Biomedical Engineering and Technology
Issue 2 (8), 2022

ISSN (Print) 2617-8974
ISSN (Online) 2707-8434

УДК 004.852 + 616-018

КОМБІНАЦІЯ ЛОКАЛЬНОЇ ПОРОГОВОЇ БІНАРИЗАЦІЇ ТА МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ПУХЛИН МОЛОЧНОЇ ЗАЛОЗИ

Добровська Людмила Миколаївна
luci.dln17@gmail.com
Бабенко Віталій Олегович
vbabenko2191@gmail.com
Іванченко Аліна Сергіївна
ivanchenko.alina@lll.kpi.ua
кафедра біомедичної кібернетики
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
м. Київ, Україна

Анотація *– Рання діагностика раку молочної залози має колосальне значення, оскільки дана патологія є одним із найбільш розповсюджених чинників летальності серед жінок по всьому світу. Чи не небезпечнішим підтипом раку молочної залози вважається інвазивна протокова карцинома. Зазвичай патологоанатоми фокусуються на областях з подібною карциномою, так як це дозволяє присвоїти оцінку агресивності усьому зразку монтування. Саме тому важливою задачею є автоматизоване виявлення карциноми при діагностиці ракових пухлин молочної залози.*

Мета даної роботи полягала у встановленні основних етапів побудови діагностичних алгоритмів класифікації типу ракової пухлини молочної залози на основі аналізу гістологічних знімків. Для цього було запропоновано алгоритм на основі методу локальної порогової бінаризації (для вилучення інформативних ознак з медичних зображень) та машинного навчання для (побудови моделей розпізнавання типу ракової пухлини молочної залози за допомогою методів класифікації).

База гістологічних знімків, яка використовувалась для дослідження, була взята з відкритого джерела Kaggle, що є онлайн ресурсом для проведення змагань з машинного навчання. Перед виконанням першого етапу дослідження, який полягав у застосуванні алгоритму локальної порогової бінаризації, вибірку зображень було розбито на робочу (75%), для навчання моделей, та екзаменаційну (25%), яка не приймала жодної участі в експериментах аж до отримання результуючої моделі.

Другий етап дослідження полягав у отриманні таких інформативних ознак як дуети (комбінації із двох пікселей) і тріо (комбінації із трьох пікселей). Вони розраховуються після застосування запропонованого методу бінаризації. На основі даних ознак були побудовані моделі наступних алгоритмів класифікації: метод групового урахування аргументів, логістична регресія, наївний Байєсівський класифікатор, метод k найближчих сусідів, а також метод випадкового лісу. Результатом моделювання є 10 моделей класифікації, найкращою з яких стала модель метода k найближчих сусідів, навчена на дуетах бінаризованих пікселей. Ця модель дала на екзаменаційній вибірці 78.5% точності класифікації, значення чутливості становило 0.803, специфічності – 0.767.

Ключові слова *– штучний інтелект, новоутворення молочної залози, класифікація, обробка зображень, машинне навчання.*

I. ВСТУП

Згідно статистики, яка була опублікована Міжнародним агентством по вивченню раку у грудні 2020 року, рак молочної залози випередив рак легень як найбільш часто діагностоване онкологічне захворювання у жінок по усьому світу [1]. За останні два десятиріччя загальна кількість людей, у яких діагностований рак, майже подвоїлась. На сьогоднішній день кожна п’ята людина в світі хворіє раком впродовж усього життя. Згідно прогнозам, у найближчі роки кількість людей, у яких діагностується рак, зросте ще сильніше, і у 2040 році їх буде майже на 50% більше. Більше ніж кожна шоста смерть викликана раком. Усі ці факти підтверджують необхідність інвестицій як у боротьбу з раком, так і у її профілактику.

