Biomedical Engineering and Technology ISSN 2617-8974
Issue 8, 2022 ISSN (online) 2707-8434

УДК 004.81 + 616-006

ДІАГНОСТИЧНІ АЛГОРИТМИ ВИЗНАЧЕННЯ ГЕНЕТИЧНИХ МУТАЦІЙ РАКУ ЗА ДОПОМОГОЮ АНАЛІЗУ МЕДИЧНИХ ТЕКСТІВ

Левчик Лілія Олександрівна
levchyk.liliia@lll.kpi.ua
Бабенко Віталій Олегович
vbabenko2191@gmail.com
Бовсуновська Катерина Сергіївна
bmk-bks-fbmi@lll.kpi.ua
Павлов Володимир Анатолійович
pavlov.volodymyr@lll.kpi.ua
Настенко Євген Арнольдович
nastenko.e@gmail.com
Кафедра біомедичної кібернетики
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
м. Київ, Україна

Анотація *- Сучасний стан аналітичних інструментів діагностики, до яких відноситься і генетичне тестування, дозволяють розраховувати, що процес діагностування онкологічних захворювань може бути автоматизованим. Однак, об'єм ручної роботи, необхідної для діагностики ракових пухлин, залишається значною перешкодою для прогресу в даній області. Процес секвенування пухлини здатний виявити тисячі генетичних мутацій, але задача полягає в тому, щоб відрізнити ті мутації, які сприяють зростанню рака (драйвери), від нейтральних мутацій (пасажирів). Ця задача потребує від клінічного патолога ручного аналізу та класифікації кожної мутації на основі інформації, отриманої з клінічної літератури. Використання комп'ютеризованих методів аналізу медичних текстів здатне автоматизувати даний етап діагностики ракових пухлин. Мета даної роботи полягала в оцінці ефективності використання методів обробки природної мови у поєднанні з машинним навчанням для автоматизованого визначення типів генетичних мутацій раку з медичних текстових даних. Наявна для використання база медичних текстових даних, що містить 3321 спостереження, і анотована по 9 типам генетичних мутацій раку провідними дослідниками та онкологами центру Меморіалу Слоуна Кеттерінга (Нью-Йорк, США). Дані були надані в рамках конкурсу по машинному навчанню спільнотою фахівців з Data Science - Kaggle. Для розв'язання багатокласової задачі класифікації використані моделі машинного навчання: мультиноміальний наївний Байєс, мультиноміальна логістична регресія, випадковий ліс, метод групового урахування аргументів, багатошаровий перцептрон, та рекурентна нейронна мережа з довгою короткостроковою пам'яттю. Модель багатошарового перцептрона виявилась найбільш ефективною для визначення типу генетичної мутації, продемонстрував точність передбачення 65.1% на тестовій вибірці, що склала 25% від загального набору даних. Друга по точності модель (випадковий ліс) досягла точність у 64.9%. Одержані результати перевершили результати учасників конкурсу Kaggle, де найвища точність класифікації (64.7%) була досягнута за допомогою лінійної моделі, заснованій на методі опорних векторів.Поєднання методів обробки природної мови та машинного навчання показує великий потенціал для застосування в медичній галузі, зокрема, у визначенні типів генетичних мутацій раку на основі текстових даних. Це надає можливість для автоматизації дій медичного персоналу в процесі діагностики. Для досягнення більш ефективних результатів планується проведення подальших досліджень.*

Ключові слова*: генетичні мутації раку, медичні текстові дані, обробка природної мови, векторизація тексту, машинне навчання, глибоке навчання.*

ISSN 2617-8974
ISSN (online) 2707-8434

I. ВСТУП

Ще донедавна більшість пацієнтів з раком певного типу і стадії отримували однакове лікування [1]. З часом виникли можливості застосування персоналізованих стратегій, що стає більш ефективним для пацієнтів, ніж універсальний підхід. Персоналізована медицина дозволяє лікарям зрозуміти генетичний профіль пацієнта і те, як буде прогресувати його пухлина. Для впровадження персоналізованої медицини проводять генетичні дослідження клітин ракової пухлини. Дослідження генетики показали, що існують варіації в генетичному коді клітин ракової пухлини [2]. Хоча для більшості видів раку зазвичай рекомендується стандартний набір методів лікування, таких як хіміотерапія або операція по видаленню пухлини, проте врахування всіх обставин та характеристик пацієнта

стає основою для визначення індивідуального плану лікування [3]. Дослідники активно шукають методи лікування і профілактичні міри проти раку, які можуть бути адаптовані до генетичного профілю людини [4]. Результати аналізуються для підбору лікування відповідно до конкретних потреб кожного пацієнта. Аналізуючи генетичні профілі пацієнта та використовуючи персоналізовану терапію раку лікарі визначають найбільш ефективну стратегію лікування, знижують ймовірність рецидиву. Персоналізована медицина – це підхід до лікування раку, який обіцяє великі перспективи [5]. Стратегії машинного навчання [6] можуть бути корисними при переході до персоналізованої медицини, оскільки існує велика кількість наукової інформації, яку клінічним спеціалістам може бути важко інтерпретувати, а аналіз даної інформації забирати багато часу. Серед такої інформації - опис результатів різних видів медичних досліджень, що стають основою для точної діагностики захворювання. Даний опис результатів клініцисти часто виконують у вигляді текстових анотацій. Для аналізу таких анотацій природньо застосувати методи обробки природної мови [7] (NLP – від natural language processing). Ці методи можуть бути використані для автоматичного аналізу текстових даних, що спростить завдання формування діагностичних висновків для фахівців.

II. АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ І ПУБЛІКАЦІЙ

Неструктурований текст, що включає медичні записи, відгуки пацієнтів та коментарі в соціальних мережах, є цінним джерелом даних для досліджень. NLP – це набір методів, які використовуються для перетворення письмового тексту в інтерпретовані набори даних, які можна проаналізувати моделями машинного навчання [8]. Застосування NLP до медичних записів може допомогти у прогнозуванні наслідків хвороб, покращенні сортування лікарень, створенні діагностичних моделей для раннього виявлення хронічних захворювань тощо [9]. Існує декілька прикладів використання NLP для розв’язання практичних медичних проблем.

Практичне застосування NLP для побудови моделі прогнозування семантики відгуку на лікарський засіб («добре» або «погано») представлене в роботі [8]. Автори використовували для цього три алгоритми машинного навчання: регуляризовану логістичну регресію, метод опорних векторів та штучну нейронну мережу.

ISSN 2617-8974
ISSN (online) 2707-8434

Найкращий результат показала модель, заснована на методі опорних векторів. На тестовій вибірці (що складала 25% від загальної) дана модель продемонструвала точність класифікації у 72%.

Subramanian et al. [10] представили систему для автоматизації вилучення доказів ефективності неракових незапатентованих препаратів на основі анотацій з PubMed. Отримані ними дані свідчать про те, що такі препарати можуть бути перспективними у лікуванні раку, а їх повторне використання здатне значно покращити результати лікування онкологічних хворих та знизити затрати в галузях медицини та охорони здоров'я. Основний вклад авторів полягає у реалізації NLP-конвеєру для отримання доказів, що включає запит, фільтрацію, вилучення сутностей типу раку, класифікацію терапевтичних асоціацій і класифікацію типів дослідження.

В роботі [11] використовувались дані Національного опитування амбулаторної медичної допомоги в лікарнях, для побудови декількох прогностичних моделей з наслідками госпіталізації або переведення в лікарню у порівнянні з випискою додому. Використовуючи NLP, автори сформували 48 головних компонент, які були використані для побудови моделей логістичної регресії та багатошарової нейронної мережі. Моделі оцінювались за допомогою показника AUC (area under curve - площа під ROC-кривою) [11]. Модель логістичної регресії досягла значення AUC 0.846 на тестовій вибірці, а нейронна мережа – 0.844.

Kolanu et al. [12] використовували NLP для проведення булевого пошуку в радіологічних звітах на предмет переломів і пов'язаних з ними термінів. Автори виявили, що модель можна навчити стилю звітів, специфічного для конкретного місця, і застосувати правила для уточнення ідентифікації (наприклад, вік більше ніж 50 років, наявність кісток тощо). В результаті, запропонована модель виявила в п'ять разів більше потенційно значущих переломів у порівнянні з ручними висновками, при цьому 97.1% знайдених моделлю переломів були підтверджені. Чутливість моделі склала 69.6%, а специфічність – 95%.

Метою роботи [13] було виявлення пацієнтів, які потребують паліативної допомоги з рідкими, але тяжкими захворюваннями за допомогою комбінації адміністративних даних і NLP. Для цього автори провели ретроспективне когортне дослідження, доповнене електронними медичними картками з мережі лікарень, які нараховували в сумі понад 2500 лікарняних ліжок. В результаті методи NLP досягли 100% специфічності, у порівнянні з 25% специфічності при використанні у якості ознак лише адміністративних кодів.

В рамках конкурсу на ресурсі Kaggle [14] автори роботи [1] розробили модель для прогнозування типів генетичних мутацій раку на основі даних текстового опису. При моделюванні використовувався метод логістичної регресії, точність класифікації на тестовій вибірці якої склала 64%. Автори припускають, що використання нейронних мереж може покращити результати роботи. Цю гіпотезу вирішено перевірити у даному дослідженні, оскільки при вдалій реалізації подібну технологію можна впровадити у діагностичну систему підтримки прийняття рішень.

III. МЕТА ДОСЛІДЖЕННЯ І ПОСТАНОВКА ЗАДАЧІ

Хоча генетичне тестування і відкриває великі перспективи для розробки більш точних та ефективних методів лікування

ISSN 2617-8974
ISSN (online) 2707-8434

раку, процес не є повність автоматизованим через значний обсяг ручної роботи, необхідної лікарям для повного розуміння геноміки пухлини. Дослідники з онкологічного центру меморіалу Слоуна Кеттерінга склали базу знань по прецизійній онкології з анотаціями експертів, яка була використана для конкурсу машинного навчання на сайті Kaggle у 2017 році [14]. База даних включає декілька тисяч анотацій на основі клінічної літератури про те, які гени підходять для клінічного застосування, а які ні. Метою конкурсу було створення класифікаційних моделей, здатних аналізувати анотації медичних статей і на основі їхнього змісту точно визначати мутаційний ефект генів (9 класів), що обговорюються в них.

