Журнал обчислювальної та прикладної математики **2023, № 2** Journal of Numerical & Applied Mathematics

УДК 519.6
MSC 35R11, 65M06

A NON-PARAMETRIC APPROACH TO EXPLAINABLE ARTIFICIAL INTELLIGENCE AND ITS APPLICATION IN MEDICINE

D. A. Klyushin, O. S. Maistrenko

Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, E-mail: dmytroklyushin@knu.ua

НЕПАРАМЕТРИЧНИЙ ПІДХІД ДО ПОЯСНЮВАЛЬНОГО ШТУЧНОГО ІНТЕЛЕКТУ ТА ЙОГО ЗАСТОСУВАННЯ В МЕДИЦИНІ

Д. А. Клюшин, О. С. Майстренко

Факультет комп'ютерних наук та кібернетики, Київський національний університет імені Тараса Шевченка, Київ, Україна, E-mail: dmytroklyushin@knu.ua

Abstract. The paper proposes a non-parametrical approach to explainable artificial intelligence based on the compactness postulate, which states that objects of one class in the feature space are, as a rule, located closer to each other than to objects of other classes. Objects are considered similar if they are located close to each other in the feature space. Meanwhile, the properties of objects in real life are often random values. Such objects are not described by a vector of features, but by a random sample or several samples of features, and the postulate of compactness should be replaced by the postulate of statistical homogeneity. Objects are considered statistically homogeneous if their features obey the same distributions. The paper describes a non-parametric measure of homogeneity and an illustration of its use in medical applications, in particular for the diagnosis of breast cancer within the framework of similarity-based explainable artificial intelligence.For comparison, the results of diagnostics of the same data set using deep learning of an artificial neural network are given. We formulate new statistical postulates of machine learning and propose to consider a machine learning algorithm as explanatory and interpretable if it satisfies these postulates.
Keywords: explained artificial intelligence, non-parametric statistics, postulates of machine learning, deep learning, convolutive neural network.

Анотація. В роботі пропонується непараметричний підхід до пояснюваного штучного інтелекту на основі постулату компактності, який стверджує, що об'єкти одного класу в просторі ознак, як правило, розташовані ближче один до одного, ніж до об'єктів інших класів. Загальноприйнято вважати об'єкти подібними, якщо

вони розташовані близько в просторі ознак. Між тим, властивості предметів у реальному житті часто є випадковими значеннями. Такі об'єкти описуються не вектором ознак, а випадковою вибіркою або кількома вибірками ознак, і постулат компактності слід замінити постулатом статистичної однорідності. Об'єкти вважаються однорідними, якщо їхні ознаки підкоряються однаковим розподілам. У розділі описується непараметрична міра однорідності та надається ілюстрація їх використання в медичних додатках, зокрема для діагностики раку молочної залози в рамках пояснюваного штучного інтелекту на основі подібності. Для порівняння наводяться результати діагностики того ж самого набору даних за допомогою глибинного навчання штучної нейронної мережі. Ми формулюємо нові статистичні постулати машинного навчання та пропонуємо вважати алгоритм машинного навчання пояснювальним та інтерпретованим, якщо він задовольняє цим постулатам.

Ключові слова: пояснюваний штучний інтелект, непараметрична статистика, постулати машинного навчання, глибинне навчання, згорткова нейронна мережа.

1. Вступ

Концепція пояснюваного штучного інтелекту (XAI — eXplainable artificial intelligence) останнім часом стала предметом інтенсивних досліджень, оскільки механічне застосування штучного інтелекту не відповідає природним вимогам, які застосовуються до систем людина–машина. В автоматичних системах одна підсистема беззастережно приймає результати роботи іншої підсистеми, оскільки ця особливість закладена в їх конструкціях. Медичні програми мають зовсім іншу природу, оскільки їх можна схематично описати як взаємодію пацієнт-–лікар—штучний інтелект—лікар—пацієнт. Як бачимо, роль людини в цій схемі набагато важливіша ролі штучного інтелекту. Замість механічної взаємодії, яка передбачає сліпу довіру та автоматизм, людина диктує штучному інтелекту властивості, якими він має володіти: довіру, причинність, можливість перенесення та інформативність [23]. На думку Ліптона, інтерпретований штучний інтелект повинен вселяти довіру до алгоритмів своєї роботи, виявляти причинно-наслідкові зв'язки між вихідними даними, проміжними результатами та кінцевим висновком, дозволяти застосування до нових даних, а також допускати вилучення нової інформації з побудованої моделі. Крім перерахованих вище критеріїв, класифікація пояснюваного штучного інтелекту також використовується на основі мети її використання. Ця класифікація базується на чотирьох критеріях: вірогідність висновків, контроль висновків, удосконалення алгоритмів і отримання нових знань [1].

В даний час різні моделі штучного інтелекту широко використовуються як в медицині загалом [3, 36, 43], так і в онкології зокрема [8, 12] тощо. Серед них варто виділити моделі пояснюваного штучного інтелекту, які

використовуються в аналізі медичних зображень і діагностиці [9,11,29] тощо. Перераховані вище роботи дають повний огляд сучасних застосувань штучного та пояснюваного штучного інтелекту в медицині. У той же час такий різновид ще не описаний єдиною математичною теорією, яка дозволяє формалізувати процес інтерпретації пояснень [42].

Особлива увага приділяється когнітивним аспектам пояснюваності алгоритмів глибинного та машинного навчання. По-перше, зауважимо, що в багатьох роботах (див., наприклад, [25]) машинне навчання безпідставно асоціюється виключно зі штучними нейронними мережами. Можливо, це пов'язано з тим, що за останні 10 років штучні нейронні мережі стали мейнстрімом, а інші алгоритми пішли в тінь. Однак машинне навчання набагато ширше, ніж вивчення нейронних мереж. По-друге, всі критерії, запропоновані для інтерпретації процесу та висновків, до яких призводить машинне навчання в штучних нейронних мережах, зводяться до спроб відповісти на питання: «Що знаходиться в чорному ящику?» У цей момент викликає когнітивний дисонанс. Якщо ви вже з самого початку вирішили скористатися чорним ящиком, то навіщо вам знати, що в ньому знаходиться, якщо він дає правильні відповіді? Кому це потрібно? Чи треба пасажиру безпілотного автомобіля знати, як працюють його механізми? Очевидно, що ні. Водночас технік, який обслуговує такий транспорт, повинен знати його конструкцію, бо інакше він не зможе його налаштувати. Однак коли алгоритми машинного навчання побудовані за принципом чорного ящика, ситуація складніша. Тут дослідник надає штучному інтелекту можливість самонастроювання, що виражається в самостійному виборі особливостей об'єктів. Той факт, що дослідник не сам будує простір ознак, а покладає відповідальність за нього на алгоритм, призводить до неприємних наслідків. По-перше, спостерігається експоненціальне зростання кількості ознак, оскільки алгоритм обмежений лише обчислювальною потужністю комп'ютера, на якому він працює, і лише автор алгоритму може поставити завдання оптимізації та вибору лише найбільш значущих ознак. По-друге, семантика самих ознак стає абсолютно незрозумілою з точки зору людини.

