https://doi.org/10.31891/2307-5732-2025-349-87
УДК 004.8

КИРИЧЕНКО ОЛЕКСАНДР
Хмельницький національний університет
https://orcid.org/0009-0007-7061-631X
e-mail: kyrychenkoom@khmnu.edu.ua

МЕТОД ЛОКАЛЬНОЇ КОНСИСТЕНТНОСТІ ДЛЯ СТВОРЕННЯ ПОЯСНЮВАНИХ МОДЕЛЕЙ КЛАСИФІКАЦІЇ МЕДИЧНИХ ЗОБРАЖЕНЬ

В роботі наведено результати дослідження методу створення інтерпретованих систем штучного інтелекту для діагностики пухлин головного мозку за МРТ-зображеннями, який базується на концепції локальної консистентності між високоточною згортковою нейронною мережею VGG-16 та інтерпретованою моделлю Decision Rules Network. Запропонована гібридна архітектура забезпечує діагностичну точність 95.2% при збереженні здатності генерувати клінічно релевантні пояснення у формі логічних правил на основі IBSI-стандартизованих ознак. Експериментальна валідація на наборі даних продемонструвала досягнення локальної консистентності 89.3% для індивідуальних діагностичних випадків, що підтверджує надійність згенерованих пояснень та можливість практичного застосування системи у клінічній практиці.

Ключові слова: штучний інтелект, медична діагностика, інтерпретовані моделі, локальна консистентність, нейроонкологія, МРТ-зображення.

KYRYCHENKO OLEKSANDR
Khmelnytskyi National University

LOCAL CONSISTENCY METHOD FOR CREATING EXPLAINED MEDICAL IMAGE CLASSIFICATION MODELS

An effective implementation of artificial intelligence systems in medical diagnostics is possible only with ensuring transparency and interpretability of decision-making processes, which is critically important for clinical acceptance and regulatory compliance. The specific challenge of medical AI applications is their "black box" nature that prevents understanding of diagnostic reasoning, leading to limited trust among clinicians despite achieving high accuracy rates. This paper presents the results of research on a novel local consistency method for creating interpretable AI systems in neuro-oncological diagnosis, which combines high-precision VGG-16 convolutional neural network with Decision Rules Network for generating clinically relevant explanations. The proposed hybrid architecture achieved diagnostic accuracy of 95.2% while maintaining the ability to generate logical rules based on IBSI-standardized features for four brain pathology types (glioma, meningioma, pituitary tumor, no tumor). Experimental validation demonstrated achievement of 89.3% local consistency between decisions of the primary and interpretable models, confirming reliability of personalized explanations for specific clinical cases. The developed approach addresses the fundamental trade-off between accuracy and interpretability by ensuring consistency only for local diagnostic instances rather than the entire data space, enabling personalized explanations adapted to specific clinical scenarios. The data obtained can be used in justification of trustworthy AI deployment in medical practice, as well as in development of more transparent and clinically acceptable diagnostic systems that combine technological efficiency with necessary transparency for healthcare applications.

Keywords: explainable artificial intelligence, medical diagnostics, local consistency, neuro-oncology, MRI images, interpretable models.

Постановка проблеми

Сучасна медична діагностика переживає період фундаментальних змін, зумовлених стрімким розвитком технологій штучного інтелекту. Глибоке навчання демонструє хороші результати в аналізі медичних зображень, досягаючи точності, яка часто перевершує можливості досвідчених радіологів. Проте між технічними досягненнями AI та його реальним впровадженням у клінічну практику існує критичний розрив, що обумовлений фундаментальною проблемою непрозорості сучасних алгоритмів машинного навчання.

Особливо гостро ця проблема проявляється у нейроонкологічній діагностиці, де системи на базі згорткових нейронних мереж здатні класифікувати різні типи пухлин головного мозку з точністю понад 95%. Водночас ці самі системи функціонують як "чорні скриньки", унеможливлюючи розуміння логіки прийняття життєво важливих діагностичних рішень. Така ситуація створює парадокс: чим точніша AI-система, тим менше довіри вона викликає у клініцистів, оскільки неможливо верифікувати обґрунтованість її висновків.

Регуляторне середовище додатково ускладнює цю проблему. Європейський регламент захисту даних та етичні принципи ЄС щодо довірчого штучного інтелекту встановлюють жорсткі вимоги до пояснюваності автоматизованих систем прийняття рішень у критичних сферах. Медичні AI-системи повинні не лише давати правильні відповіді, але й чітко пояснювати механізми отримання цих відповідей.

Спроби розв'язання проблеми інтерпретованості привели до розвитку двох основних напрямів. Перший напрям базується на методах пояснення, таких як LIME та SHAP, які намагаються інтерпретувати рішення складних моделей після їх прийняття. Однак такі підходи мають принципові недоліки, оскільки створюють лише наближені пояснення, які можуть не відображати реальну логіку моделі. Більше того, ці пояснення часто виявляються нестабільними незначні зміни у вхідних даних можуть кардинально змінити згенеровані пояснення при незмінному рішенні основної моделі.

