**FМІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**Західноукраїнський національний університет**

**Факультет комп’ютерних інформаційних технологій**

Кафедра комп’ютерної інженерії

**Будний Володимир Сергійович**

**«Методи лінійної регресії для аналізу кількісних характеристик біомедичних зображень / Linear regression methods for analysis of quantitative characteristics of biomedical images»**

спеціальність: 123 - Комп’ютерна інженерія

освітньо-професійна програма - Комп’ютерна інженерія

Кваліфікаційна робота

Виконав студент групи КІм-21

В.С. Будний

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Науковий керівник:

к.т.н., Піцун О. Й.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Кваліфікаційну роботу допущено

до захисту:

"\_\_\_" \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_\_ р.

Завідувач кафедри

\_\_\_\_\_\_\_\_\_\_\_\_\_\_О. М. Березький

**Тернопіль – 2021**

ЗМІСТ

[Вступ 7](#_Toc89967451)

[1 Аналітичний огляд методів, алгоритмів лінійної регресії для аналізу зображень 10](#_Toc89967452)

[1.1 Класифікація біомедичних зображень 10](#_Toc89967453)

[1.2 Аналіз методів та алгоритмів опрацювання зображень 16](#_Toc89967454)

[1.3 Аналіз програмних засобів опрацювання біомедичних зображень 27](#_Toc89967455)

[1.4 Постановка завдання до виконання кваліфікаційної роботи 34](#_Toc89967456)

[1.5 Висновки до розділу 1 34](#_Toc89967457)

[2 Методи та алгоритми аналізу цитологічних та гістологічних зображень на основі лінійної регресії 35](#_Toc89967458)

[2.1 Узагальнений алгоритм опрацювання біомедичних зображень 35](#_Toc89967459)

[2.2 Алгоритми виділення кількісних характеристик ядер клітин 40](#_Toc89967460)

[2.3 Класична модель лінійної регресії 47](#_Toc89967461)

[2.4 Алгоритми одномірної та багатомірної регресії 52](#_Toc89967462)

[2.5 Висновки до розділу 2 54](#_Toc89967463)

[3 Структура системи опрацювання кількісних характеристик біомедичних зображень 55](#_Toc89967464)

[3.1 Навчальна вибірка для опрацювання даних 55](#_Toc89967465)

[3.2 Порівняння алгоритмів лінійної регресії 59](#_Toc89967466)

[3.3 Реляційна база даних системи 64](#_Toc89967467)

[3.4 Тестування результатів роботи 70](#_Toc89967468)

[3.5 Висновки до розділу 3 77](#_Toc89967469)

[Висновки 78](#_Toc89967470)

[Список використаних джерел 79](#_Toc89967471)

# Вступ

Актуальність роботи. Обробка медичних зображень включає використання та дослідження наборів даних 3D-зображень людського тіла, які найчастіше отримують за допомогою сканера комп’ютерної томографії (КТ) або магнітно-резонансної томографії (МРТ) для діагностики патологій або проведення медичних втручань, таких як планування хірургічного втручання або дослідження. цілі. Обробку медичних зображень здійснюють радіологи, інженери та клініцисти, щоб краще зрозуміти анатомію окремих пацієнтів або груп населення.

Основна перевага обробки медичних зображень полягає в тому, що вона дозволяє глибоко, але неінвазивно досліджувати внутрішню анатомію. Можна створювати та вивчати 3D-моделі анатомічних структур, що цікавлять, для покращення результатів лікування пацієнта, розробки вдосконалених медичних пристроїв і систем доставки ліків або встановлення більш обґрунтованих діагнозів. Це стало одним з ключових інструментів, які використовувалися для розвитку медицини в останні роки.

Постійно покращувана якість зображень у поєднанні з передовими програмними засобами полегшує точне цифрове відтворення анатомічних структур у різних масштабах, а також із значно різними властивостями, включаючи кістки та м’які тканини. Вимірювання, статистичний аналіз і створення імітаційних моделей, які включають реальну анатомічну геометрію, дають можливість для більш повного розуміння, наприклад, взаємодії між анатомією пацієнта та медичними приладами.