У даному дослідженні було використано подібну базу даних медичних текстів для оцінки потенціалу NLP у визначенні типу генетичної мутації раку. Дослідження було побудоване відповідно до переліку наступних задач:

1. Проведення дослідницького аналізу обраної бази даних для створення попередніх гіпотез.
2. Застосування підходів NLP для попередньої обробки текстових даних і конструювання ознак.
3. Побудова прогностичних моделей машинного навчання для визначення типу генетичної мутації.

IV. ОПИС КЛІНІЧНОГО МАТЕРІАЛУ

Організатори конкурсу [14] надали два окремих набори даних – навчальний (3321 спостереження) і тестовий (5668 спостережень). Частина тестових даних була згенеровані машиною для запобігання ручного маркування, що пояснює той факт, що тестова вибірка на 70% більше ніж навчальна. Однак, оскільки тестові дані не містять заздалегідь відомих класів, для даного дослідження використовувались лише навчальні дані (рис. 1). З приведеного нижче рисунку видно, що обрана для дослідження база текстових даних містить 4 основні атрибути:

• “Text” – клінічний текстовий опис генетичної мутації раку, написаний клінічними онкологами англійською мовою, яка є найбільш поширеною у світі, і тому підходить для використання методів NLP [15].

	ID	TEXT	Gene	Variation	Class
0	0	Cyclin-dependent kinases (CDKs) regulate a var...	FAM58A	Truncating Mutations	1
1	1	Abstract Background Non-small cell lung canc...	CBL	W802*	2
2	2	Abstract Background Non-small cell lung canc...	CBL	Q249E	2
3	3	Recent evidence has demonstrated that acquired...	CBL	N454D	3
4	4	Oncogenic mutations in the monomeric Casitas B...	CBL	L399V	4
5	5	Oncogenic mutations in the monomeric Casitas B...	CBL	V391I	4
6	6	Oncogenic mutations in the monomeric Casitas B...	CBL	V430M	5

Рисунок 1 – Фрагмент бази текстових даних для дослідження

ISSN 2617-8974
ISSN (online) 2707-8434

• “Gene” – ген, в якому знаходиться конкретна генетична мутація.

• “Variation” – варіація амінокислоти для цієї мутації.

• “Class” – мітка класу, до якого було віднесено мутацію.

База даних містить 264 унікальних генів з 2996 різними варіаціями. Описова статистика даних наведена в табл. 1; автори бази даних не надали інформацію про конкретні назви класів (лише номери).

Таблиця 1. **Описова статистика бази текстових даних**

Ген		**Варіація**		**Клас**	
Назва	**Кількість**	**Назва**	**Кількість**	**Мітка**	**Кількість**
BRCA1	264 (7.95%)	Truncating Mutations	93 (2.8%)	1	568 (17.1%)
TP53	163 (4.91%)	Deletion	74 (2.23%)	2	452 (13.61%)
EGFR	141 (4.25%)	Amplification	71 (2.14%)	3	89 (2.68%)
PTEN	126 (3.79%)	Fusions	34 (1.02%)	4	686 (20.66%)
BRCA2	125 (3.76%)	Overexpression	6 (0.18%)	5	242 (7.29%)
KIT	99 (2.98%)	G12V	4 (0.12%)	6	275 (8.28%)
BRAF	93 (2.8%)	E17K	3 (0.09%)	7	953 (28.7%)
ALK	69 (2.08%)	Q61H	3 (0.09%)	8	19 (0.57%)
Залишок	2241 (67.48%)	*Залишок*	3033 (91.33%)	9	37 (1.11%)

Примітка до табл. 1:

• “Truncating Mutations” – перекладається як усікаючі мутації.

• “Deletion” – перекладається як видалення.

• “Amplification” – перекладається як ампліфікація (посилення, збільшення).

• “Fusions” – перекладається як злиття.

• “Overexpression” – перекладається як надекспресія.

Дані про гени та варіації амінокислот є інформаційними для побудови прогностичної моделі машинного навчання. Проте, ця інформація також присутня в текстовому описі генетичної мутації раку. Таким чином, атрибути “Gene” і “Variation” можуть бути видалені з бази даних без втрати інформації. В остаточному варіанті бази даних вхідний атрибут “Text” і вихідний атрибут “Class” будуть збережені.

Табл. 1 показує дисбаланс у кількості класів в базі даних. Клас «7» є найбільш розповсюдженим (майже 30%), в той час як клас «8» – найменш розповсюдженим (менше ніж 1%). Це може створити проблеми у подальшому дослідженні через те, що математична модель при навчанні опирається на ті класи, які більш часто зустрічаються. Тому застосуємо підходи для послаблення або усунення впливу дисбалансу класів. Наприклад, стратегія стратифікації [16] може бути використана для рівномірного розбиття вибірки даних на тренувальну та екзаменаційну, а метод семплування «SMOTE» [17] – для генерації штучних даних, що дозволить збільшити кількість об’єктів класів, які рідко зустрічаються.