У роботі [25] зроблено спробу розкрити зміст понять пояснюваності та інтерпретованості. Проаналізувавши велику кількість літератури, автори звели поняття пояснюваності до відповідей на три питання: 1) пояснюваність даних; 2) пояснюваність результатів; 3) пояснюваність алгоритму.

Відповідь на перше питання зводиться до пояснення того, які дані використовуються для навчання і чому. Зокрема, це вимагає формулювання робочої гіпотези, яка грунтується на даних. Наприклад, у медичних додатках в сфері онкології вхідними даними алгоритмів є інформація, зібрана від пацієнтів з різними видами раку, а також від здорових людей. Логічно припустити, що ця інформація має різне значення для діагностики. Хто і як буде відбирати необхідну інформацію для забезпечення найбільшої точності діагнозу? Класичний підхід передбачає участь у цьому процесі експерта в предметній області, який має апріорні знання про значущість тих чи інших даних. Цей експерт може, наприклад, використовувати попередні знання про те, що в організмі людини, хворої на рак, відбуваються

біохімічні реакції, які впливають на розподіл хроматину в ядрах букальних епітеліальних клітин. У цьому випадку фахівець запропонує проаналізувати фотографії клітинних ядер хворих і здорових людей. Водночас виникає друге, не менш важливе питання: наскільки ці дані об'єктивні? Наприклад, якщо фотографія містить якийсь систематичний артефакт (мітку тощо), то апріорна класифікація буде скомпрометована. Тому необхідний механізм перевірки навчальних даних, який гарантує їх неупереджений і випадковий вибір із усього набору даних.

Друге питання випливає з першого. Якщо алгоритм, який діє на основі чорного ящика, отримав вхідні дані, він може самостійно вибрати ознаки, які не мають ані біологічного, ані фізичного значення, ані будь-якого іншого значення, крім статистичного. Іншими словами, результати, до яких це призведе, будуть корельованими, а не причинно-наслідковими. В таких випадках процес розпізнавання можна описати тільки зі статистичної точки зору. Це ускладнює отримання нових знань. Так, ми правильно розділили два набори, але що це дає зі змістовної точки зору. Яких особливостей людське око не побачило на цих фотографіях і яке їхнє біологічне чи фізичне значення? Сучасна теорія машинного навчання не дає змістовних відповідей на ці питання. Як правило, ознаки — це якась функція або комбінація кількох функцій, які були обрані алгоритмом без знання змісту фотографії.

Третє питання стосується дизайну моделі. У контексті штучних нейронних мереж це зводиться до аналізу архітектури та маніпулювання шарами. Різноманітність архітектур і методів роботи з ними справляють враження сліпого жонглювання. Якщо та чи інша архітектура привела до успішного результату, здається зайвим пояснювати чому. У той же час така мережа може бути схожа на картковий будиночок, який розвалиться, як тільки в нього потраплять нові дані. Здатність таких алгоритмів до узагальнення є дуже проблематичною, а їх стабільність сумнівною. Цікава класифікація методів класифікації з точки зору їх пояснювальної здатності наведена в [5]. На думку авторів, лінійна та логістична регресія, дерева рішень, метод найближчого сусіда, алгоритми на основі правил, узагальнені адитивні моделі та байєсовські моделі не потребують пояснень, тобто за визначенням належать до пояснюваного штучного інтелекту. Водночас автори вважають метод випадкового лісу, машину опорних векторів, а також багатошарові, згорткові та рекурентні нейронні мережі частково інтерпретованими за допомогою спеціальних методів.

Здається, що така класифікація методів у сфері застосування штучного інтелекту в медицині є не зовсім точною. Очевидно, що ця класифікація оцінює пояснюваність з точки зору дослідника, тобто має сенс лише для розробника алгоритму. Водночас у схемі пацієнт–лікар-ШІ–лікар–пацієнт є ще дві людини, які належать не до категорії розробників, а до користувачів алгоритмів. Лікар хоче розуміти алгоритм і довіряти йому, а пацієнт хоче довіряти лікарю і розуміти його. Тому пояснюваність слід розглядати також з їхньої точки зору. У цих підходах є як подібності, так і відмінності.

З точки зору пояснюваності даних, лікар, як і розробник, хоче і зобов'язаний розуміти сенс і якість вхідних даних. Тут їхні інтереси збігаються.

Крім підвищення якості підготовки даних компетентним лікарем, який правильно налаштовує механіку з розумінням специфіки підготовки вхідних даних (наприклад, правильно встановить рівень освітлення, підбере правильний реагент, налаштує необхідна роздільна здатність мікроскопа), апріорне розуміння вхідних даних дозволить вчасно отримати зворотній зв'язок, якщо існує дрейф концепції, через який алгоритм може бути скомпрометований. Цей дрейф, наприклад, може спостерігатися при спробі застосувати алгоритм діагностування раку молочної залози на підставі аналізу букального епітелію до популяцій людей в різних географічних регіонах Землі.

Зрозумілість результатів також впливає на рівень довіри лікаря і пацієнта до отриманих результатів. Висновки, зроблені за алгоритмом, лікар так чи інакше перевіряє за допомогою додаткових методів (наприклад, направляючи пацієнта на обстеження іншими методами). Якщо алгоритм постійно демонструє високу точність, то лікар і пацієнт можуть приймати правильні клінічні рішення, справедливо вважаючи, що ця методика перевірена і надійна.

