[9] Spiwok, V., & Kříž, P. (2020, June 30). Time-Lagged t-Distributed Stochastic Neighbor Embedding (t-SNE) of Molecular Simulation Trajectories. Frontiers in Molecular Biosciences, 7. https://doi.org/10.3389/fmolb.2020.00132.

[10] Ayyappa. T and S. Kurse, "Fault Detection of Bearing using XGBoost Algorithm and Data Visualization using t-distributed stochastic neighbor embedding (t-SNE) Method," SSRN Electronic Journal, 2021, Published, doi: 10.2139/ssrn.3834976.

[11] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, no. 1–2, pp. 305–307, Oct. 2018, doi: 10.1007/s10710-017-9314-z.

[12] Hong, S. E. (2019, December 31). Exploring Independent Component Analysis Based on Ball Covariance. The Korean Data Analysis Society, 21(6), 2721–2735. https://doi.org/10.37727/jkdas.2019.21.6.2721.

[13] ADNI | Study Documents. (n.d.). https://adni.loni.usc.edu/methods/documents/;

[14] UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/102/thyroid+disease.

Надійшла до редколегії 02.08.2023

Перова Ірина Геннадіївна, доктор технічних наук, професор, професор кафедри системотехніки ХНУРЕ, м. Харків, Україна, e-mail: iryna.perova@nure.ua. ORCID: https://orcid.org/0000-0003-2089-5609.
Мірошниченко Неля Сергіївна, аспірант кафедри системотехніки ХНУРЕ, м. Харків, Україна, e-mail: nelia.miroshnychenko@nure.ua. ORCID: https://orcid.org/0000-0002-3846-1668.

УДК 004.89:61 DOI: 10.20837/0135-1710.2023.179.050

І.Ю. ПАНФЬОРОВА, А.С. БУЦЬКА

ДОСЛІДЖЕННЯ АНСАМБЛЮВАННЯ МОДЕЛЕЙ MACHINE LEARNING В МЕДИЧНІЙ ДІАГНОСТИЦІ

Розглянуто застосування моделей машинного навчання (machine learning – ML) в медичній діагностиці. Представлено основні виклики та цілі останніх досліджень у сфері медичного прогнозування. Основну увагу зосереджено на порівнянні існуючих моделей ML. Проведено аналіз вже існуючих рішень для розробки ансамблю моделей ML. Розраховано ключові характеристики моделей ML: точність, чутливість, специфічність та AUC-ROC. Запропоновано варіант об'єднання цих моделей в ансамбль для покращення точності – основної характеристики прогнозування діагнозу.

1. Вступ

Медична діагностика є критично важливим аспектом охорони здоров'я, оскільки її рівень безпосередньо впливає на точність результатів діагностування і, як наслідок, на вірність рішення щодо лікування пацієнтів. Нова ера в медичній діагностиці розпочалася з появою машинного навчання (machine learning – ML) , яке надало розширені інструменти для аналізу великих обсягів медичних даних, включаючи історії хвороби пацієнтів, зображення та генетичну інформацію. Використання моделей ML може сприяти виявленню прихованих закономірностей і аномалій, а також прогнозуванню ризиків розвитку того чи іншого захворювання, що робить їх цінними інструментом для медичних працівників.

Один з напрямів сучасних досліджень в галузі застосування ML у медичній діагностиці полягає у ліквідації недоліків існуючих моделей ML шляхом їх комбінації. Цей різновид ML, який забезпечує покращення продуктивності і точності прогнозів за рахунок комбінації кількох моделей ML, отримав назву «ансамблеве навчання» [1]. Мета досліджень за цим напрямом полягає у порівнянні існуючих моделей ML, пропозиції множини варіантів та виборі найкращого варіанту вирішення задачі створення ансамблю для прогнозування діагнозу.

Ансамблеве навчання, на жаль, не вільне від деяких недоліків. Так непростою задачею є інтеграція моделей ML, адже до того, як впровадити моделі ML у вже існуючі медичні

інформаційні системи (МІС), слід провести значну роботу з даними, починаючи від їх збору, попередньої обробки та закінчуючи інтерпретацією отриманих результатів. Ансамблі моделей ML потребують більшого об'єму даних, аніж окремі моделі.

