Biomedical Engineering and Technology
Issue 15(3), 2024 ISSN (Online) 2707-8434

УДК 004.85:616.9:616.24-07

ЕФЕКТИВНІСТЬ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ЗМІН СТРУКТУРИ ЛЕГЕНЬ У ПОСТКОВІДНИХ ТА ГОСТРИХ СТАДІЯХ COVID-19

Лутченко Вікторія Геннадіївна[1]
viktoriialutchenko@gmail.com
Бабенко Віталій Олегович[1]
vbabenko2191@gmail.com
Настенко Євген Арнольдович[1,2]
nastenko.e@gmail.com
Линник Микола Іванович[3]
nicklinnik1957@gmail.com

[1]Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
м. Київ, Україна,
[2]ДУ «Національний інститут серцево-судинної хірургії
імені М.М. Амосова НАМН України»
м. Київ, Україна,
[3]ДУ «Національний інститут фтизіатрії і пульмонології
імені Ф.Г. Яновського НАМН України»
м. Київ, Україна

Анотація. *Комп'ютерна томографія (КТ) є важливим інструментом для діагностики змін структури легень завдяки своїй високій точності та чутливості у визначенні патологічних змін на тканинах. Деталізація тканин легень є основною причиною ефективності даного методу у виявленні як гострих стадій захворювання COVID-19, так і ускладнень, що виникли на тлі постковідного періоду (починається через три місяці після гострої стадії). Однак, важливу роль відіграють медичні фахівці, які працюють зі знімками КТ. Застосування комп'ютерних алгоритмів машинного навчання може сприяти покращенню медичної практики, від чого виграють як фахівці, отримуючи підтримку у прийнятті діагностичних рішень, так і пацієнти, отримуючи своєчасне та ефективне лікування. В даному дослідженні використовувалась база зрізів КТ легень, що була надана фахівцями ДУ «Національний інститут фтизіатрії та пульмонології імені Ф.Г. Яновського НАМН України» у рамках співпраці з КПІ ім. Ігоря Сікорського. Усього база містила 10 031 зріз КТ легень, які були взяті у 36 знеособлених пацієнтів. З них, 5 213 (52%) зрізів містили ознаки гострої фази COVID-19, а 4 818 (48%) – ознаки постковіду. Перед процесом моделювання було прийнято рішення розподілити пацієнтів в тренувальну (80.6%) і тестову (19.4%) вибірки. Завдяки такому розподіленню співвідношення між зрізами КТ у двох вибірках складало 3:1. Для вилучення інформативних ознак з наданих зрізів застосовувались методи текстурного аналізу, такі як: GLCM, GLRLM, GLDS та LBP. Використовуючи отримані ознаки, була проведена класифікація стану легень наступними алгоритмами машинного навчання: Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost, LightGBM, а також випадковий ліс дерев оптимальної складності (ВЛДОС). За допомогою мір точності, чутливості та специфічності було визначено, що виявляти зміни структури легень, як у постковідних пацієнтів, так і у пацієнтів з гострою стадією COVID-19, здатні моделі XGBoost LightGBM та ВЛДОС з показниками точності на тестовій вибірці 85%, 86% та 89% відповідно. Подібні результати дозволяють надавати необхідну для медичних фахівців підтримку у прийнятті діагностичних рішень, що сприяє кращому виявленню змін структури легень. В майбутніх дослідженнях планується розробка програмного забезпечення з можливостями пояснювального інтелекту, а також подальша інтеграція застосунку в чинну медичну інфраструктуру.*

Ключові слова: *COVID-19, постковід, алгоритми класифікації, комп'ютерна томографія, текстурний аналіз зображень.*

I. ВСТУП

Збудником COVID-19 є одноланцюговий РНК вірус SARS-CoV2, який дістається дихальними шляхами до альвеол, де зв'язується з рецепторами ACE2 (ангіотензин перетворюючого ферменту 2) [1], що міститься в альвеолярних війчастих і келихоподібних клітинах II типу [2]. Сприяє цьому проникненню основний S-глікопротеїн трансмембранної серинової протеази людини 2-го типу (TMPRSS2). Після цього вірус проникає в цитоплазму клітини, вивільняє свою РНК для реплікації за допомогою механізмів клітини-хазяїна, і таким чином розмножується та поширюється на решту клітин організму. У міру розвитку цього процесу в альвеолах починає накопичуватися рідина, що викликає сухий кашель та може ускладнювати дихання [3].

