НОВІ ЗАСОБИ КІБЕРНЕТИКИ, ІНФОРМАТИКИ, ОБЧИСЛЮВАЛЬНОЇ ТЕХНІКИ ТА СИСТЕМНОГО АНАЛІЗУ

УДК 004.048+616-079.4

В.О. БАБЕНКО
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна, e-mail: *vbabenko2191@gmail.com*.

Є.А. НАСТЕНКО
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна, e-mail: *nastenko.e@gmail.com*.

В.А. ПАВЛОВ
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна, e-mail: *pavlov.vladimir264@gmail.com*.

О.К. ГОРОДЕЦЬКА
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна, e-mail: *o.nosovets@gmail.com.*

І.М. ДИКАН
Інститут ядерної медицини і променевої діагностики, Київ, Україна,
e-mail: *irinadykan@gmail.com*.

Б.А. ТАРАСЮК
Інститут ядерної медицини і променевої діагностики, Київ, Україна,
e-mail: *btarasyuk13@gmail.com*.

В.В. ЛАЗОРИШИНЕЦЬ
Національний інститут серцево-судинної хірургії імені М.М. Амосова, Київ, Україна,
e-mail: *lazorch@ukr.net*.

КЛАСИФІКАЦІЯ ПАТОЛОГІЙ ЗА МЕДИЧНИМИ ЗОБРАЖЕННЯМИ АЛГОРИТМОМ ВИПАДКОВОГО ЛІСУ ДЕРЕВ ОПТИМАЛЬНОЇ СКЛАДНОСТІ

Анотація. Запропоновано підхід до побудови класифікаторів у класі алгоритмів Random Forest. Для визначення оптимального поєднання та складу ансамблів ознак під час побудови дерев лісу застосовано генетичний алгоритм. Для оптимізації структури дерев використано принципи методу групового урахування аргументів. Оптимізацію процедури голосування дерев у лісі реалізовано із застосуванням методу аналізу ієрархій. Наведено приклади використання запропонованого алгоритму для виявлення патологій на медичних зображеннях, а також результати класифікації порівняно з іншими відомими аналогами.

Ключові слова: класифікація патологій, медичні зображення, Random Forest, генетичний алгоритм, метод групового урахування аргументів, метод аналізу ієрархій.

ВСТУП

Розвиток інформаційно-обчислювальних технологій зумовив перспективи розв’язання задач ідентифікації та прогнозування стану об’єктів різного масштабу та складності в широкому колі предметних галузей. Ресурси сучасних комп’ютерних систем забезпечують синтез моделей штучного інтелекту та алгоритмів машинного навчання з високим рівнем якості розв’язання прикладних задач [1]. Поява надвеликих наборів даних зумовила стабільний зростаючий попит на застосування цих алгоритмів завдяки їхній здатності до ефективного оброблення даних у задачах інтелектуального аналізу.

Нині для синтезу моделей штучного інтелекту найбільш ефективним є використання таких підходів: алгоритмів ансамблевого навчання (бустинг [2, 3], випадкового лісу [4, 5], стекінгу [6]) та алгоритмів глибокого навчання (згорткових [7, 8] і рекурентних [9] нейронних мереж). Ці підходи конкурують між собою, маючи свої переваги та недоліки. Ансамблеве навчання є перспективним для аналізу, оскільки вибрані методології створення множини моделей та згортки ансамблів підказують відповідні технології інтерпретації одержуваних рішень. До того ж, як свідчить практика конкурсів прогнозування даних від сервісу Kaggle [10], ансамблеві алгоритми випадкового лісу (Random Forest) є одними з найбільш ефективних для розв'язання більшості задач [11].

Метою пропонованої роботи є вдосконалення підходів ансамблевого навчання у класі алгоритмів Random Forest для досягнення більшої точності результатів розв'язання задач класифікації. Це особливо актуально у задачах розпізнавання патології на медичних зображеннях [5, 12, 13], що забезпечує підтримку прийняття рішення під час діагностики пацієнта. Крім того, розроблення високопродуктивних алгоритмів класифікації є актуальним у разі застосування у неінвазивних підходах діагностики патологій [13, 14]. Розвиток методів неінвазивної діагностики є суттєво важливим для пацієнтів, оскільки інвазивні підходи [14], хоча і вважаються більш точними, можуть завдавати значної шкоди людському організму.

АНАЛІЗ ДОСЛІДЖЕНЬ, ЩО РОЗВИВАЮТЬ КОНЦЕПЦІЮ АЛГОРИТМІВ RANDOM FOREST

Згідно з принципом ансамблевого навчання, щоб розв'язати певну проблему, застосовують не одну модель, а множину моделей, які називають «слабкими учнями».

Основна гіпотеза полягає в тому, що шляхом оптимального поєднання ансамблю моделей можна отримати більш точні та надійні результати прогнозування. Мета-алгоритм ансамблевого навчання містить три основні процедури, зосереджені на об'єднанні «слабких учнів»: бегінг, бустинг та стекінг.

Випадковий ліс є найбільш яскравим представником бегінгу, де основним принципом є застосування однорідних «слабких учнів». Навчають їх паралельно та незалежно один від одного, далі одержані класифікатори об'єднують за допомогою детермінованого процесу усереднення. Цей алгоритм був винайдений Тін Кам Хо [4] у 1995 р. та досі є актуальним у сучасних застосуваннях.

Проте випадковий ліс має певні недоліки, які свого часу намагалися усунути різні вчені. У роботі [15] для створення вдосконаленого алгоритму випадкового лісу запропоновано свій варіант оптимізації багатовимірних лінійних порогових функцій як окремих функцій дерева прийняття рішень. Автори зазначають, що введений критерій приросту інформації, використаний для створення дерев, є перервним і його важко оптимізувати. Проте вони пропонують скористатися для оптимізації його безперервною верхньою межею. Це забезпечує значне покращення результатів класифікації порівняно з базовим випадковим лісом.

У дослідженні [16] описано комбінацію випадкового лісу з методом оцінювання атрибутів і методом фільтрації екземплярів. Пропонований підхід покращив результати багатокласової задачі класифікації та забезпечив точність близько ста відсотків. Надійність результатів підтверджено використанням алгоритму на п'яти контрольних наборах даних, де в кожному випадку варіант авторів мав переваги над класичним варіантом лісу.