Процес обробки медичних зображень починається з отримання необроблених даних із зображень КТ або МРТ та їх реконструкції у формат, придатний для використання у відповідному програмному забезпеченні. Тривимірне растрове зображення інтенсивності сірого, що містить сітку вокселів (3D-пікселів), створює типовий вхід для обробки зображень. Інтенсивність відтінків сірого КТ залежить від поглинання рентгенівського випромінювання, тоді як у МРТ вона визначається силою сигналів від протонних частинок під час релаксації та після застосування дуже сильних магнітних полів. Отож, очефидним є той факт, що використання у сучасних умовах алгоритмів цифрової обробки зображень в системах медичної діагностики є чи не обов’язковим фактором. А тому постає завдання по розробці алгоритмів опрацювання зобаржень та їх подальшого аналізу з метою встановлення залежностей між окремими характеристиками. Тому задача реалізації алгоритму аналізу взаємозв’язку між двома характеристиками є актуальною.

Метою роботи є розробка алгоритму аналізу кількісних характеристик об’єктів зображення на основі лінійної регресій.

Для досягнення даної мети ставились наступні завдання:

* провести класифікацію існуючих типів біомедичних зображень;
* проаналізувати методи та алгоритми опрацювання зображень;
* провести аналіз програмних засобів опрацювання біомедичних зображень;
* провести аналіз та порівняння алгоритмів одномірної та багатомірної регресії;
* сформувати навчальну вибірку для навчання згорткової нейронної мережі;
* розробити реляційну базу даних системи опрацювання зображень та провести її тестування.

Об’єкт дослідження – процес обробки цифрових зображень.

Предмет дослідження – методи і алгоритми лінійної регресії для оцінки кількісних характеристик об’єктів.

Наукова новизна одержаних результатів визначається наступним чином:

* проведено аналітичний огляд та класифікацію методів та алгоритмів опрацювання зображень та досліджено їх вплив на кінцевий результат отриманих результатів;
* розроблено алгоритм оцінки взаємозалежності двох параметрів об’єктів на основі лінійної регресії, що дозволило реалізувати програмне зебезпечення аналізу та оцінки біомедичних зображень.

Практична цінність одержаних результатів полягає в тому, що:

* розроблено та проведено експерементальне дослідження існуючих алгоритмів опрацювання зображень, а семе алгоритмів сегментації та виділення об’єктів в результаті чого було запропоновано узагальнений алгоритм опрацювання зображень;
* реалізовано базу даних та програмний додаток опрацювання зображень, проведено їх тестування та порівняння з аналогами, що дозволило провести оцінку взаємозалежності двох параметрів об’єкту.

Публікації та апробація до випускної кваліфікаційної роботи. За результатами наукових досліджень, проведених у випускній кваліфікаційні роботі, підготовлено тези доповіді «Розробка методу опрацювання цитологічних зображень» обсягом 1 сторінка на V Науково-практичній конференції молодих вчених і студентів «Інтелектуальні комп’ютерні системи та мережі», а також «Аналіз зображень» обсягом 1 сторінка на V Науково-практичній конференції молодих вчених і студентів «Інтелектуальні комп’ютерні системи та мережі».

# 1 Аналітичний огляд методів, алгоритмів лінійної регресії для аналізу зображень

## 1.1 Класифікація біомедичних зображень

Медична візуалізація – це техніка та процес візуалізації внутрішньої частини тіла для клінічного аналізу та медичного втручання, а також візуального зображення функцій деяких органів або тканин (фізіологія). Медичне зображення спрямоване на виявлення внутрішніх структур, прихованих шкірою та кістками, а також на діагностику та лікування захворювання. Медична візуалізація також створює базу даних нормальної анатомії та фізіології, що дозволяє виявляти відхилення. Хоча візуалізація видалених органів і тканин може проводитися за медичними показаннями, такі процедури зазвичай вважаються частиною патології, а не медичною візуалізацією.

Як дисципліна і в її найширшому сенсі, вона є частиною біологічної візуалізації та включає радіологію, яка використовує технології візуалізації рентгенівської рентгенографії, магнітно-резонансної томографії, ультразвуку, ендоскопії, еластографії, тактильної візуалізації, термографії, медичної фотографії, ядерної медицини функціональні методи візуалізації, як позитронно-емісійна томографія та однофотонна емісійна комп’ютерна томографія.