Успішне впровадження інформаційно-комунікативних технологій в медичну практику є важливою запорукою в оновленні системи медицини і охорони здоров’я, а точніше, у лікуванні раку. Застосування машинного навчання [2] на великих масивах даних (Big Data) революціонізувало область об’єму даних,

ISSN (Print) 2617-8974
ISSN (Online) 2707-8434

завдяки чому підвищилась їхня цінність. Алгоритми машинного навчання виконали значні зміни в області Business Intelligence (BI) [3], аналізуючи великий об'єм неструктурованих, неоднорідних, нестандартних та неповних даних про охорону здоров'я. Вони не лише прогнозують, але і допомагають у прийнятті рішень, і все частіше відмічаються як прорив у неперервному прогресі для підвищення якості обслуговування пацієнтів і зниження вартості лікування.

Найбільш розповсюдженими цілями моделювання алгоритмів машинного навчання є класифікація і прогнозування. Загалом дослідницьке суспільство [4] виділяє п'ять основних методів моделювання: метод опорних векторів (SVM) [5], випадковий ліс [6], логістична регресія [7], дерева прийняття рішень C4.5 [8], та *k* найближчих сусідів (KNN) [9].

Актуальність роботи полягає у дослідженні можливостей машинного навчання в задачі ідентифікації злоякісної пухлини молочної залози на основі зображень медичного класу. Для цього була взята база даних гістологічних знімків із відкритого джерела *Kaggle* [10]. Крім того, планується дослідити можливості текстурного аналізу [11] для винайдення оригінального способу візуальної діагностики гістологічних знімків. Це дасть можливість для медичних спеціалістів не бути повністю залежними від моделей машинного навчання, які володіють властивостями «чорного ящику» [12].

II. ПОСТАНОВКА ЗАДАЧІ

Метою даної роботи було встановлення основних етапів побудови діагностичних алгоритмів класифікації типу ракової пухлини молочної залози на основі аналізу гістологічних знімків, а саме, за допомогою методу локальної порогової бінаризації, а також машинного навчання.

III. МАТЕРІАЛИ І МЕТОДИ ДОСЛІДЖЕННЯ

3.1. Опис клінічного матеріалу

Для реалізації поставленої мети була взята база даних медичних зображень із відкритого джерела Kaggle [10]. Вона містить 6000 гістологічних знімків молочної залози, як з доброякісною раковою пухлиною, так і зі злоякісною (а якщо точніше, то інвазивною протоковою карциномою). Розмір кожного знімку становить 50x50 пікселей (рис. 1).

а)  б) 

Рис. 1. Приклад гістологічного знімка із: а) доброякісною раковою пухлиною, б) злоякісною
Джерело: сайт *Kaggle* [10]

Співвідношення між доброякісними і злоякісними зразками становить 50:50, тобто можна сказати, що вибірка даних зображень є збалансованою, і немає необхідності у використанні методів збалансування класів [13].

3.2. Аналіз останніх досліджень

За останні десятиріччя було винайдено чимало методів штучного інтелекту і машинного навчання для інтелектуального аналізу зображень з метою побудови моделей прогнозування/класифікації. Наразі найбільшою популярністю користуються методи глибокого навчання, а якщо конкретніше, то нейронні мережі [14]. Вони дозволяють напряму працювати зі зображеннями, без використання інформативних мета-ознак, якими знімки як об'єкти не володіють.

Найбільш відомим прикладом використання нейронних мереж для класифікації зображень є робота Кріжевського та ін. [15]. Дослідниками була розроблена згорткова нейронна мережа для класифікації 22 тисяч категорій на базі знімків ImageNet, що налічує більше 15 мільйонів розмічених зображень високої якості. В результаті їм вдалося досягти 84.7% точності класифікації з використанням 7 моделей згорткової нейронної мережі.

Крім властивості класифікації, згорткові нейронні мережі здатні сегментувати зображення, таким чином виділяючи на них

ISSN (Print) 2617-8974
ISSN (Online) 2707-8434

основні зони інтересу. В роботі [16] автори, використовуючи згорткову нейронну мережу під назвою ResNet 101, змогли досягти 86.4% точності сегментації гістологічних зображень раку молочної залози.

Також згорткові нейронні мережі можуть бути використані для вилучення ознак зі зображень. Подібний трюк було спробовано зробити Хао та ін. в роботі [17], де, використовуючи подібну методику, були отримані результати точності більше 90 відсотків.

