КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет комп'ютерних наук та кібернетики
Кафедра обчислювальної математики

Кваліфікаційна робота

на здобуття ступеня магістра

за спеціальністю 113 Прикладна математика

на тему:

МЕТОДИ МАШИННОГО НАВЧАННЯ В ЗАДАЧАХ ДІАГНОСТИКИ РАКУ

Виконав студент 2-го курсу
Кравець Олексій Павлович _____

Науковий керівник:
Професор, доктор фіз.-мат. наук
Клюшин Дмитро Анатолійович _____

Засвідчую, що в цій роботі немає запозичень з праць інших авторів без відповідних посилань.

Студент _____

Роботу розглянуто й допущено до захисту на засіданні кафедри обчислювальної математики
«___» _____ 202_ р.,
протокол № ___
Завідувач кафедри
С. І. Ляшко _____

Київ — 2022

ЗМІСТ

1 Вступ

Згідно World Cancer Research Fund International рак молочної залози – це найпоширеніший рак серед жінок та найбільш росповсіджений рак в цілому. Тільки в 2020 році було діагностовано більше 2.26 випадків захворювання раком молочної залози серед жінок. Тому постає важливе питання створення методу ранньої діагностики раку молочної залози. Такий метод повинен бути простим, дешевим, точним та не травматичним. Розповсюджені в даний момент методи діагностики, такі як: клінічний огляд, мамографія та аспіраційна біопсія, далеко не завжди показують високу точність, а також можуть бути травматичними.

Моя робота зосереджена на розгляді методів машинного навчання для діагностики раку молочної залози, використовуючи метод отримання ознак, що отримуються за допомогою фрактального аналізу розподілу хроматина у забарвлених за Фельгеном зображеннях ядер букального епітелію [1].

2 Огляд літератури

Існує багато сучасних підходів до задачі діагностики раку молочної залози. Так, наприклад, роботи [2], [3], [4] зосереджені на використанні згорткових нейронних мереж (CNN) та рекурентних нейронних мереж (RNN) для диагностики раку молочної залози на основі гістопатологічних зображеннь, що зібрані методом біопсії. Причому в роботах [3], [4] приведені методи, що можна інтерпретувати, що важливо для медицинської діагностики. Іншим прикладом може бути робота [5], що використовує для діагностики раку штучну нейронну мережу поєднану з бджолиним алгоритмом, для вибору найкращих ознак і підбору параметрів, а також датасет з зображеннями, що були отримані методом тонкоголкової біопсії. Робота [6] використовує ті самі дані, що і [5], але для діагностики використовує методи Adaboost та Random forest. Інша робота [7] також є прикладом діагностики раку молочної залози на основі багатопараметричної магнітно-резонансна томографії. З зображень за допомогою згорткової нейронної мережі добуваються ознаки, а потім проводиться класифікація з використанням методу опорних векторів. Робота [8] зосереджена на розгляді штучних нейронних мереж, що донавчаються на зображеннях мамографії. Також прикладом може бути робота [9], в якій для діагностики раку молочної залози за допомогою зображень мамографії використовуються нейронні мережі та методи глибокого навчання. Усі приведені в якості прикладів роботи показують високу точність. Але ці роботи ґрунтуються на даних, що зібрані небезпечними та дорогими методами діагностики. Так біопсія та мамографія високої чіткості можуть травматичними або шкідливими для здоров'я через радіоактивне опромінення. Тому постає питання пошуку більш безпечного та дешевого методу раньої діагностики раку молочної залози.

Організм людини має здатність реагувати на зміни внутрішньої і зовнішньої середи своїми системами. Так, наприклад, в 1960-ті H. Nieburgs [10] виявили зміни у розподілі хроматина у клітинах букального епітелію ракових пацієнтов. Це питання піднімається і в сучасних дослідженнях, так у роботі [22] досліджуються зміни в розподілі хроматину у випадку раку молочної залози. Пізніше в 2009 Lieberman-Aiden [11] виявив, що згортка ДНК у клітини має фрактальні властивості. Після цього почалося активне дослідження та використання фрактального аналізу в діагностиці захворювань. Так сучасні дослідження показують наявність фрактальних властивостей у мозку [12], серці [13], легенях [14], кістках [15],

очах [16], плаценті [17]. Фрактальна розмірність має широку сферу використання. Так в роботі [23] фрактальна розмірність клітини розглядається як міра гетерогенності клітин складної ендометріоїдної гіперплазії та диференційованої ендометріоїдної карциноми. Також сучасні дослідження [18] розглядають фрактальний аналіз хроматина як потенційних індикатор впливу на людину іонізуючого випромінювання.

Ще в 1994 році в роботі [24] було виявлено, що цитологічні дослідження вмісту ДНК і текстури хроматину можна використовувати для детекції злоякісних новоутворень. Також було показано, що виявлені зміни в клітинах з'являються не тільки в області пухлини, але і на деякій відстані від неї. Такий висновок також підтверджує сучасне дослідження [19] в якому для діагностики раку носоглотки використовувалися гістологічно нормальні клітини поверхневого епітелію носоглотки. Інша робота [21] показує використання фрактального розмірності клітин крові для діагностики лейкемії.

В роботах [1], [20] діагностика раку молочної залози проводиться на основі фрактального аналізу розподілу хроматину в ядрах букального епітелія, що були взяти з ротової порожнини. Отримані результати показують закономірність між наявність раку молочної залози та змінами в інтерфазному ядрі букального епітелію.

Проте в роботах [1], [20] для класифікації пацієнтів була використана статистика Петуніна [25] в якості міри близькості між вибірками. Моя мета полягає у пошуку інших підходів та методів машинного навчання для класифікації пацієнтів з раком молочної залози.

3 Опис даних

В роботі використовуються вже оброблені дані та готові ознаки, що були отримані з аналізу контрольної групи (29 людей), групи пацієнтів з раком молочнох залозі другої стадії (68 людей) та групи пацієнтів з фиброаденоматозом (доброякісна пухлина) (33 людини). Всі діагнози були підтверджені гістологічно. Дані базуються на морфологічному датасеті, що складається з 20256 зображень інтерфазного ядра букального епітелію (6752 адер клітин було просканировано у трьох варіантах: без фільтру, з жовтим фільтром, з фіолетовим фільтром). Морфологічним матеріалом є мазки епітеліоцитів слизової оболонки порожнини рота.

Для отримання ознак, що базуються на властивостях розподілу хроматина в ядрі клітини, з морфологічного датасету було використано лише зображення без фільтру контрольної групи та пацієнтів з раком молочної залози. Перед обрахуванням фрактальної розмірності зображення була проведена передобробка зображень. Кожне зображення було бінарізовано за методом Отцу [31], після цього було знайдено та видалено артефакти. Таким чином на зображеннях залишилися лише ядра клітин на білому фоні.