Головна проблема ансамблевого навчання полягає у виборі реалізації ансамблю з окремих моделей ML, оскільки не всі моделі добре поєднуються між собою, а невірний вибір може погіршити продуктивність ансамблю. У конкретних МІС використання окремих засобів ML показало їх перспективність, проте поєднання сильних сторін МІС та засобів ML може призвести не тільки до підвищення надійності прогнозів діагностики, а й до зниження швидкодії МІС та збільшення витрат для отримання бажаного результату. Ця проблема також породжує проблему розрахунку обчислювальної складності та ресурсів, необхідних для реалізації відразу декількох моделей ML. Тому проведення досліджень, спрямованих на вирішення зазначених вище проблем визнається необхідним як з теоретичної, так і з практичної точок зору.

2. Аналіз літературних джерел та визначення проблеми дослідження

У 2019 році було проведено огляд тенденцій у галузі використання моделей ML для діагностики [2]. Цей огляд розділив моделі на дві великі категорії: моделі, керовані даними, і моделі, керовані знаннями. Кожна з цих категорій додатково поділяється на контрольоване та неконтрольоване навчання. Моделі, керовані даними, зосереджені на аналізі та використанні великих обсягів даних для навчання моделей, в той час, як моделі, керовані знаннями, використовують збагачені бази експертних знань для вирішення конкретних завдань. Обидва підходи можуть взаємодіяти. Ансамблювання може використовувати моделі керовані даними для автоматичного вивчення залежності у великих обсягах історичних даних, а потім використовувати експертні знання для уточнення та покращення результатів діагностування.

Дослідження, опубліковані в [2], продемонстрували потенціал ML в медичній діагностиці. Моделі, керовані даними, такі, як згорткові нейронні мережі (CNN) продемонстрували виняткову продуктивність в аналізі медичних зображень, включаючи виявлення пухлин на радіологічних зображеннях. Ключова перевага моделі CNN перед своїми попередниками полягає в тому, що вона визначає основні функції без потреби втручання людини [3]. Зазвичай CNN використовує зображення як вхідні дані. Така модель призначає вагові коефіцієнти різним частинам зображення для отримання диференціації пікселів та їх просторового розташування, що дозволяє виявляти різноманітні аспекти та характеристики, такі як границі об'єктів, текстури та інші ключові елементи. CNN застосовують операцію згортки принаймні на одному рівні замість простого множення матриць [4], що відрізняє CNN від інших моделей.

Контрольована модель навчання «Машина векторів підтримки» (Support Vector Machine –SVM) принесла користь багатьом напрямам прогнозування, таким, наприклад, як прогнозування захворювань на основі даних, отриманих за допомогою магнітно-резонансної томографії (МРТ). Дослідники використовували SVM для діагностики нервово-м'язових розладів [5]. Обробка природної мови (NLP) була застосована для вилучення цінної інформації з неструктурованих клінічних записів, що допомагає класифікувати захворювання та оцінювати ризик. Рекурентні нейронні мережі (RNN) здатні розпізнавати послідовності. Існує багато варіантів реалізації RNN, таких як LSTM, BLSTM, MDLSTM і HLSTM. Недоліком RNN є виникнення значних ускладнень у його застосуванні у випадках зникнення градієнта. Крім того, RNN для своєї експлуатації потребують великих наборів даних, зокрема, наборів даних медичних зображень. Але такі набори занадто обмежені в порівнянні з наборами даних для загальних проблем комп'ютерного зору, які зазвичай коливаються від кількох сотень тисяч до мільйонів анотованих фотографій [6].

Загальним недоліком моделей ML є те, що через складність структур даних навчання моделей може стати дороговартісним завданням. Іноді для обробки цих структур даних

необхідні потужні графічні процесори та сотні комп'ютерів, що призводить до збільшення вартості для кінцевих користувачів [7].