Окремі версії алгоритмів моделювання випадкового лісу можуть потребувати налаштування основних параметрів: кількості дерев рішень (що утворюють ансамбль) та кількості порогів розділення в кожному вузлі. За припущення щодо існування підкласів можна задавати вектори, що представляють кластери у кожному класі. Значна розмірність простору пошуку цих параметрів спонукала E. Elyan та M.M. Gaber [17] скористатися генетичним алгоритмом для оптимізації рішень. Успіх підходу підтверджено покращеною точністю результатів для кількох наборів даних з різних галузей застосування.

Автори цього дослідження запропонували новий спосіб удосконалення у класі алгоритмів випадкового лісу [18, 19]. Він ґрунтується на використанні

принципів методу групового урахування аргументів (МГУА) [20] у навчанні дерев рішень, що дає змогу отримати у такий спосіб дерева оптимальної складності. Алгоритм був застосований для розпізнавання патології на зображеннях ультразвукового дослідження (УЗД) печінки, де була досягнута точність класифікації в межах від 96 до 99 %. Проте пропонований варіант мав низку недоліків, які в цій роботі усунено.

АЛГОРИТМ ВИПАДКОВОГО ЛІСУ ДЕРЕВ ОПТИМАЛЬНОЇ СКЛАДНОСТІ

Відмітною особливістю алгоритму випадкового лісу є використання принципу бутстреп-агрегації (бегінгу): паралельного навчання різних дерев незалежно одне від одного з подальшою агрегацією моделей для одержання кінцевого результату.

Метою застосування бегінгу є часткова нейтралізація ефекту змінності моделі лісу внаслідок варіативності вибірки вхідних даних. Використовуючи бутстреп, можна зменшити помилку класифікації тестових даних. Завдяки цьому модель набуває властивостей не апроксимації, а узагальнення результатів.

Структура лісу (кількість дерев) обмежується помилкою тестової вибірки даних. Ці механізми у класичній реалізації випадкового лісу є основними у запобіганні перенавчанню моделі. Нижче запропоновано низку механізмів оптимізації параметрів, структури та агрегації дерев лісу задля покращення результатів задач класифікації.

Побудова дерева оптимальної складності. Зазвичай у стандартних версіях випадкового лісу під час моделювання дерев у кожному вузлі застосовують деяке значення порогу незалежної змінної, яке є оптимальним за такою функцією вартості, як індекс Джіні [21] або ентропія [22].

Автори пропонують для визначення порогів розбиття діапазонів значення ознак застосовувати коефіцієнт кореляції Метьюза (ККМ) [23], як найбільш загальну форму критерію якості класифікації, що враховує всі особливості задачі, забезпечуючи збалансовану метрику оцінювання класифікації у разі несиметричного наповнення класів [23].

Визначення структури (змінних у вузлах) та параметрів (порогів) дерева полягає в ітеративному переборі незалежних змінних та їхніх порогів, що забезпечують максимум функції

$$W = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \qquad (1)$$

де TP (True Positives) — кількість правильних прогнозів першого класу, TN (True Negatives) — кількість правильних прогнозів другого класу, FP (False Positives) — кількість неправильних прогнозів першого класу (α-помилка або помилка першого роду), FN (False Negatives) — кількість неправильних прогнозів другого класу (β-помилка або помилка другого роду).

Процедуру синтезу дерева рішень запропоновано реалізовувати за принципами МГУА [20] з використанням двох вибірок: A (навчання) і B (перевірка). Цей індуктивний підхід дає змогу отримувати моделі оптимальної складності, що максимізує точність прогнозу на перевірочній вибірці B.

У вузлі дерева оптимальний поріг розраховують на вибірці A для кожної незалежної змінної, а вибір змінної здійснюють на вибірці B. Отже, використання W (1) як функції вартості та індуктивного підходу МГУА для навчання дерева забезпечує отримання дерев оптимальної складності (рис. 1).

На рис. 1 наведено приклад одного з дерев оптимальної складності, одержаного під час розв'язання задачі розпізнавання патології за ультразвуковими зображеннями печінки [18, 19]. Точність класифікації на тестовій вибірці становила від 94 до 97 %.

Формування лісу дерев оптимальної складності. Під час формування лісу дерев оптимальної складності застосовують принцип бегінгу (рис. 2). Ідею алгоритму бегінгу запропонував статистик Л. Брайман [24, 25] задля покращення якості прогнозування за рахунок комбінування (агрегування) прогнозів, одержаних на згенерованих у випадковий спосіб різних тренувальних наборах даних. Брайман також довів [26], що бегінг дає змогу знаходити більш ефек-