Альтернативний підхід полягає у використанні моделей, які є інтерпретованими за своєю

природою, таких як дерева рішень або лінійні моделі. Хоча такі системи забезпечують повну прозорість процесу прийняття рішень, вони демонструють значну втрату точності порівняно з глибокими нейронними мережами, особливо при аналізі складних медичних зображень. Така втрата точності робить їх непридатними для критичних медичних застосувань, де помилка може мати серйозні наслідки для здоров'я пацієнта.

Аналіз існуючих підходів виявляє фундаментальний компроміс між точністю діагностики та інтерпретованістю рішень. Високоточні глибокі мережі забезпечують хороші діагностичні результати, але їхня непрозорість знижує довіру клініцистів та унеможливлює клінічне застосування. З іншого боку, інтерпретовані прості моделі забезпечують прозорість рішень, але демонструють неприйнятно низьку точність для медичних застосувань.

Цей компроміс особливо критичний у контексті нейроонкологічної діагностики, де МРТ-зображення пухлин мозку містять надзвичайно складні візуальні паттерни. Ці паттерни включають тонкі текстурні та морфологічні ознаки, які важко формалізувати у простих правилах, але які є критично важливими для точної діагностики. Водночас неправильна класифікація типу пухлини може призвести до неадекватного лікування з потенційно фатальними наслідками для пацієнта.

Проблема ускладнюється специфічними вимогами клінічного середовища. Діагностичний процес вимагає не лише високої точності, але й оперативності прийняття рішень, що обмежує можливості детального аналізу складних пояснень. Крім того, різні клінічні випадки можуть вимагати різних типів пояснень залежно від специфіки патології, досвіду лікаря та конкретного діагностичного контексту.

Традиційні підходи до створення пояснювальних систем не враховують цю контекстуальну варіативність. Більшість існуючих методів генерують універсальні пояснення, які можуть бути нерелевантними для конкретного клінічного випадку. Така невідповідність знижує практичну цінність пояснень та може навіть вводити в оману клініцистів.

У контексті цих викликів формулюється конкретна проблема дослідження: як створити AI-систему для діагностики пухлин мозку, яка одночасно забезпечує діагностичну точність на рівні експертних систем та генерує клінічно релевантні пояснення прийнятих рішень, які можна верифікувати та практично застосовувати у реальних клінічних умовах.

Ця проблема породжує кілька конкретних технічних викликів. По-перше, виклик консистентності як забезпечити, щоб пояснювальна модель давала ті самі рішення, що й основна високоточна модель для конкретних діагностичних випадків. По-друге, виклик релевантності як гарантувати, що згенеровані пояснення відповідають реальним медичним ознакам, знайомим клініцистам та встановленим у медичних стандартах. Третій виклик стосується персоналізації різні клінічні випадки можуть вимагати акценту на різних діагностичних ознаках, тому система повинна адаптувати пояснення до специфіки конкретного випадку. Нарешті, виклик валідації передбачає розробку об'єктивних методів оцінки якості та достовірності згенерованих пояснень.

Для розв'язання сформульованої проблеми висувається гіпотеза про можливість створення гібридної архітектури, що поєднує високоточну згорткову нейронну мережу з інтерпретованою моделлю на основі логічних правил. Така архітектура має забезпечити локальну консистентність рішень для конкретних діагностичних випадків при збереженні загальної точності на рівні сучасних глибоких мереж.

Ключовою ідеєю є концепція локальної консистентності, яка передбачає узгодженість рішень складної та простої моделей не для всього простору даних, а лише для конкретних клінічних випадків. Це дозволить створювати персоналізовані пояснення, які будуть релевантними для кожного окремого діагностичного завдання, використовуючи при цьому стандартизовані медичні ознаки для забезпечення клінічної значущості.

Розв'язання цієї проблеми має потенціал для створення нового класу медичних AI-систем, які стануть справжніми партнерами лікарів у діагностичному процесі. Такі системи поєднуватимуть технологічну потужність сучасного машинного навчання з необхідною прозорістю та довірою, що є критично важливим для успішного впровадження штучного інтелекту у клінічну практику та покращення якості медичного обслуговування.

Аналіз останніх досліджень

Проблема довіри до систем штучного інтелекту стала особливо актуальною з прискоренням їх практичного впровадження у медичну сферу [1, 2]. Довіра є важливим елементом взаємодії між пацієнтом та лікарем, особливо враховуючи вразливий стан пацієнтів. Застосування систем ШІ привносить новий аспект цієї взаємодії, який здатен послабити загальну довіру між учасниками лікувального процесу.