При цьому ані лікаря, ані пацієнта не повинно цікавити, скільки шарів використовується в архітектурі нейронної мережі. Ця інформація виходить за межі його компетенції і не має нічого спільного з пояснюваністю з точки зору лікаря чи пацієнта. Тому зрозумілість алгоритму для лікаря і пацієнта не має значення.

Перейдемо до інтерпретації, другого аспекту XAI. Згідно з [25], можливість інтерпретації є синонімом зрозумілості моделі (алгоритму та даних) для спостерігача. У нашій схемі взаємодії пацієнт–лікар–ШІ–лікар–пацієнт неявно присутня третя особа, а саме розробник алгоритму. Розглянемо ролі зацікавлених сторін відповідно до парадигми, запропонованої в [5, 23, 25]. Необхідно проаналізувати, наскільки модель має бути зрозумілою кожній зацікавленій особі. Цілком природно, що модель повинна бути абсолютно зрозумілою її розробнику; інакше він не зможе гарантувати його високу якість. Чи повинна вона бути зрозуміло лікарю? Щоб відповісти на це питання, необхідно з'ясувати, що ми розуміємо під моделлю в цій парадигмі. З нашої точки зору, модель це комбінація алгоритму та даних. Алгоритм налаштований на певний тип вхідних даних і створює певний тип вихідних даних. Це означає, що алгоритми та дані нероздільні. При цьому слід визнати, що лікар повинен правильно інтерпретувати вхідні дані, але не зобов'язаний заглиблюватися в структуру алгоритму. Інтереси лікаря та розробника тут частково перетинаються. Конструкція моделі також не важлива для пацієнта (як і конструкція будь-якого медичного обладнання). Для нього важлива точність, з якою працює модель. У цьому випадку можливість інтерпретації моделі з точки зору пацієнта може підвищити довіру до неї. Якщо ми повернемося до критеріїв пояснюваності, перелічених на початку розділу, то можна погодитися, що для пацієнта важлива можливість інтерпретувати результати, а не модель.

З концепцією інтерпретованості тісно пов'язана концепція прозорості, яка зводиться до трьох можливостей: імітації, декомпозиції та алгоритмічного навчання [23]. В [5] дерева рішень класифікуються як повністю прозорі, оскільки правила висновку сформульовані мовою, зрозумілою людині, а метод найближчого сусіда, алгоритми на основі правил, узагальнені адитивні моделі та байєсівські моделі класифікуються як методи, для розуміння яких необхідні математичні знання. У той же час автори вважають абсолютно непрозорими метод випадкового лісу, машину опорних векторів, а також багатошарові, згорткові та рекурентні нейронні мережі. Імітаційність моделі означає, що її просто і легко відтворити на нових вхідних даних для отримання очікуваних результатів. Ми вже неявно використовували можливість декомпозиції вище, коли розділяли модель на дані та алгоритми їх обробки. Алгоритмічність означає зрозумілість процесу навчання алгоритму. В алгоритмах, заснованих на інтуїтивних поняттях (наприклад, поняттях близькості, лінійної роздільності тощо), ступінь алгоритмічності навчання досить висока. Це можна сказати, наприклад, про лінійну і логістичну регресії, методи kNN і SVM. Звичайно, процес навчання нейронної мережі зрозумілий лише фахівцям високого рівня, а для необізнаного спостерігача (лікарів і пацієнтів) він неминуче виглядає як маніпуляція з чорним ящиком.

Детальний і широкий аналіз, проведений у роботах [5] і [25], приводить до висновку, що на даний момент ще не розроблено формальний та строго обґрунтований математичний апарат, який би дозволив оцінити пояснюваність алгоритмів навчання машинних систем. Для кожного конкретного алгоритму дається певна суб'єктивна оцінка, залежно від точки зору автора.

Варті уваги змістовний аналіз Сінтії Рудін [33] і [34]. Ці публікації відображають альтернативну та добре задокументовану точку зору на можливість інтерпретації машинного навчання. Рудін чітко розрізняє машинне навчання, яке можна інтерпретувати, і штучний інтелект, який можна пояснити. За словами Рудін, інтерпретаційне машинне навчання не є частиною теорії пояснюваного штучного інтелекту, оскільки мета цієї теорії полягає в тому, щоб пояснити модель чорного ящика шляхом її наближення до більш простих і зрозумілих моделей, і мета інтерпретованого машинного навчання полягає у виборі початково інтерпретованої моделі, яка забезпечує високу точність [34]. На думку Рудін, слід не пояснювати роботу чорної скриньки, а будувати прозорі, інтерпретовані моделі. Це особливо важливо для медичних застосувань ШІ, де ризик помилки пов'язаний з високою ціною.

Однак, не маючи можливості відкрити чорний ящик і втрутитися в його роботу, ми можемо спробувати дослідити, наскільки добре він відповідає критеріям пояснюваності та інтерпретованості. Цьому присвячена наша робота, в якій ми показуємо, за якими критеріями можна оцінити інтерпретабельність чорного ящика, якщо його довелося застосовувати, і як побудувати прозору інтерпретативну за Рудіном модель діагностики раку

молочної залози, яка також відповідає цим критеріям. Таким чином, ми пропонуємо компроміс між цими двома альтернативами.

Як ми покажемо нижче, побудова формальної теорії XAI, заснованої на строгому математичному підході, призводить до висновку, що пояснення роботи та висновків штучного інтелекту неможливе без застосування постулатів машинного навчання.

Мета цього розділу — запропонувати новий формалізм XAI на основі альтернативних статистичних постулатів машинного навчання.

2. Математичний формалізм

Надійність, логічність, можливість узагальнення та інформативність є основними властивостями будь-кого методу машинного навчання. Щоб довести цю тезу, розглянемо загальну схему будь-якого методу машинного навчання.

Нехай X — набір об'єктів, Y — набір міток класу, а $f : X \to Y^{-}i$, яка отримує значення на елементах навчальної вибірки, взятої з X. Мета навчального алгоритму — розширити функцію $f : X \to Y$ на всю множину X для побудови розв'язувальної функції $g : X \to Y$, що відображає всі об'єкти на їхні мітки та мінімізує функцію ризику (наприклад, кількість помилок).

Оскільки сама постановка проблеми машинного навчання включає мінімізацію помилки та визначення функції на всій множині даних X, точність і можливість узагальнення гарантуються за замовчуванням. Водночас поняття причинності та інформативності мають неформальний характер, який важко описати за допомогою математичного апарату. Однак важко не означає неможливо.