Методи вимірювання та запису, які в першу чергу не призначені для отримання зображень, такі як електроенцефалографія (ЕЕГ), магнітоенцефалографія (МЕГ), електрокардіографія (ЕКГ) та інші, представляють інші технології, які надають дані, прийнятні до представлення, як графік параметрів у порівнянні з часом або карти, що містять дані про місця вимірювань. У обмеженому порівнянні ці технології можна розглядати як форми медичної візуалізації в іншій дисципліні.

Станом на 2010 рік у всьому світі було проведено 5 мільярдів медичних досліджень. Радіаційне опромінення від медичних зображень у 2006 р. Становило близько 50% загального опромінення іонізуючого випромінювання в США. Медичне обладнання для візуалізації виробляється з використанням технологій напівпровідникової промисловості, включаючи мікросхеми CMOS, силові напівпровідникові пристрої, такі датчики, датчики зображення (зокрема CMOS-датчики) та біосенсори, а також такі процесори, як мікроконтролери, мікропроцесори, цифрові сигнали, процесори медіа та пристрої на чіпі. Станом на 2015 рік щорічні поставки чіпів медичної візуалізації становлять 46 мільйонів одиниць та 1,1 мільярда доларів.

Медична візуалізація часто сприймається як позначення набору методів, які неінвазивно створюють зображення внутрішнього аспекту тіла. У цьому обмеженому розумінні медичну візуалізацію можна розглядати як вирішення математичних обернених задач. Це означає, що причина (властивості живої тканини) виводиться з наслідку (спостережуваний сигнал). У разі медичного ультразвуку зонд складається з ультразвукових хвиль та їх відлуння, що виникає при проходженні через різні тканини, що дозволяє показати їх внутрішню структуру. У разі проекційної рентгенографії зонд використовує рентгенівське випромінювання, яке з різною швидкістю поглинається різними типами тканин, такими як кістка, м’язи та жир.

Термін "неінвазивний" використовується для позначення процедури, при якій жоден інструмент не вводиться в тіло пацієнта, що має місце для більшості використовуваних методів візуалізації.

Зображення медичного типу не завжди можуть забезпечувати необхідну інформацію яка потрібна для однозначної ідентифікації захворювань. Поділ медичних зображень за критеріями є умовним.

Медичні зображення можна класифікувати по таким критеріям:

1. за способом отримання зображення:

а) цифрова радіологія;

б) комп’ютерна томографія;

в) ядерний магнітний резонанс;

г) ультразвук;

д) еластографія;

e) мікроскопія;

1. за типом зображення:

а) RGB зображення (мікроскопія, еластографія, термографія);

б) зображення в градаціях сірого (рентгенографія, томографія);

в) чорно–біле (ультразвукове дослідження);

1. за розмірністю:

а) 2D (всі зображення незалежно від способу отримання);

б) 3D (послідовність радіологічних зображень, томографічне зобра­ження динамічного об'єкта).

Рентгенографічні зображення – це 2D зображення в форматі градаціях сірого, що способом отримання відносяться до класу цифрової радіології (рисунок 1.1 (а)). Якість зображення буде залежати від роздільної здатності та інтенсивності. Роздільна здатність – це здатність зображення відображати близько розташовані структури в об’єкті як окремі об’єкти на зображенні, тоді як інтенсивність – це степінь затемнення зображення. Різкість рентгенографічного зображення сильно залежить від джерела рентгенівського випромінювання. Велике джерело фотонів призводить до більшої розмитості у кінцевому зображенні та погіршується зі збільшенням відстані до формування зображення. Це розмиття можна виміряти як внесок у передавальну функцію модуляції системи формування зображення.

Прилад для магнітно–резонансної томографії (МРТ-сканер) використовує потужні магніти для поляризації та збудження ядер водню (тобто одиничних протонів) молекул води в тканинах людини, продукуючи розпізнаваний сигнал, що при просторовому декодуванні формує 2D або 3D зображення тіла. Перевагою використання 2D МРТ зображень (рисунок 1.1 (б)) є вища швидкість формування зображень у порівнянні з 3D. Також 2D зображення краще підходять для дослідження довгих судин. Перевагою використання 3D зображень є висока просторова роздільна здатність у порівняні з 2D та відображення об’єкту у тривимірному просторі.