Внаслідок враження великої кількості альвеол спостерігаються такі ускладнення, як гострий респіраторний дистрес-синдром (ARDS), або ж висотний набряк легень (HAPE). У найважчих випадках частою причиною смерті від COVID-19 є синдром системної запальної відповіді (SIRS). Багата на пептиди рідина з легень (альвеол) потрапляє в кров, що призводить до септичного шоку та поліорганної недостатності [3].

Згідно з інформацією, що міститься в одному з досліджень, найпоширенішими ознаками ковіду є: плямисті/пунктатні затемнення типу матового скла (85.7%), одиничні вузлики типу матового скла (22.2%), плямиста консолідація (19.0%), фіброзні смуги (17.5%) та нерегулярні тверді вузли (12.7%) [4]. Щодо ознак постковіду, то результати показали, що у близько третини пацієнтів (35%) спостерігаються фіброзні зміни в легенях, які проявляються як заміщення нормальної тканини на рубцеву. Ці зміни включають тракційну бронхоектазію, паренхімальні смуги та/або гроновидні зміни (honeycombing) [5].

На сьогодні відомо чимало підходів, що застосовуються для аналізу певних змін у структурі легень та визначення COVID-19. Втім, з початку всесвітньої пандемії найкраще себе показала саме комп'ютерна томографія (далі – КТ). Дана технологія дозволяє максимально детально переглянути структуру легень та всі вищезазначені зміни, що були або можуть бути викликані внаслідок COVID-19, з чутливістю 98% [6, 7]. Жоден з інших доступних методів не володіє подібними особливостями та не має настільки високої чутливості.

II. МЕТА РОБОТИ

Удосконалення діагностики та класифікації патологічних змін у легенях, спричинених COVID-19, за допомогою аналізу КТ-знімків із застосуванням технологій обробки зображень і машинного навчання.

III. ОГЛЯД ІСНУЮЧИХ РІШЕНЬ

Оскільки пандемія коронавірусної хвороби була новим викликом для галузі медицини, на сьогодні наявна недостатня кількість статей, які спеціалізуються на розпізнаванні COVID-19 за допомогою текстурного аналізу структури легень по зображеннях комп'ютерної томографії. Проте, певні рішення існують.

Перше дослідження було опубліковано в журналі «Scientific Reports» і ґрунтувалось на використанні рентгенологічних ознак зображень КТ для того, щоб розпізнати де на знімках пацієнти з COVID-19, а де представлені інші види пневмонії. Автори дослідження використовували особливості матриць GLCM і GLRLM для розрізнення COVID-19. Серед запропонованих класифікаторів були: метод опорних векторів (SVM), метод випадкового лісу (Random Forest), метод k-найближчих сусідів (K-nearest neighbors) та екстремальне градієнтне підсилювання (XGBoost). Загалом, авторам вдалось досягти високих показників точності (89.83%), чутливості (94.22%) та специфічності (85.44%) [8].

Інше дослідження, яке було опубліковано в журналі «SN Computer Science», акцентувало на використанні особливостей матриці GLCM для скринінгу COVID-19. Ця робота продемонструвала ефективність використання таких методів класифікації, як Random Forest та SVM для GLCM, що підтвердили результати загальної точності класифікації – 99.94% [9].

IV. МАТЕРІАЛИ І МЕТОДИ

База зрізів КТ легень, що використовувалась у даному дослідженні, була надана в рамках наступних договорів про співпрацю:

1. №Д/0002.01/3400.02/5/2023 від 05 січня 2023 року між КПІ ім. Ігоря Сікорського та ДУ «Національний інститут фтизіатрії і пульмонології імені Ф.Г. Яновського НАМН України».

2. №Д/0002.01/4047.01/34/2023 від 02 лютого 2023 року між КПІ ім. Ігоря Сікорського та ДУ «Національний інститут серцево-судинної хірургії імені М.М. Амосова НАМН України».