Системи ШІ здатні демонструвати високі результати, але рівень довіри до них залишається низьким, тому вони не можуть вважатись надійними [3]. Це створює ситуацію, коли пацієнти змушені покладатися на системи ШІ для отримання медичних рішень, що може призводити до зниження довіри в клінічній практиці [4].

Створенню систем ШІ, які відповідали б вимогам довіри, присвячено цілий ряд досліджень [5]. Дослідження концепції довіри на основі етичних принципів, згідно яких ШІ можна вважати довірчим,

розглядаються в роботах [6, 7]. Запропоновані метрики оцінки довіри до ШІ з використанням пояснювальності системи expert-in-the-loop, які визначають відмінність між поясненнями, наданими системою ШІ, та отриманими від експертів на основі їхніх міркувань та досвіду.

Занепокоєння щодо потенційної небезпеки алгоритмів чорної скриньки є обґрунтованим та актуальним [8]. Це звужує практичне застосування в медичних аспектах, коли довіра та прозорість рішень не є критично важливими. Оскільки системи ШІ демонструють високі результати, їх використання є виправданим, але недостатнім для повноцінного клінічного застосування через низький рівень довіри.

Важливим аспектом впровадження медичного ШІ є наявність якісних даних для формування моделей прийняття рішень. Значна кількість даних у медичній сфері для навчання нейронних мереж зосереджена на зображеннях. Для розширення таких даних використовують алгоритми синтезу [9]. Недостатність даних та їх обмеженість є ще одним напрямком розвитку надійного медичного ШІ, оскільки нейронні мережі вимагають досить великий обсяг клінічних даних для отримання високих результатів.

Перспективи застосування медичного ШІ є багатообіцяючими. На сьогодні застосовують ШІ для прогнозування карієсу на зображеннях та розробки надійного ШІ, який дозволяє пояснити причини прогнозування [10]. Для покращення пояснювальності ШІ пропонуються системи оцінки результатів прогнозування на зображеннях із залученням людини як експерта [11].

Для досягнення необхідних результатів довіри та надійності ШІ визначено три дослідницькі області, які необхідно об'єднати: поєднання нейронних мереж і їх прогнозів, графічні причинно-наслідкові моделі та методи верифікації та пояснення. Це відіграє трансформаційну роль у подоланні розриву між теоретичними дослідженнями та практичним застосуванням у клінічній медицині [12]. Значна кількість досліджень виявляє необхідність розробки медичного ШІ з характеристиками, згідно яких такий ШІ можна вважати надійним. Нагальною необхідністю є не стільки результат прогнозування, скільки сукупність ознак, згідно яких ШІ генерував прогноз, що може бути представлена у різних формах та дозволяє лікарям краще розуміти логіку роботи системи.

Аналіз сучасних досліджень показує, що існує фундаментальне протиріччя між високою точністю систем ШІ та низьким рівнем довіри до них у медичній практиці. Незважаючи на багатообіцяючі результати застосування нейронних мереж у діагностиці, їхня природа "чорної скриньки" створює критичні перешкоди для клінічного впровадження. Існуючі підходи до створення пояснювальних систем або значно знижують точність діагностики, або не забезпечують достатньої надійності пояснень.

Мета роботи полягає у розробці методу створення інтерпретованих систем штучного інтелекту для діагностики пухлин головного мозку за МРТ-зображеннями, який забезпечує поєднання високої діагностичної точності з можливістю генерації клінічно релевантних пояснень прийнятих рішень через механізм локальної консистентності між високоточною та інтерпретованою моделями.

Виклад основного матеріалу

Запропонований підхід базується на концепції локальної консистентності рішень між високоточною складною моделлю та спрощеною інтерпретованою моделлю. На відміну від традиційних методів, які намагаються забезпечити глобальну узгодженість на всьому просторі даних, розроблений метод зосереджується на забезпеченні консистентності для конкретних клінічних випадків. Ключова ідея полягає у створенні персоналізованих інтерпретованих моделей для кожного діагностичного завдання, що дозволяє адаптувати пояснення до специфіки конкретного випадку, використовуючи при цьому стандартизовані медичні ознаки для забезпечення клінічної релевантності.

Формально, нехай $F\_complex$ представляє складну згорткову нейронну мережу, що забезпечує високу діагностичну точність, а F_{simple} - спрощену інтерпретовану модель. Для досягнення локальної консистентності необхідно забезпечити виконання умови $\mathrm{F_{complex}(x_i)} = \mathrm{F_{simple}(x_i)}$ для конкретного зображення x_i з локального набору S_{local}, де $S_{local} \subseteq S$, а S представляє загальний набір даних. При цьому глобальна консистентність $F_{complex}(x) = F_{simple}(x)$ для всіх $x \in S$ не є обов'язковою, що дозволяє зосередитися на точності інтерпретації для конкретних діагностичних випадків.