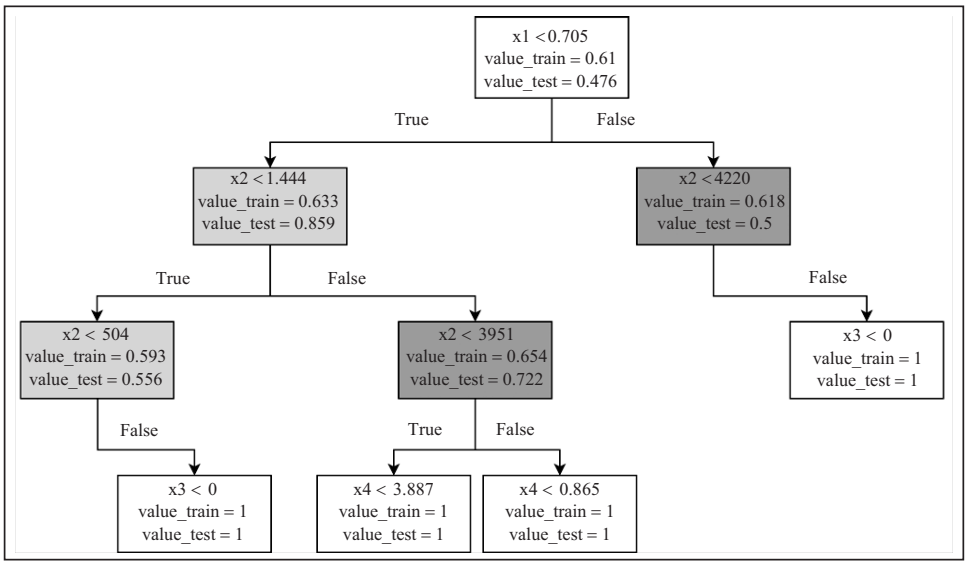

Рис. 1. Приклад дерева оптимальної складності

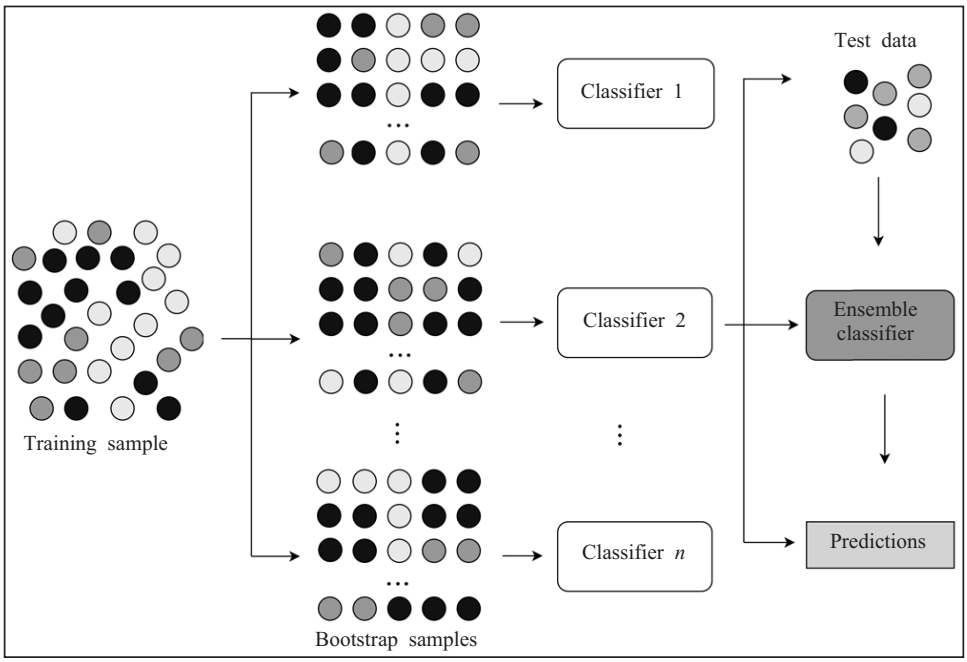

Рис. 2. Ілюстративний приклад роботи бегінгу

тивні результати для нестабільних процедур, до яких відносять нейронні мережі та дерева прийняття рішень.

Ключовим компонентом бегінгу є «мудрість натовпу»: результат формується як колективне рішення, а не як рішення окремого експерта. Перевагою цього підходу є те, що, виконуючи спільне прогнозування, моделі підстраховують одна одну допоки не помиляться на одних і тих самих об’єктах.

Необхідною умовою ефективності цього принципу є різноманітність і спеціалізація моделей, що у контексті дерев прийняття рішень забезпечується навчанням на різних вибірках даних. Дерева є чутливими до варіації вхідних даних, тому зміни у складі вибірки призводять до іншої структури дерева.

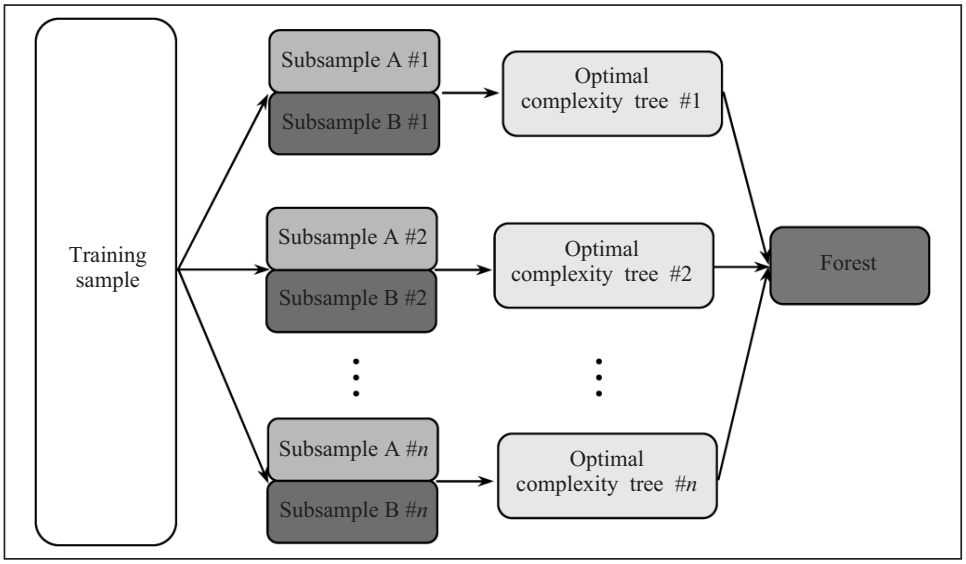

Рис. 3. Ліс дерев оптимальної складності

Для побудови дерев оптимальної складності бегінг формує n навчальних вибірок, які за принципом МГУА поділяються на підвибірки A і B (рис. 3).

Так само як метод випадкового лісу, пропонований алгоритм спрямований на формування ансамблю дерев прийняття рішень, кожне з яких одержано на підвибірках A_i та B_i, та які (A_i з A_j та B_i з B_j) лише частково перетинаються для $i \neq j$. Застосовуючи принципи МГУА з використанням підходу бегінгу, формують дерева класифікації та в результаті отримують випадковий ліс дерев оптимальної складності (ВЛДОС).

Ще однією особливістю випадкового лісу є застосування для кожного дерева генерації випадкового підпростору ознак [4], що певною мірою підвищує ефективність прогнозування. Проте виникає питання оптимальності побудованого дерева рішень, оскільки у разі повторного запуску алгоритму отримують інші результати. Таку задачу в цій роботі розв'язано аналогічно до [17] за допомогою генетичного алгоритму [27, 28].

Послідовність дій з розв'язання проблеми містить такі кроки.

1. Для заданих підвибірок A_i та B_i у випадковий спосіб генерують k (заданих користувачем) підмножин незалежних змінних.