V. КОНСТРУЮВАННЯ ОЗНАК

Текстові дані є складними для аналізу, оскільки у необробленому вигляді не містять корисної інформації для прогнозування. В NLP існують інструменти, які дозволяють отримати інформативні ознаки з тексту [7]. Проте, перш ніж

ISSN 2617-8974
ISSN (online) 2707-8434

переходити до конструювання ознак, необхідно провести попередню обробку текстових даних, яка складається з наступних кроків:

1. Приведення тексту до нижнього регістру для усунення залежності від букв, що знаходяться у верхньому регістрі.

2. Видалення «стоп-слів», які є загальними словами в будь-яких текстових даних, і також можуть завадити аналізу. Приклади стоп-слів в англійській мові включають: “a”, “the”, “is”, “are”, і т.д.

3. Лематизація – процес перетворення слова в його кореневу форму, наприклад, “learning” після процесу лематизації перетворюється в “learn”. Це дозволяє знизити кількість унікальних слів в тексті.

Оскільки машини не можуть розуміти символи та текст, необхідно перетворення їх у числові дані за допомогою векторизації тексту. Даний метод підраховує кількість усіх можливих слів для кожного об’єкту текстових даних. Після цього обраховується метрика, яку називають «частота терміну – зворотна частота документа» (term frequency–inverse document frequency, або коротко TF-IDF) [18], де під документом розуміється текст. TF-IDF – це числова статистика, яка відображає важливість слова в тексті. Розрахунок виконується у 3 етапи:

1. Обчислення частоти слова (term frequency, або коротко TF) за формулою:

$$tf(t,d) = \frac{n_t}{\sum_k n_k} \qquad (1)$$

де: n_t – число входжень слова t в текст, а в знаменнику – загальна кількість слів в тексті.

TF (1) дозволяє оцінити важливість слова ti в межах окремого тексту.

2. Обчислення зворотної частоти документу (inverse document frequency, або коротко IDF). Це інверсія частоти, з якою деяке слово зустрічається в масиві текстів. Параметр розраховується за формулою:

$$idf(t,D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|} \qquad (2)$$

де: |D| – число текстів в масиві, в знаменнику стоїть число текстів із масиву D, в яких зустрічається слово t (коли nt не дорівнює нулю).

Врахування IDF зменшує вагу широко застосованих слів. Для кожного унікального слова з масиву текстів обраховується лише одне значення IDF.

3. Значення TF-IDF обраховується як:

$$tf - idf(t,d,D) = tf(t,d) \times idf(t,D) \qquad (3)$$

В результаті даного розрахунку, слова, що мають високу частоту в межах конкретного тексту і низьку частоту вживань в інших текстах, отримають більшу вагу. Сформовані параметри можна використовувати як набір інформативних ознак для побудови прогностичних моделей машинного навчання.

Враховуючи, що кількість унікальних слів з усіх текстових даних обраної бази [14] може сягати сотні тисяч, необхідно обрати найбільш інформативні з них для моделювання. Для відбору ознак був застосований кореляційний критерій [19], який оцінює підмножину незалежних ознак на основі гіпотези про те, що «хороші підмножини містять ті ознаки, які сильно корелюють з залежною змінною, але некорельовані одне з одним». Ця гіпотеза направлена на розв’язання проблеми мультиколінеарності [20].

Застосування кореляційного критерію передбачає знаходження підмножини k

ISSN 2617-8974
ISSN (online) 2707-8434

незалежних ознак, яке дає максимум значення, що задається формулою:

$$S_k = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (4)$$

де: k – кількість ознак в підмножині, S_k – кореляційний критерій оцінки підмножини ознак, $\overline{r_{cf}}$ – середнє значення модулів кореляцій усіх незалежних ознак з залежною змінною, $\overline{r_{ff}}$ – середнє значення модулів кореляцій усіх ознак між собою.

VI. РЕЗУЛЬТАТИ

На основі відібраних даних були побудовані прогностичні моделі на основі аналітичних моделей класифікації та моделі нейронних мереж. При використанні класичних алгоритмів, для визначення оптимальної підмножини (k = 100) інформативних ознак було застосовано критерій (4). При розрахунку кореляцій використовувалась метрика Спірмена [21], оскільки вона інваріантна до типу та розподілу даних.

Застосовані класичні алгоритми:

1. Мультиноміальний наївний Байєс [22] – є представником сімейства ймовірнісних алгоритмів, заснованих на теоремі Байєса [23], з «наївним» припущенням про умовну незалежність між вхідними ознаками. Даний алгоритм є одним з найбільш застосовуваних для побудови прогностичних моделей на основі текстових даних [24].

2. Мультиноміальна логістична регресія [25] – метод класифікації, який розширює принципи простої логістичної регресії у простір багатокласових задач. Як альтернатива «наївному» Байєсу, цей метод не припускає статистичної незалежності вхідних ознак.

3. Випадковий ліс [26] – широко відомий алгоритм навчання ансамблю моделей. Застосування даного алгоритму аргументовано тим, що він демонструє високу продуктивність в подібних задачах [27].

4. Український вчений Олексій Івахненко запропонував метод групового урахування аргументів (МГУА) [28] як індуктивний підхід до моделювання. Попри свою специфічність, даний алгоритм, часто дає сильні прогностичні моделі [29-30].