Єдиною проблемою нейронних мереж є їхня характеристика чорного ящику [12]. Це означає, що незважаючи на свої вражаючі результати прогнозування в різних задачах штучного інтелекту і машинного навчання, навіть досвідчені спеціалісти не здатні до кінця пояснити яким чином подібні результати отримуються. Це є критичним для спеціалістів зі сфери медицини, оскільки від цього в першу чергу залежить людське життя.

Класичні методи машинного навчання є більш простими і зрозумілими, однак вони не здатні видавити подібну ефективність у прогнозуванні. Тим не менш, у поєднанні з текстурним аналізом [11] можна розробити новий підхід до прогнозування на основі медичних зображень.

3.3. Алгоритм розпізнавання типу ракової пухлини

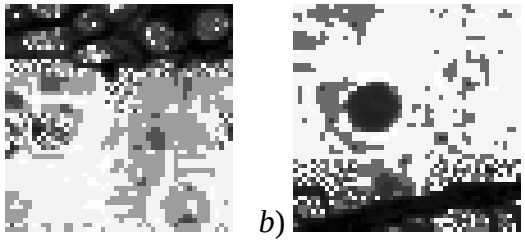
До основних кроків алгоритму розпізнавання типу пухлини молочної залози на основі гістологічних знімків [10] належать такі:

Крок 1. Попередня обробка зображення: зображення проходить різні фільтри (серед них усереднюючий).

Крок 2. Конструювання ознак (Feature Engineering): 1) спочатку локалізуються внутрішні і зовнішні межі області (метод локальної порогової бінаризації); 2) потім виконується еквалізація, яка фактично є покращеним варіантом нормалізації зображення; 3) наостанок виокремлюються ознаки (значущої інформації), які є необхідними для моделювання.

Крок 3. Побудова моделі класифікації (або зіставлення із елементами БД), що робиться на основі класичних алгоритмів машинного навчання.

Усі зображення є кольоровими, що трохи ускладнює мету дослідження. Тому, щоб спростити виконання задачі, знімки були спочатку переведені у сірошкальний варіант, а потім гістограмно перетворені за допомогою методу еквалізації (рис. 2).



a) *b*)

Рис. 2. Приклад еквалізованого знімка із:
а) доброякісною раковою пухлиною,
б) злоякісною

Еквалізація, фактично, є покращеним варіантом нормалізації. Цей метод не лише приводить знімки до однієї шкали вимірювань, але й «вирівнює» гістограму градацій сірого (від 0 до 255) по всій площині, таким чином виділяючи ті деталі, які на оригінальному знімку могли бути непомітні.

Процес еквалізації виконується за допомогою формули (1):

$$I'(x,y) = \min + \frac{(\max - \min)}{N} \bullet \sum_{i=\min}^{k} n_i \qquad (1)$$

де N – кількість усіх пікселів зображення, k – рівень яскравості, що дорівнює значенню градації сірого, n_i – кількість усіх пікселів, які мають i-й рівень яскравості.

Для подальшого дослідження вибірка зображень була поділена на робочу (75%) та екзаменаційну (25%). Екзаменаційна вибірка не буде використовуватись впродовж кінця дослідження, і необхідна для об’єктивної перевірки гіпотези.

Відомо, що існує *порогова бінаризація*, яка переводить усі значення градацій сірого на знімку лише у два: 0 (чорний колір) і 255 (білий колір). Тобто, якщо значення пікселю зображення менше порогу t, то піксель набуває чорного кольору, в іншому випадку – білого. Це можна використовувати в діагностичних цілях.

ISSN (Print) 2617-8974
ISSN (Online) 2707-8434

Припустимо, що знімкам з доброякісною пухлиною притаманні темні відтінки сірого, а злоякісною – світлі. При порозі t = 128 усі темні відтінки стають чорними, а світлі, відповідно, білими. Тоді, за згаданим припущенням знімки з доброякісною пухлиною будуть зафарбовані, загалом, чорними пікселями, а зі злоякісною – білими.