2. Отримують k навчених дерев оптимальної складності.

3. Дерева оцінюють на валідаційній вибірці (що не брала участь у навчанні) за значенням цільової функції W (1) генетичного алгоритму.

4. Виконують перевірку умови зупинки алгоритму (досягнуто максимум цільової функції або перевищено заданий ліміт епох).

4.1. Якщо умова зупинки виконується успішно, то отримують оптимальну підмножину ознак i-го дерева.

4.2. В іншому випадку виконують генетичні оператори (селекція, кросовер/мутація), повертаючи нове покоління k підмножин незалежних ознак.

5. Повторюють пп. 2, 3 доки не будуть досягнуті умови зупинки алгоритму в п. 4.

Цей алгоритм виконують для кожного i-го дерева. У результаті отримують ВЛДОС, що складається з n дерев з оптимальними підвибірками незалежних змінних.

Удосконалення функції голосування лісу. Природним способом підвищення ефективності колективного механізму прийняття рішень є використання оптимізаційних процедур для формування функції голосування. Вдосконалення інтеграції первинних результатів запропоновано здійснювати шляхом зважено-

го голосування за допомогою багатокритерійного методу прийняття рішень — методу аналізу ієрархій Сааті [28, 29].

Сенс пропозиції полягає у присвоєнні кожному дереву ВЛДОС вагового коефіцієнта, який надає пріоритет кращим моделям під час процедури голосування. Вагові коефіцієнти отримують за допомогою механізму попарного порівняння (табл. 1) пріоритетів критеріїв (у цьому випадку — дерев ВЛДОС).

Тут v_i є порядковим номером у списку критеріїв, ранжованих за якістю моделі.

Таблиця 1. Загальний вигляд попарного порівняння критеріїв

Критерії	y_1	y_2	…	y_n
y_1	$\frac{v_1}{v_1}$	$\frac{v_2}{v_1}$	…	$\frac{v_n}{v_1}$
y_2	$\frac{v_1}{v_2}$	$\frac{v_2}{v_2}$	…	$\frac{v_n}{v_2}$
…	…	…	…	…
y_n	$\frac{v_1}{v_n}$	$\frac{v_2}{v_n}$	…	$\frac{v_n}{v_n}$

Після підставлення v_i у таблицю попарних порівнянь в i-му рядку таблиці обчислюють середнє геометричне значення, кожне з яких нормують через ділення на суму усіх значень, зводячи їх у такий спосіб до інтервалу від 0 до 1.

Якість моделі можна визначити зі значень функції W (1) на валідаційній вибірці. Наведемо такий приклад. Нехай є побудований ВЛДОС із 11 дерев. На валідаційній вибірці значення функції W є такими: {0.821, 0.766, 0.749, 0.747, 0.766, 0.745, 0.575, 0.808, 0.749, 0.821, 0.823}. Ранжований ряд має вигляд {2, 4, 5, 6, 4, 7, 8, 3, 5, 2, 1}. Обчисливши нормовані середні геометричні методом аналізу ієрархій, отримуємо список вагових коефіцієнтів: {0.14, 0.093, 0.07, 0.047, 0.093, 0.047, 0.023, 0.116, 0.07, 0.14, 0.163}. Ці коефіцієнти будуть вагами дерев оптимальної складності під час голосування.

Остаточна класифікація випливає з функції адитивної згортки F_{ac}:

$$F_{ac} = w_1 y_1 + w_2 y_2 + \ldots + w_n y_n, \qquad (2)$$

де y_i — результат класифікації (–1 або 1), отриманий i-м деревом ВЛДОС.

Значення функції F_{ac} варіюються в інтервалі від –1 до 1. Для $F_{ac} < 0$ результатом класифікації буде перший клас, в іншому разі — другий клас.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Розроблений алгоритм ВЛДОС (технічно реалізований мовою програмування Python) застосовано для класифікації на наборі даних медичних зображень у задачах виявлення патологій. Використано дві бази даних знімків, які більш детально описані нижче.

Розпізнавання патології печінки. За замовленням державної установи «Інститут ядерної медицини та променевої діагностики НАМН України» розроблено систему підтримки прийняття рішень під час визначення стану «норма–патологія» печінки пацієнта.

Спеціалістами Інституту надано базу даних ультразвукового дослідження (УЗД) печінки (рис. 4), що складається зі 163 зображень печінки у нормальному стані та 154 знімків патології.

Патології представлено зображеннями ознак аутоімунного гепатиту (53 знімки), хвороби Вільсона (50 знімків), гепатиту Б (чотири знімки), гепатиту С (11 знімків), стеатозу (п'ять знімків) і цирозу (15 знімків); 12 знімків не були ідентифіковані як такі, що свідчать про конкретну патологію, але вони також були використані у дослідженні. Зображення одержано за допомогою двох різних типів датчиків:

- з використанням конвексного — 152 зображення ознак: 89 норми та 63 патології (18 знімків аутоімунного гепатиту, 23 знімки хвороби Вільсона, п'ять знімків гепатиту С, п'ять знімків стеатозу, три знімки цирозу і п'ять знімків невідомої патології);
- з використанням лінійного — 94 зображення ознак: 50 норми та 44 патології (26 знімків аутоімунного гепатиту, сім знімків хвороби Вільсона, два знімки гепатиту С, шість знімків цирозу і три знімки невідомої патології). Для

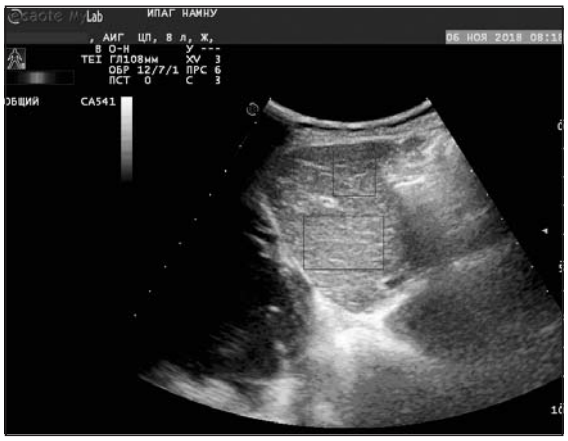

Рис. 4. Приклад знімку УЗД печінки (знімки є знеособленими)

більшої наочності фахівці взяли 71 зображення УЗД, отримане за допомогою лінійного датчика у посиленому режимі: 24 норми та 47 патології (дев'ять знімків аутоімунного гепатиту, 20 знімків хвороби Вільсона, чотири знімки гепатиту Б, чотири знімки гепатиту С, шість знімків цирозу і чотири знімки невідомої патології).