Наданий набір даних містив 10 031 зріз КТ легень (рис. 1), які були взяті у 36 знеособлених пацієнтів.

Рисунок 1 – Приклад КТ зображення легень з наданої бази

З них 5 213 (52%) зрізів містили гострі ознаки COVID-19, а 4 818 (48%) – ознаки постковіду. Постковідним вважався стан пацієнтів через 3 та більше місяців після гострої стадії COVID-19.

4.1. Сегментовані маски зображень

Для кожного із зображень було створено сегментовану маску для виділення області інтересу (ОІ) (рис. 2) за допомогою бібліотеки *«lungmask»*, написаної мовою програмування *Python*. Вона використовує глибокі нейронні мережі для сегментації різних областей легень, зокрема правої та лівої легені, їхніх долей, враховуючи при цьому патології. Серед трьох доступних моделей було обрано саме мережу U-net (R231CovidWeb), яка навчена на додаткових даних і може вдало визначати специфічні характеристики та зміни, що викликані COVID-19 [10].

Рисунок 2 – Приклад сегментованої маски з виділеною ОІ

Варто зазначити, що усі зображення та їхні сегментовані маски були збережені у форматі PNG. Перевагою даного формату є те, що він застосовує стиснення без втрат, що дозволяє зберегти якість зображення. Це забезпечує видимість усіх важливих деталей для роботи із медичними зображеннями [10].

4.2. Текстурний аналіз зображень

Наступним етапом було виконання текстурного аналізу за допомогою спеціалізованої бібліотеки *«pyfeats»*, також написаної мовою програмування *Python*. Для роботи із зображеннями було відібрано 4 наступних методи: матриця спільної зустрічальності рівнів сірого (GLCM), статистика різниць рівнів сірого (GLDS), матриця довжин пробігу рівнів сірого (GLRLM) та локальні бінарні шаблони (LPB) [10].

GLCM визначає текстурний зв'язок між парою пікселів, виконуючи операцію на основі статистики другого порядку в зображеннях [11]. Властивості GLCM записуються у вигляді матриці, що має однакову кількість рядків і стовпців (кількість відповідає градації сірого в зображенні). Потім відбувається аналіз того, наскільки часто пари пікселів з певними значеннями зустрічаються один з одним, і з цього вже

робляться висновки по різним аспектам текстури зображення: контраст, кореляція, енергія та однорідність. Усе це сприяє розрізненню здорових тканин і тканин з патологіями [12].

GLRLM також є інструментом для витягування текстурних ознак із зображень. GLRLM призначений для роботи з ознаками вищого порядку, а сірий рівень пробігу описується як ряд пікселів, що розглядається у певному визначеному напрямку та має однакові значення інтенсивності.

Для опису терміну довжини пробігу рівня сірого використовується кількість пікселів у ряді, а пробігом називають певну кількість пікселів у ряді, що задані в одному напрямку та мають однакове значення інтенсивності сірого [13].

Щодо статистики різниць рівнів сірого, то вона бере за основу статистичні властивості різниць між значеннями сірого у сусідніх пікселях, що дозволяє визначати вищі порядкові статистичні ознаки текстури [14].

Даний метод містить кілька основних ознак, а саме: середнє значення різниць, стандартне відхилення різниць, ентропія, енергія й контраст.

Останнім з використаних методів текстурного аналізу медичних зображень є LBP. Це текстурні дескриптори, що застосовуються у реалізації комп'ютерного зору. LBP засновується на описі локальної структури зображення. Крім того, він є інваріантним до змін освітлення – зміни в умовах освітлення не впливають на його ефективність.

LBP діє шляхом порівняння інтенсивності центрального пікселя з пікселями, що його оточують. Під час даного процесу кожному пікселю-сусіду привласнюється бінарне значення залежно від того, чи інтенсивність цього сусіднього пікселя менше або більше ніж інтенсивність центрального. Потім отримані бінарні значення формують бінарне число, яке позначає текстуру даного сусідства [15].