Кожна модель було навчено на тренувальній вибірці (що складала 75% від загальної), а потім оцінено на екзаменаційній вибірці (25%). Дані містять 9 різних класів, тому ефективність прогнозування моделей оцінювалась за двома наступними метриками:

- точність класифікації – відсоток правильно класифікованих об’єктів;
- зважена F міра (F-score) [31].

Результати моделювання наведені в табл. 2.

Таблиця 2. **Результати класифікації класичними методами машинного навчання**

Алгоритм класифікації	Тренувальна вибірка (75%)		Екзаменаційна вибірка (25%)	
	Точність	F-score	Точність	F-score
Наївний Байєс	41.6%	0.432	41.9%	0.436
Логістична регресія	58.5%	0.592	52.7%	0.534
Випадковий ліс	89.2%	0.896	64.9%	0.647
МГУА	41.3%	0.434	41%	0.434

ISSN 2617-8974
ISSN (online) 2707-8434

Для того, щоб запобігти випадків перенавчання моделей, використовувалась крос-валідація [32]. За допомогою даного підходу було виявлено, що оптимальною кількістю дерев випадкового лісу є 75, а для побудови моделі МГУА необхідно використовувати швидкий комбінаторний алгоритм.

Застосовані моделі нейронних мереж: при їх формуванні використовувались функція softmax [33] і похибка крос-ентропії [34]. Нейронні мережі були реалізовані з використанням бібліотеки Tensorflow версії 2 та API Keras [35].

Перша модель побудована з використанням алгоритму багатошарового перцептрона [36]. Крос-валідація показала, що оптимальна кількість епох навчання для даної моделі складає 50, а розмір батчу має становить 128.

Рис. 2 ілюструє ітеративний процес навчання моделі багатошарового перцептрона. Можна бачити, що точність класифікації на екзаменаційних даних демонструє тенденцію до збільшення, попри наявність високої похибки крос-ентропії. Це свідчить про адекватність даного варіанту моделі.

Друга модель нейронної мережі була побудована з використанням рекурентної нейронної мережі з довгою короткостроковою пам'яттю (LSTM – від long short-term memory) [37]. Даний тип нейронної мережі відомий високою ефективністю при розв'язанні прогностичних задач, пов'язаних з текстовими даними. Однак, її складна архітектура потребує значної кількості часового ресурсу для тренування ефективної моделі.

На рис. 3 показаний процес навчання отриманої моделі LSTM. Як показано на рисунку нижче, точність і похибка крос-ентропії на екзаменаційній вибірці виявились значно гіршими у порівнянні з моделлю багатошарового перцептрона. Такий результат пояснюється технічними обмеженнями, які не дозволили виділити більше часу на тренування моделі, в результаті чого прогностичний потенціал LSTM не був повністю реалізований. Це обмеження планується виправити у майбутніх роботах.

ISSN 2617-8974
ISSN (online) 2707-8434

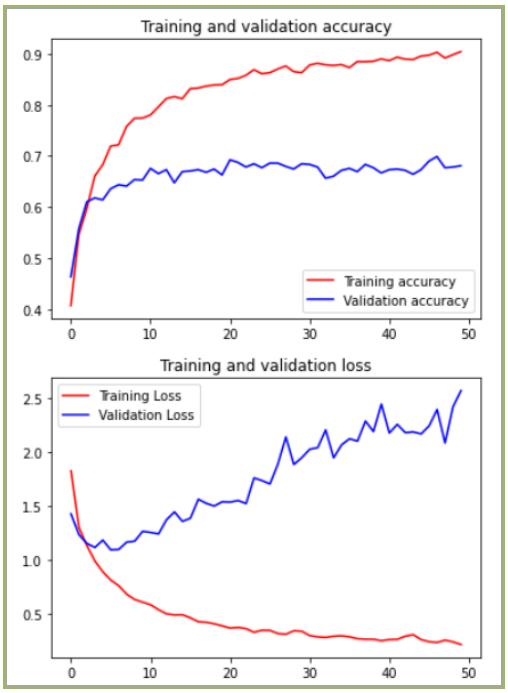

Рисунок 2 – Результати навчання моделі багатошарового перцептрону

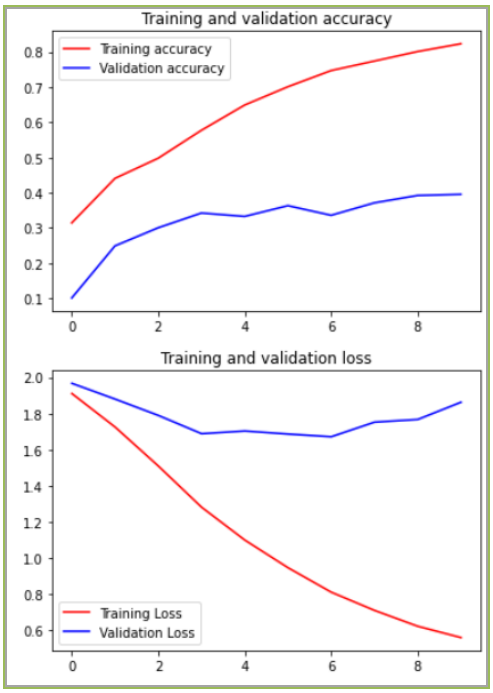

Рисунок 3 – Результати навчання моделі LSTM

VII. ОБГОВОРЕННЯ

В табл. 3 представлене зведення результатів класифікації типу генетичної мутації раку усіх використаних моделей.