Кожен знімок УЗД був вручну сегментований експертами Інституту. Ділянки сегментації (або ділянки інтересу) вважаються найбільш інформативними у діагностуванні патології. Їх далі було використано як об'єкти дослідження та навчання системи класифікації.

Всього було використано:

1) 304 об'єкти конвексного датчика (197 зображень норми та 107 — патології);

2) 154 об'єкти лінійного датчика стандартного режиму (80 зображень норми та 74 — патології);

3) 124 об'єкти лінійного датчика посиленого режиму (35 зображень норми та 89 — патології).

Отже, формуються три вибірки даних, для кожної з яких окремо виконується задача бінарної класифікації («норма–патологія»). Слід звернути увагу на наявну незбалансованість класів на вибірках конвексного датчика і лінійного датчика посиленого режиму, яку потрібно враховувати під час побудови моделей класифікації.

Одержані ділянки сегментації зображень не були супроводжені метаданими, які можна було б використати як ознаки класів «норма–патологія» печінки. Для формування множини ознак авторами застосовано методи текстурного аналізу (опис технології аналізу можна знайти в [13, 18, 19]) усіх об'єктів дослідження.

Обчислені текстурні ознаки використано у таких алгоритмах класифікації:

- логістична регресія [30];
- адаптивний бустинг (AdaBoost) [31];
- випадковий ліс [4, 5];
- авторський алгоритм ВЛДОС.

Проблему масштабованості ознак розв'язано шляхом приведення до єдиної шкали від 0 до 1 за допомогою max-min-нормалізації.

Найбільш узагальнені моделі за всіма алгоритмами класифікації отримано шляхом поділу кожної вибірки на навчальні (80 % від загальної вибірки), валідаційні (10 %) і тестові (10 %). Мета застосування валідаційної та тестової вибірок є такою: на валідаційній вибірці знаходять оптимальні гіперпараметри кожного класифікатора, на тестовій здійснюють незалежне оцінювання отриманих моделей.

Одержані моделі оцінено за критеріями точності (частка правильно класифікованих об'єктів), F-score (середнє гармонійне значення чутливості і специфічності) та ККМ (1). Результати бінарної класифікації для об'єктів дослідження кожного датчика представлено в табл. 2. У дужках наведено результати застосування ВЛДОС із голосуванням за більшістю.

Розпізнавання ішемічної хвороби серця. Задача розпізнавання ішемічної хвороби серця надійшла від державної установи «Національний інститут серцево-судинної хірургії ім. М.М. Амосова НАМН України». Фахівцями установи надано 154 відеозаписи спекл-трекінг ехокардіографії (СТЕ) для 56 пацієнтів з

Таблиця 2. Результати класифікації ділянок інтересу

Алгоритм класифікації	Вибірка	Метрика оцінювання моделі		
		Точність	F-score	ККМ
Конвексний датчик				
Логістична регресія	Навчальна	0.835	0.821	0.645
	Валідаційна	0.733	0.729	0.464
	Тестова	0.71	0.659	0.333
AdaBoost	Навчальна	0.996	0.995	0.991
	Валідаційна	0.667	0.641	0.342
	Тестова	0.71	0.635	0.319
Випадковий ліс	Навчальна	1	1	1
	Валідаційна	0.8	0.792	0.614
	Тестова	0.742	0.631	0.441
ВЛДОС	Навчальна	1	1	1
	Валідаційна	1	1	1
	Тестова	0.903 (0.867)	0.886 (0.788)	0.795 (0.693)
Лінійний датчик стандартного режиму				
Логістична регресія	Навчальна	0.894	0.894	0.789
	Валідаційна	0.667	0.661	0.327
	Тестова	0.875	0.873	0.775
AdaBoost	Навчальна	1	1	1
	Валідаційна	0.867	0.866	0.732
	Тестова	0.75	0.75	0.5
Випадковий ліс	Навчальна	1	1	1
	Валідаційна	0.867	0.866	0.732
	Тестова	0.875	0.873	0.775
ВЛДОС	Навчальна	1	1	1
	Валідаційна	1	1	1
	Тестова	1 (0.875)	1 (0.873)	1 (0.775)
Лінійний датчик посиленого режиму				
Логістична регресія	Навчальна	1	1	1
	Валідаційна	0.917	0.874	0.775
	Тестова	0.923	0.902	0.822
AdaBoost	Навчальна	1	1	1
	Валідаційна	0.583	0.556	0.192
	Тестова	0.615	0.575	0.158
Випадковий ліс	Навчальна	1	1	1
	Валідаційна	0.917	0.874	0.775
	Тестова	0.615	0.381	-0.192
ВЛДОС	Навчальна	1	1	1
	Валідаційна	1	1	1
	Тестова	1 (0.923)	1 (0.902)	1 (0.822)

підозрою на ішемічну хворобу серця (у 16 з них під час обстеження не було виявлено відхилень кінематики міокарда). Відео СТЕ (рис. 5) слугує демонстрацією серцевого циклу людини для визначення типу деформації міокарда.