Один із найбільш значущих математичних формалізмів інтерпретованого штучного інтелекту був запропонований у [42]. Цей формалізм базується на байєсівському підході та використовує концепцію зрозумілого висновку та універсальний закон Шеппарда, який стверджує, що ймовірність узагальнення реакції з одного стимулу на інший є функцією подібності між двома стимулами в психологічному просторі. Вводячи метричний простір замість психологічного простору близькості та інтерпретуючи ймовірність узагальнення реакції на новий стимул як імовірність розпізнавання нового об'єкта, схожого на об'єкт навчання, автори пропонують міру зрозумілості машинного навчання, подаючи результати у формі функції ймовірності.. Вірогідність запропонованого підходу продемонстровано експериментом.

Слід зазначити, що цей формалізм має суттєвий недолік, який пояснюється байєсівським характером всього методу. Він вимагає знання апріорної ймовірності розпізнаваних класів, яку мають надати експерти. Автори також пропонують розрахувати функцію подібності Сломана як міру косинуса між векторами, створеними з допомогою відповідної карти експертами, які беруть участь в експерименті. Очевидно, що все це робить методологію локальною та суб'єктивною. Її правильність залежить від компетентності конкретних експертів.

Ми пропонуємо використовувати більш строгий та об'єктивний підхід, який базується не на суб'єктивних оцінках експертів, а на об'єктивно визначених значеннях міри однорідності об'єкта з іншими об'єктами та його статистичної глибини. Коротше кажучи, ми пропонуємо вважати результати роботи штучного інтелекту такими, що піддаються інтерпретації з точки зору лікаря, якщо вони задовольняють постулат статистичної компактності. З іншого боку, ми пропонуємо вважати ці результати такими, що можна інтерпретувати з точки зору пацієнта, якщо вони дозволяють визначити індивідуальний ризик пацієнта. Точність будь-якого алгоритму машинного навчання в медичній діагностиці розглядається за Фішером. Припустимо, що чутливість алгоритму становить 95%. Це означає, що зі 100 випадково відібраних пацієнтів при багаторазовому повторенні цього вибору діагноз встановлюється правильно в середньому у 95 пацієнтів, а 5 пацієнтів отримують помилковий діагноз. Якщо алгоритм має специфічність 95%, це означає, що зі 100 випадково відібраних здорових людей, якщо цей вибір повторити багато разів, 5% отримають діагноз «хворий». Ці оцінки не відповідають на природні запитання кожного пацієнта: «Яка ймовірність того, що я дійсно хворий? До якої групи я потрапляю: 95 правильних діагнозів чи 5 помилкових?» На це питання можна відповісти за допомогою двох понять: статистичної однорідності та статистичної глибини. З одного боку, порівнюючи дані пацієнта з еталоном хворої людини, можна оцінити ступінь подібності між ними, з іншого боку, оцінюючи статистичну глибину даних, можна оцінити типовість поставленого діагнозу для конкретного пацієнта в діапазоні від вірогідного до сумнівного.

3. Близькість

Машинне навчання базується на двох постулатах, які ми вже неявно використовували вище: 1) постулат представлення об'єкта як вектора ознак у векторному просторі і 2) постулат компактності Авер'янова та Бравермана [28].

Перший постулат відображає природне прагнення спеціалістів машинного навчання використовувати апарат алгебри, геометрії та методів оптимізації. По суті, цей постулат дозволяє звести проблему машинного навчання до задачі оптимізації, тобто до мінімізації чи максимізації певної функції за певних обмежень.

Другий постулат менш очевидний, але також інтуїтивно зрозумілий. Він стверджує, що вектори ознак об'єктів, що належать до одного класу, розташовані в просторі ознак ближче один до одного, ніж до об'єктів з іншого класу. Цей постулат часто супроводжується вимогою, щоб ці набори векторів у просторі ознак були розділені досить простою функцією. Яскравими прикладами методів, побудованих на основі цих постулатів, є лінійні дискримінантні методи Фішера, метод опорних векторів і метод найближчих сусідів.

Незважаючи на очевидний успіх перерахованих вище методів, не можна заперечувати, що постулати векторного простору і компактності у сформульованому вище вигляді застосовні не до всіх задач. У багатьох

медико-біологічних дослідженнях окремому пацієнту відповідає не вектор ознак, тобто впорядкований набір чисел, що характеризують його різні властивості, а випадкова вибірка, тобто невпорядкований набір випадкових результатів вимірювань певних параметрів (наприклад, площа ядра, оптична щільність ядра тощо). Такі вибірки виникають, наприклад, при аналізі зішкрібів, взятих у пацієнта. Ці зішкріби зазвичай містять кілька десятків клітин, тому пацієнту відповідає не одна точка у векторному просторі, а множина точок, причому ці точки ніяк не впорядковані. Звичайно, цю задачу можна спростити, розрахувавши середні вибіркові значення (або квантілі, як буде показано нижче) та застосувавши стандартні постулати, але цілком очевидно, що в цьому випадку втрачається важлива частина інформації про розподіл вимірюваних параметрів.

Ми пропонуємо альтернативні статистичні постулати: 1) об'єкти можуть бути представлені вибірковими значеннями їх параметрів; 2) параметри об'єктів, що належать до одного класу, мають однакові розподіли, а параметри об'єктів, що належать до різних класів, мають різні розподіли.

Такий підхід дозволяє звести проблему оцінки подібності між об'єктами до перевірки гіпотези про однорідність двох і більше вибірок. Метод, який ми пропонуємо використовувати для оцінки подібності між об'єктами, наведено нижче. На відміну від класичних методів, таких як статистика Колмогорова–Смирнова та статистика Манна–Уїтні–Вілкоксона, він має статистичну універсальність, тобто працює однаково добре як для вибірок із різними середніми значеннями та однаковими стандартними відхиленнями, так і для вибірок із однаковими середніми значеннями, але з різними стандартними відхиленнями.

Розглянемо вибірку розміром n, що складається з неперервних випадкових змінних із симетрично залежним розподілом.

Теорема Хілла [15,16] стверджує: ймовірність того, що випадкові значення з того самого розподілу є більшими за i-ту і меншими за j-ту порядкові статистики вибірки ($i < j$), дорівнює $\frac{j-i}{n+1}$.