Після цього для кожного зображення було обраховано його фрактальну розмірність для аналізу розподіла хроматина в ядрі клітини. Причому для забеспечення інваріантності фрактальної розмірності відносно поворота зображення була використана крива Пеано (крива, що заповнює простір) [32], а саме крива Серпінського, що проходить через кожну точку зображення. Таким чином зображення з трьома каналами (RGB) було перетворене в три вектора кожен з яких відповідає одному з каналів зображення. Для обчислення фрактальної розмірності використовується експонента Херста [33].

Таким чином датасет з ознаками, що використовуються в роботі – це дані 97 пацієнтів (68 пацієнтів з раком молочної залози, 29 людей з контрольної групи), кожен з яких представлений трьома вибірками (по кожному з каналів RGB) фрактальних розмірностей для кожного зображення інтерфазного ядра букального епітелію пацієнта.

3.1 Модифікація та аналіз даних

Для того щоб використати звичайні методи машинного навчання додамо ще декілька ознак, це необхідно оскільки кожен пацієнт представлений лише вибіркою з фрактальних розмірностей, що ускладнює процес класифікації.

Основна ідея розширення кількості ознак полягає у додаванні до даних різних середніх та статистичних величин, що обчислені по кожній вибірці. Отже по кожній вібірці кожного пацієнта обрахуємо наступні велечини

- середнє арифметичне: $x_{\text{mean}} = \frac{\sum_i^n x_i}{n}$
- середнє геометричне: $x_{\text{gmean}} = (\prod_i^n x_i)^{\frac{1}{n}}$
- середнє гармонічне: $x_{\text{hmean}} = \frac{1}{\frac{1}{n}\sum_i^n \frac{1}{x_i}}$
- медіана
- стандартне відхилення: $x_{\text{std}} = \sqrt{\frac{1}{n}\sum_i^n (x_i - \overline{x})^2}$

Проведемо попередній аналіз отриманих даних. Розглянемо рисунки (3.1, 3.2, 3.3).

Аналізуючи вигляд графіків (3.1, 3.2, 3.3) можна висунути теорію, що ознаки отримані з синього каналу найбільш інформативні.

Рис. 3.1: Порівняльний графік для ознак, що були створені з вибірки синього каналу. Графік демонструє розподіл даних де осі ординат та абсцис – це відповідні ознаки. Діагональні графіки розподілу даних по відповідній ознаці. Параметр **BC** рівний 0 – це контрольна група, **BC** рівний 1 – група хворих на рак грудей

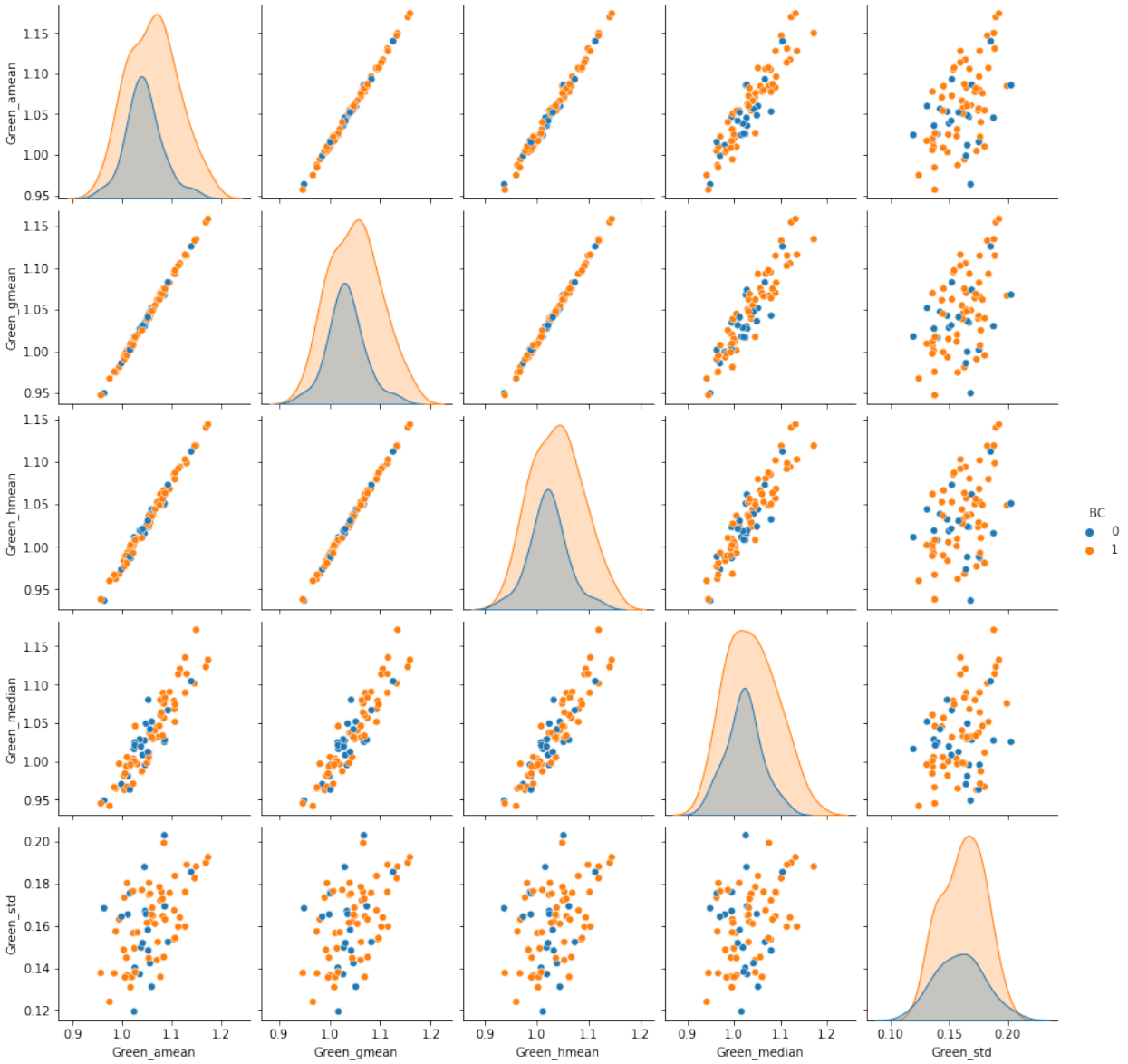

Рис. 3.2: Порівняльний графік для ознак, що були створені з вибірки зеленого каналу. Графік демонструє розподіл даних де осі ординат та абсцис – це відповідні ознаки. Діагональні графіки розподілу даних по відповідній ознаці. Параметр **BC** рівний 0 – це контрольна група, **BC** рівний 1 – група хворих на рак грудей