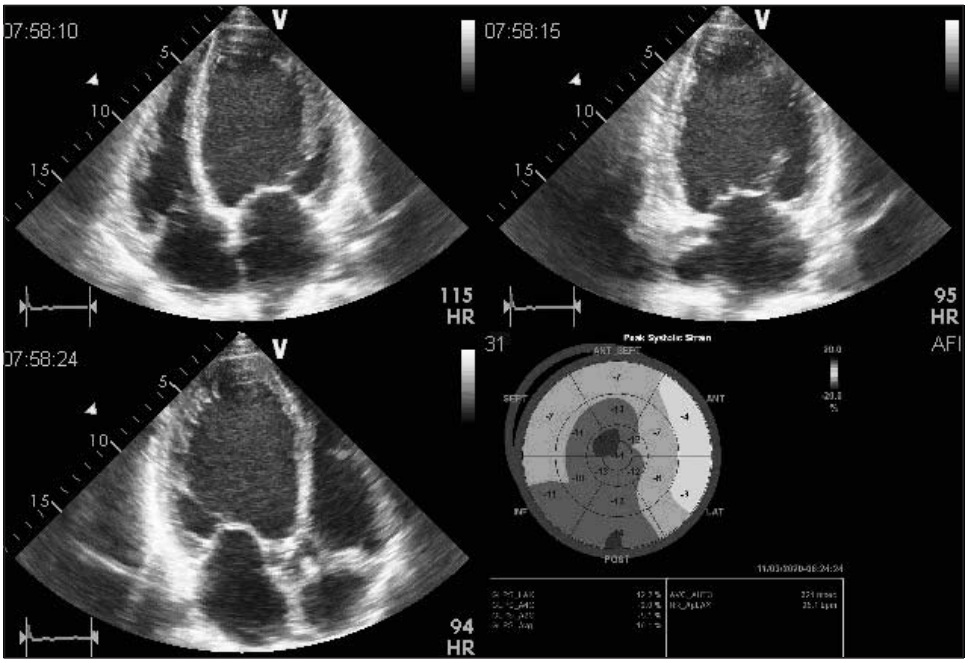

Рис. 5. Приклад наданого СТЕ

За допомогою цієї технології можна записувати ехокардіограму серця у *B*-режимі у трьох інформативних проєкціях (рис. 5): чотирикамерній позиції, двокамерній позиції та трикамерній позиції (поздовжня вісь). Кожна позиція відповідає певній комбінації сегментів лівого шлуночка серця, які розташовані в басейнах магістральних коронарних артерій. СТЕ проводилось на пацієнтах без застосування добутамінової проби (56 осіб) та із застосуванням добутамінової проби разом з ехострестестом (38 осіб) у тих випадках, коли відхилень у стані спокою не було виявлено. Дози добутаміну застосовувалися під наглядом анестезіолога за згодою пацієнтів, і використання проби неодмінно припинялось у разі появи щонайменших ознак дискомфорту та/або порушень серцевої діяльності.

Метою завдання є розроблення класифікаторів за кожною проєкцією ехокардіографії задля розпізнавання ішемічної хвороби серця за наданими відеопотоками даних. Для застосування даних кожне відео поділено на фрейми (кадри). Оскільки відео ехокардіографії відображає повний серцевий цикл, логічно припустити, що першим фреймом є систола серця, а останнім — діастола. На думку експертів, ці фрейми можуть бути найбільш інформативними для діагностики серця. Виходячи з цього, було прийнято рішення використати їх для подальшої задачі класифікації (загальна кількість таких фреймів становила 308). Надалі з кожного фрейму були окремо вирізані три позиції серця у *B*-режимі. Отже, сформовано три вибірки, для кожної з яких окремо розв'язано задачі бінарної класифікації. Розподіл об'єктів за класами «норма–патологія» є таким: 116 фреймів норми і 192 фрейми патології (чотирикамерна позиція), 130 фреймів норми і 178 фреймів патології (двокамерна позиція), 92 фрейми норми і 216 фреймів патології (поздовжня вісь). Наведені дані використано також у роботах [5, 32, 33].

Для розв'язання поставленої задачі застосовано такий набір алгоритмів: логістична регресія, адаптивний бустинг (AdaBoost), випадковий ліс, авторський алгоритм ВЛДОС. Результати розрахунків представлено в табл. 3. У дужках наведено результати застосування ВЛДОС із голосуванням за більшістю.

Таблиця 3. Результати класифікації фреймів

Алгоритм класифікації	Вибірка	Метрика оцінювання моделі		
		Точність	F-score	ККМ
Чотирикамерна позиція				
Логістична регресія	Навчальна	0.733	0.721	0.442
	Валідаційна	0.733	0.683	0.373
	Тестова	0.742	0.735	0.477
AdaBoost	Навчальна	0.984	0.983	0.966
	Валідаційна	0.8	0.744	0.489
	Тестова	0.774	0.765	0.533
Випадковий ліс	Навчальна	0.996	0.996	0.992
	Валідаційна	0.967	0.959	0.921
	Тестова	0.839	0.832	0.667
ВЛДОС	Навчальна	1	1	1
	Валідаційна	1	1	1
	Тестова	0.833	0.842	0.667
Двокамерна позиція				
Логістична регресія	Навчальна	0.785	0.784	0.574
	Валідаційна	0.6	0.583	0.172
	Тестова	0.742	0.735	0.47
AdaBoost	Навчальна	0.996	0.996	0.992
	Валідаційна	0.7	0.67	0.342
	Тестова	0.71	0.698	0.398
Випадковий ліс	Навчальна	1	1	1
	Валідаційна	0.8	0.785	0.569
	Тестова	0.774	0.765	0.533
ВЛДОС	Навчальна	1	1	1
	Валідаційна	1	1	1
	Тестова	0.846 (0.806)	0.889 (0.801)	0.735 (0.603)
Поздовжня вісь				
Логістична регресія	Навчальна	0.749	0.722	0.453
	Валідаційна	0.667	0.625	0.313
	Тестова	0.71	0.676	0.367
AdaBoost	Навчальна	0.976	0.971	0.943
	Валідаційна	0.767	0.689	0.38
	Тестова	0.774	0.748	0.513
Випадковий ліс	Навчальна	1	1	1
	Валідаційна	0.867	0.83	0.671
	Тестова	0.903	0.868	0.766
ВЛДОС	Навчальна	1	1	1
	Валідаційна	1	1	1
	Тестова	0.903 (0.871)	0.886 (0.8)	0.775 (0.71)

ОБГОВОРЕННЯ РЕЗУЛЬТАТІВ

Авторський алгоритм ВЛДОС продемонстрував кращу ефективність класифікації порівняно з іншими більш відомими алгоритмами як у першій, так і у другій задачах.