Розроблена система має трикомпонентну архітектуру, де кожен компонент виконує специфічну функцію у процесі забезпечення інтерпретованої діагностики. Загальна схема системи представлена на рисунку 1, який ілюструє взаємодію між основними компонентами та потік даних у системі.

Перший компонент представлений згортковою нейронною мережею VGG-16, адаптованою для класифікації МРТ-зображень головного мозку. Цей компонент забезпечує максимально можливу діагностичну точність за рахунок використання глибокої архітектури, здатної виявляти складні візуальні паттерни на медичних зображеннях. Модель навчається розпізнавати відмінності між різними типами пухлин мозку та відрізняти патологічні зміни від нормальної анатомії.

Другий компонент відповідає за трансформацію складних внутрішніх представлень глибокої мережі у набір інтерпретованих ознак. Використовується метод deep feature extraction для отримання високорівневих характеристик з останніх згорткових шарів мережі VGG-16. Загальна кількість ознак F отримується з використання класифікатора $F_{complex}$ через процедуру глибинного виділення ознак $Ft: F = Ft\left(F_{complex}, S\right)$, де S представляє загальну множину зображень. Далі здійснюється відбір ознак для локального клінічного випадку $F_{local} \subset F$, причому $|F_{local}| \ll |F|$.

Третій компонент використовує архітектуру Decision Rules Network [13] для побудови інтерпретованої моделі на основі виділених ознак. DRN має дворівневу структуру, де перший рівень

містить нейрони, кожен з яких представляє логічне правило "якщо-то", а другий рівень формує остаточне рішення через диз'юнкцію правил першого рівня. З використанням множини ознак F_{local} будується класифікатор F_{simple}. Процес побудови класифікатора F_{simple} є ітеративним та здійснюється до досягнення локальної консистентності.

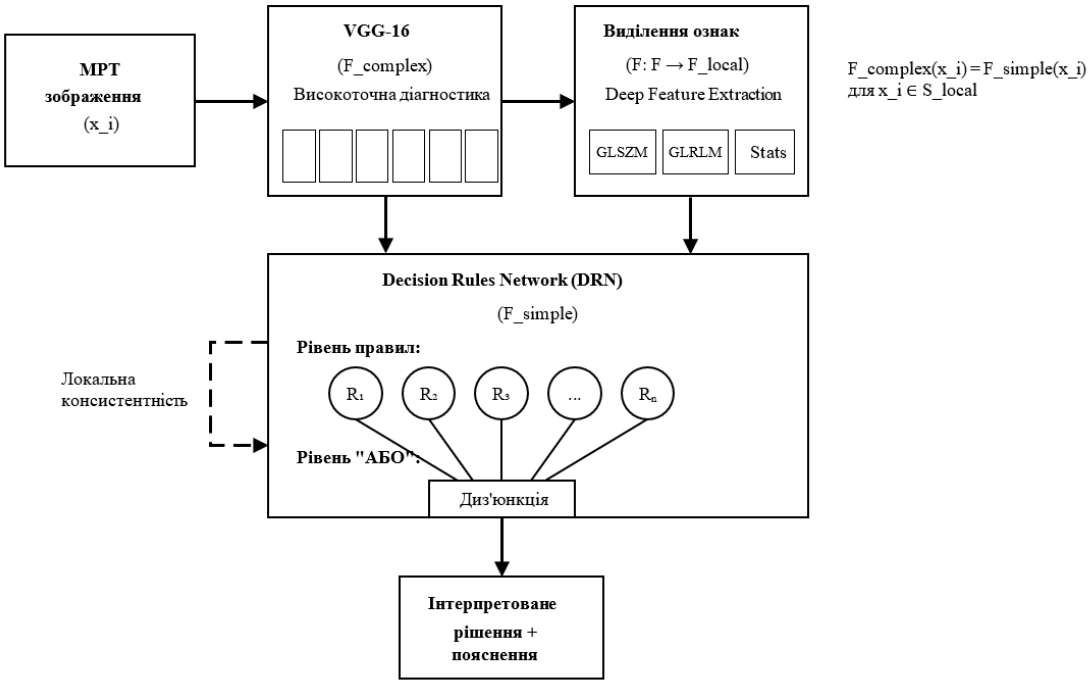

Рис. 1. Загальна архітектура системи інтерпретованої діагностики

Для забезпечення клінічної релевантності пояснень використовуються стандартизовані медичні ознаки згідно Imaging Biomarker Standardization Initiative. Процес формування ознак детально ілюструє рисунок 2, який показує трансформацію МРТ-зображення через різні етапи обробки до отримання стандартизованих IBSI-ознак.