Для того, аби обрахувати бінарні значення для сусідніх пікселів використовується наступна формула:

$$\mathrm{LBP}(x_c, y_c) = \sum_{p=0}^{P-1} s(i_p - i_c) \cdot 2^p$$

де: $s(x) = 1$, якщо $x \geq 0$, і $s(x) = 0$, якщо $x < 0$; i_c – інтенсивність центрального пікселя; i_p – інтенсивність сусіднього пікселя; P – кількість сусідніх пікселів; (x_c, y_c) – координати центрального пікселя.

Винятковою відмінністю саме цього методу є його висока дискримінативність для текстурного аналізу. Проте, серед недоліків слід зазначити чутливість локальних бінарних шаблонів до шуму, певні обмеження щодо захоплення глобальної текстурної інформації, а також відсутність ротаційної інформації [15].

Застосування усіх цих методів дозволяє виконати повний аналіз текстурних характеристик КТ-знімків легень, що є надзвичайно важливим аспектом у діагностиці та класифікації змін, які спричинені захворюванням COVID-19.

Окрім застосування методів текстурного аналізу, також було автоматично встановлено мітки на зображеннях, відповідно до директорій, де ті розміщуються (covid19 або postcovid). Таким чином, зображенням з ознаками COVID-19 було привласнено мітку 0, а з ознаками постковіду – 1.

4.3. Методологія класифікації змін структури легень

За допомогою мови програмування *Python* було реалізовано комплексний алгоритм класифікації, який включає кілька етапів: підготовка даних, створення та налаштування моделей, та оцінка їх продуктивності.

Першим етапом було розподілення пацієнтів в тренувальну (80.6%) і тестову (19.4%) вибірки. Відповідно співвідношення зрізів КТ у двох вибірках становило 3:1. Даний підхід дозволяє уникнути «витоку даних» (англ. data leakage), або іншими словами – інфляції продуктивності моделі, що виникає у випадку використання інформації, яка не може бути доступною під час навчання моделі (приклад подібної ситуації детально описаний в [16]).

Наступним етапом було завантаження у програму попередньо підготовлених даних, що містять набори ознак та міток, у вигляді файлів формату *NumPy*. Для стандартизації даних їх було масштабовано за допомогою StandardScaler. Завдяки цьому вплив всіх

ознак на моделі машинного навчання є рівномірним.

Крім того, оскільки наявний певний дисбаланс між класами, то було використано метод синтетичного надзразкового розширення даних SMOTE (Syntetic Minority Over-sampling Technique). За допомогою нього створюються нові синтетичні зразки, які точно відображають розподіл даних менш представлених класів, таким чином балансуючи їх, що покращує результати моделей.

Безпосередньо для класифікації було застосовано шість алгоритмів: Random Forest, SVM, KNN, XGBoost, LightGBM та ВЛДОС (його більш детально буде описано у наступному підрозділі). Для перших п'яти виконувався підбір оптимальних параметрів, використовуючи GridSearchCV з п'ятикратною стратифікованою крос-валідацією – StratifiedKFold. Такий підхід дає можливість систематично перевіряти комбінації гіперпараметрів, щоб підвищити точність класифікації.

Оцінка моделей проводилася за допомогою таких параметрів, як точність (accuracy) та детальний звіт класифікації. Останній містить такі міри класифікації, як: accuracy, precision, recall та F1-score.

Окрім того, було прийнято рішення додати розрахунок часу, за який один алгоритм виконує класифікацію, щоб побачити наскільки швидко вони працюють у реальному часі та порівняти їхню ефективність.

4.4. Алгоритм випадкового лісу дерев оптимальної складності

Алгоритм випадкового лісу дерев оптимальної складності, або ж ВЛДОС, містить декілька етапів, що спрямовані на покращення результатів класифікації шляхом оптимізації параметрів, структури та агрегації дерев. Ключовою відмінністю ВЛДОС є його використання бутстреп-агрегації (бегінгу). Даний підхід полягає у тому, що відбувається паралельне незалежне навчання кількох різних дерев з подальшою агрегацією моделей задля отримання кінцевого результату. Тобто основною ідеєю є часткова нейтралізація ефекту мінливості моделі лісу, яке викликане варіативністю вхідної вибірки даних, що надає змогу мінімізувати помилку класифікації тестових даних і забезпечує моделі властивість узагальнення результатів.