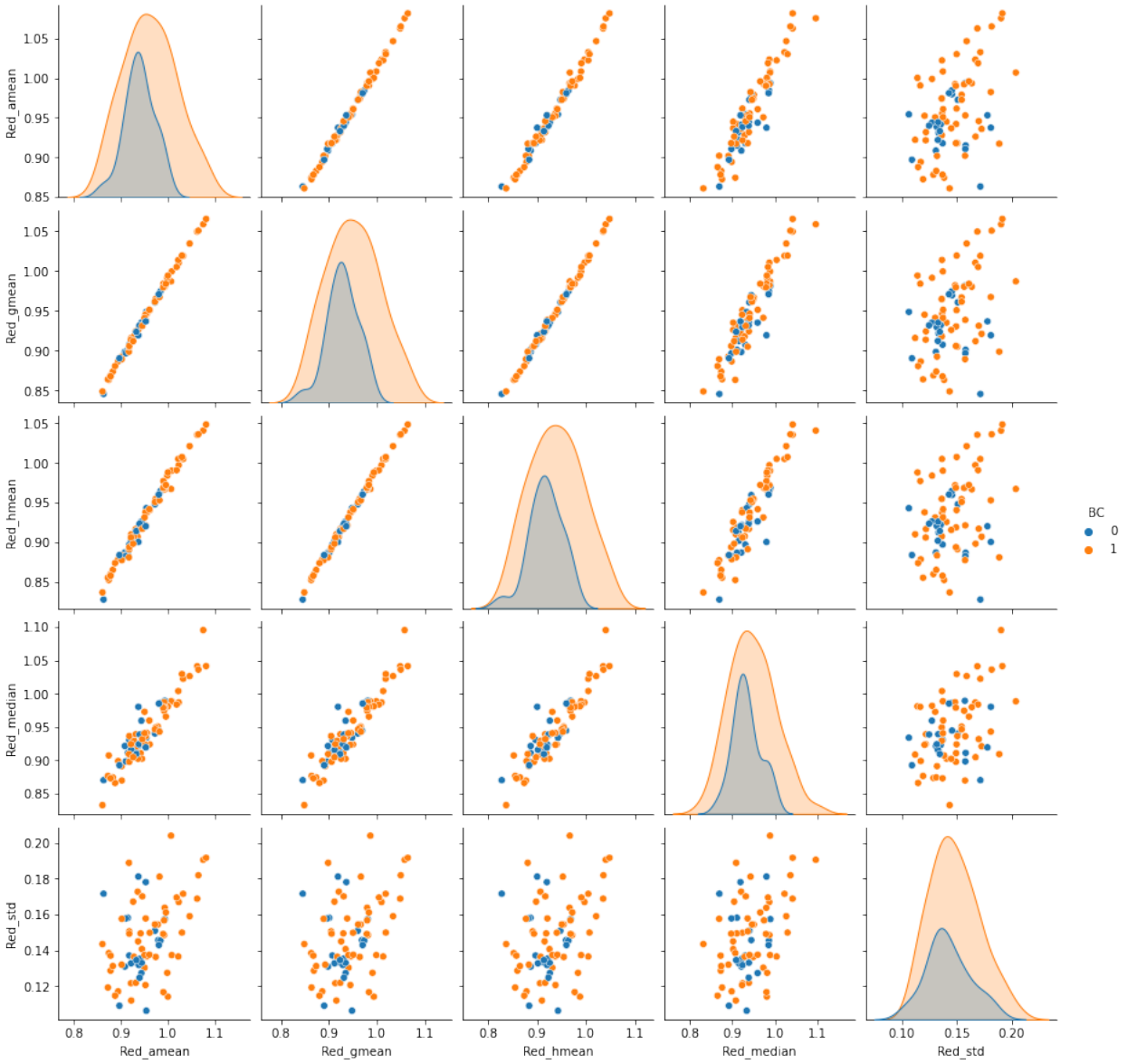

Рис. 3.3: Порівняльний графік для ознак, що були створені з вибірки червоного каналу. Графік демонструє розподіл даних де осі ординат та абсцис – це відповідні ознаки. Діагональні графіки розподілу даних по відповідній ознаці. Параметр **BC** рівний 0 – це контрольна група, **BC** рівний 1 – група хворих на рак грудей

4 Adaboost

Розглянемо алгоритм Adaboost [27] для класифікації пацієнтів з раком молочної залози. Це потужний алгоритм машинного навчання, що заснований на послідовному об'єднанні простих алгоритмів класифікації, кожен наступний з яких буде виправляти помилку попередніх.

4.1 Опис алгоритму Adaboost

Нехай ми маємо наступні дані для навчання алгоритму: $(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots (\mathbf{x}_n, c_n)$ де $\mathbf{x}_i \in R^p$ – це вхід моделі, а $c_i \in \{1, 2, \ldots, K\}$ – вихід моделі, K – кількість класів.

Мета алгоритму полягає у пошуку правила класифікації $C(\mathbf{x})$ на основі навчальних даних. При новому вхідному $\mathbf{x}$ $C(\mathbf{x})$ повертає мітку класу з $\{1, 2, \ldots K\}$.

Алгоритм Adaboost-SAMME

1. Ініціалізувати ваги спостереження $w_i = \frac{1}{n},\ i = 1, 2, \ldots, n.$
2. Починая з $m = 1$ до M:
   а) Навчаємо простий класифікатор $T^{(m)}(x)$ на навчальних даних використовуючи ваги w_i.
   б) Обраховуємо похибку отриманого класифікатору
   $$err^{(m)} = \sum_{i=1}^{n} w_i \mathbf{I}\left(c_i \neq T^{(m)}(x_i)\right) / \sum_{i=1} n w_i$$
   в) Обраховуємо
   $$\alpha^{(m)} = \log \frac{1 - err^{(m)}}{err^{(m)}} + \log(K - 1)$$
   г) Оновлюємо ваги спостереження
   $$w_i \leftarrow w_i \cdot \exp\left(\alpha^{(m)} \cdot \mathbf{I}\left(c_i \neq T^{(m)}(x_i)\right)\right)$$
   for $i = 1, \ldots, n$
   д) Нормалізуємо нові ваги w_i
3. В результаті отримуємо таку функцію класифікації
   $$C(\mathbf{x}) = \arg\max_k \sum_{m=1}^{M} \alpha^{(m)} \cdot \mathbf{I}\left(T^{(m)}(\mathbf{x}) = k\right)$$

де $\mathbf{I}$ – індикаторна функція.

Decision Tree

В якості простого класифікатору у алгоритмі Adaboost зазвичай використовується дерево рішень (Decision Tree) [28]. Цей алгоритм машинного навчання відомий легкістю інтерпретації. Алгоритм будує дерево в кожному вузлі якого проходить розщеплення даних по обраній алгоритмом змінній. Кожний лист дерева належить одному з класів на як проводиться класифікація.