Оскільки при використанні декількох моделей у ансамблевому навчанні складність обчислень підвищується, виникає високе обчислювальне навантаження, що негативно впливає на продуктивність процесу навчання моделей. Для подолання проблеми зникання градієнта і перетримки використовуються вдосконалені функції активації, архітектура функції вартості та методи видалення, описані в [8]. Вирішити проблему великої обчислювальної завантаженості допомагає використання ресурсів такого формату як графічні процесори та пакетна нормалізація. ML добуває інформацію з великої кількості даних і генерує результати, які можна використовувати для індивідуального прогнозування та прийняття клінічних рішень [9]. Цей підхід відкриває шлях до розвитку персоналізованої медицини, де враховуються генетичні фактори та спосіб життя людини для профілактики захворювань, лікування та прогнозування розвитку захворювання [10].

3. Мета і задачі дослідження

Метою даного дослідження є розробка ансамблевої моделі для вирішення задачі прогнозування діагнозу за результатами медичних дослідів та порівняння ансамблевої моделі з існуючими одиничними моделями ML, котрі вже активно використовуються для прогнозування в медицині. Досягнення цієї мети дає поштовх для покращення існуючих рішень в медичній діагностиці та подальшого розвитку MIC шляхом вдосконалення існуючих рішень задачі прогнозування діагнозу.

Для досягнення поставленої мети пропонується вирішити такі задачі:

- розробити схематичне зображення роботи ансамблевої моделі;
- зібрати достатній обсяг даних з медичних досліджень;
- провести дослідження на окремих моделях ML та запропонованій ансамблевій моделі;
- порівняти отримані результатів.

4. Матеріали і методи дослідження

Як зазначено в [1] у загальному вигляді схему ансамблевого навчання моделей ML може бути представлено у вигляді комбінації кроків, у якій дані можна тренувати на різних базових класифікаторах, а вихідні дані об’єднуються для отримання остаточного прогнозу (див. рис. 1).

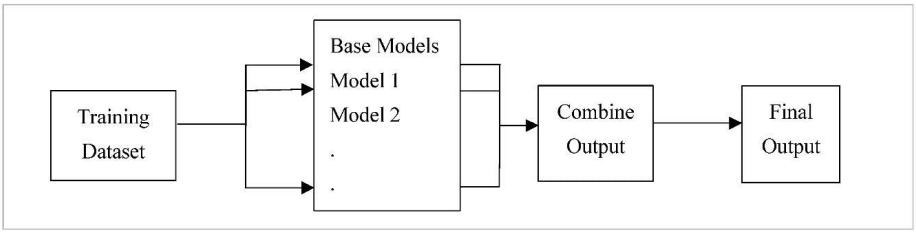

Рис. 1. Схема ансамблевого навчання моделей ML

Рис. 1 ілюструє базове представлення реалізації ансамблювання моделей в ML, де різні моделі можуть навчатися на вхідних даних, а на виході отримується об’єднаний результат. Існують різні підходи до створення ансамблевих моделей. Одними з найпопулярніших таких підходів є stacking, bagging, boosting та voting.

Stacking передбачає використання декількох різнорідних слабких моделей, які навчаються незалежно одна від одної, а потім об'єднуються для створення прогнозу на основі результатів кожної з моделей. Bagging, з іншого боку, передбачає навчання однорідних

моделей на різних наборах даних і їх об'єднання. Прогноз отримується шляхом усереднення прогнозів кожної з моделей. Boosting – це підхід послідовного навчання декількох однорідних моделей, при якому кожна наступна модель виправляє помилки попередніх моделей. Voting – це підхід, який агрегує прогнози з численних незалежних моделей (базових оцінювачів), щоб зробити остаточний прогноз [11].

Як наслідок, в результаті ансамблювання кілька моделей ML поєднуються для створення точніших прогнозів, ніж ті, що реалізовані за допомогою одного класифікатора [12]. Кожен тип ансамблювання детально розглянуто в [1]

Для моделі ансамблю запропоновано обрати три моделі (SVM, CNN і LSTM), оскільки вони продемонстрували високу продуктивність та мають попит в MIC (рис. 2).

