.3047	6.139
Число випадків захворювань системи кровообігу	3814.174	2794.433	1789.727	571.018	573.004	1766.470	1752.416	1620.676	441.889
Число всіх зареєстрованих випадків раку	1400.357	1050.690	367.381	156.387	210.157	336.419	338.953	272.465	226.096

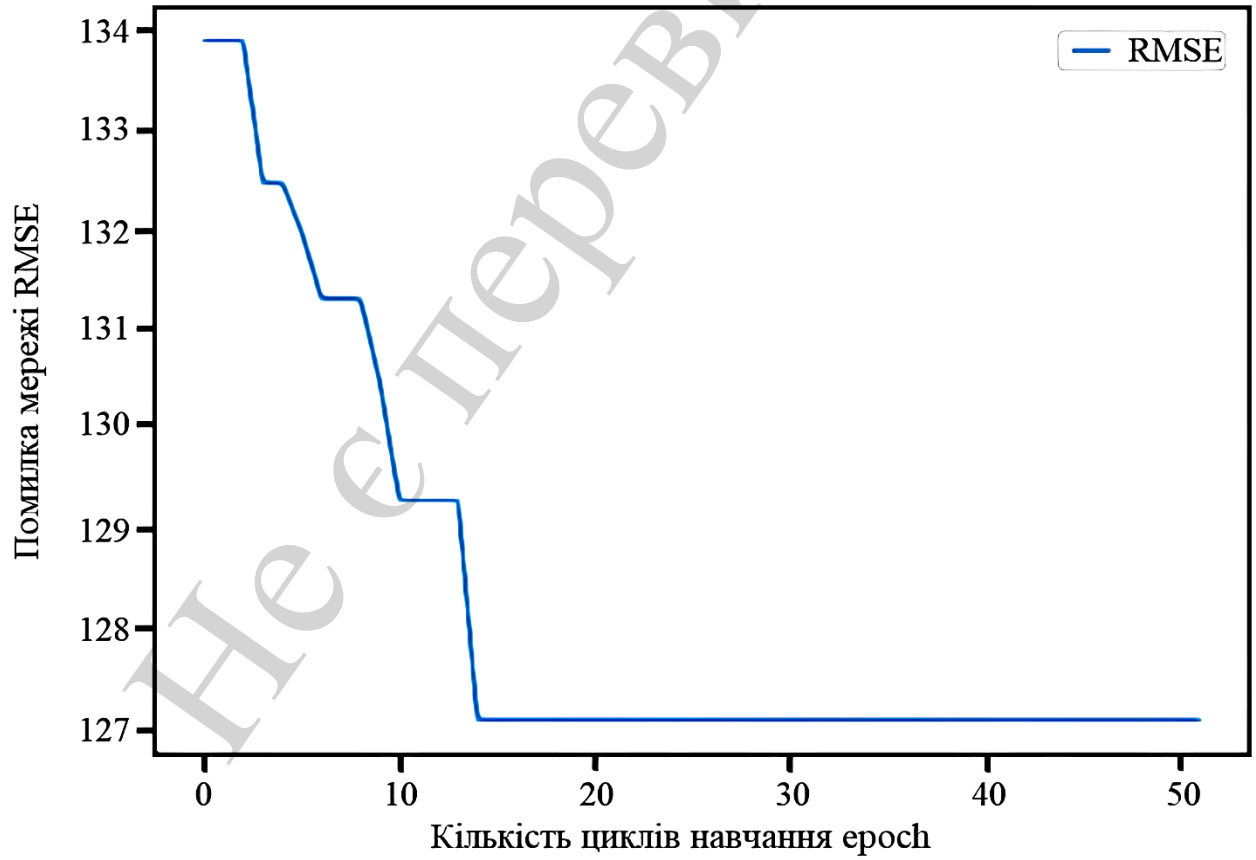

Рис. 9. Результат роботи алгоритму рою часток для оптимізації мережі довгої короткочасної пам’яті

Таким чином результати досліджень показали, що у якості моделі залежності показників здоров'я від обсягів викидів забруднюючих речовин, можна використовувати модель на основі штучної нейронної мережі довгої короткочасної пам'яті. Для підбору таких параметрів, як кількість вузлів LSTM шару, методу оптимізації мережі, розміру підвибірки і кількості епох навчання мережі бажаним є використання модифікованого генетичного методу.