Нехай нашу навчальні дані містять N спостережень: тобто $(\mathbf{x}_i, y_i)$, де $i = 1, 2, \ldots, N$, а $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Мітки класів $y_i \in \{1, 2, \ldots K\}$, де K – кількість класів. Нехай ми маємо розподіл на M регіонів $R_1, R_2, \ldots, R_M$, тоді в вузлі m, що представляє регіон R_m з N_m навчальним прикладами всередині, позначимо

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} \mathbf{I}(y_i = k)$$

пропорцію класу k у вузлі m. Ми класифікуємо спостереження у вузлі m у класс $k(m) = \arg\max_k \hat{p}_{mk}$. Потрібно зазначити, що при використанні вагів спостереження будуть змінюватися саме пропорції класів у вузлах.

В в експериментах з алгоритмом Adaboost для вибору оптимальної змінної для розщеплення використовується Індекс Джині

$$Q_m(T) = \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

де T – побудоване дерево. Для розщеплення вибирається та змінна, що мінімізує індекс Джині.

Інший критерій розщеплення – це крос-ентропія

$$Q_m(T) = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

У дерева рішень є багато критеріїв зупинки розщеплення, але в алгоритмі Adaboost зазвичай використовуються дерева глибини 1 або дерева максимальної глибини 2.

4.2 Тестування Adaboost

Для проведення тестування я розділив датасет на дві частини навчальну вибірку (80%) та на тестову вибірку (20%) на якій буде оцінюватися отримана модель.

Для алгоритму Adaboost я проводив тестування з різними наборами параметрів, а саме кількість класифікаторів та швидкість навчання. Потрібно зазначити, що параметр швидкості навчання – це ваги, що застосовуються до кожного класифікаторі на кожній ітерації бустингу. Це означає, що після побудови m класифікаторів, вага першого класифікатора $T^{(1)}(x)$ завдяки буде рівною lr^{m-1}, де lr – відповідно швидкість навчання. Аналогіно вага класифікатора $T^{(2)}(x)$ буде lr^{m-1}. Таким чином при швидкості навчання більше одиниці збільшується внесок згенерованих раніше класифікаторів у кінцевий результат, а при швидкості навчання меньше одиниці – навпаки зменшується. При такому підході потрібно притримуватися компромісу між швидкістю навчання та кількістю класифікаторів.

Також для кожного набору параметрів я проводив кроссвалідацію з розбиттям навчальних даних на п’ять частин.

Вибір оптимального набору параметрів я робив за усередненою по кроссвалідації метрикою accuracy:

$$\text{accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

Після вибору оптимальних параметрів моделі проводиться її навчанання на всій навчальній вибірці. Остаточна оцінка моделі проводиться на тестовій вибірці за метриками:

- $\text{accuracy} = \frac{\text{True Positive+True Negative}}{\text{True Positive+True Negative+False Positive+False Negative}}$
- $\text{sensitivity} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$
- $\text{specificity} = \frac{\text{True Negative}}{\text{True Negative+False Positive}}$

Для алгоритму Adaboost з класифікатором Decision Tree з максимальною глибиною 1 виявилися найкращими наступні параметри: швидкість навчання 0.2, кількість класифікаторів - 150. Процес пібору параметру можно побачити на рисунку (4.1).

Отримана модель дає такі значення метрик на тестовій вибірці:

- $\text{accuracy} = 0.95$
- $\text{sensitivity} = 0.92$

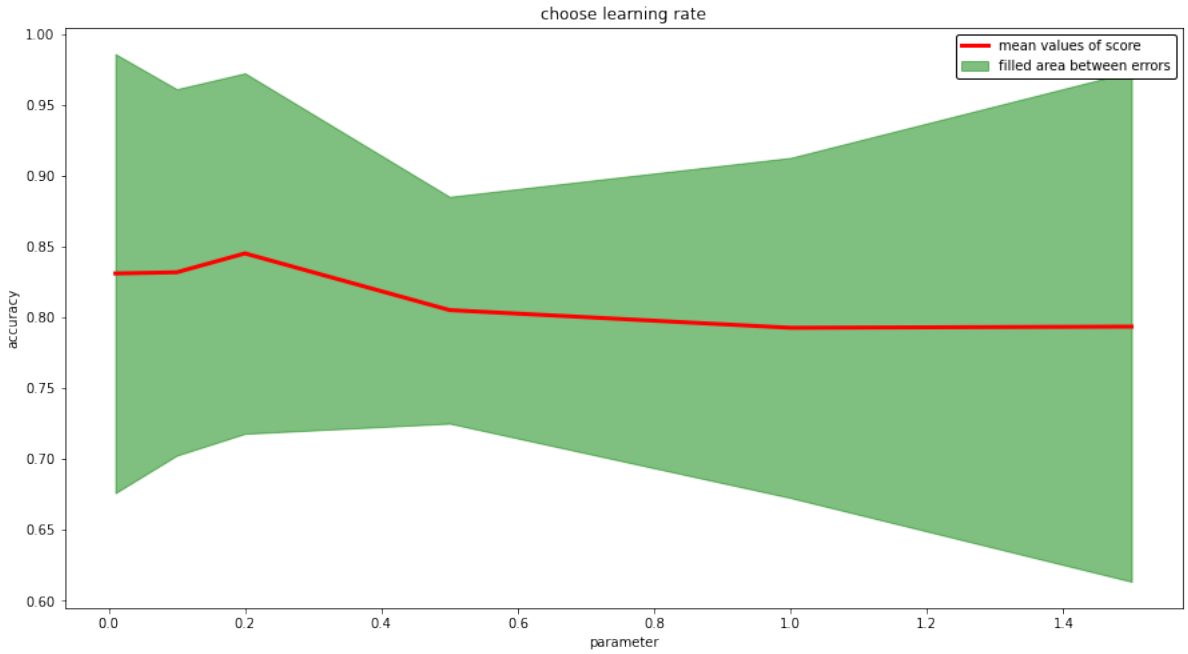

Рис. 4.1: Графік підбору оптимальної швидкості навчання для алгоритму Adaboost з Decision Tree with depth 1 при 150 класифікаторів. На осі абсцис позначено значення параметру швидкості навчання, на осі ординат – значення метрики accuracy. Червоним позначено середнє значення метрики accuracy, зеленим – відхилення значення метрики від середнього.

- specificity $= 1.0$

Для алгоритму Adaboost з класифікатором Decision Tree з максимальною глибиною 2 виявилися найкращими наступні параметри: швидкість навчання 1.5, кількість класифікаторів - 150. Процес пібору параметру можно побачити на рисунку (4.2).