Позитрон-емісійна томографія (ПЕТ) (рисунок 1.1 (в)) – метод, заснований на застосуванні радіофармпрепаратів, мічених ізотопами, що випромінюють позитрони.

Медичний ультразвук використовує високочастотні широкосмугові звукові хвилі, які відбиваються тканинами в різному ступені для отримання 2D або 3D зображень. Ультразвук також використовується як популярний інструмент дослідження для збору необроблених даних, які можуть бути доступні за допомогою ультразвукового дослідницького інтерфейсу, з метою встановлення характеристик тканин та впровадження нових методів обробки зображень. 2D зображення – це двовимірне чорно-біле зображення (рисунок 1.1 (г)). Результат дослідження з використанням 3D зображень є об'ємне, тривимірне RGB зображення.

Еластографія – це відносно новий метод візуалізації, який відображає еластичні властивості м’яких тканин. Еластографія корисна в медичних діагнозах, оскільки еластичність може відрізнити здорову від нездорової тканини певних органів/новоутворень. Наприклад, ракові пухлини часто бувають твердішими за навколишні тканини, а хворі печінки жорсткіші за здорові. Існує кілька еластографічних методів, заснованих на використанні ультразвуку, магнітно-резонансної томографії та тактильної візуалізації. Широке клінічне застосування ультразвукової еластографії є результатом впровадження технології в клінічних апаратах ультразвуку.

Ехокардіографія - метод ультразвукової діагностики, що спрямований на дослідження морфологічних і функціональних змін клапанного апарату серця (рисунок 1.1 (г)). Ехокардіографія використовує 2D, 3D та доплерографію для створення картини стану серця та візуалізації потоку крові, яка проходить через кожен із чотирьох його клапанів. Трансторакальне ультразвукове дослідження виявилося безпечним для пацієнтів будь-якого віку, від немовлят до літніх людей, без ризику шкідливих побічних ефектів або опромінення, що відрізняє його від інших методів візуалізації. Ехокардіографія є однією з найпоширеніших у світі методів візуалізації завдяки своїй портативності та можливості використання у багатоьох сферах діяльності. Тактильна візуалізація (ТВ) (рисунок 1.1 (е)) є медичним методом візуалізації, який переводить дотик в цифрове зображення.

Імуногістохімія включає процес вибіркової ідентифікації антигенів (білків) у клітинах тканинного зрізу шляхом використання принципу антитіл, що зв'язуються конкретно з антигенами в біологічних тканинах. Візуалізація взаємодії антитіло-антиген (рисунок 1.1 (є)) може бути здійснена декількома способами, найпоширенішими з яких є хромогенна імуногістохімія, де антитіло кон'юговане з ферментом, таким як пероксидаза, та імунофлуоресценція, де антитіло помічається на флуорофор, наприклад флуоресцеїн або родамин. Імуногістохімічне фарбування широко використовується для діагностики аномальних клітин, таких як ті, що знаходяться в ракових пухлинах. Специфічні молекулярні маркери характерні для окремих клітинних подій, таких як проліферація або загибель клітин (апоптоз). Імуногістохімія також широко використовується в фундаментальних дослідженнях для розуміння розподілу та локалізації біомаркерів та диференційно виражених білків у різних частинах біологічної тканини.

Гістологія – це вивчення мікроанатомії клітин, тканин і органів, які можна побачити в мікроскоп. Гістологія, також відома як мікроскопічна анатомія або мікроанатомія. В клінічній медицині гістопатологія відноситься до дослідження біопсії або хірургічного зразка патологоанатомом після того, як зразок був оброблений і гістологічні зрізи були поміщені на предметне скло. Для візуалізації різних компонентів тканини під мікроскопом зрізи фарбують одним або кількома барвниками. Метою фарбування є виявлення клітинних компонентів; для забезпечення контрасту використовуються контрбарвники.