7. Обговорення результатів дослідження модифікованого генетичного методу

Дивлячись на порівняльний аналіз створених моделей (табл. 1) можна побачити, що найкращі результати значення середньої абсолютної помилки для прогнозування числа нових випадків туберкульозу показала мережа довгої короткочасної пам'яті. А саме MAE складає 6.139, що є прийнятним показником у порівнянні з методом опорних векторів, помилка якого складає 40.271. Для прогнозування кількості випадків захворювань системи кровообігу найкращі результати були отримані при використанні мережі довгої короткочасної пам'яті (MAE складає 441.889). При прогнозуванні кількості всіх зареєстрованих випадків раку було отримано найменшу помилку у випадковому лісі, що становить 156.387, у порівнянні з логістичною регресією, яка складає 1400.357.

Результати аналізу стабільності роботи модифікації генетичного алгоритму наведені на рис. 1–4. Проводилося 20 запусків алгоритму з різним числом ітерацій. З наведених графіків видно, що впродовж тренування мережі значення абсолютної помилки зменшуються та наприкінці тренування спостерігається збіг мережі, що призводить до досягнення локального мінімуму с подальшим виходом з нього. Під кінець тренування (рис. 4) виникає відсутність подальшого зменшення значення помилки моделі, тому можна вважати, що під час тренування досягнуто глобального мінімуму помилки, а мережа вважається тренованою. Крім цього, завдяки використанню дропауту (Dropout), крива тренування втрачає плавність і перетворюється на ломану.

Як видно з рис. 5, під час роботи алгоритму рою часток для оптимізації мережі довгої короткочасної пам'яті було отримано найменше значення помилки (RMSE) – 127.08, що є прийнятним показником.

Таким чином, запропонований модифікований генетичний метод дозволяє підвищити точність прогнозування та зменшити час навчання при синтезі моделей залежності показників здоров'я населення від обсягів викидів забруднюючих речовин у повітря. Це досягається за рахунок використання у розроблених модифікованих методах нових евристичних процедур, зокрема запропоновано використання диплоїдного набору хромосом популяції, яка еволюціонує. Така модифікація робить залежність фенотипу особини від генотипу менш детермінованою і, таким чином, сприяє збереженню різноманітності генофонду популяції і варіабельності ознак фенотипу впродовж виконання алгоритму. Крім цього, запропоновано модифікацію генетичного оператору мутації. На відміну від класичного методу, особини, які піддаються дії оператору мутації, обираються не випадковим чином, а у відповідності до їх мутаційної стійкості, що відповідає значенню функції прис-

тосованості особини. Це дозволило підвищити показник точності у порівнянні з базовою версією генетичного алгоритму.

Недоліком запропонованого модифікованого генетичного методу, розробленого та дослідженого у цій роботі, є необхідність витрачання великого часу при обробці великих масивів даних, що при розв'язанні деяких практичних завдань є неприпустимим. Таким чином, обмеженнями на використання запропонованого модифікованого генетичного методу є невеликі обсяги оброблюваних даних.

Розвиток даного дослідження може бути пов'язаний з усуненням зазначених недоліків, обумовлених практичним порогом використання запропонованого модифікованого генетичного методу для побудови моделей на основі нейронної мережі довгої короткочасної пам'яті. Для цього доцільно розробити його паралельну реалізацію, що дасть можливість суттєво (в рази) збільшити швидкість роботи методу. Супутні проблеми, які можуть виникнути при розробленні паралельних модифікацій генетичного методу для побудови моделей на основі нейронної мережі довгої короткочасної пам'яті, пов'язані з необхідністю планування ресурсів паралельної комп'ютерної системи. Вони призводять до збільшення вимог до апаратного забезпечення, задіяного в процесі генетичної оптимізації.

8. Висновки

1. Розроблено моделі залежності показників здоров'я від обсягів викидів забруднюючих речовин на основі штучних нейронних мереж. Перша побудована модель складається з одного прихованого шару. Під час випробування даної моделі було отримано значення середньої абсолютної похибки, що складає 708.78. Далі в роботі було створено модель з двома прихованими шарами. Друга модель під час випробування показала середнє значення абсолютної похибки, що складає 721.01. Також створено модель з двома прихованими шарами і дропаутами. Під час випробування даної моделі було отримано значення середньої абсолютної похибки, що складає 638.5. Потім було побудовано модель з використанням довгої короткочасної пам'яті з 50 вузлами ДКЧП шару. Під час випробування даної моделі отримано такі значення середньої абсолютної похибки 647.13. Порівнюючи отримані показники з відомими методами, такими як логістична регресія, методи опорних векторів, метод найкращих квадратів, можна побачити, що розроблені моделі в роботі дають кращий результат.