У задачі розпізнавання патології печінки за ультразвуковими зображеннями точність класифікації моделей ВЛДОС варіювалася від 90.3 до 100 % на тестовій вибірці. Серед аналогів найкращий результат показали моделі логістичної регресії, точність яких варіювалася від 71 до 92.3 %. Із табл. 2 видно, що макси-

мальна точність досягнута на вибірках лінійних датчиків обох типів завдяки зваженому голосуванню з використанням методу аналізу ієрархій. Досягнення максимальної точності на таких вибірках можна пояснити їхніми малими розмірами. Це підтверджується тим, що на більшій вибірці об'єктів дослідження конвексного датчика найкращий результат на тесті (отриманий алгоритмом ВЛДОС) становив 90.3 %.

У задачі розпізнавання ішемічної хвороби серця за відеоданими ехокардіографії точність класифікації моделей ВЛДОС варіювалася від 83.3 до 90.3 % на тестовій вибірці. Другі за ефективністю результати показали моделі випадкового лісу із варіацією точності від 77.4 до 90.3 %. Логічно припустити, що погіршення результатів порівняно з попередньою задачею пов'язано з розмірами вибірок, тому на майбутнє заплановано доопрацювання алгоритму ВЛДОС та адаптування класифікаторів до розширеної бази даних. Із табл. 3 видно, що підхід зваженого голосування (за винятком вибірки даних чотирикамерної позиції ехокардіографії) виявився стабільно кращим ніж голосування за більшістю.

Результати класифікації отримано для таких параметрів ВЛДОС.

• Розмір ансамблю ознак під час формування дерев оптимальної складності дорівнював квадратному кореню від загальної кількості використаних ознак. Подібний еквівалент рекомендовано задавати у класичному алгоритмі випадкового лісу [11].

• Побудовані моделі ВЛДОС складалися з 11 дерев. Цю кількість обрано через найкращі в середньому значення точності прогнозування (тестування також проведено на лісах із п'яти, 15 і 21 дерева оптимальної складності). Рекомендовано обирати непарну кількість дерев, аби не виникало спірних рішень під час голосування [4].

Врахування оптимізації зазначених параметрів є предметом подальшого удосконалення цього алгоритму задля підвищення його ефективності.

ВИСНОВКИ

За результатами дослідження запропоновано новий алгоритм ансамблевого навчання «Випадковий ліс дерев оптимальної складності» (ВЛДОС), який поєднує у собі підходи різних відомих методів, як-от: випадковий ліс, метод групового урахування аргументів, генетичний алгоритм та метод аналізу ієрархій. Алгоритм використано для розв'язання задач класифікації патологій на медичних зображеннях. Порівняння ефективності алгоритму ВЛДОС з відомими аналогами свідчить про те, що він забезпечує більш високі результати якості класифікації. Алгоритм не є специфічною розробкою для класифікації зображень, має універсальний характер та може бути застосований у різноманітних прикладних галузях.

СПИСОК ЛІТЕРАТУРИ

1. Sarker I.H. Machine learning: algorithms, real-world applications and research directions. *SN Computer Science.* 2021. Vol. 2, N 3. P. 160–160. https://doi.org/10.1007/s42979-021-00592-x.
2. Mayr A., Binder H., Gefeller O., Schmid M. The evolution of boosting algorithms. *Methods of Information in Medicine.* 2014. Vol. 53, N 06. P. 419–427. https://doi.org/10.3414/ME13-01-0122.
3. Osman A.H., Aljahdali H.M.A. An effective of ensemble boosting learning method for breast cancer virtual screening using neural network model. *IEEE Access*. 2020. Vol. 8. P. 39165–39174. https://doi.org/10.1109/ACCESS.2020.2976149.
4. Ho T.-K. Random decision forests. *Proc. 3rd International Conference on Document Analysis and Recognition* (14–16 August 1995, Montreal, QC, Canada). Montreal, 1995. Vol. 1. P. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.
5. Nastenko I., Maksymenko V., Potashev S., Pavlov V., Babenko V., Rysin S., Matviichuk O., Lazoryshinets V. Random forest algorithm construction for the diagnosis of coronary heart disease based on echocardiography video data streams. *Innovative Biosystems and Bioengineering*. 2021. Vol. 5, N 1. P. 61–69. https://doi.org/10.20535/ibb.2021.5.1.225794.