Як бачимо, ця ймовірність залежить лише від номера порядкової статистики та розміру вибірки. Використовуючи цей факт, можна перевірити гіпотезу про однорідність двох вибірок. Для цього ми сортуємо за зростанням першу вибірку, отримуючи її порядкову статистику, і підраховуємо відносну частоту події, коли елемент з другої вибірки є більшим за i-ту та меншим за j-ту порядкову статистику першої вибірки. Маючи ці відносні частоти, ми можемо побудувати довірчий інтервал для біноміальної пропорції в розглянутій узагальненій схемі Бернуллі з заданим рівнем значущості (наприклад, інтервал Вільсона). Потім ми перевіряємо той факт, що цей довірчий інтервал містить $\frac{j-i}{n+1}$ для кожної пари номерів i та j, де $i < j$.

Таким чином, підраховуючи відносну частоту цієї події, ми отримуємо так звану p-статистику. Нарешті, ми будуємо довірчий інтервал для p-статистики з заданим рівнем значущості α. Ми відхиляємо нульову гіпотезу про однорідність двох вибірок, якщо довірчий інтервал, побудований на цих вибірках, не містить 1–α [17].

4. Статистична глибина

Детальний огляд різноманітних концепцій статистичної глибини наведений в [27]. Метою цієї концепції є упорядкування багатовимірних випадкових величин.

Розглянемо функцію розподілу F в R^n Функцією статистичної глибини називається функція $D_F(x)$, яка упорядковує точки x з розподілу F, монотонно спадаючи при збільшенні відстані від центру. Статистична глибина точки $x \in F$ — це значення, яке набуває функція $D_F(x)$ в точці x [44]. Центром розподілу може бути медіана, центроїд, геометричний центр множини тощо. Функція статистичної глибини повинна мати наступні властивості [44]: 1) не залежати від координатних систем і афінних перетворень (афінна інваріантність); 2) досягати максимуму в центрі розподілу (ця точка називається найбільш глибокою, тобто найбільш ймовірною); 3) монотонно спадати від найбільш на найменш глибокої точки (монотонність); 4) Якщо відстань від точки x до центру розподілу прямує до нескінченості, функція $D_F(x)$ має прямувати до нуля (гранична властивість).

Якщо замість повної інформації про функцію розподілу F є лише вибірка n точок з розподілу F, маємо вибіркову функцію статистичної глибини $D_n(x)$. Розглянемо декілька прикладів обчислення статистичної глибини.

1. Глибина Тьюкі [38]. Спочатку введемо необхідні поняття. Центр вибірки — це така точка, що кожна гіперплощина, що проходить через неї, ділить вибірку на дві майже рівні підмножини. Якщо ця точка є елементом вибірки, то вона є медіаною вибірки. Глибина Тьюкі елемента вибірки — це мінімальна кількість елементів вибірки, що лежать по один бік випадкової гіперплощини, що проходить через неї.

2. Пілінг опуклої оболонки [7]. Опуклою оболонкою множини точок є мінімальний многокутник, що містить дані точки.

Пілінг опуклої оболонки — це процедура послідовного пошуку та видалення замкнутих опуклих оболонок. Вершини однієї опуклої оболонки мають однакову статистичну глибину.

3. Глибина Ойя [30]. Глибина Ойа вибіркового елемента — це середній об'єм симплекса, побудованого по d випадкових точках вибірки

$$d(x|x_1, x_2, ..., x_n)$$

і точці x.

4. Симплеціальна глибина [24]. Симплеціальна глибина елемента вибірки x — це кількість симплексів, побудованих на основі випадкової вибірки точок, яка містить x.

5. Зоноїдна глибина [20]. Зоноїдна глибина вибіркового елемента — це число $d(x|x_1, x_2, ..., x_n) = \sup \alpha : y \in D_\alpha(x_1, ..., x_n)$ де

$$D_\alpha(x_1, ..., x_n) = \left\{ \sum_{i=1}^{n} \lambda_i x_i : \sum_{i=1}^{n} \lambda_i = 1, \;\; 0 \le \lambda_i, \forall i : \alpha\lambda_i \le \frac{1}{n} \right\}.$$

6. Глибина Махаланобіса [44]. Глибина Махаланобіса є узагальненням відстані Махаланобіса за формулою

$$MHD_F(x) = (1 + d^2(x, E(F)))^{-1},$$

де $d^2(x, y) = (x - y)^T \Sigma_F^{-1}(x - y)$, де $E(F)$ — математичне сподівання розподілу F, а Σ_F — коваріаційна матриця.

7. Еліптична статистична глибина [26]. Еліптична статистична глибина — це функція, яка відображає точки вибірки на ранги, що зростають, використовуючи довірчі еліпсоїди Петуніна [31]. Ці еліпсоїди є концентричними і покривають вибіркові випадкові точки. Таким чином, ми маємо послідовність еліпсоїдів $E_1 \subset E_2 \subset ... \subset E_n$. Кожна точка вибірки лежить на поверхні лише одного еліпсоїда, і ймовірність того, що випадкова точка з F лежить в E_n дорівнює $\frac{n-1}{n+1}$. Таким чином, еліптична статистична глибина є монотонною функцією, яка досягає максимуму в найглибшій точці та зменшується від центру назовні.

8. Області, упорядковані по глибині [10] — це набір точок, де статистична глибина більша або дорівнює заданому значенню

$$D_\alpha(F) = \left\{ x \in R^d : D_F(x) \geq \alpha \right\},$$

де $D_F(x)$ — статистична глибина точки x з розподілу. Області, упорядковані по глибині, є афінно-еквіваріантними, вкладеними, монотонними, компактними та субадитивними. Очевидно, що еліпсоїди Петуніна є областями, упорядкованими по глибині.

5. Приклад застосування альтернативних постулатів компактності

Фрактальний аналіз став дуже поширеною технікою в дослідженнях структури пухлинних клітин [6, 13, 35, 41]. У той же час наявність змін, пов'язаних зі злоякісним новоутворенням, у нормальних клітинах стала загальновідомим фактом [19, 40]. Морфологічний аналіз нормальних клітин у хворих на рак зазвичай виконується поблизу пухлини, враховуючи, що ракові клітини безпосередньо впливають на сусідні тканини [22, 32]). Однак нормальні клітини, розташовані далеко від пухлини, такі як клітини букального епітелію, також реагують на наявність пухлини в організмі [37,39]. Наше дослідження [4], яке ми представляємо як ілюстрацію використання альтернативних статистичних постулатів для діагностики раку, перевірило гіпотезу гетерогенності розподілу фрактальної розмірності хроматину в ядрах букального епітелію у жінок з раком молочної залози, жінок з фіброаденоматозом та здорових жінок.