Щоб перевірити це на практиці, виконаємо порогову бінаризацію з порогом t = 128, порахувавши на кожному бінаризованому зображенні долю білого (відношення білих пікселей до усіх). Потім, знайшовши оптимальний поріг долі білого можна визначити, які знімки належать до яких класів. Оптимальний поріг знаходився ітераційним шляхом, і повинен давати максимум коефіцієнту кореляції Метьюза (2):

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (2)$$

Коефіцієнт кореляції Меттьюза (або ККМ) – це міра бінарної класифікації, що враховує усі можливі випадки прогнозування двох класів. За рахунок цього дана міра вважається збалансованою, і не залежить від співвідношення спостережень обох класів. Також ККМ називають кореляцією між реальними і спрогнозованими даними, значення якого варіюються від -1 (повна невідповідність) до 1 (ідеальна відповідність). Значення «0» виходить у випадку, коли класифікатор прогнозує класи випадковим чином.

Знайшовши поріг долі білого 0.49, були отримані наступні значення мір класифікації:

- ККМ – 0.317;
- точність – 65.2%;
- чутливість – 0.51;
- специфічність – 0.795.

Слід також уточнити, що перед бінаризацією кожне еквалізоване зображення буде відфільтроване усереднюючим фільтром (3) розміром 3x3, щоб усунути ризик наявності на зображеннях непотрібних шумів:

$$K = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad (3)$$

Проте, поріг t = 128 і усереднюючий фільтр 3x3 не є єдиними параметрами для виконання подібної задачі. На рис. 3 показані значення ККМ на різних значеннях порогів та розмірах фільтрів.

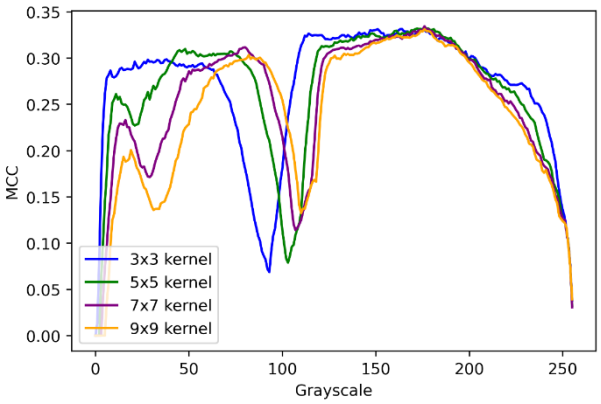


Рис. 3. Прогонка різних варіантів порогів бінаризації та усереднюючих фільтрів з різними розмірами

Найкращий результат було отримано на порозі 176 з усереднюючим фільтром розміром 7x7, а саме:

- поріг долі білого – 0.565;
- ККМ – 0.335;
- точність – 61.6%;
- чутливість – 0.257;
- специфічність – 0.976.

Як можна побачити, хоча точність і впала у порівнянні з порогом 128, тим не менш значення специфічності становить тепер близько 1. Це може означати, що хоча знімкам з доброякісною пухлиною і притаманний чорний колір, однак знімкам зі злоякісною пухлиною притаманні обидва кольори, через що важко знайти баланс долі білого.

Дану проблему можна спробувати виправити наступним чином: знаходити оптимальний поріг не для усього знімку, а для кожної позиції. Тобто, нехай після усереднюючого фільтрування 3x3 отримуємо 4800 знімків розміром 50x50. Візьмемо позицію значення пікселей усіх знімків на позиції (0, 0). Після цього знайдемо ітераційним методом оптимальний поріг відтінка сірого для позиції (0, 0) і так повторюємо для усіх 2500

ISSN (Print) 2617-8974
ISSN (Online) 2707-8434

позицій. Даний процес буде називатися **локальна порогова бінаризація**, оскільки вона виконується не по усій області знімка, а локально на кожній позиції.

IV. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Оптимальний варіант локальної порогової бінаризації було отримано після застосування усереднюючого фільтру розміром 9x9:

- поріг долі білого – 0.509;
- ККМ – 0.352;
- точність – 62.9%;
- чутливість – 0.289;
- специфічність – 0.969.