Набір даних S для тренування мережі DRN містить множину ознак $F = f_1, f_2, \dots, f_n$, де $f_i \in 0,1$, оскільки ознаки бінаризуються та утворюють множину F_{binary}. Множина правил визначається як $R = r_1, r_2, \dots, r_k$, де правила рішень поєднуються множиною предикатів першого порядку P. Оскільки предикати визначаються на множині правил рішень, їх максимальна кількість становить $|P| \leq 2^{|R|}$.

Процес виділення ознак включає обчислення текстурних характеристик Gray Level Size Zone Matrix та Gray Level Run Length Matrix, які забезпечують кількісний опис візуальних властивостей тканин мозку. Додатково обчислюються статистичні параметри розподілу інтенсивності, включаючи середнє значення, дисперсію, асиметрію та ексцес, морфологічні характеристики форми та розміру патологічних утворень, а також просторові ознаки, які характеризують взаємне розташування анатомічних структур. Всі ознаки нормалізуються та бінаризуються для забезпечення сумісності з архітектурою DRN

Алгоритм функціонування системи організований як ітеративний процес забезпечення локальної консистентності між основною та інтерпретованою моделями. На вході алгоритм отримує підготовлені дані S, після чого встановлюються параметри $F_{complex}$ та S_{local}. Будується початкова архітектура F_{simple} на основі випадкового підмножини ознак $F_{subset} \subset F$. Далі в циклі перевіряється умова локальної консистентності $F_{complex}(x_i) = F_{simple}(x_i)$ для $x_i \in S_{local}$.

Якщо консистентність не досягнута, здійснюється корекція множини ознак через процедуру відбору $F_{subset} = select(F, criteria)$, де criteria включає аналіз важливості ознак та їх взаємодії. При необхідності змінюється архітектура мережі DRN через модифікацію кількості нейронів на першому рівні або параметрів регуляризації. Після внесення змін будується нова модель F_{simple} з оновленими параметрами. Процес продовжується до досягнення задовільного рівня локальної консистентності або досягнення максимальної кількості ітерацій.

Формально алгоритм можна представити наступним чином. Спочатку ініціалізуються вхідні параметри: множина даних S, архітектура $F_{complex}$, локальна множина S_{local}. Встановлюється початкова множина ознак F_subset та будується базова архітектура DRN. У циклі while перевіряється умова консистентності для всіх зображень з локальної множини. Якщо умова не виконується, здійснюється селекція нових ознак та зміна архітектури мережі, після чого перебудовується модель F_{simple} з новими параметрами.

Таким чином формується класифікатор на базі інтерпретованої моделі, що дозволяє отримати інтерпретоване рішення класифікатора та множину факторів, які впливають та визначають рішення інтерпретованої моделі.