Навчальна вибірка пацієнтів складається з 130 жінок: 68 випадків раку молочної залози, 33 випадки фіброаденоматозу і 29 здорових жінок. У кожної жінки брали зішкріб букального епітелію, який в середньому містив близько 50 клітин. Зішкріб обробляли згідно з методом, описаним у [18], і фарбували за Фьольгеном. Після фотозйомки під мікроскопом було більше 20 000 фотографій клітин, зроблених через зелений, жовтий і фіолетовий фільтри, а також без фільтра в колірній моделі RGB і в градаціях сірого.

Кожна фотографія представляла собою матрицю розміром 160x160 пікселів.

Розглядалися дві групи пацієнтів: ракові (BC) та неракові (Control). До першої групи увійшли жінки з раком молочної залози, до другої — жінки з фіброаденоматозом і здорові жінки. Групи були достатньо збалансованими з 68 раковими і 62 нераковими пацієнтами. Діагностика полягала в обчисленні фрактальної розмірності кожного ядра, формуванні вибірок, що містять фрактальні розмірності ядер із зішкрібу, і проведенні багаторазової перехресної перевірки за допомогою методу 1NN, навчання алгоритму на $p\breve{}1$ підвибірках і багаторазової оцінки його точності на довільно вибрані підвибірки. Для оцінки однорідності вибірок використовувалась p-статистика. Перехресну перевірку було виконано для всіх 11 пар (фільтр, канал). Рішення прийнято простим голосуванням. Фільтри також оцінювали простим голосуванням. Фрактальну розмірність було розраховано за Мінковським за допомогою модифікованого алгоритму box-counting [21].

Голосування за методикою 1NN проводилося окремо між парами (фільтр, канал) та між фільтрами. Крім того, було проведено єдину перехресну перевірку з різними розмірами контрольної вибірки (5%, 10%, 20% і за винятком одного). Переможцем голосування стала сіра шкала в жовтому та фіолетовому фільтрах (точність = 99,28%, специфічність = 100%) у голосуванні «один проти всіх». Для кожного пацієнта була оцінена міра однорідності з найближчим сусідом із групи раку та з групи нормального стану. Мірою однорідності в цій задачі є статистична глибина пацієнта. Чим вище цей показник, тим вище ступінь вірогідності діагнозу. Статистична глибина оцінювала індивідуальний ризик пацієнта, на відміну від показника точності, який застосовується до всіх пацієнтів в цілому. Значення показників однорідності показано на рис. 1 і 2. Як ми бачимо, внутрішньогрупові розподіли показників однорідності ракових (BC–BC) і нормальних (Control–Control) груп є компактними на відміну від міжгруповими розподілами у парах BC–Control та Control–BC.

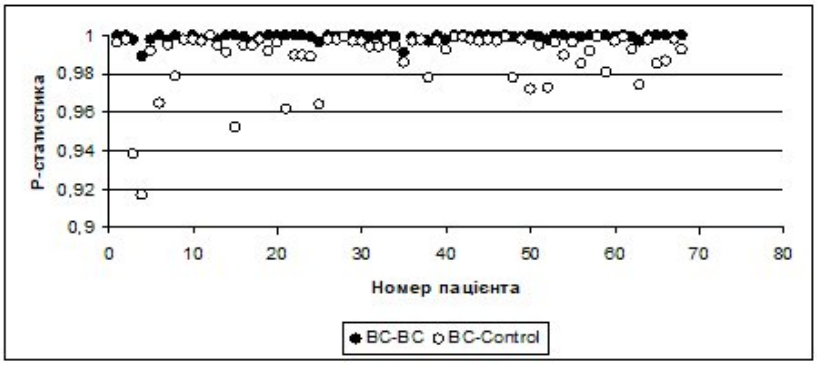

Рис 1. Однорідність вибірок всередині групи BC та між групами BC та Control

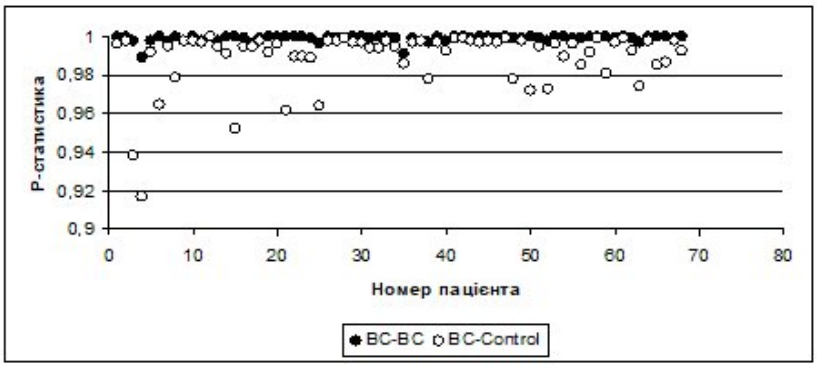

Рис 2. Однорідність вибірок всередині групи Control та між групами Control та BC

З точки зору теорії пояснюваного штучного інтелекту цей алгоритм відповідає всім вимогам. По-перше, довіру фахівців і пацієнтів забезпечує його висока точність, чутливість і специфічність. По-друге, алгоритм використовує причинно-наслідкові зв'язки між вхідними даними та результатами. Він не генерує випадково вибрані характеристики та виявляє не кореляцію між випадковими вибірками, а природну закономірність, що проявляється у пов'язаних із пухлиною змінах у фрактальному розмірі фотографій забарвленого за Фьольгеном хроматину ядер букального епітелію у пацієнтів з раком молочної залози. По-третє, він дозволяє узагальнювати висновки на нові дані, оскільки базується на біологічних закономірностях, а не на випадковій подібності. По-четверте, це дозволяє отримати нові знання, оскільки за допомогою нього можна діагностувати будь-який тип раку, а не тільки рак грудей.

Очевидно, всі ці властивості забезпечуються тим, що ми відмовилися від концепції чорного ящика і апріорно забезпечили можливість пояснити та інтерпретувати метод за рахунок біологічно обґрунтованого вибору вхідних даних, використання інтуїтивного методу класифікації, використання методів непараметричної статистики та обчисленні статистичної глибини кожного пацієнта.