Результат дещо покращився, однак недостатньо. Можна також спробувати порахувати комбінації пікселей чорного та білого кольору. Наприклад, «дуети» – буде пораховано кількість пар пікселей, таких як: [0, 0], [0, 255], [255, 0], [255, 255].

Подібні ознаки можна використовувати для побудови моделей класифікації із використанням таких алгоритмів:

- метод групового урахування аргументів (МГУА);
- логістична регресія;
- наївний байєсів класифікатор;
- метод *k* найближчих сусідів;
- метод випадкового лісу.

За кожним із алгоритмів будуть побудовані моделі, необхідні для прогнозування типу ракових пухлин молочної залози на основі гістологічних знімків.

Результати моделювання, яке оцінювалось за трьома метриками (точність, чутливість і специфічність), показані в табл. 1-2.

Таблиця 1. **Результати моделей класифікації на основі дуетів (комбінації з двох пікселей)**

Вибірка	Точн.	Чутл.	Спец.
МГУА			
Робоча (75%)	72.5%	0.788	0.664
Екзаменаційна (25%)	75.2%	0.827	0.672
Логістична регресія			
Робоча (75%)	58.6%	0.564	0.607
Екзаменаційна (25%)	62%	0.607	0.633
Наївний Байєс			
Робоча (75%)	57.2%	0.521	0.624
Екзаменаційна (25%)	61.3%	0.573	0.653

Продовження таблиці 1. **Результати моделей класифікації на основі дуетів (комбінації з двох пікселей)**

Вибірка	Точн.	Чутл.	Спец.
KNN			
Робоча (75%)	75.1%	0.773	0.729
Екзаменаційна (25%)	78.5%	0.803	0.767
Випадковий ліс			
Робоча (75%)	99.4%	0.992	0.995
Екзаменаційна (25%)	71.7%	0.722	0.712

Таблиця 2. **Результати моделей класифікації на основі тріо (комбінації з трьох пікселей)**

Вибірка	Точн.	Чутл.	Спец.
МГУА			
Робоча (75%)	73.3%	0.766	0.701
Екзаменаційна (25%)	75.3%	0.807	0.695
Логістична регресія			
Робоча (75%)	59.4%	0.566	0.622
Екзаменаційна (25%)	63.3%	0.612	0.655
Наївний Байєс			
Робоча (75%)	57.2%	0.521	0.624
Екзаменаційна (25%)	61.3%	0.573	0.653
KNN			
Робоча (75%)	74.1%	0.752	0.73
Екзаменаційна (25%)	77.8%	0.792	0.763
Випадковий ліс			
Робоча (75%)	100%	1	1
Екзаменаційна (25%)	72.9%	0.74	0.718

Експерименти проводилися у один етап: розпізнавання на основі моделей машинного навчання (МГУА, логістична регресія, наївний Байєс, KNN, випадковий ліс). В якості ознак були пораховані дуети та тріо комбінацій бінаризованих пікселей кожного знімку.

Результати експерименту показали, що найкращий результат на тестовій вибірці дала модель KNN на основі дуетів, де точність становила 78.5%. Однак у майбутньому потрібно провести ще ряд додаткових експериментів.

VII. ВИСНОВКИ

В результаті виконання даного дослідження був запропонований та розроблений алгоритм локалізації та розпізнавання типу пухлини молочної залози на основі гістологічних знімків.

Точність розпізнавання на основі моделі KNN на множині тестування становила 78.5%.

Основна перевага запропонованого методу у порівнянні з існуючими – збільшення точності на множині тестування.