Цитологія, з іншого боку, пов’язана з вивченням клітин з точки зору структури, функції та хімічного складу і реакцій. У результаті найменш інвазивних біопсій цитологічні зображення найчастіше використовуються як для скринінгу захворювання, так і для визначення цілі біопсії. Крім того, характеристики цитологічних зображень, а саме наявність ізольованих клітин і кластерів клітин на зображеннях, а також відсутність більш складних структур, таких як залози, полегшують аналіз цих зразків порівняно з гістопатологією.

|  |  |  |
| --- | --- | --- |
| а) рентгенографічнезображення | б) МРТ-зображення | в) ПЕТ-зображення |
| г) УЗД-зображення | д) ехокардіографічне зображення | е) тактильна візуалізація |
| є) імуногістохімічне зображення | ж) гістологічне зображення | з) цитологічне зображення |

Рисунок 1.1 – Приклади медичних зображень

Дані зображення найчастіше використовуються в процесі автоматичної обробки біомедичних зображень оскільки саме вони в значній мірі використовуються для швидкого діагностування.

## 1.2 Аналіз методів та алгоритмів опрацювання зображень

Комп'ютерна обробка зображень передбачає обробку цифрових зображень за допомогою комп'ютерів або спеціалізованих пристроїв, побудованих на цифрових сигнальних процесорах. При цьому під обпрацюванням зображень розуміється не лише покращення зорового сприйняття зображень, а й класифікація об'єктів, що виконується під час аналізу зображень. Область застосування цифрової обробки в даний час значно розширилась, витіснивши аналогові методи обробки сигналів зображення. Методи обробки цифрових зображень широко застосовуються у промисловості, мистецтві та медицині. Вони застосовуються при управлінні процесами, автоматизації виявлення та супроводу об'єктів, розпізнавання образів та у багатьох інших задачах.

До технологій опрацювання зображень належить: попередня обробка зображень, сегментація, контурний та текстурний аналіз, класифікація і розпізнавання зображень.

Дослідники як в області аналізу зображень, так і в області патології відзначають важливість кількісного аналізу зображень. Оскільки більшість сучасної діагностики базується на суб’єктивній (але освіченій) думці спеціалістів, очевидна потреба в кількісній оцінці цифрових слайдів на основі зображень. Цей кількісний аналіз цифрової патології важливий не тільки з діагностичної точки зору, але й для того, щоб зрозуміти основні причини постановки конкретного діагнозу (наприклад, специфічна структура хроматину в ракових ядрах, яка може вказувати на певні генетичні аномалії).

Обробку зображень розділяють на рівні представлені на рисуноку 2.1

Таблиця 1.1 – Рівні опрацювання зображень

|  |  |  |
| --- | --- | --- |
| Низький рівень | Середній рівень | Високий рівень |
| * Просторові методи покращення зображення
* Частотні методи покращення зображення
* Геометричні перетворення
* Вейвлет–перетворення
* Морфологічні перетворення
 | * Сегментація
* Контурний аналіз
* Текстурний аналіз
 | * Алгебраїчний метод
* Статистичний метод
* Структурний метод
* Штучні нейронні мережі
 |

Крім того, кількісна характеристика зображень патології важлива не тільки для клінічних застосувань (наприклад, щоб зменшити/усунути розбіжності в діагностиці між інтра- та інтер-діагностами), але й для дослідницьких роботах (наприклад, щоб зрозуміти біологічні механізми процесу захворювання). Так, просторовий аналіз гістопатологічних зображень став основою більшості автоматизованих методів аналізу медичних зображень. Незважаючи на прогрес, досягнутий у цій області, він все ще залишається сферою відкритих досліджень через різноманітність методів візуалізації та специфічних характеристик захворювання.

Попередня обробка зображень – це термін для операцій над зображеннями на найнижчому рівні абстракції. Ці операції збільшують інформативність зображення, але дозволяють зменшити його ентропію. Метою попередньої обробки є вдосконалення даних зображення, яке пригнічує небажані спотворення або покращує деякі функції зображення, що мають значення для подальшої обробки та аналізу.

Існує 4 різні типи методів попередньої обробки зображень:

а) перетворення яскравості пікселів/ Корекції яскравості;

б) геометричні перетворення;

в) фільтрація та сегментація зображень;

г) перетворення Фур'є та відновлення зображення.

Перетворення яскравості змінюють яскравість пікселя, і перетворення залежить від властивостей самого пікселя. Приклади таких операторів включають коригування яскравості та контрастності, а також корекцію кольору та перетворення. Підвищення контрастності є важливою областю обробки зображень як для людського, так і для комп’ютерного зору. Воно широко використовується для обробки медичних зображень.