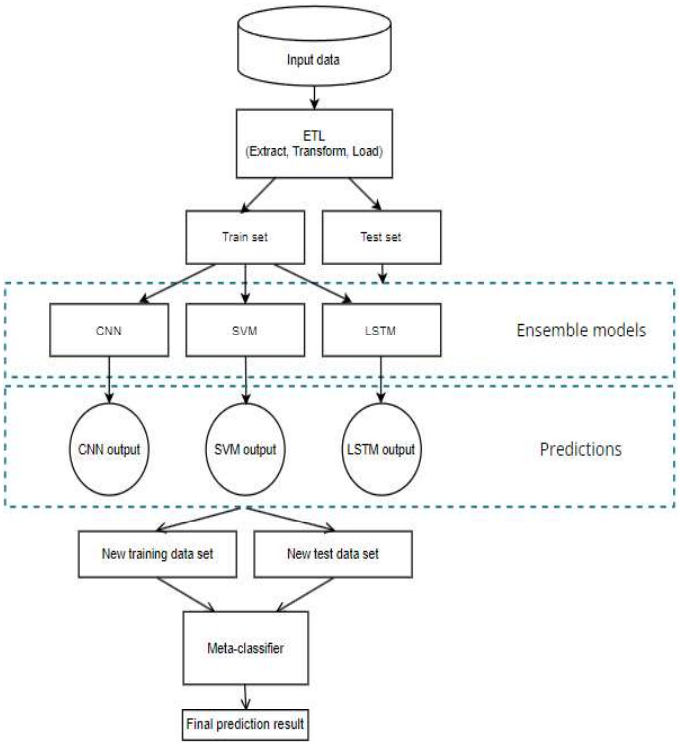

Рис. 2. Схема архітектури моделі ансамблевого навчання з використанням підходу stacking на основі моделей SVM, CNN та LSTM

Дослідження передбачає навчання та оцінку обраних моделей на наборах даних, зібраних з реальних медичних закладів. Ці набори складаються з радіологічних зображень, клінічних записів, генетичних даних та історій пацієнтів, охоплюють широкий спектр захворювань і забезпечують необхідну різноманітність для комплексного аналізу.

Тому дослідження пропонується проводити за допомогою даних із тестового набору даних з відкритого репозиторію даних Kaggle, що належить корпорації Google [13].

5. Результати дослідження

Серед поширених захворювань, які можуть значно вплинути на здоров'я пацієнтів, зазвичай виділяють діабет, рак шкіри, захворювання нирок, печінки та серцеві захворювання У цьому дослідженні зроблено акцент на діагностуванні раку шкіри, а саме, створенні класифікатора меланоми, який зможе з достатньою точністю розрізняти доброякісні (неракові) та злоякісні (ракові) ділянки шкіри.

Для того, щоб дослідити запропоновану модель ансамблю, було використано тестові та навчальні зразки зображень. Для доповнення існуючих зображень, пов'язаних із захворюваннями шкіри, була використана текстова інформація щодо атрибутів цих зображень, наведена у файлі формату .csv [13]. Досліди проводилися на портативному комп'ютері з процесором Intel Core(TM) i5-6300HQ @ 2,30 ГГц x 4, 16 Гб оперативної пам'яті DDR3 і графічним процесором NVIDIA GeForce GTX 960M 4 Гб DDR5.

У цьому дослідженні для вирішення проблеми, пов'язаної з тим, що деякі вхідні дані були неповними або необробленими, використано додаткову інформацію з матриці стохастичних

пертурбацій. Матриця стохастичних пертурбацій, кожен елемент якої являє собою випадкове число, використовується для введення випадкових значень в дані під час тренування моделі. Тому матриця доповнення формується шляхом додавання до вхідних даних матриці стохастичних пертурбацій:

$$E = x_{i\times j} + D_{i\times j}, \tag{1}$$

де E – розширена матриця; x – вхідні дані; D – матриця стохастичних пертурбацій.

В даному випадку значення елементів матриці стохастичних пертурбацій мають бути значно меншими за вихідні дані, щоб уникнути зміни характеристик необроблених даних. У той же час розмірність стохастичної матриці пертурбацій має відповідати розмірності матриці вхідних даних.