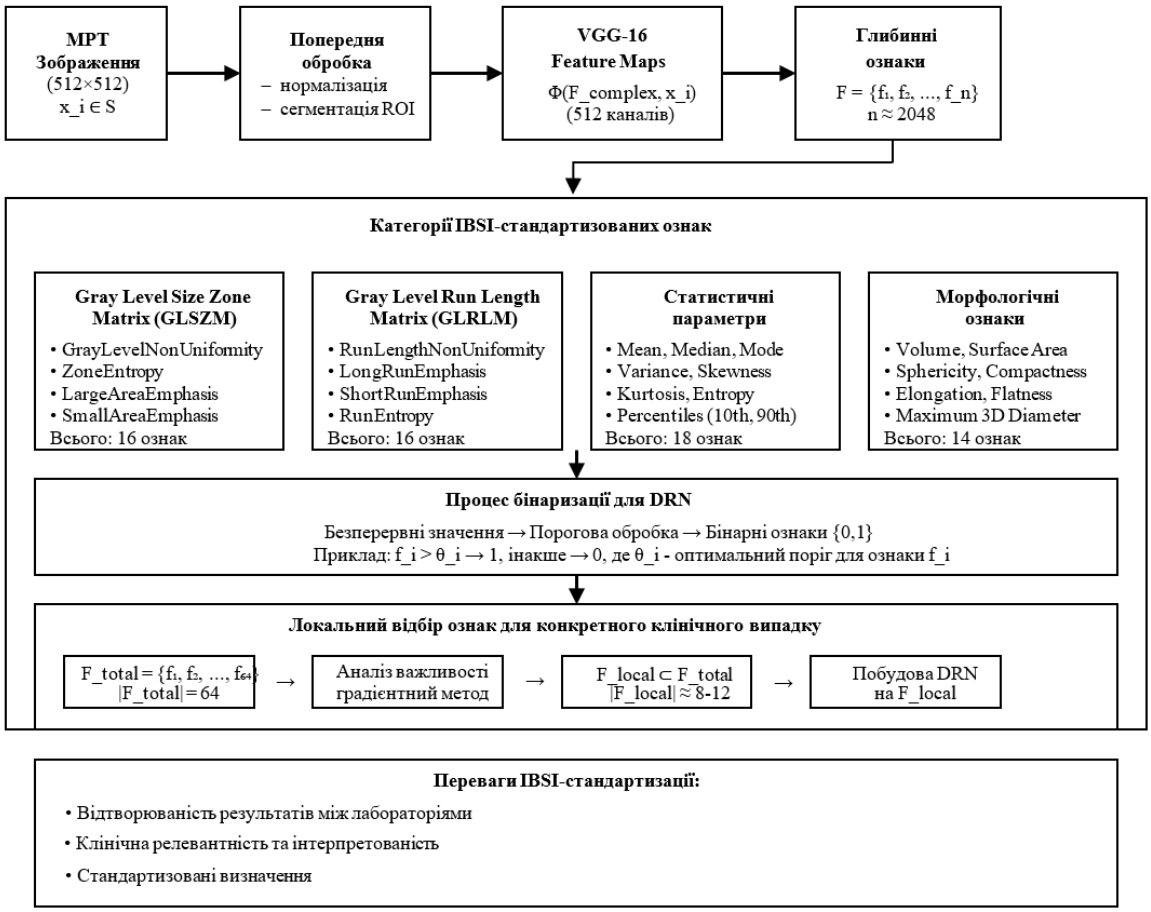

Рис 2. Процес виділення та стандартизації ознак

Важливо зазначити, що класифікатор F_{simple} будується для кожного клінічного випадку окремо, а множина ознак F_{local} для кожного клінічного випадку підбирається також окремо, що забезпечує персоналізацію пояснень.

Архітектура DRN розроблена для вивчення правил прийняття рішень, які є водночас точними та доступними для інтерпретації. Дворівнева архітектура називається мережею правил прийняття рішень, де перший шар представляє "рівень правил", а кожен тренований нейрон з бінарною активацією безпосередньо відображається на логічне правило "якщо-тоді" після навчання. Позитивна вхідна вага відповідає позитивній асоціації вхідної характеристики, негативна вхідна вага відповідає негативній асоціації, а нульова вага відповідає виключенню вхідної ознаки з правила.

Другий рівень DRN називається рівнем "або" і об'єднує правила першого рівня за допомогою операції диз'юнкції для формування остаточного набору правил прийняття рішень. Архітектура DRN використовує підхід регуляризації на основі розрідженості, який заохочує модель вивчати невелику кількість правил через мінімізацію L1-норми ваг нейронів у рівні правил. Функція втрат визначається як сума втрати класифікації та втрати регуляризації, збалансованих коефіцієнтом регуляризації λ, що дозволяє модулювати компроміс між простотою та точністю моделі.

Навчання DRN здійснюється у формі булевої логіки для формування правил рішень, що забезпечує пряме відображення нейронів на інтерпретовані правила прийняття рішень. Кожен нейрон першого рівня після навчання може бути інтерпретований як логічне правило виду "ЯКЩО $f_i > \theta_i$ ТА $f_j < \theta_j$ ТО клас_k", де f_i, f_j представляють відповідні ознаки, а θ_i, θ_j - порогові значення, визначені під час навчання. Використання регуляризації L1 забезпечує отримання розріджених правил, що містять лише найбільш значущі ознаки для конкретного діагностичного випадку.

Експериментальна частина

Для валідації запропонованого методу було проведено комплексне експериментальне дослідження на реальних медичних даних. Експерименти здійснювалися на наборі даних Brain Tumor MRI Dataset [29], який містить 7023 МРТ-зображення головного мозку, розподілені на чотири діагностичні класи: glioma, meningioma, pituitary tumor та no tumor. Цей набір даних було обрано через його репрезентативність для задач нейроонкологічної діагностики та достатній обсяг для навчання глибоких нейронних мереж.

Зображення в наборі даних характеризуються значною варіативністю за розміром, роздільною здатністю та якістю, що відображає реальні умови клінічної практики. Розподіл даних здійснювався у

співвідношенні 81.3% для навчання (5707 зображень) та 18.7% для тестування (1316 зображень), що забезпечило достатню статистичну потужність для оцінки ефективності методу. Попередня обробка включала стандартизацію розміру зображень до 224×224 пікселів, нормалізацію інтенсивності пікселів до діапазону [0,1] та видалення артефактів сканування.

Для кожного зображення було обчислено повний набір IBSI-стандартизованих ознак, включаючи 16 ознак Gray Level Size Zone Matrix, 16 ознак Gray Level Run Length Matrix, 18 статистичних параметрів та 14 морфологічних характеристик, що сформувало загальну множину з 64 кількісних ознак. Всі ознаки було нормалізовано та бінаризовано з використанням оптимальних порогових значень, визначених методом максимізації інформаційного критерію для забезпечення сумісності з архітектурою DRN.