Щоб побудувати дерева оптимальної складності для початку визначаються пороги розбиття діапазонів значень коефіцієнтом кореляції Метьюза (ККМ). Після цього, аби забезпечити максимальне значення цільової функції, застосовується ітеративний перебір незалежних змінних та їхніх порогів. Це дозволяє отримати оптимальну структуру дерев для більш точного прогнозування.

Саме на етапі формування лісу дерев оптимальної складності і використовується принцип бегінгу (рис. 3), що був згаданий раніше.

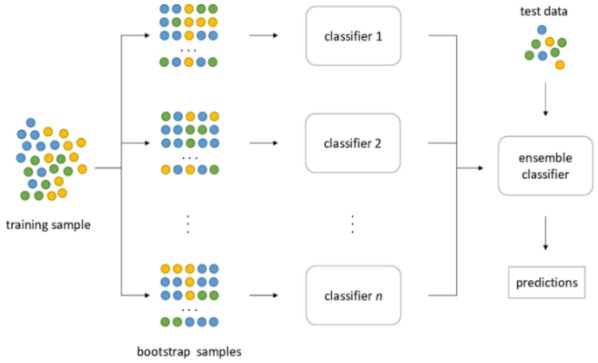

Рисунок 3 – Демонстрація роботи бегінгу [17]

Оптимізація рішень для максимізації точності класифікації відбувається за допомогою генетичного алгоритму.

Аби ефективність колективного механізму прийняття рішень була вищою, використовується зважене голосування. Даний метод реалізується за допомогою методу ієрархій Сааті й надає можливість присвоювати кожному з дерев власний ваговий коефіцієнт. Це сприяє покращенню точності класифікації завдяки тому, що кращі моделі мають вищу пріоритетність під час голосування. Детальну інформацію про розробку можна знайти у статті [17].

V. РЕЗУЛЬТАТИ

Результати проведеної класифікації за допомогою методів Random Forest, Support Vector Machine, K-Nearest Neighbors, XGBoost, LightGBM та ВЛДОС можна переглянути у табл. 1.

Таблиця 1
Результати класифікації методами машинного навчання

Метод	Тестова вибірка (24.5%)		
	accuracy	**recall**	**specificity**
Random Forest	0.77	0.78	0.76
SVM	0.74	0.71	0.77
KNN	0.74	0.68	0.81
XGBoost	0.85	0.88	0.82
LightGBM	0.86	0.88	0.84
ВЛДОС	0.89	0.99	0.75

Аналіз отриманих результатів на тестовій вибірці КТ-знімків проводився спираючись на показники точності (accuracy), чутливості (recall) та специфічності (specificity). Найкращі результати продемонстрував метод ВЛДОС, тоді як найгірші – K-Nearest Neighbors (KNN). У табл. 2 наведено скільки часу потребує кожен із наведених алгоритмів для підбору оптимальних гіперпараметрів та проведення класифікації.

Таблиця 2
Результати часу навчання моделей

Назва методу	Час	Точність
Random Forest	7 хв 4 с	0.77
SVM	9 хв 30 с	0.74
KNN	0 хв 48 с	0.74
XGBoost	0 хв 49 с	0.85
LightGBM	4 хв 52 с	0.86
ВЛДОС	1 год 48 хв 8 с	0.89

Проаналізувавши дану інформацію, найкращий баланс між точністю та часом виконання продемонстрували моделі XGBoost та LightGBM. Модель ВЛДОС хоч і має високі значення точності, проте може поступатися попереднім двом результатам у задачах, де час на виконання може бути обмеженим.

Модель KNN хоч і має найменший час виконання, проте не є гарним вибором через параметри оцінки даної моделі.

VI. ВИСНОВКИ

Реалізовані алгоритми класифікації цілком відповідають поставленій задачі, а також є ефективними для класифікації змін структури легень у постковідних та гострих стадіях COVID-19.