Найпоширенішими операціями перетворення яскравості пікселів є:

а) гамма-корекція або перетворення степеневого закону;

б) сигмоїдальне перетворення;

в) вирівнювання гістограми.

Двома часто використовуваними точковими процесами є множення та додавання з константою.

g(x)=αf(x)+β.

Параметри α>0 і β називають параметрами посилення і зміщення, а іноді кажуть, що ці параметри контролюють контраст і яскравість відповідно.

Гамма-корекція – це нелінійне коригування значень окремих пікселів. У той час як у нормалізації зображення виконуються лінійні операції над окремими пікселями, такі як скалярне множення та додавання/віднімання, гамма-корекція виконує нелінійну операцію над вихідними пікселями зображення і може викликати сатурацію змінюваного зображення. Приклад гама-корекції наведено на рисуноку 1.2.

$$O=\left(\frac{I}{255}\right)^{γ}×255$$

Тут відношення між вихідним зображенням і гамою є нелінійним.



Рисунок 1.2 – Гамма-корекція

Сигмовидна функція є безперервною нелінійною функцією. Статистики називають цю функцію логістичною функцією.

$f\left(x\right)=\frac{1}{1+e^{-tx}}$*,*

$g\left(x,y\right)=\frac{1}{1-e^{\left(c×\left(th-fs\left(x,y\right)\right)\right)}}$*,*

де *g(x,y)* – розширене значення пікселя;

 *c* – коефіцієнт контрасту;

 *th* – порогове значення;

 *fs(x,y)* – вихідне зображення.

Налаштувавши коефіцієнт контрасту *c* і порогове значення, можна налаштувати кількість освітлення та затемнення, щоб контролювати загальне підвищення контрастності.

Гістограмне вирівнювання є добре відомим методом посилення контрасту завдяки його ефективності майже на всіх типах зображень. Вирівнювання гістограми надає методи для зміни динамічного діапазону та контрасту зображення шляхом зміни цього зображення так, щоб його гістограма інтенсивності мала потрібну форму. Оператори моделювання гістограм можуть використовувати нелінійні та немонотонні функції передачі для відображення між значеннями інтенсивності пікселів у вхідних і вихідних зображеннях.

Метою використання фільтрів є зміна або покращення властивостей зображення та/або отримання корисної інформації з зображень, наприклад країв, кутів і крапель. Фільтр визначається ядром, який являє собою невеликий масив, що застосовують до кожного пікселя та його сусідів у зображенні.

Деякі з основних методів фільтрації:

а) Низькочастотна фільтрація (згладжування): фільтр низьких частот є основою для більшості методів згладжування. Зображення згладжується шляхом зменшення розбіжності між значеннями пікселів шляхом усереднення сусідніх пікселів

б) Високочастотна фільтрація (виявлення країв, різкість): фільтр високих частот можна використовувати, щоб зробити зображення чіткішим. Ці фільтри підкреслюють дрібні деталі зображення – у протилежність фільтру низьких частот. Фільтрація високих частот працює так само, як і фільтрація низьких частот; він просто використовує інше ядро згортки.

в) Направлена фільтрація: направлений фільтр – це детектор країв, який можна використовувати для обчислення перших похідних зображення. Перші похідні (або нахили) найбільш очевидні, коли між сусідніми значеннями пікселів відбувається велика зміна. Направлені фільтри можуть бути розроблені для будь-якого напрямку в межах даного простору.

г) Фільтрація Лапласа: фільтр Лапласа – це детектор країв, який використовується для обчислення других похідних зображення, вимірюючи швидкість зміни перших похідних. Це визначає, чи відбувається зміна значень сусідніх пікселів від краю чи безперервної прогресії. Ядра фільтра Лапласа зазвичай містять негативні значення в перехресному шаблоні, по центру масива. Кути мають або нульові, або додатні значення. Центральне значення може бути як негативним, так і додатним.

Сегментація зображень – це техніка обробки та аналізу цифрових зображень для поділу зображення на кілька частин або областей, часто на основі характеристик пікселів у зображенні. Сегментація зображення може включати відокремлення переднього плану від фону або групування областей пікселів на основі подібності кольору або форми.