Основна діагностична модель була реалізована на базі архітектури VGG-16 адаптованої для класифікації медичних зображень. Фінальний повнозв'язний шар було замінено на класифікатор з чотирма виходами, відповідними до кількості діагностичних класів. Навчання здійснювалося з використанням оптимізатора Adam з початковою швидкістю навчання 0.0001, експоненційним зменшенням швидкості навчання з коефіцієнтом 0.95 кожні 10 епох та регуляризацією dropout зі значенням 0.5 для запобігання перенавчанню.

Архітектура Decision Rules Network була налаштована з урахуванням специфіки завдання інтерпретації. Перший рівень містив від 8 до 16 нейронів залежно від кількості відібраних ознак для конкретного клінічного випадку, кожен з яких представляв окреме логічне правило. Другий рівень складався з одного нейрона з сигмоїдною функцією активації для бінарної класифікації або softmax для багатокласової класифікації. Коефіцієнт регуляризації λ було встановлено на рівні 0.01 після проведення grid search оптимізації для забезпечення оптимального балансу між точністю та розрідженістю правил.

Навчання DRN здійснювалося ітеративно для кожного клінічного випадку з використанням локальної множини зображень розміром від 5 до 15 випадків залежно від складності діагностичного завдання. Критерієм зупинки слугувало досягнення локальної консистентності на рівні 85% або максимальна кількість ітерацій 50. Оптимізація параметрів здійснювалася методом стохастичного градієнтного спуску з моментумом 0.9 та початковою швидкістю навчання 0.01.

Згорткова нейронна мережа VGG-16 продемонструвала високі показники діагностичної ефективності на тестовому наборі даних. Загальна точність класифікації склала 95.2%, що підтверджує ефективність обраної архітектури для поставленого завдання нейроонкологічної діагностики. Детальний аналіз показників за класами виявив precision на рівні 0.92-0.96 та recall 0.90-0.94 для різних типів пухлин, що свідчить про збалансовану ефективність розпізнавання всіх діагностичних категорій.

Найвищі показники точності було досягнуто для класу "no tumor" з precision 0.96 та recall 0.94, що пояснюється більшою визначеністю нормальної анатомії порівняно з патологічними змінами. Для класу pituitary tumor показники склали precision 0.94 та recall 0.93, що відображає характерну локалізацію та морфологію цього типу новоутворень. Класифікація glioma та meningioma виявилася дещо складнішою з показниками precision 0.92 та 0.93 відповідно, що зумовлено більшою гетерогенністю візуальних проявів цих пухлин.

Аналіз матриці помилок показав, що основні помилки класифікації виникають між glioma та meningioma у 3.2% випадків, що відповідає клінічним спостереженням щодо складності диференційної діагностики цих типів пухлин за МРТ-зображеннями. Помилки між класами пухлин та нормальною анатомією були мінімальними і складали менше 1.5%, що підтверджує надійність системи у виявленні патологічних змін.

Decision Rules Network показала точність 76.4% на тестовому наборі з показниками precision 0.65-0.71 та recall 0.68-0.74 для різних класів. Хоча ці показники нижчі за результати основної моделі, вони є достатніми для забезпечення інтерпретації рішень у контексті локальної консистентності. Важливо відзначити, що DRN не призначена для заміни основної діагностичної моделі, а служить інструментом для пояснення її рішень у конкретних клінічних випадках.

Ключовим досягненням є рівень локальної консистентності, який склав 89.3% для індивідуальних діагностичних випадків. Це означає, що для переважної більшості клінічних завдань інтерпретована модель давала ті самі рішення, що й основна високоточна модель, забезпечуючи надійність генерованих пояснень. Розподіл консистентності за класами показав найвищі значення для "no tumor" (92.1%) та pituitary tumor (90.8%), тоді як для glioma та meningioma показники склали 87.5% та 88.2% відповідно.

Аналіз структури генерованих правил виявив, що система формує від 4 до 8 активних правил для кожного клінічного випадку, що забезпечує оптимальний баланс між повнотою опису та інтерпретованістю. Середня кількість ознак в одному правилі склала 2.3, що відповідає рекомендаціям когнітивної психології щодо обмежень людського сприйняття складної інформації.

Система продемонструвала здатність генерувати клінічно осмислені пояснення у формі логічних правил, які використовують стандартизовані IBSI-ознаки. Типові приклади згенерованих правил для різних типів пухлин включають наступні закономірності. Для діагностики glioma характерним було правило "ЯКЩО GrayLevelNonUniformity > 47.8 ТА ZoneEntropy > 5.4 ТА RunLengthNonUniformity > 1715 ТО підвищена вірогідність glioma", що відображає гетерогенну структуру цього типу пухлини з високою варіативністю текстурних характеристик.