Зокрема, методи XGBoost, LightGBM та ВЛДОС продемонстрували найвищі показники точності класифікації на тестовій вибірці, а саме 85% 86% та 89% відповідно. Отримані значення точності, чутливості, специфічності й інших параметрів оцінки свідчать про оптимальність використання даних методів у задачах класифікації КТ зображень. Невеликий час виконання у моделей XGBoost та LightGBM робить їх зручними та ефективними у використанні на практиці, коли наявні обмеження щодо швидкості обчислення.

Моделі Random Forest та Support Vector Machine демонструють дещо гірші результати. Хоча чутливість (78.35%) та специфічність (75.79%), отримані за допомогою Random Forest на тестових даних є збалансованими, що свідчить про здатність моделі правильно ідентифікувати як позитивні, так і негативні випадки, проте дана модель демонструє незначне перенавчання, що потребує коригування надалі. Удосконаливши даний недолік та враховуючи час навчання можна стверджувати, що Random Forest буде ефективним методом для застосувань, де потрібен баланс між точністю та часом обчислення.

Щодо SVM, то дана модель не виявилась найефективнішою, проте є стабільною, на що вказують її результати класифікації. Невелика різниця між точністю на тренувальних і тестових даних свідчить про відсутність перенавчання, а результати крос-валідації підтверджують здатність моделі до узагальнення.

KNN показав найгірші результати класифікації з точністю 74% через перенавчання. Чутливість даного методу на тестових даних становить (68.33%), є найнижчою серед інших розглянутих алгоритмів та вказує на недостатню здатність моделі ідентифікувати позитивні випадки. Водночас специфічність на рівні 81.15% свідчить про те, що модель краще справляється з визначенням негативних випадків.

Серед можливих причин погіршення результатів класифікації є доволі висока кореляція між ознаками COVID-19 та постковіду. Багато ознак гострої стадії ковіду,

такі як граунд-глас затемнення (GGO), консолідації тощо мають здатність зберігатися або ж змінюватися на фіброзні утворення та бронхоектази постковідної стадії, тим самим еволюціонуючи в ознаки хронічного характеру або ж персистуючі.

Також наявний певний дисбаланс між класами тестової вибірки, що може негативно впливати на класифікацію міноритарного класу.

Усе це створює певні виклики для проведення машинного навчання на таких даних. Мультиколінеарність, що спричинена високою кореляцією класів, робить інтерпретацію моделей складнішою та знижує їх стабільність до виконання правильних прогнозів. Це призводить до перенавчання, що спостерігається у результатах деяких моделей, оскільки модель чудово підлаштовується під тренувальну вибірку КТ-знімків, у тому числі й ознаки, що корелюють, проте вона не може успішно розпізнавати та узагальнювати нові дані. Окрім того, ознаки, які персистують, також можуть бути помилково інтерпретовані моделлю, що призведе до зниження точності класифікації.

Надалі планується врахування даних особливостей, а також мінімізація їхнього впливу в майбутніх дослідженнях на дану тематику. Крім того, планується додати методику пояснювального інтелекту, що дозволить побудованим моделям не лише прогнозувати стан легень, а й надавати низку причин.

Результати даного дослідження можуть виявитися корисними для створення систем підтримки прийняття рішень для лікарів, оскільки це дозволить покращити та пришвидшити діагностику змін структури легень у пацієнтів хворих на коронавірусну хворобу та у постковідних станах. Дані для реалізації даного дослідження було використано в рамках наступних договорів про співпрацю:

1. №Д/0002.01/3400.02/5/2023 від 05 січня 2023 року між КПІ ім. Ігоря Сікорського та ДУ «Національний інститут фтизіатрії і пульмонології ім. Ф.Г. Яновського НАМН України».

2. №Д/0002.01/4047.01/34/2023 від 02 лютого 2023 року між КПІ ім. Ігоря Сікорського та ДУ «Національний інститут серцево-судинної хірургії імені М.М. Амосова НАМН України».

Фінансування. Дане дослідження не отримувало зовнішнього фінансування.

Конфлікт інтересів. Автори заявляють про відсутність конфлікту інтересів.

Згода на публікацію. Усі пацієнти, що мають відношення до рукопису дали згоду на публікацію даної роботи.

ORCID ID та внесок авторів.