ISSN (Print) 2617-8974
ISSN (Online) 2707-8434

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. Ludmila Dobrovska (40%) – 0000-0002-4055-6834
2. Vitalii Babenko (10%) – 0000-0002-8433-3878
3. Alina Ivanchenko (50%) – 0000-0002-5508-2328

ПЕРЕЛІК ПОСИЛАНЬ

[1] WHO | Breast cancer', WHO. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020)
[2] Tavasoli, S. (2021). The Importance of Machine Learning for Data Scientists. https://www.simplilearn.com/importance-of-machine-learning-for-data-scientists-article
[3] Spil, T. A. M., Stegwee, R. A., & Teitink, C. J. A. (2002). Business intelligence in healthcare organizations. In Proceedings of the Annual Hawaii International Conference on System Sciences (Vol. 2002-January, pp. 9–17). IEEE Computer Society. DOI: 10.1109/HICSS.2002.994108
[4] Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015
[5] Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2), 121–167. DOI: 10.1023/A:1009715923555
[6] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. Stata Journal, 20(1), 3–29. DOI: 10.1177/1536867X20909688
[7] Connelly, L. (2020). Logistic regression. MEDSURG Nursing, 29(5), 353–354. DOI: 10.46692/9781847423399.014C4.5
[8] Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. Edik Informatika, 2(2), 213–219. DOI: 10.22202/ei.2016.v2i2.1465
[9] Xing, W., & Bei, Y. (2020). Medical Health Big Data Classification Based on KNN Classification Algorithm. IEEE Access, 8, 28808–28819. DOI: 10.1109/ACCESS.2019.2955754
[10] Kaggle. Breast Histopathology Images Dataset. https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images
[11] Kunimatsu, A., Yasaka, K., Akai, H., Sugawara, H., Kunimatsu, N., & Abe, O. (2022). Texture Analysis in Brain Tumor MR Imaging. Magnetic Resonance in Medical Sciences. Japanese Society for Magnetic Resonance in Medicine. DOI: 10.2463/mrms.rev.2020-0159
[12] Loyola-Gonzalez, O. (2019). Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access. Institute of Electrical and Electronics Engineers Inc. DOI: 10.1109/ACCESS.2019.2949286
[13] Yan, S., Kao, H. T., & Ferrara, E. (2020). Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. In International Conference on Information and Knowledge Management, Proceedings (pp. 1715–1724). Association for Computing Machinery. DOI: 10.1145/3340531.3411980
[14] Chen, Y., Xie, Y., Song, L., Chen, F., & Tang, T. (2020, March 1). A Survey of Accelerator Architectures for Deep Neural Networks. Engineering. Elsevier Ltd. DOI: 10.1016/j.eng.2020.01.007
[15] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. DOI: 10.1145/3065386
[16] Khan, F. S., Mohd, M. N. H., Khan, M. D., & Bagchi, S. (2020). Breast Cancer Histological Images Nuclei Segmentation using Mask Regional Convolutional Neural Network. In 2020 IEEE Student Conference on Research and Development, SCOReD 2020 (Vol. 2020-January). Institute of Electrical and Electronics Engineers Inc. DOI: 10.1109/SCOReD50371.2020.9383186
[17] Hao Y, Zhang L, Qiao S, Bai Y, Cheng R, Xue H, et al. (2022) Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix. PLoS ONE 17(5): e0267955. DOI: 10.1371/journal.pone.0267955

REFERENCES

[1] WHO | Breast cancer', WHO. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020)
[2] Tavasoli, S. (2021). The Importance of Machine Learning for Data Scientists. https://www.simplilearn.com/importance-of-machine-learning-for-data-scientists-article
[3] Spil, T. A. M., Stegwee, R. A., & Teitink, C. J. A. (2002). Business intelligence in healthcare organizations. In Proceedings of the Annual Hawaii International Conference on System Sciences (Vol. 2002-January, pp. 9–17). IEEE Computer Society. DOI: 10.1109/HICSS.2002.994108
[4] Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015
[5] Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2), 121–167. DOI: 10.1023/A:1009715923555
[6] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. Stata Journal, 20(1), 3–29. DOI: 10.1177/1536867X20909688
[7] Connelly, L. (2020). Logistic regression. MEDSURG Nursing, 29(5), 353–354. DOI: 10.46692/9781847423399.014C4.5
[8] Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. Edik Informatika, 2(2), 213–219. DOI: 10.22202/ei.2016.v2i2.1465
[9] Xing, W., & Bei, Y. (2020). Medical Health Big Data Classification Based on KNN Classification Algorithm. IEEE Access, 8, 28808–28819. DOI: 10.1109/ACCESS.2019.2955754
[10] Kaggle. Breast Histopathology Images Dataset. https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images
[11] Kunimatsu, A., Yasaka, K., Akai, H., Sugawara, H., Kunimatsu, N., & Abe, O. (2022). Texture Analysis in Brain Tumor MR Imaging. Magnetic Resonance in Medical Sciences. Japanese Society for Magnetic Resonance in Medicine. DOI: 10.2463/mrms.rev.2020-0159
[12] Loyola-Gonzalez, O. (2019). Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access. Institute of Electrical and Electronics Engineers Inc. DOI: 10.1109/ACCESS.2019.2949286
[13] Yan, S., Kao, H. T., & Ferrara, E. (2020). Fair Class Balancing: Enhancing Model Fairness without Observing