З неї видно, модель багатошарового перцептрона дала найкращі результати: точність класифікації 90.6% на навчальній вибірці та 65.1% на екзаменаційній.

Другою за ефективністю стала модель випадкового лісу, точність якої склала 89.2% на навчальній вибірці та 64.9% на тестовій.

Таблиця 3. **Зведення результатів класифікації усіх моделей**

Алгоритм класифікації	**Навчальна вибірка (75%)**		**Екзаменаційна вибірка (25%)**	
	Точність	**F-score**	**Точність**	**F-score**
Наївний Байєс	41.6%	0.432	41.9%	0.436
Логістична регресія	58.5%	0.592	52.7%	0.534
Випадковий ліс	89.2%	0.896	64.9%	0.647
МГУА	41.3%	0.434	41%	0.434
Багатошаровий перцептрон	90.6%	0.908	65.1%	0.657
LSTM	81.3%	0.818	30.9%	0.343

Хоча модель LSTM зайняла третє місце по точності на навчальній вибірці, вона показала найгірший результат класифікації на екзаменаційній, не досягнув порогу у 40%. Такий низький результат можна пояснити технічними обмеженнями, які не дозволили виділити більше часу на повноцінне навчання доволі складної архітектури нейронної мережі.

ISSN 2617-8974
ISSN (online) 2707-8434

Результати класифікації, отримані у даному дослідженні, перевершили результати, продемонстровані учасниками конкурсу Kaggle [14]. Серед цих робіт слід виділити: модель LSTM, яка досягла точність у 36.2% на екзаменаційній вибірці [38]; модель випадкового лісу з 57.8% точності класифікації [39]; лінійну модель, засновану на принципі метода опорних векторів, що в результаті продемонструвала 64.7% точності на екзамені [40].

Кращі результати у даній роботі можна пояснити використанням методів стратифікації та семплування SMOTE для збалансування класів медичних текстових даних, а також застосуванням кореляційного критерію відбору інформативних ознак для запобігання проблеми мультиколінеарності.

VIII. ВИСНОВКИ

Дослідження методів обробки природної мови у комбінації з машинним навчанням для розв'язання багатокласової задачі класифікації типу генетичної мутації раку на основі медичних текстових даних дозволило формулювати наступні висновки: для побудови прогностичних моделей було обрано 4 моделі класичного машинного навчання і 2 моделі глибокого навчання (на основі нейронних мереж). Моделлю, що показала найкращі результати класифікації, стала модель багатошарового перцептрона, точність якої становила 90.6% на тренувальній вибірці і 65.1% на тестовій.

Обрані стратегії попередньої обробки даних перед моделюванням виявились успішними, оскільки допомогли досягти кращих результатів, ніж ті, які продемонстрували учасники конкурсу Kaggle. Отримані моделі можна використовувати, як базовий рівень для побудови діагностичної системи підтримки прийняття рішень. Подібна система допоможе клініцистам і онкологам в діагностичних процедурах ракових пухлин.

Фінансування. Дане дослідження не отримувало зовнішнього фінансування.

Конфлікт інтересів. Автори заявляють про відсутність конфлікту інтересів.

ORCID ID та внесок авторів:
0000-0003-4913-6323 (C, D) Lilia Levchyk
0000-0002-8433-3878 (B) Vitalii Babenko
0000-0003-0936-2246 (A) Kateryna Bovsunovska
0000-0002-3293-5308 (E) Volodymyr Pavlov
0000-0002-1076-9337 (F) Ievgen Nastenko

A - Концепція роботи та дизайн, B - аналіз стандартів текстового опису генетичних мутацій раку, C – Проектування програмного коду, D - Написання статті, E - Критичний огляд, F - Остаточне схвалення статті.

ПЕРЕЛІК ПОСИЛАНЬ

1. Martinez, A., López, G., Bola nos, C., Alvarado, D., Solano, A., López, M., Mora, R. (2017). Building a Personalized Cancer Treatment System. Journal of Medical Systems, 41(2). DOI: 10.1007/s10916-016-0678-z.
2. National Institutes of Health (US); Biological Sciences Curriculum Study. NIH Curriculum Supplement Series [Internet]. Bethesda (MD): National Institutes of Health (US); 2007. Understanding Human Genetic Variation. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20363/
3. Verma M. Personalized medicine and cancer. J Pers Med. 2012 Jan 30;2(1):1-14. DOI: 10.3390/jpm2010001. PMID: 25562699; PMCID: PMC4251363.
4. Cialdella-Kam L, Sabado P, Bispeck MK, Silverman S, Bernstein L, Krawiec V, Hawk E, O'Donnell JF. Implementing cancer prevention into clinical practice. J Cancer Educ. 2012 May;27(2 Suppl):S136-43. doi: 10.1007/s13187-012-0331-6. PMID: 22367592; PMCID: PMC4126604.
5. Maughan, T. (2017). The Promise and the Hype of 'Personalised Medicine.' New Bioethics, 23(1), 13–20. DOI: 10.1080/20502877.2017.1314886
6. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. New England Journal of