1. Viktoriia Lutchenko (A, B, C, D) - 0009-0004-9684-7659

2. Vitalii Babenko (B, C) – 0000-0002-8433-3878

3.Ievgen Nastenko (E, G) – 0000-0002-1076-9337

4. Mykola Linnik (F) – 0000-0002-0011-7482

A – Огляд та аналіз пов’язаних робіт.

B – Реалізація класифікаційних алгоритмів, що наведені в роботі.

C – Реалізація алгоритму через написання програмного коду.

D – Написання статті.

E – Критичний огляд статті.

F – Критичний огляд медичних аспектів статті.

G – Остаточне схвалення статті.

VI. ПЕРЕЛІК ПОСИЛАНЬ

[1] Passaro, A., Addeo, A., Von Garnier, C., Blackhall, F., Planchard, D., Felip, E., et al. (2020). ESMO management and treatment adapted recommendations in the COVID-19 era: Lung cancer. ESMO Open, 5, e000820. https://doi.org/10.1136/esmoopen-2020-000820

[2] Bonomi, M., Maltese, M., Brighenti, M., Muri, M., & Passalacqua, R. (2021). Tocilizumab for COVID-19 pneumonia in a patient with non-small-cell lung cancer treated with chemoimmunotherapy. Clinical Lung Cancer, 22(1), e67-e69. https://doi.org/10.1016/j.cllc.2020.08.002

[3] Desjardins, J. (2020, April 2). Visualizing what COVID-19 does to your body. Visual Capitalist. Retrieved from https://www.visualcapitalist.com/visualizing-what-covid-19-does-to-your-body/

[4] Pan, Y., Guan, H., Zhou, S., Wang, Y., Li, Q., Zhu, T., Hu, Q., & Xia, L. (2020). Initial CT findings and temporal changes in patients with the novel coronavirus pneumonia 73 (2019-nCoV): a study of 63 patients in Wuhan, China. European Radiology. Retrieved from https://link.springer.com/article/10.1007/s00330-020-06731-x

[5] Han, X., Fan, Y., Alwalid, O., Li, N., Jia, X., Yuan, M., Li, Y., Cao, Y., Gu, J., Wu, H., & Shi, H. (2021). Six-month follow-up chest CT findings after severe COVID-19 pneumonia. Radiology, 299(1), E177-E186. Retrieved from https://pubmed.ncbi.nlm.nih.gov/33497317/

[6] Wong, H. Y. F., Lam, H. Y. S., Fong, A. H. T., Leung, S. T., Chin, T. W. Y., Lo, C. S. Y., et al. (2019). Frequency and distribution of chest radiographic findings in COVID-19 positive patients. Radiology, 27, 201160. https://doi.org/10.1148/radiol.2020201160

[7] Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., et al. (2020). Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases. Radiology, 26, 200642. https://doi.org/10.1148/radiol.2020200642

[8] Ning, Y., Liu, F., Li, C., Xiao, W., Xie, S., Yuan, S., Zuo, W., Ma, X., & Jiang, G. (2021). Diagnostic classification of coronavirus disease 2019 (COVID-19) and other pneumonias using radiomics features in CT chest images. Scientific Reports, 11, 17885. https://doi.org/10.1038/s41598-021-97497-9

[9] Godbin, A. B., & Jasmine, S. G. (2023). Screening of COVID-19 based on GLCM features from CT images using machine learning classifiers. SN Computer Science, 4, 133. https://doi.org/10.1007/s42979-022-01583-2

[10] Lutchenko, V. H. Program application for the classification of post-COVID signs of lung structure changes compared to the acute phase of coronavirus: Bachelor's thesis, Computer Science, No. 122. Kyiv, 2024, 77 p. https://ela.kpi.ua/handle/123456789/67667

[11] Albregtsen, F., Nielsen, B., &Danielsen, H. E. (2000). Adaptive gray level run length features from class distance matrices. In Pattern Recognition, 2000. Proceedings. 15th International Conference on (Vol. 3, pp. 738-741). IEEE

[12] Nabizadeh, N., &Kubat, M. (2015). Brain tumors detection and segmentation in MR images: Gabor wavelet vs. statistical features. Computers & Electrical Engineering, 45, 286-301.