2. Розроблено метод побудови нейромережевих моделей на основі довгої короткочасної пам'яті. Запропонований метод використовує генетичний підхід для параметричного синтезу нейромоделей на основі довгої короткочасної пам'яті. Принципова відмінність запропонованого генетичного алгоритму від існуючих модифікацій полягає у використанні диплоїдного набору хромосом в особин популяції, яка еволюціонує. Така модифікація робить залежність фенотипу особини від генотипу менш детермінованою і, врешті, сприяє збереженню різноманітності генофонду популяції і варіабельності ознак фенотипу впродовж виконання алгоритму. Результатом такої модифікації є підтримання досить високої варіабельності ознак (генів) в популяції (генофонду популяції) під час еволюції, яка, в той же час, може мати невеликий вплив на фенотип особин. У запропонованому методі використовується модифікований генетичний опера-

тор мутацій, в якому, на відміну від існуючих підходів до реалізації таких операторів, особини, які піддаються дії мутації, обираються не випадковим чином, а у відповідності до їх мутаційної стійкості, що відповідає значенню функції пристосованості особини. Таким чином, мутують «слабкіші» особини, а геном «сильних» особин залишається незмінним. У цьому випадку зменшується вірогідність втрати досягнутого впродовж еволюції екстремуму функції внаслідок дії оператора мутацій, а перехід до нового екстремуму здійснюється у випадку накопичення достатньої питомої ваги «кращих» ознак в популяції. Така модифікація оператора дозволяє проводити пошук оптимальних значень, виключаючи втрату надбаних під час пошуку кращих рішень.

3. Виконано експериментальне дослідження запропонованого генетичного методу при синтезі нейромережевих моделей залежності показників здоров'я населення. Результати досліджень показали, що розроблена модель дає найменшу помилку передбачення числа нових випадків туберкульозу, що становить 6.139 та числа захворювань системи кровообігу, що становить 441.889. Під час створення і тренування моделі на основі мережі довгої короткочасної пам'яті було досліджено можливість використання методу рою часток для оптимізації параметрів мережі. За допомогою алгоритму рою часток було отримано найменше значення помилки (RMSE) – 127.08, що є прийнятним показником. Практична значущість роботи полягає у тому, що розв'язано практичне завдання синтезу моделей залежності показників здоров'я населення на основі штучних нейронних мереж, що дозволить своєчасно коригувати плановані лікувально-діагностичні, профілактичні заходи, завчасно визначати необхідні ресурси для локалізації та ліквідації захворювань з метою збереження здоров'я населення.

Література

1. Paustenbach, D. (Ed.) (2002). Paustenbach Human and ecological risk assessment. Theory and practice. New York, 635.
2. Викиди забруднюючих речовин в атмосферне повітря. Державна служба статистики України. URL: https://ukrstat.org/uk/operativ/operativ2009/ns_rik/ns_u/dvsr_u2008.html
3. Стан забруднення природного середовища на території України. URL: http://cgo-sreznevskyi.kiev.ua/index.php?fn=u_zabrud&f=ukraine
4. Tuberculosis. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/tuberculosis
5. Ghazvini, K., Yousefi, M., Firoozeh, F., Mansouri, S. (2019). Predictors of tuberculosis: Application of a logistic regression model. Gene Reports, 17, 100527. doi: https://doi.org/10.1016/j.genrep.2019.100527
6. Mei, B., Xu, Y. (2019). Multi-task least squares twin support vector machine for classification. Neurocomputing, 338, 26–33. doi: https://doi.org/10.1016/j.neucom.2018.12.079
7. Rubal, Kumar, D. (2018). Evolving Differential evolution method with random forest for prediction of Air Pollution. Procedia Computer Science, 132, 824–833. doi: https://doi.org/10.1016/j.procs.2018.05.094

8. Dembinski, H., Schmelling, M., Waldi, R. (2019). Application of the iterated weighted least-squares fit to counting experiments. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 940, 135–141. doi: https://doi.org/10.1016/j.nima.2019.05.086

9. Soebiyanto, R. P., Kiang, R. K. (2000). Modeling Influenza Transmission Using Environmental Parameters. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science, XXXVIII, 330–334.

10. Yi, H.-C., You, Z.-H., Zhou, X., Cheng, L., Li, X., Jiang, T.-H., Chen, Z.-H. (2019). ACP-DL: A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation. Molecular Therapy - Nucleic Acids, 17, 1–9. doi: https://doi.org/10.1016/j.omtn.2019.04.025

11. Speiser, J. L., Miller, M. E., Tooze, J., Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications, 134, 93–101. doi: https://doi.org/10.1016/j.eswa.2019.05.028

12. Alam, S., Dobbie, G., Koh, Y. S., Riddle, P., Ur Rehman, S. (2014). Research on particle swarm optimization based clustering: A systematic review of literature and techniques. Swarm and Evolutionary Computation, 17, 1–13. doi: https://doi.org/10.1016/j.swevo.2014.02.001

13. Kumar, J., Goomer, R., Singh, A. K. (2018). Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters. Procedia Computer Science, 125, 676–682. doi: https://doi.org/10.1016/j.procs.2017.12.087

14. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.