ISSN 2617-8974
ISSN (online) 2707-8434

Medicine, 380(14), 1347–1358. DOI: 10.1056/nejmra1814259
7. Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., & Kitchen, G. B. (2021, June 1). Natural language processing in medicine: A review. Trends in Anaesthesia and Critical Care. Churchill Livingstone. DOI: 10.1016/j.tacc.2021.02.007
8. Harrison, C.J., Sidey-Gibbons, C.J. Machine learning in medicine: a practical introduction to natural language processing. BMC Med Res Methodol 21, 158 (2021). DOI: 10.1186/s12874-021-01347-1
9. Saskia Locke, Anthony Bashall, Sarah Al-Adely, John Moore, Anthony Wilson, Gareth B. Kitchen. Natural language processing in medicine: A review. Trends in Anaesthesia and Critical Care, Volume 38, 2021, Pages 4-9. ISSN 2210-8440. DOI: 10.1016/j.tacc.2021.02.007
10. S. Subramanian, I. Baldini, S. Ravichandran, D.A. Katz-Rogozhnikov, K.N. Ramamurthy, P. Sattigeri, et al. A Natural Language Processing System for Extracting Evidence of Drug Repurposing from Scientific Publications. Proceedings of the AAAI Conference on Artificial Intelligence (2020), 34(08), pp. 13369-13381. DOI: 10.1609/aaai.v34i08.7052
11. X. Zhang, J. Kim, R.E. Patzer, S.R. Pitts, A. Patzer, J.D. Schrager. Prediction of emergency department hospital admission based on natural language processing and neural networks. Methods Inf. Med., 56 (2017), pp. 377-389. DOI: 10.3414/ME17-01-0024
12. N. Kolanu, A.S. Brown, A. Beech, Jacqueline Center, C.P. White. OR29-02 natural language processing of radiology reports improves identification of patients with fracture. J Endocr Soc, 4 (2020). DOI: 10.1210/jendso/bvaa046.1619
13. B. Udelsman, I. Chien, K. Ouchi, K. Brizzi, J.A. Tulsky, C. Lindvall. Needle in a haystack: natural language processing to identify serious illness. J. Palliat. Med., 22 (2019), pp. 179-182. DOI: 10.1089/jpm.2018.0294
14. Kaggle. Personalized Medicine: Redefining Cancer Treatment. 2017. URL: https://www.kaggle.com/competitions/msk-redefining-cancer-treatment/data
15. Ma, L., Wiggans, G. R., Wang, S., Sonstegard, T. S., Yang, J., Crooker, B. A., … Da, Y. (2012). Effect of sample stratification on dairy GWAS results. BMC Genomics, 13(1). DOI: 10.1186/1471-2164-13-536
16. Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018, April 1). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research. AI Access Foundation. DOI: 10.1613/jair.1.11192
17. 9. Zhang, Y., Zhang, Y., Qi, P., Manning, C. D., & Langlotz, C. P. (2021). Biomedical and clinical English model packages for the Stanza Python NLP library. Journal of the American Medical Informatics Association, 28(9), 1892–1899. DOI: 10.1093/jamia/ocab090
18. Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications, 181(1), 25–29. DOI: 10.5120/ijca2018917395
19. Hall, M. A. (1999). “Correlation-Based Feature Selection for Machine Learning”, 109 p.
20. Kim, J. H. (2019). Multicollinearity and misleading statistical results. Korean Journal of Anesthesiology, 72(6), 558–569. DOI: 10.4097/kja.19087
21. Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. Statistical Methods and Applications, 19(4), 497–515. DOI: 10.1007/s10260-010-0142-z
22. Anggraeni, M., Syafrullah, M., & Damanik, H. A. (2019). Literation Hearing Impairment (I-Chat Bot): Natural Language Processing (NLP) and Naïve Bayes Method. In Journal of Physics: Conference Series (Vol. 1201). Institute of Physics Publishing. DOI: 10.1088/1742-6596/1201/1/012057
23. Berrar, D. (2018). Bayes’ theorem and naive bayes classifier. In Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics (Vol. 1–3, pp. 403–412). Elsevier. DOI: 10.1016/B978-0-12-809633-8.20473-1
24. Susanti, A. R., Djatna, T., & Kusuma, W. A. (2017). Twitter’s sentiment analysis on GSM services using Multinomial Naïve Bayes. Telkomnika (Telecommunication Computing Electronics and Control), 15(3), 1354–1361. DOI: 10.12928/TELKOMNIKA.v15i3.4284
25. El-Habil, A. M. (2012). An application on multinomial logistic regression model. Pakistan Journal of Statistics and Operation Research, 8(2), 271–291. DOI: 10.18187/pjsor.v8i2.234
26. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. DOI: 10.1023/A:1010933404324
27. Antony Vijay, J., Anwar Basha, H., & Arun Nehru, J. (2021). A dynamic approach for detecting the fake news using random forest classifier and nlp. In Advances in Intelligent Systems and Computing (Vol. 1257, pp. 331–341). Springer Science and Business