[13] Wan Kairuddin, W. N. H., & Wan Mahmud, W. M. H. (2017). Texture Feature Analysis for Different Resolution Level of Kidney Ultrasound Images. Department of Electronic Engineering, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia.

[14] Giakoumoglou, N. G. (Year). PyFeats: Open-source software for image feature extraction. PyPI. Retrieved from https://pypi.org/project/pyfeats/

[15] Ihalapathirana, A. (Year). Understanding the Local Binary Pattern (LBP): A Powerful Method for Texture Analysis in Computer Vision. Medium. Retrieved from https://aihalapathirana.medium.com/understanding-the-local-binary-pattern-lbp-a-powerful-method-for-texture-analysis-in-computer-4fb55b3ed8b8

[16] Rosenblatt, M., Tejavibulya, L., Jiang, R. et al. Data leakage inflates prediction performance in connectome-based machine learning models. Nat Commun 15, 1829 (2024). https://doi.org/10.1038/s41467-024-46150-w

[17] Babenko, V. O., Nastenko, I. A., Pavlov, V. A., Nosovets, O. K., Dykan, I. M., Tarasyuk, B. A., & Lazoryshinets, V. V. (n.d.). Pathologies Classification from Medical Images by Random Forest of Optimal Complexity Trees algorithm. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institut

Biomedical Engineering and Technology
Issue 15(3), 2024 ISSN (Online) 2707-8434

UDC 004.85:616.9:616.24-07

EFFICIENCY OF MACHINE LEARNING ALGORITHMS FOR CLASSIFYING LUNG STRUCTURAL CHANGES IN POST-COVID-19 AND ACUTE STAGES OF COVID-19

Viktoriia Lutchenko[1]
viktoriialutchenko@gmail.com
Vitalii Babenko[1]
vbabenko2191@gmail.com
Ievgen Nastenko[1,2]
nastenko.e@gmail.com
Mykola Linnik[3]
nicklinnik1957@gmail.com

[1]National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine,
[2]Amosov National Institute of Cardiovascular Surgery
Kyiv, Ukraine,
[3]Yanovsky National Institute of Phthisiology and Pulmonology
Kyiv, Ukraine

Abstract *– Computed tomography (CT) is an important tool for diagnosing changes in lung structure due to its high accuracy and sensitivity in detecting pathological changes in tissues. The detail of lung tissue is the main reason for the effectiveness of this method in detecting both the acute stages of COVID-19 and complications that occurred during the post-COVID period (beginning three months after the acute stage). However, medical specialists working with CT scans play an important role. The use of computer machine learning algorithms can help improve medical practice, which benefits both specialists by supporting them in making diagnostic decisions and patients by providing timely and effective treatment. This study used a database of CT lung slices provided by the specialists of the F.G. Yanovsky National Institute of Phthisiology and Pulmonology of the National Academy of Medical Sciences of Ukraine in cooperation with Igor Sikorsky Kyiv Polytechnic Institute. In total, the database contained 10,031 CT lung scans taken from 36 anonymized patients. Of these, 5,213 (52%) contained signs of the acute phase of COVID-19, and 4,818 (48%) contained signs of post-COVID. Before the modeling process, it was decided to divide the patients into training (80.6%) and test (19.4%) samples. Due to this distribution, the ratio between CT slices in the two samples was 3:1. Texture analysis methods were used to extract informative features from the provided slices, such as GLCM, GLRLM, GLDS, and LBP. Using the obtained features, the following machine learning algorithms were used to classify the lung condition: Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost, LightGBM, and Random Forest of Optimal Complexity (RFOC). Using the accuracy, sensitivity, and specificity measures, it was determined that XGBoost LightGBM and ROCT models can detect changes in lung structure in both post-COVID and acute COVID-19 patients with accuracy rates of 85%, 86%, and 89%, respectively. Such results allow us to provide the necessary support for medical professionals in making diagnostic decisions, which contributes to better detection of changes in lung structure. In future research, it is planned to develop software with explanatory intelligence capabilities, as well as further integrate the application into the existing medical infrastructure.*

Keywords *– COVID-19, post-COVID, classification algorithms, computer tomography, texture image analysis.*