Для ілюстрації того факту, що запропонований підхід не заперечує, а лише розширює класичні підходи, ми провели класифікацію вибірок за допомогою методів глибинного навчання згорткових нейронних мереж. Діагностика полягала в обчисленні фрактальної розмірності кожного ядра, формуванні вибірок, що містять фрактальні розмірності ядер із зішкрібу і застосуванні штучної нейронної мережі. Фрактальну розмірність було розраховано за допомогою коефіцієнта Хьорста і кривої Гільберта [18]. Для забезпечення сумісності вхідних даних із вимогами нейронних мереж на вхід подавалися не вибірки фрактальної розмірності, а вектори квантилів цих вибірок.

Внаслідок проведених експериментів була розроблена згорткова нейронна мережа для класифікації даних, яка включає ряд компонентів: згорткові

шари Conv1d, шари пулінгу, повністю з'єднані шари та шари нормалізації та регуляризації. У прямому проході вхідні дані пройшовши згортковий шар, проходять через шари пулінгу та повністю з'єднані шари, забезпечуючи комплексну обробку і адаптацію до особливостей даних. Conv1d є ключовим компонентом архітектури мережі і призначений для виявлення локальних патернів у вхідних даних. Згортковий шар Conv1d дозволяє моделі ефективно аналізувати просторові особливості у вхідних даних, використовуючи ядро згортки та враховуючи обсяг інформації завдяки параметру padding.

Для виділення прихованих ознак в навчальних даних було проведено декілька експериментів з початковими шарами нейронної мережі. Найкращі результати вдалося отримати з одномірним згортковим шаром на вході моделі. За допомогою згортки було виділено k прихованих послідовностей менших по довжині за вхідну, які далі передавалися підмережі, що складається з декількох повністю зв'язаних шарів.

Модель з одновимірними згортковими шарами виявляється дуже ефективною у класифікації даних завдяки здатності мережі Conv1d виділяти локальні особливості. Повністю з'єднані шари та шари нормалізації додають гнучкості та стабільності цій моделі. За рахунок специфічності даних та задачі, потребувалися додаткові підходи для забезпечення стабільної роботи моделі. Цієї цілі було досягнуто за рахунок сучасних підходів в оптимізації нейромереж, зокрема batch normalization та dropout.

Для оцінки результатів моделі було проведено декілька експериментів. Тренувальний датасет не був рівномірним – кількість елементів в одному з класів переважала – тому за основні метрики було обрано чутливість (precision) та специфічність (recall). Щоб встановити базовий рівень для метрик, спочатку було використано випадкову модель. Далі було порівняно дві нейроні мережі, остання з яких мала згортковий шар, що підвищило чутливість та специфічність (табл. 1.)

Табл. 1. Чутливість і специфічність різних моделей

	Випадковий генератор	FCN	FCN + Conv1d
Чутливість	0,56	0,83	0,94
Специфічність	0,49	0,8	0,91

Як бачимо, результати, отримані за допомогою штучної нейронної мережі за чутливістю та специфічністю є співставними із результатами, отриманими за допомогою оцінки статистичної однорідності вибірок. Але з токи зору можливості пояснення штучна мережа програє, оскільки, окрім чутливості та специфічності, ми нічого не можемо сказати про індивідуальний ризик конкретного пацієнта та не маємо змоги інтерпретувати роботу алгоритму (чи оцінює від компактність множини вибірок тощо). Таким чином, це типовий чорний ящик, який, тим не менше, має високу точність.

Автори висловлюють щиру подяку доктору медичних наук Бородан Н.В. та кандидату технічних наук Голубєвій К.М. за надані вихідні матеріали та

Анастасії Андрійчук [4] і Дмитру Шерварли [18] за допомогу в обчисленні фрактальної розмірності.

6. Висновок

Оцінка критеріїв пояснюваного штучного інтелекту за існуючими методами суб'єктивна. Ми пропонуємо об'ктивний підхід і універсальний критерій пояснюваності XAI: штучний інтелект вважається пояснюваним, якщо результати його застосування задовольняють двом статистичним постулатам: 1) об'єкти можуть бути представлені вибірковими значеннями їх параметрів; 2) параметри об'єктів, що належать до одного класу, мають однакові розподіли, а параметри об'єктів, що належать до різних класів, мають різні розподіли. Крім того, XAI має передбачати оцінку статистичної глибини результатів (індивідуальний ризик у контексті охорони здоров'я). Таким чином, запропонована модель машинного навчання, проілюстрована в роботі на прикладі діагностики раку молочної залози, відповідає всім вимогам інтерпретованого штучного інтелекту.

Література

1. Adadi A., Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 2018. Vol. 6. P. 52138–52160.
2. Alabdulhadi M., Coolen-Maturi T., Coolen F. Nonparametric predictive inference for comparison of two diagnostic tests. *Communications in Statistics – Theory and Methods* 2021. Vol. 50. P. 4470–4486.
3. Amann, J. et al. To explain or not to explain? — Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*. 2022. Vol. 1(2). P. e0000016.
4. Andreichuk A. V., Boroday N. V., Golubeva K. M., Klyushin D. A. Artificial Intelligence System for Breast Cancer Screening Based on Malignancy-Associated Changes in Buccal Epithelium. In: Enabling AI Applications in Data Science. Part of the Studies in Computational Intelligence book series (SCI, 2022, volume 911) Springer, 2022, pp. 267–285.
5. Arrieta A.B. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible A. *Information Fusion*. 2020. Vol. 58. P. 82–115.
6. Bakalis E. et al. (2022) Universal Markers Unveil Metastatic Cancerous Cross-Sections at Nanoscale. *Cancers*. 2022. Vol. 14. No. 15. P. 3728.
7. Barnett V. (1976) The ordering of multivariate data. *Journal of the Royal Statistical Society. Series A (General)*. Vol. 139. No. 3. P.318–355.
8. Bi, W. et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer Journal for Clinicians*. 2019. Vol. 69. No. 2). P. 127–157.
9. Borys K. et al. Explainable AI in medical imaging: An overview for clinical practitioners — Saliency-based XAI approaches. *European Journal of Radiology*. 2023. Vol. 162. P. 110787.
10. Cascos, I. (2007) Depth function as based of a number of observation of a random vector. Working Paper 07-29, Statistic and Econometric Series 07, 2:1–28.
11. Chaddad A, Peng J, Xu J, Bouridane A. Survey of Explainable AI Techniques in Healthcare. *Sensors*. 2023. Vol. 23. No. 2. P. 634.