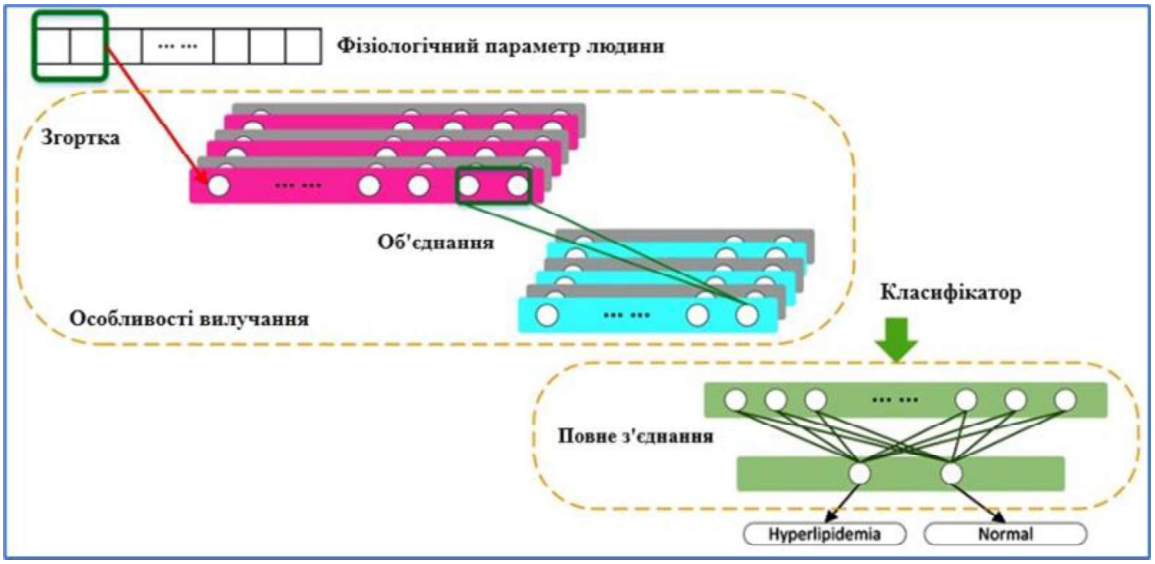

Рис. 3. Розширення алгоритму навчання

Було також проведено дослідження впливу кількості і точності інформації на результативність моделі в дослідах. На етапі кореляції даних значення фізіологічних параметрів з негативними результатами були замінені значеннями, близькими до нуля, але не рівними нулю.

Для одночасного навчання моделей ML у дослідженні було використано необроблені та розширені дані. Рис. 3 демонструє основний принцип розширення алгоритму навчання. Експеримент проводився у середовищі Ubuntu 16.04 з використанням програмного забезпечення Keras, яке базується на Tensorflow.

Для порівняння результатів моделювання було використано такі характеристики.

Точність *Accuracy* характеризує загальну правильність передбачень як відношення кількості правильних прогнозів до загальної кількості передбачень.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \tag{2}$$

де TP (True Positives) – кількість правильно передбачених позитивних класів; TN (True Negatives) – кількість правильно передбачених негативних класів; FP (False Positives) – кількість хибнопозитивних передбачень; FN (False Negatives) – кількість хибнонегативних передбачень.

Чутливість *Recall* характеризує частку справді позитивних випадків, правильно ідентифікованих, як відношення справжніх позитивних випадків до суми справжніх позитивних та помилкових негативних випадків.

$$Recall = \frac{TP}{(TP + FN)}. \tag{3}$$

Специфічність *Specificity* характеризує частку справді негативних випадків, правильно визначених, як відношення справжніх негативних випадків до суми справжніх негативних та помилкових позитивних випадків.

$$Specificity = \frac{TN}{(TN + FP)}. \tag{4}$$

AUC-ROC характеризує здатність моделі розрізняти класи, причому вищі значення вказують на кращу продуктивність. Точний розрахунок AUC-ROC вимагає побудови кривої ROC та інтегрування площі під нею, що зазвичай виконується за допомогою чисельних методів або спеціалізованого програмного забезпечення.

З результатами досліджень можна ознайомитись в табл. 1.

Таблиця 1

Результати, отримані в ході дослідження різних моделей

Модель	Точність	Чутливість	Специфічність	AUC-ROC
LSTM	0.88	0.92	0.86	0.91
SVM	0.82	0.88	0.78	0.86
CNN	0.89	0.93	0.87	0.92
CNN-SVM-LSTM	0.95	0.96	0.93	0.95

Значення в табл. 1 було обчислено за допомогою формул (2)-(4) на основі прогнозів моделі та справжніх міток класу в тестовому наборі даних.