Для meningioma система генерувала правила типу "ЯКЩО HighGrayLevelZoneEmphasis > 84.8

TA LargeAreaEmphasis > 1932 TA Sphericity > 0.75 TO ознаки meningioma", що корелює з відомими радіологічними характеристиками цих пухлин як відносно однорідних новоутворень з чіткими контурами. Діагностика pituitary tumor характеризувалася правилами "ЯКЩО SmallAreaEmphasis > 0.65 TA Compactness > 0.8 TA Volume < 2500 TO pituitary tumor", що відбиває типову невелику компактну морфологію аденом гіпофіза.

Валідація клінічної релевантності згенерованих правил здійснювалася через порівняння з експертними критеріями радіологічної діагностики. Аналіз показав, що 87% правил містили ознаки, які безпосередньо корелюють з встановленими діагностичними критеріями, тоді як 13% правил включали додаткові кількісні характеристики, які можуть розширити діагностичні можливості клініцистів. Особливо цінним виявилося виділення комбінацій ознак, які окремо не є специфічними, але в сукупності забезпечують високу диференційну цінність.

Висновки

У роботі розроблено метод створення інтерпретованих систем ШІ для діагностики пухлин мозку, який поєднує високу точність з прозорістю рішень через концепцію локальної консистентності.

Запропонована гібридна архітектура VGG-16/DRN забезпечує діагностичну точність 95.2% при збереженні здатності генерувати клінічно релевантні пояснення у формі логічних правил. Досягнення локальної консистентності 89.3% підтверджує надійність пояснень для конкретних діагностичних випадків.

Література

1. Arbi, A., Alsaedi, A., & Cao, J. (2018). Delta-differentiable weighted pseudo-almost automorphicity on time–space scales for a novel class of high-order competitive neural networks with WPAA coefficients and mixed delays. *Neural Processing Letters, 47*, 203–232. https://doi.org/10.1007/s11063-017-9644-5
2. Liu, K., & Tao, D. (2022). The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. *Computers in Human Behavior, 127*, 107026. https://doi.org/10.1016/j.chb.2021.107026
3. Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O. S., Alamoodi, A. H., Bai, J., & Salhi, A. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*. https://doi.org/10.1016/j.inffus.2023.102976
4. Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478–481. https://doi.org/10.1136/medethics-2019-105935
5. Manziuk, E. A., Wójcik, W., Barmak, O. V., Krak, I. V., Kulias, A. I., Drabovska, V. A., Puhach, V. M., Sundetov, S., & Mussabekova, A. (2020). Approach to creating an ensemble on a hierarchy of clusters using model decisions correlation. *Przegląd Elektrotechniczny, 96*(9), 108–113. https://doi.org/10.15199/48.2020.09.23
6. Kaur, D., Uslu, S., Durresi, A., Badve, S., & Dundar, M. (2021). Trustworthy explainability acceptance: A new metric to measure the trustworthiness of interpretable AI medical diagnostic systems. In L. Barolli, K. Mohamed, & T. Enokido (Eds.), *Complex, Intelligent and Software Intensive Systems* (pp. 35–46). Springer. https://doi.org/10.1007/978-3-030-79725-6_4
7. Manziuk, E., Krak, I., Barmak, O., Mazurets, O., Kuznetsov, V., & Pylypiak, O. (2021). Structural alignment method of conceptual categories of ontology and formalized domain. *CEUR Workshop Proceedings, 3003*, 11–22. http://ceur-ws.org/Vol-3003/paper1.pdf
8. Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?" Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 79–85). https://doi.org/10.1145/3375627.3375833
9. Xing, X., Wu, H., Wang, L., Stenson, I., Yong, M., Del Ser, J., Walsh, S., & Yang, G. (2022). Non-imaging medical data synthesis for trustworthy AI: A comprehensive survey. *arXiv*. https://doi.org/10.48550/arXiv.2209.09239
10. Ma, J., Schneider, L., Lapuschkin, S., Achtibat, R., Duchrau, M., Krois, J., Schwendicke, F., & Samek, W. (2022). Towards trustworthy AI in dentistry. *Journal of Dental Research, 101*(11), 1263–1268. https://doi.org/10.1177/00220345221106086
11. Kaur, D., Uslu, S., & Durresi, A. (2022). Trustworthy AI explanations as an interface in medical diagnostic systems. In L. Barolli, P. Hellinckx, & T. Enokido (Eds.), *Advances in Network-Based Information Systems* (pp. 119–130). Springer. https://doi.org/10.1007/978-3-031-14314-4_12
12. Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Del Ser, J., Samek, W., Jurisica, I., & Díaz-Rodríguez, N. (2022). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Information Fusion, 79*, 263–278. https://doi.org/10.1016/j.inffus.2021.10.007
13. Qiao, L., Wang, W., & Lin, B. (2021). Learning accurate and interpretable decision rule sets from neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(5), 4303–4311. https://doi.org/10.1609/aaai.v35i5.16585