15. Викторова, Е. В. (2012). Применение нечетких нейронных сетей для технической диагностики дорожных машин. Вестник Харьковского национального автомобильно-дорожного университета, 56, 98–102.

16. McClure, N. (2017). TensorFlow Machine Learning Cookbook. Packt Publishing, 370.

17. Колесніков, К. В., Карапетян, А. Р., Царенко, Т. А. (2013). Генетичні алгоритми для задач багатокритеріальної оптимізації в мережах адаптивної маршрутизації даних. Вісник Нац. техн. ун-ту "ХПІ", 56 (1029), 44–50.

18. Oliinyk, A., Fedorchenko, I., Stepanenko, A., Rud, M., Goncharenko, D. (2018). Evolutionary Method for Solving the Traveling Salesman Problem. 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T). doi: https://doi.org/10.1109/infocommst.2018.8632033

19. Lin, B., Sun, X., Salous, S. (2016). Solving Travelling Salesman Problem with an Improved Hybrid Genetic Algorithm. Journal of Computer and Communications, 04 (15), 98–106. doi: https://doi.org/10.4236/jcc.2016.415009

20. Haupt, R. L., Haupt, S. E. (2003). Practical Genetic Algorithms. John Wiley & Sons. doi: https://doi.org/10.1002/0471671746

21. Shkarupylo, V., Skrupsky, S., Oliinyk, A., Kolpakova, T. (2017). Development of stratified approach to software defined networks simulation. Eastern-European Journal of Enterprise Technologies, 5 (9 (89)), 67–73. doi: https://doi.org/10.15587/1729-4061.2017.110142

22. Fedorchenko, I., Oliinyk, A., Stepanenko, A., Zaiko, T., Shylo, S., Svyrydenko, A. (2019). Development of the modified methods to train a neural network to solve the task on recognition of road users. Eastern-European Journal of Enterprise Technologies, 2 (9 (98)), 46–55. doi: https://doi.org/10.15587/1729-4061.2019.164789

23. Oliinyk, A., Zaiko, T., Subbotin, S. (2014). Training sample reduction based on association rules for neuro-fuzzy networks synthesis. Optical Memory and Neural Networks, 23 (2), 89–95. doi: https://doi.org/10.3103/s1060992x14020039

24. Fedorchenko, I., Oliinyk, A., Stepanenko, A., Zaiko, T., Korniienko, S., Burtsev, N. (2019). Development of a genetic algorithm for placing power supply sources in a distributed electric network. Eastern-European Journal of Enterprise Technologies, 5 (3 (101)), 6–16. doi: https://doi.org/10.15587/1729-4061.2019.180897

25. Fedorchenko, I., Oliinyk, A., Stepanenko, A., Zaiko, T., Shylo, S., Svyrydenko, A. (2019). Development of the modified methods to train a neural network to solve the task on recognition of road users. Eastern-European Journal of Enterprise Technologies, 2 (9 (98)), 46–55. doi: https://doi.org/10.15587/1729-4061.2019.164789

26. Oliinyk, A. O., Zayko, T. A., Subbotin, S. O. (2014). Synthesis of Neuro-Fuzzy Networks on the Basis of Association Rules. Cybernetics and Systems Analysis, 50 (3), 348–357. doi: https://doi.org/10.1007/s10559-014-9623-7

27. Oliinyk, A., Fedorchenko, I., Stepanenko, A., Rud, M., Goncharenko, D. (2019). Combinatorial Optimization Problems Solving Based on Evolutionary Approach. 2019 IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM). doi: https://doi.org/10.1109/cadsm.2019.8779290

28. Sharifzadeh, M., Sikinioti-Lock, A., Shah, N. (2019). Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and Gaussian Process Regression. Renewable and Sustainable Energy Reviews, 108, 513–538. doi: https://doi.org/10.1016/j.rser.2019.03.040

29. Buduma, N., Locascio, N. (Eds.) (2017). Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms. O'Reilly Media, 298.

30. Lapan, M. (2018). Deep Reinforcement Learning Hands-On: Apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO, AlphaGo Zero and more. Packt Publishing, 546.

31. Aggarwal, C. C. (2018). Neural Networks and Deep Learning. Springer. doi: https://doi.org/10.1007/978-3-319-94463-0