Аналіз наведених в табл.1 результатів, отриманих в ході дослідження моделей ML, показав, що окремі моделі (LSTM, SVM, CNN) демонструють добрі показники для характеристик, котрі наведені в табл. 1, але інтегрована модель, яка об'єднує CNN, SVM і LSTM, показує загальне покращення точності діагностики. Використання ансамблювання моделей дозволило досягти значення оцінки AUC-ROC в 0.95, що перевершує значення оцінки індивідуальної моделі діагностування за допомогою тільки CNN моделі ML, яке дорівнює 0.89.

Запропонована ансамблева модель також надала інтерпретовані пояснення своїх рішень, що сприяло клінічному прийняттю результатів моделювання. Таке сприйняття важливе для медичних працівників, щоб зрозуміти прогнози та довіряти їм.

6. Обговорення результатів дослідження та висновки

Проведені дослідження показали, що впровадження ансамблевих моделей в MIC сприяє подоланню розриву між теорією ML та практичним застосуванням цих моделей в сфері

охорони здоров'я. Адже поєднання кількох моделей ML забезпечує кращу продуктивність не лише при вирішенні задачі автоматичного діагностування, а й в інших сферах повсякденного життя, про що свідчить, наприклад, робота [14]. Для вирішення задачі прогнозування діагнозу за результатами медичних дослідів запропоновано об'єднати моделі, які показали добрі результати при вирішенні задач класифікації, обробки візуальної інформації та обробки тексту.

В роботі було запропоновано ансамблеву модель, яка є результатом об'єднання моделей CNN, SVM та LSTM. Розроблено схему архітектури моделі ансамблевого навчання з використанням підходу stacking на основі моделей SVM, CNN та LSTM. Застосування цієї схеми сприяє підвищенню точності діагностики, що може сприяти ранньому виявленню захворювання та підвищенню ефективності лікування.

Для експериментальної перевірки отриманих рішень використаний набір даних з медичних досліджень, з яким можна ознайомитися в [13]. Проведено попередню обробку даних для забезпечення статистичної значущості та забезпечення репрезентативності вибірки.

Проведено дослідження на окремих моделях ML та запропонованій ансамблевій моделі та виконано порівняння отриманих результатів. Підтверджено перспективність поєднання існуючих моделей ML для підвищення точності медичних діагнозів. Ансамблева модель CNN-SVM-LSTM випереджає окремі моделі за точністю, чутливістю, специфічністю, а також забезпечує можливість інтерпретації результатів прогнозування.

В результаті дослідження було отримано порівняльну оцінку обраних моделей ML. За результатами оцінювання можна зробити висновок, що запропонована ансамблева модель досягла задовільних результатів і може бути запропонована до використання при автоматизації процесу діагностування, адже точність ансамблевої моделі досягла значення в 95%, що свідчить про надійність моделі.

Основними проблемами залишаються: відсутність стандартизованих протоколів збору даних і анотацій, що може призвести до упередженості та обмежити можливість узагальнення моделей; складність у отриманні достатнього обсягу даних про стан пацієнта; проблеми з конфіденційністю даних; проблеми з інтеграцією ML в клінічні робочі процеси; неоднорідність даних з різних джерел, що може вимагати різних етапів попередньої обробки та інтеграції даних для ефективного поєднання.

У цілому ансамблювання моделей машинного навчання є перспективним напрямком досліджень, і в майбутньому буде використовуватися для вирішення складніших завдань та забезпечувати точніші прогнози і класифікації в різних галузях. Майбутні дослідження в цьому напрямку повинні бути зосереджені на зазначених вище проблемах. Однак слід враховувати, що ансамблювання може бути більш обчислювально вимогливим, оскільки вимагає навчання та прогнозування більшої кількості моделей, що може збільшити час та ресурси, необхідні для обробки даних.