Отримана модель дає такі значення метрик на тестовій вибірці:

- accuracy $= 0.85$
- sensitivity $= 0.92$
- specificity $= 0.71$

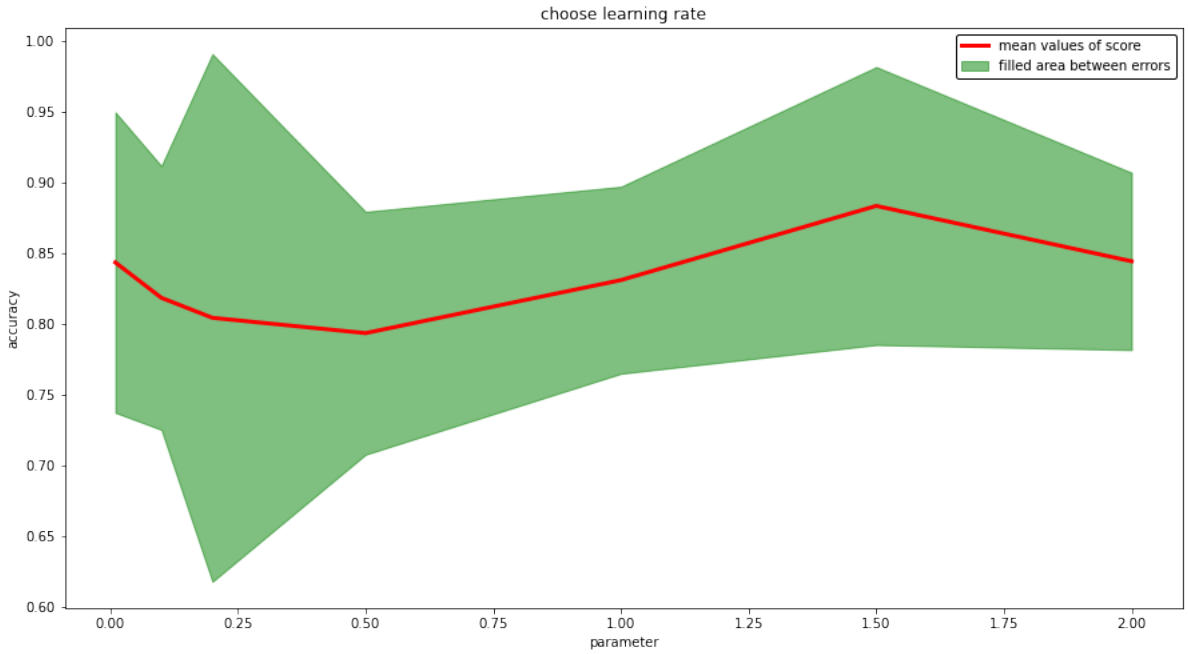

Рис. 4.2: Графік підбору оптимальної швидкості навчання для алгоритму Adaboost з Decision Tree with depth 2 при 150 класифікаторів. На осі абсцис позначено значення параметру швидкості навчання, на осі ординат – значення метрики accuracy. Червоним позначено середнє значення метрики accuracy, зеленим – відхилення значення метрики від середнього.

4.3 Висновки Adaboost

Алгоритм Adaboost з 150 деревами рішень глибиною 1 та швидкістю навчання 0.2 показав найкращі результати. Тестування показує, що алгоритм Adaboost гарно працює на поставленій задачі. Отримані accuracy (95%), sensitivity (92%) та specificity (100%) перевершують результати деяких інших методів діагностики. Так, наприклад, sensitivity мамографії приблизно 87% (Breast Cancer Surveillance Consortium, 2017), а specificity вагається між 93% та 88% [26].

Проте алгоритм Adaboost не має зрозумілої та прозорої інтерпретації, що важливо у задачах діагностики захворювань.

5 Random Forest

Розглянемо алгоритм Random Forest [29] для задачі діагностики раку молочної залози. Також буде перевірена висунута раніше теорія про більшу інформативність синього каналу даних.

Ідея алгоритму Random forest полягає у побудові ансамблю Decision Tree, що треновані на різних підвибірках з навчальних даних. Кінцевим результатом класифікації є усереднення всіх результатів побудованих дерев рішень.

5.1 Опис алгоритму Random Forest

Навчання алгоритму Random forest полягає у застосуванні техніки Boostrap aggregation (Bagging) до алгоритмів Decision Tree. Тобто, нехай ми маємо навчальні дані $X = \mathbf{x}_1, \dots \mathbf{x}_n$ з мітками класів $Y = y_1, \dots y_n$, Boostrap aggregation повторно (B разів) вибирає випадкову підвибрку з навчальної вибірки та навчає алгоритм Decision Tree на новій отриманих вибірці.

Для $b = 1, \dots, B$:

1. Створюємо підвибірку X_b, Y_b з навчальної вибірки X, Y
2. Тренуємо Decision Tree f_b на вибірці X_b, Y_b

Після тренування всіх алгоритмів Decision Tree, для нових даних $\mathbf{x}'$ робиться усереднення результатів усіх класифікаторів Decision Tree. Тобто буде вибраний той класс, який вибрала найбільша кількість Decision Tree. Такий метод ще називається методом простого голосування.

Для перевірки гіпотези про більшу інформативність даних отриманих з синього каналу в роботі був використаний метод Випадкового підпростору (Random subspace method) [30] для вибору підвибірки для навчання алгоритмів Decision Tree. Ідея методу полягає у проектуванні навчальної вибірки на підпростори (простору навчальної вибірки), таким чином отримуються підвибірки із меншою розмірністю на яких навчаються класифікатори. Для своєї задачі я розподілюю простір ознак на три підпростори по кожному з каналів (синій, зелений, червоний).

Також крім методу простого голосування була використана логістична регресія для алгоритму Random forest з методом Випадкового підпростору. Таким чином логістична регресія навчалася на результатах трьох Decision Tree, що були

навченні на кожному з каналів даних. Це дає можливість за допомогою перевірки вагів логістичної регресії знайти найбільш інформативний канал даних.

Опис Logistics regression

Логістична регресія [28] – це статистичний алгоритм класифікації, що повертає ймовірність того, що введені дані належать тому чи іншому класу.

Нехай маємо бінарну класифікацію (тобто класи 0 – контрольна група 1 – пацієнти з раком молочної залози), та вхідні дані $\mathbf{x}$. Тоді ймовірність того, що $\mathbf{x}$ належить другому класу буде:

$$P(G = 1|X = \mathbf{x}) = \sigma(\beta_0 + \beta_1^T \mathbf{x})$$

де G – це відповідний клас, а функція сигмоїд має такий вигляд

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Тренування логістичної регресії виконується за допомогою мінімізації відповідної функції втрат $L(\beta)$ градієнтними методами.

$$L(\beta) = \sum_{i=1}^{N} \{y_i \log p(x_i;\ \beta) + (1 - y_i)p(x_i;\ \beta)\}$$

де N – кількість тренувальних даних, $\mathbf{x}$ – вхідні дані, y – мітки класів, $p(x_i;\ \beta) = P(G = 1|X = \mathbf{x}_i;\ \beta)$

5.2 Тестування Random Forest

Для проведення тестування я розділив датасет на дві частини навчальну вибірку (80%) та на тестову вибірку (20%) на якій буде оцінюватися отримана модель.

Було розглянуто декілька підходів з алгоритмом Random Forest.