ISSN (Print) 2617-8974
ISSN (Online) 2707-8434

Sensitive Attributes. In International Conference on Information and Knowledge Management, Proceedings (pp. 1715–1724). Association for Computing Machinery. DOI: 10.1145/3340531.3411980

[14] Chen, Y., Xie, Y., Song, L., Chen, F., & Tang, T. (2020, March 1). A Survey of Accelerator Architectures for Deep Neural Networks. Engineering. Elsevier Ltd. DOI: 10.1016/j.eng.2020.01.007

[15] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. DOI: 10.1145/3065386

[16] Khan, F. S., Mohd, M. N. H., Khan, M. D., & Bagchi, S. (2020). Breast Cancer Histological Images Nuclei Segmentation using Mask Regional Convolutional Neural Network. In 2020 IEEE Student Conference on Research and Development, SCOReD 2020 (Vol. 2020-January). Institute of Electrical and Electronics Engineers Inc. DOI: 10.1109/SCOReD50371.2020.9383186

[17] Hao Y, Zhang L, Qiao S, Bai Y, Cheng R, Xue H, et al. (2022) Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix. PLoS ONE 17(5): e0267955. DOI: 10.1371/journal.pone.0267955

ISSN (Print) 2617-8974
ISSN (Online) 2707-8434

UDC 004.852 + 616-018

COMBINATION OF LOCAL THRESHOLD BINARIZATION AND MACHINE LEARNING FOR BREAST TUMOR CLASSIFICATION

Ludmila Dobrovska
luci.dln17@gmail.com
Vitalii Babenko
vbabenko2191@gmail.com
Alina Ivanchenko
ivanchenko.alina@lll.kpi.ua
Department of Biomedical Cybernetics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine

Abstract – *Early diagnosis of breast cancer is of great importance, as this pathology is one of the most common causes of mortality among women around the world. Invasive ductal carcinoma is the most dangerous subtype of breast cancer. Typically, pathologists focus on areas with similar carcinoma, as this allows an aggressiveness score to be assigned to the entire mount specimen. That is why the automated detection of carcinoma in the diagnosis of cancerous tumors of the mammary gland is an important task.*

The purpose of this work was to establish the main stages of building diagnostic algorithms for the classification of the type of breast cancer tumor based on the analysis of histological images. For this, an algorithm was proposed based on the local threshold binarization method (for extracting informative features from medical images) and machine learning (building breast cancer tumor type recognition models using classification methods).

The database of histological images used for the study was taken from the open-source Kaggle, an online resource for running machine learning competitions. Before performing the first stage of the research, which consisted of the application of the local threshold binarization algorithm, the sample of images was divided into working (75%), for model training, and examination (25%), which did not participate in any experiments until obtaining the resulting model.

The second stage of the research consisted in obtaining such informative features as duets (combinations of two pixels) and trios (combinations of three pixels). They are calculated after applying the proposed binarization method. Models of the following classification algorithms were built based on these features: group method of data handling, logistic regression, naive Bayesian classifier, the method of k nearest neighbors, and random forest method. The result of the modeling is 10 classification models, the best of which was the k-nearest neighbors model, trained on binarized pixel pairs. This model gave 78.5% classification accuracy on the exam sample, the sensitivity value was 0.803, and the specificity value was 0.767.

Keywords *– Artificial Intelligence, Breast Neoplasms, Classification, Image Processing, Machine Learning.*