Перелік посилань

1. Mahajan P., Uddin S., Hajati F., Moni M. A. Ensemble learning for disease prediction: A Review. *Healthcare.* 2023. Vol. 11. No. 12. P 1808. doi:10.3390/healthcare11121808

2. Ali R., Hardie R. C., Narayanan Narayanan B., De Silva S. Deep learning ensemble methods for skin lesion analysis towards melanoma detection. *2019 IEEE National Aerospace and Electronics Conference (NAECON).* 2019. doi:10.1109/naecon46414.2019.9058245.

3. Jain G., Mittal D., Thakur D., Mittal M. K. A deep learning approach to detect covid-19 coronavirus with X-ray images. *Biocybernetics and Biomedical Engineering.* 2020. Vol. 40. No. 4. P. 1391–1405. doi:10.1016/j.bbe.2020.08.008.

4. Noor M.B. et al. Application of deep learning in detecting neurological disorders from Magnetic Resonance Images: A survey on the detection of alzheimer's disease, parkinson's disease and schizophrenia. 2020. P. 11. doi:10.1186/s40708-020-00112-2.

5. Jeena R.S., Kumar S. Stroke prediction using SVM. *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*. 2016. doi:10.1109/iccicct.2016.7988020.

6. Apostolopoulos I. D., Mpesiana T. A. COVID-19: Automatic detection from X-ray images utilizing transfer learning with Convolutional Neural Networks. *Physical and Engineering Sciences in Medicine*. 2020. Vol. 43. No. 2. P. 635–640. doi:10.1007/s13246-020-00865-4.

7. Ballin A, Karlinsky L, Alpert S, Hasoul S, Ari R, Barkan E. A region based convolutional network for tumor detection and classification in Breast Mammography. Deep Learning and Data Labeling for Medical Applications. 2016. P. 197–205. doi:10.1007/978-3-319-46976-8_21.

8. Anavi Y., Kogan I., Gelbart E., Geva O., Greenspan H. Visualizing and enhancing a deep learning framework using patients age and gender for chest X-ray image retrieval. *Medical Imaging 2016: Computer-Aided Diagnosis*. 2016. doi:10.1117/12.2217587.

9. Hassan M., Ali S., Alquhayz H., Safdar K.. Developing Intelligent Medical Image Modality Classification system using Deep Transfer Learning and LDA. *Scientific Reports*. 2020. Vol. 10. No.1. doi:10.1038/s41598-020-69813-2.

10. Abbas A., Abdelsamea M. M., Gaber M. M. Classification of covid-19 in chest x-ray images using DeTraC deep convolutional neural network. *Applied Intelligence*. 2020. Vol. 51. No. 2. P. 854–864. doi:10.1007/s10489-020-01829-7.

11. Jani R., Shariful Islam Shanto Md., Mohsin Kabir Md., Saifur Rahman Md., Mridha M. F. Heart disease prediction and analysis using ensemble architecture. *2022 International Conference on Decision Aid Sciences and Applications (DASA)*. 2022. doi:10.1109/dasa54658.2022.9765237.

12. Suganyadevi S., Seethalakshmi V., Balasamy K. A review on deep learning in Medical Image Analysis. *International Journal of Multimedia Information Retrieval*. SpringerLink. https://link.springer.com/article/10.1007/s13735-021-00218-1 (accessed Oct. 16, 2023).

13. Siim-ISIC melanoma classification. Kaggle. https://www.kaggle.com/competitions/siim-isic-melanoma-classification/discussion/162486 (accessed Oct. 16, 2023).

14. Livieris I. E., Pintelas E., Stavroyiannis S., Pintelas P. Ensemble deep learning models for forecasting cryptocurrency time-series. *Algorithms*. 2020. Vol. 13. No. 5. P. 121. doi:10.3390/a13050121.

Надійшла до редколегії 17.10.2023

Панфьорова Ірина Юріївна, кандидат технічних наук, доцент, професор кафедри ІУС ХНУРЕ, м. Харків, Україна, e-mail: iryna.panforova@nure.ua. ORCID: https://orcid.org/0000-0001-7032-9109.

Буцька Анастасія Сергіївна, здобувачка вищої освіти гр. ІУСТм-22-1, факультет комп'ютерних наук ХНУРЕ, м. Харків, Україна, e-mail: anastasiia.levchenko@nure.ua.