На початку було використано метод Випадкового підпростору (Random subspace method) для створення трьох підпросторів навчальних даних по кожному з каналів. Після цього на даних кожного з каналів даних був тренований алгоритм Decision Tree. Для кожного з трьох дерев я проводив кроссвалідацію з розбиттям навчальних даних на п’ять частин та підбором параметрів максимальної глибини дерева та критерію розщеплення алгоритма (параметри Decision Tree вибиралися за метрикою accuracy). В результаті були отримані наступні результати для кожного з каналів даних.

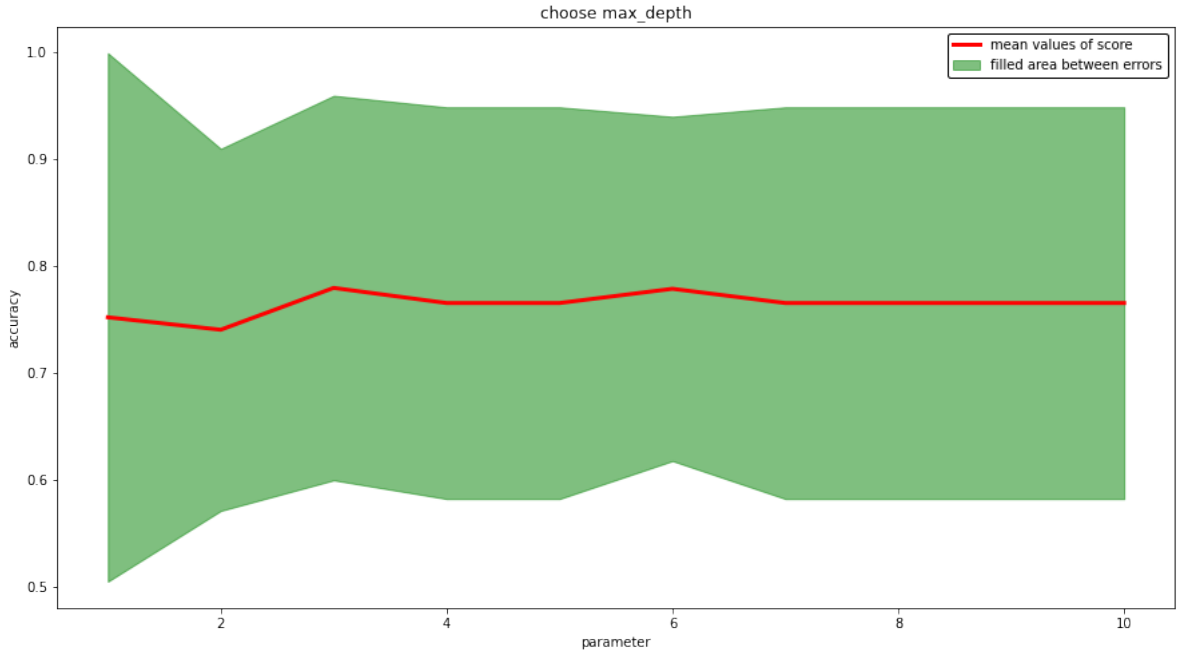

Рис. 5.1: Графік підбору оптимальної глибини для алгоритму Decision Tree з критерієм розщеплення Гіні, що тренується на **даних синього каналу**. На осі абсцис позначено значення параметру швидкості навчання, на осі ординат – значення метрики accuracy. Червоним позначено середнє значення метрики accuracy, зеленим – відхилення значення метрики від середнього.

- Decision Tree треноване на даних з синього каналу (рис. 5.1). Має глибину 3 та критерій розщеплення Індекс Джині. При тестуванні були отримані такі

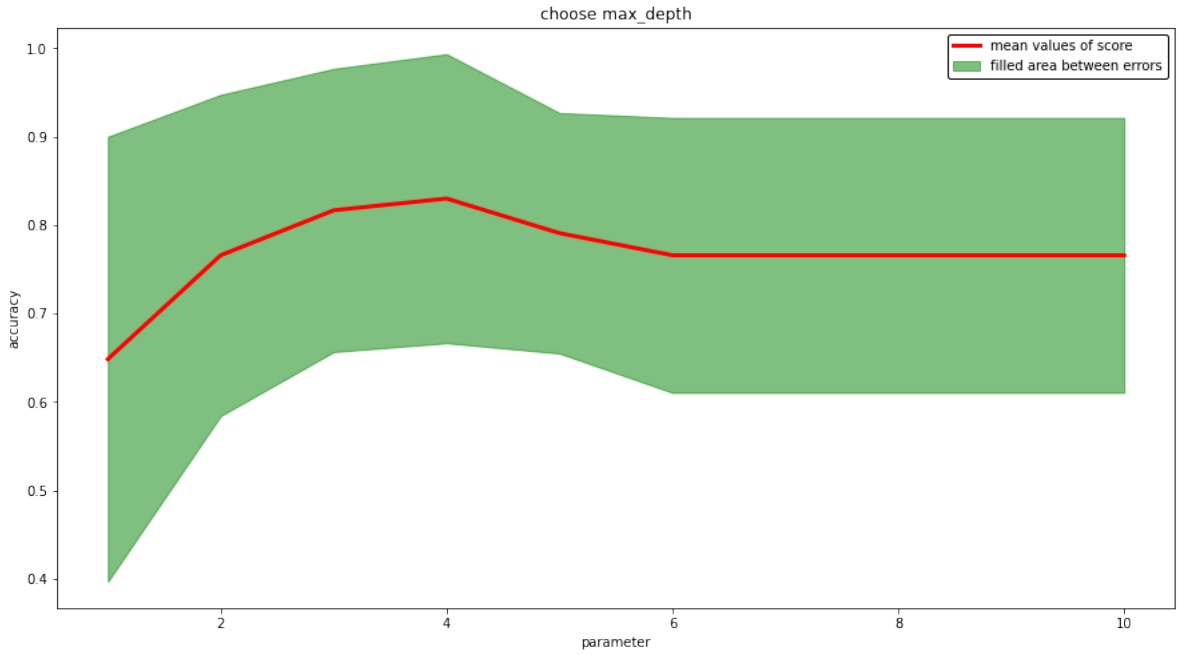

Рис. 5.2: Графік підбору оптимальної глибини для алгоритму Decision Tree з критерієм розщеплення Гіні, що тренується на **даних зеленого каналу**. На осі абсцис позначено значення параметру швидкості навчання, на осі ординат – значення метрики accuracy. Червоним позначено середнє значення метрики accuracy, зеленим – відхилення значення метрики від середнього.