ISSN 2617-8974
ISSN (online) 2707-8434

Media Deutschland GmbH. DOI: 10.1007/978-981-15-7907-3_25
28. Vaishnav, V., & Vajpai, J. (2020). Assessment of impact of relaxation in lockdown and forecast of preparation for combating COVID-19 pandemic in India using Group Method of Data Handling. Chaos, Solitons and Fractals, 140. DOI: 10.1016/j.chaos.2020.110191
29. Настенко, Є., Максименко, В., Поташев, С., Павлов, В., Бабенко, В., Рисін, С., … Лазоришинець, В. (2021). ЗАСТОСУВАННЯ МЕТОДУ ГРУПОВОГО УРАХУВАННЯ АРГУМЕНТІВ ДЛЯ ПОБУДОВИ АЛГОРИТМІВ ДІАГНОСТИКИ ІШЕМІЧНОЇ ХВОРОБИ СЕРЦЯ. Біомедична Інженерія і Технологія, (5), 1–9. DOI: 10.20535/2617-8974.2021.5.227141
30. 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. Innov Biosyst Bioeng [Internet]. 2021Sep.10;5(3):153-66. Available from: http://ibb.kpi.ua/article/view/234990
31. Vujović, Ž. (2021). Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599–606. DOI: 10.14569/IJACSA.2021.0120670
32. Berrar, D. (2018). Cross-validation. In Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics (Vol. 1–3, pp. 542–545). Elsevier. DOI: 10.1016/B978-0-12-809633-8.20349-X
33. Peng, H., & Yu, S. (2021). Beyond softmax loss: Intra-concentration and inter-separability loss for classification. Neurocomputing, 438, 155–164. DOI: 10.1016/j.neucom.2020.11.030
34. Ho, Y., & Wookey, S. (2020). The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. IEEE Access, 8, 4806–4813. DOI: 10.1109/ACCESS.2019.2962617
35. Joseph, F. J. J., Nonsiri, S., & Monsakul, A. (2021). Keras and TensorFlow: A Hands-On Experience. In EAI/Springer Innovations in Communication and Computing (pp. 85–111). Springer Science and Business Media Deutschland GmbH. DOI: 10.1007/978-3-030-66519-7_4
36. Murtagh, F. (1991). Multilayer perceptrons for classification and regression. Neurocomputing, 2(5–6), 183–197. DOI: 10.1016/0925-2312(91)90023-5
37. Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2222–2232. DOI: 10.1109/TNNLS.2016.2582924
38. Doc2Vec with Keras URL: https://www.kaggle.com/code/alyosama/doc2vec-with-keras-0-77
39. Basic NLP: Bag of Words, TF-IDF, Word2Vec, LSTM. URL: https://www.kaggle.com/code/reiinakano/basic-nlp-bag-of-words-tf-idf-word2vec-lstm
40. Redefining Cancer Treatment - Linear SVC. URL: https://www.kaggle.com/code/bhuvaneshwaran/redefining-cancer-treatment-linear-svc

UDC 004.81 + 616-006

DIAGNOSTIC ALGORITHMS FOR DETERMINING GENETIC MUTATIONS OF CANCEROUS TUMORS BY MEDICAL TEXT ANALYSIS METHODS

Lillia Levchyk
levchyk.liliia@lll.kpi.ua
Vitalii Babenko
vbabenko2191@gmail.com
Kateryna Bovsunovska
bmk-bks-fbmi@lll.kpi.ua
Volodymyr Pavlov
pavlov.volodymyr@lll.kpi.ua
Ievgen Nastenko
nastenko.e@gmail.com
Department of Biomedical Cybernetics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine

Abstract *– Analytical diagnostic tools, including genetic testing, have advanced to the point where an increasing proportion of disease diagnoses, such as cancer, can be automated. However, the manual work required for the diagnosis of cancerous tumors remains a significant hurdle to progress in this area. The sequencing process for a cancerous tumor can reveal thousands of genetic mutations, but the challenge is to identify the mutations that contribute to cancer growth (drivers) versus those that are neutral (passengers). This task requires a clinical pathologist to manually review and classify each mutation based on information obtained from clinical literature, a time-consuming process. The use of computerized methods for analyzing medical texts has the potential to alleviate the burden of diagnosing cancerous tumors. The aim of this study was to determine the utility of natural language processing and machine learning in automatically identifying cancer genetic mutation types from medical text data. A publicly accessible database of medical text data, containing 3321 observations and annotated with 9 types of cancer genetic mutations by leading researchers and oncologists at the Memorial Sloan Kettering Cancer Center (New York, USA), is available for use. This data was provided as part of a machine learning competition on Kaggle. To address the multi-class classification problem, various machine learning models were employed, including multinomial naive Bayes multinomial logistic regression, random forest, group method of data handling, multilayer perceptron, and a recurrent neural network with long short-term memory. The multilayer perceptron model was found to be the most effective approach for determining the type of genetic mutation, demonstrating a 65.1% prediction accuracy on the test sample (25% of the total dataset). A random forest model also performed well, achieving a 64.9% accuracy. These results outperformed those of the Kaggle contestants, where the highest classification accuracy, 64.7%, was achieved using a linear model based on the support vector method. The combination of natural language processing and machine learning techniques shows great potential for application in the medical field, particularly in the identification of cancer genetic mutation types based on text data. This has the potential to significantly facilitate the work of clinicians and oncologists in the diagnostic process. Further research is planned to achieve more effective results.*

Key words *– cancer genetic mutations, medical text data, natural language processing, text vectorization, machine learning, deep learning.*