12. Chen Z. et al. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Communications* 2021. Vol. 41. No. 11. P. 1100–1115.
13. Elkington L., Adhikari P., Pradhan P. Fractal Dimension Analysis to Detect the Progress of Cancer Using Transmission Optical Microscopy. *Biophysica*. 2022. Vol. 2. No. 1. P. 59–69.
14. Hacking S., Yakirevich E., Wang Y. From Immunohistochemistry to New Digital Ecosystems: A State-of-the-Art Biomarker Review for Precision Breast Cancer Medicine. *Cancers*. 2022. Vol. 14. No. 14/ P. 3469.
15. Hill B. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of American Statistical Association*. 1968. Vol. 63. P. 677–691.
16. Hill B. De Finetti's theorem, induction, and A(n) or Bayesian nonparametric predictive inference (with discussion). In: D. V. Lindley, J. M. Bernardo, M. H. DeGroot, and A. F. M. Smith (Eds.), Bayesian statistics (1988, Vol. 3, pp. 211–241). Oxford: Oxford University Press.
17. Klyushin D. A., Petunin Yu.I. A Nonparametric Test for the Equivalence of Populations Based on a Measure of Proximity of Samples. *Ukrainian Mathematical Journal*. 2003. Vol. 55. No. 2. P. 181–198.
18. Klyushin D., Golubeva K.. Boroday N., Shervarly D. Breast cancer diagnosis using machine learning and fractal analysis of malignancy-associated changes in buccal epithelium. Chapter in: Artificial Intelligence, Machine Learning, and Data Science Technologies Future Impact and Well-Being for Society 5.0 N. Mohan, R. Singla, P. Kaushal, and S. Kadry (Eds.), CRC Press, 2021, P. 1–18.
19. Koopaie M., Kolahdooz S., Fatahzadeh M., Manifar S. Salivary biomarkers in breast cancer diagnosis: A systematic review and diagnostic meta-analysis. *Cancer Medicine*. 2022. Vol. 11. No. 13. P. 2644–2661.
20. Koshevoy G., Mosler K. Zonoid trimming for multivariate distributions. *Annals of Statistics* 1997. Vol. 25. P. 1998–2017.
21. Li J., Du Q., Sun C. An improved box-counting method for image fractal dimension estimation. *Pattern Recognition*. Vol. 42. No. 11. P. 2460–2469.
22. Liang W. et al. Cancer cells corrupt normal epithelial cells through miR-let-7c-rich small extracellular vesicle-mediated downregulation of p53/PTEN. *Intertational Journal of Oral Science*. Vol. 14. No. 36.
23. Lipton Z. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*. 2018. Vol. 16. No. 3. P. 31–57.
24. Liu R.J. On a notion of data depth based on random simplices. *Annals of Statistics*. 1990. Vol. 18. P. 405–414.
25. Love P. et al. Explainable Artificial Intelligence (XAI): Precepts, Methods, and Opportunities for Research in Construction. arXiv:2211.06579v2, 2022.
26. Lyashko S. Klyushin D., Alexeyenko V. Mulrivariate ranking using elliptical peeling. *Cybernetic and Systems Analysis*. 2013. Vol. 49. No. 4. P. 511–516.
27. Mosler K., Mozharovskyi P. Choosing among notions of multivariate depth statistics. *Statistical Science*. Vol. 37. No. 3. P. 348–368.
28. Mottl V., Seredin O., Krasotkina O. Compactness Hypothesis, Potential Functions, and Rectifying Linear Space in Machine Learning. In: International Conference Commemorating the 40th Anniversary of Emmanuil Braverman's Decease, Boston, MA, USA, April 28-30, 2017, Invited Talks.

29. Nazir S., Dickson D., Akram M. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine*. 2023. Vol. 156. P. 106668.
30. Oja H. Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*. 1983. Vol. 1. P. 327–332.
31. Petunin Yu., Rublev B. Pattern recognition using quadratic discriminant functions. *Numerical and Applied Mathematics*. 1996. Vol. 80. P. 89–104.
32. Polverini P., N?r F., N?r J. Crosstalk between cancer stem cells and the tumor microen-vironment drives progression of premalignant oral epithelium. *Frontiers in Oral Health*. 2023. Vol. 3. No. 1095842.
33. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019. No. 1. P. 206–215.
34. Rudin C. et al. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistical Surveys*. 2022. Vol. 16. P. 1—85.
35. S?nchez J., Mart?n-Landrove M. Morphological and Fractal Properties of Brain Tumors. *Frontiers in Physiology*. 2022. Vol. 13. No. 878391.
36. Sheu R.-K., Pardeshi M. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. *Sensors*. 2022. Vol. 22. No. 8068.
37. Subramanian H. et al. Procedures for risk-stratification of lung cancer using buccal nanocytology. *Biomedical Optics Express*. (2016). Vol. 7. No. 9. P. 3795–3810.
38. Tukey J. Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematician (1975, pp. 523–531). Montreal, Canada.
39. Us-Krasovec M. et. al. Malignancy associated changes in epithelial cells of buccal mucosa: a potential cancer detection test. *Analytical and Quantitative Cytology and Histology*. 2005. Vol. 27. No. 5. P. 254–262.
40. Wu C. et al. Cancer-Associated Adipocytes and Breast Cancer: Intertwining in the Tumor Microenvironment and Challenges for Cancer Therapy. *Cancers*. 2023. Vol. 15. No. 3. P. 726.
41. Xu C. et al. Modeling and analysis fractal order cancer model with effects of chemotherapy. *Chaos, Solitons and Fractals*. 2022. Vol. 161. No. 112325.
42. Yang S., Folke T., Shafto P. A psychological theory of explainability. In: Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, 2022, PMLR 162.
43. Zhang Y., Weng Y., Lund J. Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. *Diagnostics*. 2022. Vol. 12. No. 2. P. 237.
44. Zuo Y., Serfling R. General notions of statistical depth function. *Annals of Statistics*. Vol. 28. P. 461–482.

Надійшла: 20.11.2023 / Прийнята: 1.12.2023