значення метрик:

- accuracy $= 0.75$
- sensitivity $= 0.85$
- specificity $= 0.57$

- Decision Tree треноване на даних з зеленого каналу (рис. 5.2). Має глибину 4 та критерій розщеплення Індекс Джині. При тестуванні були отримані такі значення метрик:
  - accuracy $= 0.60$
  - sensitivity $= 0.69$
  - specificity $= 0.43$
- Decision Tree треноване на даних з червоного каналу (рис. 5.3). Має глибину 5 та критерій розщеплення крос-ентропія. При тестуванні були отримані такі значення метрик:
  - accuracy $= 0.50$
  - sensitivity $= 0.77$
  - specificity $= 0.0$

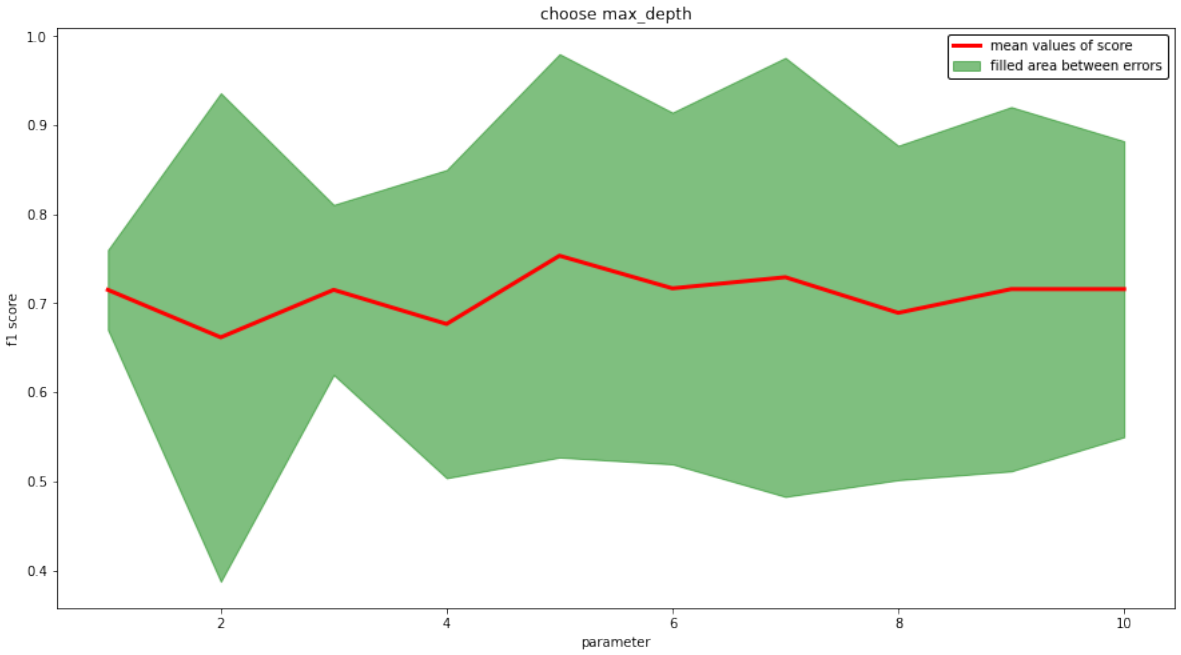

Рис. 5.3: Графік підбору оптимальної глибини для алгоритму Decision Tree з критерієм розщеплення крос-ентропії,що тренується на **даних червоного каналу**. На осі абсцис позначено значення параметру швидкості навчання, на осі ординат – значення метрики accuracy. Червоним позначено середнє значення метрики accuracy, зеленим – відхилення значення метрики від середнього.

Після цього для об'єднання отриманих алгоритмів Decision Tree був використаний метод простого голосування (усереднення відповіді). А саме

$$P(y = 1) = \sum_{i=1}^{3} T_i(\mathbf{x})$$

де $P(y = 1)$ – ймовірність того, що при вхідних даних $\mathbf{x}$ результатом Random Forest буде класс 1 (пацієнт хворий на рак молочної залози). T_i – це алгоритми Decision Tree, що видають ймовірність класу 1. При тестуванні такого варіанту алгоритму Random Forest були отримані наступні результати:

- accuracy $= 0.70$
- sensitivity $= 0.85$
- specificity $= 0.42$

Також на основі отриманих раніше дерев була застосована логістична регресія. Алгоритм логістичної регресії на вхід приймає ймовірності класу 1 (пацієнт хворий на рак молочної залози), що були отримані з кожного з трьох дерев рішень. Навчання логістичної регресії триває доки максимальна компонента функції втрат

більше за 10^{-4}. Така модифікація алгоритму Random Forest дала наступні результати

- accuracy $= 0.75$
- sensitivity $= 0.69$
- specificity $= 0.86$

Логістична регресія має наступні ваги $10.74, 8.9, 7.47$ для Decision Tree тренованих на даних з синього, зеленого, червоного каналів даних відповідно.

Крім того був протестований алгоритм Random Forest з технікою Bootstrap aggregation розподілення даних на 100 підвибірок для тренування відповідної кількості Decision Tree з критерієм розщеплення Гіні. Дерева розщеплюються доки в кожному листі не будуть навчальні приклади лише одного класу, або в листі буде не більше двох прикладів. Такий варіант Random forest дав наступні результати:

- accuracy $= 0.90$
- sensitivity $= 0.92$
- specificity $= 0.85$

Висновки Random Forest

Було протестовано алгоритм Random Forest та отримано гарні результати. Найкращий результат показав алгоритм Random forest з bootstrap aggregation та 100 деревами рішень: accuracy (90%), sensitivity (92%) та specificity (85%).

Ідея побудови Random forest на основі трьох Decision Tree, кожне з яких треновано на своєму (синьому, зеленому, червоному) каналі даних, не показала свою ефективність. Decision Tree побудоване на даних з синього каналу показує таку ж точність, що і Random Forest побудований на цих деревах з використанням логістичної регресії.

В той же час дерева побудовані на зеленому та червоному каналах даних показують дуже низькі результати. Беручи до уваги, що ваги логістичної регресії для дерева рішень, тренованого на даних синього каналу, найбільші, можна зробити висновок, що синій канал даних найбільш інформативний. Проте не варто викидати з розгляду інші канали даних, їх важливість підтверджує висока точність алгоритму Random Forest із 100 деревами, що навчався одразу на всіх каналах даних.

6 Висновки

У роботі було запропоновано методи машинного навчання для діагностики раку молочної залози на основі фрактальної розмірності ядер букального епітелію. Було отримано метод діагностики раку з точністю 95%, чутливістю 92% та специфічністю (100%)

Було перевірено значимість каналів даних (синій, зелений, червоний) за допомогою алгоритму Random Forest та логістичної регресії. Та показано, що синій канал даних має більшу значимість ніж два інших канала. Однак при використанні усіх каналів даних було отримано кращі результати.

Подальший розвиток роботи може полягати у розгляді зображень з жовтим та фіолетовим фільтрами (в даній роботі використовувалися лише зображення без фільтрів), що додасть 6 додаткових розмірностей до навчальних даних. Також відкритим є питання класифікації пацієнтів з фиброаденоматозом, що не були розглянуті в цій роботі.