6. Pavlyshenko B. Using stacking approaches for machine learning models. *Proc. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)* (21–25 August 2018, Lviv, Ukraine). Lviv, 2018. P. 255–258. https://doi.org/10.1109/DSMP.2018.8478522.
7. Indolia S., Goswami A.K., Mishra S.P., Asopa P. Conceptual understanding of convolutional neural network — a deep learning approach. *Procedia Computer Science*. 2018. Vol. 132. P. 679–688. https://doi.org/10.1016/j.procs.2018.05.069.
8. Gu J., Wang Z., Kuen J., Ma L., Shahroudy A., Shuai B., Liu T., Wang X., Wang G., Cai J., Chen T. Recent advances in convolutional neural networks. *Pattern Recognition*. 2018. Vol. 77. P. 354–377. https://doi.org/10.1016/j.patcog.2017.10.013.
9. Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena.* 2020. Vol. 404. P. 132306–132306. https://doi.org/10.1016/j.physd.2019.132306.
10. Bojer C.S., Meldgaard J.P. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting.* 2021. Vol. 37, Iss. 2. P. 587–603. https://doi.org/10.1016/j.ijforecast.2020.07.007.
11. Gururaj T., Vishrutha Y.M., Uma M., Rajeshwari D., Ramya B.K. Prediction of lung cancer risk using random forest algorithm based on Kaggle data set. *International Journal of Recent Technology and Engineering.* 2020. Vol. 8, Iss. 6. P. 1623–1630. https://doi.org/10.35940/ijrte.F7879.038620.
12. Litjens G., Kooi T., Bejnordi B.E., Setio A.A.A., Ciompi F., Ghafoorian M., van der Laak J.A.W.M., van Ginneken B., Sánchez C.I. A survey on deep learning in medical image analysis. *Medical Image Analysis*. 2017. Vol. 42. P. 60–88. https://doi.org/10.1016/j.media.2017.07.005.
13. Настенко Є., Павлов В., Носовець О., Круглий В., Гончарук М., Карлюк А., Грішко Д., Трофименко О., Бабенко В. Застосування текстурного аналізу у вирішенні задачі класифікації медичних зображень. *Біомедична інженерія і технологія*. 2020. № 4. С. 69–82. https://doi.org/10.20535/2617-8974.2020.4.221876.
14. Cosgun Y., Yildirim A., Yucel M., Karakoc A.E., Koca G., Gonultas A., Gursoy G., Ustun H., Korkmaz M. Evaluation of invasive and noninvasive methods for the diagnosis of helicobacter pylori infection. *Asian Pacific Journal of Cancer Prevention.* 2016. Vol. 17, N 12. P. 5265–5272. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5454669/
15. Norouzi M., Collins M.D., Fleet D.J., Kohli P. CO2 Forest: improved random forest by continuous optimization of oblique splits. 2015. P. 1–8. arXiv preprint arXiv:1506.06155.
16. Chaudhary A., Kolhe S., Kamal R. An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*. 2016. Vol. 3, Iss. 4. P. 215–222. https://doi.org/10.1016/j.inpa.2016.08.002.
17. Elyan E., Gaber M.M. A genetic algorithm approach to optimising random forests applied to class engineered data. *Information Sciences*. 2017. Vol. 384. P. 220–234. https://doi.org/10.1016/j.ins.2016.08.007.
18. Nastenko I., Maksymenko V., Dykan I., Nosovets O., Tarasiuk B., Pavlov V., Babenko V., Kruhlyi V., Soloduschenko V., Dyba M., Umanets V. Liver pathological states identification in diffuse diseases with self-organization models based on ultrasound images texture features. *Proc. 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT)* (23–26 September 2020, Zbarazh, Ukraine). Zbarazh, 2020. Vol. 2. P. 21–25. https://doi.org/10.1109/CSIT49958.2020.9321999.
19. Nastenko I., Maksymenko V., Galkin A., Pavlov V., Nosovets O., Dykan I., Tarasiuk B., Babenko V., Umanets V., Petrunina O., Klymenko D. Liver pathological states identification with self-organization models based on ultrasound images texture features. In: Advances in Intelligent Systems and Computing V. Shakhovska N., Medykovskyy M.O. (Eds.). 2021. Vol. 1293. P. 401–418. https://doi.org/10.1007/978-3-030-63270-0_26.
20. Anastasakis L., Mort N. The development of self-organization techniques in modelling: A review of the group method of data handling (GMDH). Research Report. ACSE Research Report 813. University of Sheffield, Department of Automatic Control and Systems Engineering. 2001. URL: https://gmdhsoftware.com/GMDH_%20Anastasakis_and_Mort_2001.pdf.
21. Furman E., Kye Y., Su J. Computing the Gini index: A note. *Economics Letters*. 2019. Vol. 185. P. 108753–108753. https://doi.org/10.1016/j.econlet.2019.108753.
22. Dong X., Qian M., Jiang R. Packet classification based on the decision tree with information entropy. *The Journal of Supercomputing*. 2020. Vol. 76, Iss. 6. P. 4117–4131. https://doi.org/10.1007/s11227-017-2227-z.

23. Chicco D., Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020. Vol. 21, N 1. 6. https://doi.org/10.1186/s12864-019-6413-7.
24. Breiman L. Bagging predictors. Technical Report 421. Berkeley: University of California, Department of Statistics, 1994.
25. Breiman L. Random forests. *Machine Learning*. 2001. Vol. 45, Iss. 1. P. 5–32. https://doi.org/10.1023/A:1010933404324.
26. Breiman L. Bagging predictors. *Machine Learning*. 1996. Vol. 24, Iss. 2. P. 123–140. https://doi.org/10.1007/BF00058655.
27. Goldberg D.E. Genetic algorithms in search, optimization & machine learning. Boston: Addison-Wesley Longman Publishing Co., Inc., 1989. 432 p.
28. Nosovets O., Babenko V., Davydovych I., Petrunina O., Averianova O., Zyonh L.D. Personalized clinical treatment selection using genetic algorithm and analytic hierarchy process. *Advances in Science, Technology and Engineering Systems Journal.* 2021. Vol. 6, Iss. 4. P. 406–413. https://doi.org/10.25046/aj060446.
29. Saaty T.L. Decision making for leaders: The analytic hierarchy process for decisions in a complex world. Pittsburgh: RWS Publications, 1990. 292 p.
30. Sperandei S. Understanding logistic regression analysis. *Biochemia Medica.* 2014. Vol. 24, N 1. P. 12–18. https://doi.org/10.11613/BM.2014.003.
31. Žižka J., Dařena F., Svoboda A. Adaboost. In: Text Mining with Machine Learning. 2019. P. 201–210. https://doi.org/10.1201/9780429469275-9.
32. Petrunina O., Shevaga D., Babenko V., Pavlov V., Rysin S., Nastenko I. Comparative analysis of classification algorithms in the analysis of medical images from speckle tracking echocardiography video data. *Innovative Biosystems and Bioengineering*. 2021. Vol. 5, N 3. https://doi.org/10.20535/ibb.2021.5.3.234990.
33. Настенко Є., Максименко В., Поташев С., Павлов В., Бабенко В., Рисін С., Матвійчук О., Лазоришинець В. Застосування методу групового урахування аргументів для побудови алгоритмів діагностики ішемічної хвороби серця. *Біомедична інженерія і технологія.* 2021. № 5. С. 1–9. https://doi.org/10.20535/2617-8974.2021.5.227141.

V. Babenko, Ie. Nastenko, V. Pavlov, O. Horodetska, I. Dykan, B. Tarasiuk, V. Lazoryshinets

PATHOLOGY CLASSIFICATION FROM MEDICAL IMAGES BY THE ALGORITHM OF RANDOM FOREST OF OPTIMAL-COMPLEXITY TREES

Abstract. The authors propose an approach to the construction of classifiers in the class of random forest algorithms. A genetic algorithm is used to determine the optimal combination and composition of features' ensembles in the construction of forest trees. The principles of the group method of data handling are used to optimize the trees' structure. Optimization of the tree voting procedure in the forest is implemented by the analytic hierarchy process. Examples of using the proposed algorithm to identify pathologies in medical images and the classification results as compared with other known analogs are presented.

Keywords: pathology classification, medical images, random forest, genetic algorithm, group method of data handling, analytic hierarchy process.

Надійшла до редакції 29.07.2022