Бібліографія

[1] Dmitriy Klyushin, Kateryna Golubeva1, Natalia Boroday, Dmytro Shervarly. Random Convolution of Feulgen-Stained Images and Breast Cancer Diagnosis. Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Ukraine, 03680, Kyiv.

[2] R. Yan et al., ”A Hybrid Convolutional and Recurrent Deep Neural Network for Breast Cancer Pathological Image Classification,”2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 957-962, doi: 10.1109/BIBM.2018.8621429.

[3] H. Yang, J. -Y. Kim, H. Kim and S. P. Adhikari, ”Guided Soft Attention Network for Classification of Breast Cancer Histopathology Images,”in IEEE Transactions on Medical Imaging, vol. 39, no. 5, pp. 1306-1315, May 2020, doi: 10.1109/TMI.2019.2948026.

[4] A. Patil, D. Tamboli, S. Meena, D. Anand and A. Sethi, ”Breast Cancer Histopathology Image Classification and Localization using Multiple Instance Learning,”2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 2019, pp. 1-4, doi: 10.1109/WIECON-ECE48653.2019.9019916.

[5] Punitha S., Fadi Al-Turjman, Thompson Stephan, An automated breast cancer diagnosis using feature selection and parameter optimization in ANN, Computers and Electrical Engineering, Volume 90, 2021, 106958, ISSN 0045-7906

[6] D. Yifan, L. Jialin and F. Boxi, ”Forecast Model of Breast Cancer Diagnosis Based on RF-AdaBoost,”2021 International Conference on Communications, Information System and Computer Engineering (CISCE), 2021, pp. 716-719, doi: 10.1109/CISCE52179.2021.9445847.

[7] Hu, Q., Whitney, H.M. and Giger, M.L. A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. Sci Rep 10, 10536 (2020). https://doi.org/10.1038/s41598-020-67441-4

[8] Khamparia, A., Bharati, S., Podder, P. et al. Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. Multidim Syst Sign Process 32, 747–765 (2021). https://doi.org/10.1007/s11045-020-00756-7

[9] Kavitha, T., Mathai, P.P., Karthikeyan, C. et al. Deep Learning Based Capsule Neural Network Model for Breast Cancer Diagnosis Using Mammogram Images. Interdiscip Sci Comput Life Sci 14, 113–129 (2022). https://doi.org/10.1007/s12539-021-00467-y

[10] Nieburgs H.F., Herman B.E., Reisman H. 1962. Buccal cell changes in patients with malignant tumors. Laboratory Investigation. 11, 1, pp..80-88.

[11] Lieberman-Aiden E. et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human Genome. Science. 326, 5959, pp. 289–193. doi: 10.1126/science.1181369.

[12] Gutierrez Aceves GA, Celis López MA, Alonso Vanegas M, et al. Fractal anatomy of the hippocampal formation. Surg Radiol Anat. 2018;40:1209–1215.

[13] Bohara G, Lambert D, West BJ, et al. Crucial events, randomness, and multifractality in heartbeats. Phys Rev E. 2017;96:062216.

[14] Hwang J, Oh Y-M, Lee M, et al. Low morphometric complexity of emphysematous lesions predicts survival in chronic obstructive pulmonary disease patients. Eur Radiol. 2019;29:176–185.

[15] Rabelo GD, Roux JP, Portero-Muzy N, et al. Cortical fractal analysis and collagen crosslinks content in femoral neck after osteoporotic fracture in postmenopausal women: comparison with osteoarthritis. Calcif Tissue Int. 2018;102:644–650.

[16] Zhu T, Ma J, Li J, et al. Multifractal and lacunarity analyses of microvascular morphology in eyes with diabetic retinopathy: a projection artifact resolved optical coherence tomography angiography study. Microcirculation. 2018;27:e12519.

[17] Shah RG, Girardi T, Ma X. Fractal dimensions and branching characteristics of placental chorionic surface arteries. Placenta. 2018;70:4–6.

[18] Xavier AISF, Cavalcanti MB, Silva EB, et al. Fractal analysis of chromatin as a potential indicator of human exposures to ionizing radiation. Sci Plena. 2018;14:020901.

[19] Jabalee J, Carraro A, Ng T, et al. Identification of malignancy-associated changes in histologically normal tumor-adjacent epithelium of patients with HPV-positive oropharyngeal cancer. Anal Cell Pathol (Amst). 2018;11:1607814.

[20] Andreichuk, A.V., Boroday, N.V., Golubeva, K.M., Klyushin, D.A. (2021). Artificial Intelligence System for Breast Cancer Screening Based on Malignancy-Associated Changes in Buccal Epithelium. In: Hassanien, AE., Taha, M.H.N., Khalifa, N.E.M. (eds) Enabling AI Applications in Data Science. Studies in Computational Intelligence, vol 911. Springer, Cham.

[21] Hassanluie TY, Rezaie MR and Rostami Z. Diagnosis of B-CLL leukemia using fractal dimension. J Kerman Univ Med Sci. 2017;243:229–236.

[22] Swinstead, E. E., Paakinaho, V., and Hager, G. L. (2018). Chromatin reprogramming in breast cancer, Endocrine-Related Cancer, 25(7), R385-R404. Retrieved May 17, 2022, from

[23] Bikou O et al. 2016. Fractal Dimension as a Diagnostic Tool of Complex Endometrial Hyperplasia and Well- differentiated Endometrioid Carcinoma. In Vivo. 30, pp. 681–690.

[24] Palcic B. 1994 Nuclear texture: can in be used as a suurrogate endpoind biomarker? Journal of Cellular Biochemistry. 19, 1, pp. 40–46

[25] Клюшин Д.А., Петунин Ю.И. Непараметрический критерий эквивалентности генеральных совокупностей, основанный на мере близости между выборками // Український математичний журнал – 2003 – с. 147-163.

[26] Nelson H.D., Fu R., Cantor A., Pappas M., Daeges M., Humphrey L. 2016. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. Preventive Services Task Force Recommendation. Annals of Internal Medicine. 164, 4, pp. 244-255. doi: 10.7326/M15-0969.

[27] Zhu, H. Zou, S. Rosset, T. Hastie, “Multi-class AdaBoost”, 2009.

[28] T. Hastie, R. Tibshirani and J. Friedman. “Elements of Statistical Learning”, Springer, 2009.

[29] Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.

[30] Ho TK (1998). ”The Random Subspace Method for Constructing Decision Forests”. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844. doi:10.1109/34.709601

[31] N. Otsu, ”A Threshold Selection Method from Gray-Level Histograms,”in IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.

[32] Sagan H. Space-filling curves. — Springer-Verlag: New York–Berlin, 1994. ISBN 978-0-387-94265-0.

[33] Butakov V., Grakovskiy A. 2005. Evaluation of arbitrary time series stochastic level by Hurst parameter. Computer Modelling and New Technologies. 9, 2, pp. 27–32.