Наразі немає універсально прийнятного методу для сегментації зображень, і тому вибір методу сегментації залишається проблемою в обробці зображень і комп'ютерного зору.

На основі різних технологій підходи до сегментації зображень можна розділити на наступні категорії на основі властивостей зображення.

Виявлення розривів – поділ зображення на основі різких змін інтенсивності, сюди входять алгоритми сегментації зображень, як-от виявлення країв.

Виявлення подібності – поділ зображення на подібні області за набором заздалегідь визначених критеріїв, сюди входять алгоритми сегментації, такі як пошук порогових значень, нарощування регіонів, розщеплення та злиття регіонів.

а) Виділення країв – метод, що використовує перепад яскравості на границях областей для виокремлення окремих областей. Цей метод намагається вирішити сегментацію зображення шляхом виявлення краю або пікселі між різними областями, які мають швидкій перехід за інтенсивністю витягуються і зв'язуються у замкнену форму межі об’єкта. Результатом є бінарне зображення. Існує два основні методи сегментації на основі країв –метод на основі гістограми та градієнтний. Результат техніки виявлення країв на основі сірої гістограми залежить головним чином від вибору порогового значення *T*, і дуже важко відшукати максимальну і мінімальну інтенсивність сірого, оскільки сіра гістограма нерівномірна для впливу шуму, таким чином ми приблизно співставляємо криві об'єкта і фон з двома конічними кривими Гаусса перетин яких є долиною гістограми. Поріг *Т* є сіре значення точки перетину цієї долини. Градієнт є першою похідною для зображення *f*(*x*, *y*), якщо є різка зміна інтенсивності біля краю і невелике зображення з мінімальними шумом, градієнтний метод показує хороші результати.

в) Сегментація зображень за допомогою порогового значення є простим, але продуктивним підходом до сегментації зображень зі світлими об’єктами на темному фоні. Техніка визначення порогу заснована на областях простору зображень, тобто на характеристиках зображення. В результаті проходить перетворення багаторівневого зображення в двійкове зображення, тобто при виборі порогу T зображення розділяється на регіони та відокремлюються об’єкти від фону. Будь-який піксель (x, y) вважається частиною об'єкта, якщо його інтенсивність більше або дорівнює пороговому значенню, тобто f(x,y) ≥T, інакше піксель належить до фону.

Порівняно з методом виявлення країв, алгоритмами сегментації на основі областей відносно прості та більш стійкі до шумів. Методи на основі границь розбивають зображення швидкої зміни інтенсивності поблизу країв, тоді як методи на основі регіонів розділяють зображення на області, подібні відповідно до набору із заздалегідь визначених критеріїв. Алгоритми сегментації на основі регіону в основному включають методи нарощування областей – це така процедура, що групує пікселі в зображенні на підобласті або більші регіони на основі попередньо визначених критеріїв, та метод поділу і злиття регіонів –користувач може розділити зображення на набір довільних незв'язаних областей, а потім об'єднати регіони у спробі задовольнити умови.

д) Методи сегментації на основі PDE (частково диференціальне рівняння) Використавши методи на основі PDE та розв’язавши рівняння за числовою схемою можна сегментувати зображення. Сегментація зображення на основі PDE в основному здійснюється активною контурною моделлю або “змійкою”.

Активні контури або “змійки” – це згенеровані комп'ютером криві які рухаються всередині зображення, щоб знайти межі об’єкта під впливом внутрішніх і зовнішніх сил. Обчислювальна складність алгоритму досить висока. Щоб вирішити цю проблему, ряд дослідників проводять різні вдосконалення базової моделі, але недолік “змія” досі не подолано принципово.

Рівневі моделі, які використовуються в обробці зображень, засновані на переміщенні кривих та поверхонь зі швидкостями на основі кривизни. У цій області. Основна ідея – представляти криві або поверхні як набір нульового рівня більш високого розміру як гіперповерхню. Ця техніка не тільки забезпечує більш точні чисельні реалізації, але також дозволяє якісно обробляти топологічні зміни. Метод має ряд переваг; його стабільність і неактуальність топології, демонструє хороші результати для вирішення проблем створення кутових точок, розрив кривих і комбінування тощо.

Основна ідея моделі C-V полягає в пошуку певного розділу даного зображення на дві області, одна з яких представляє об'єкти, які потрібно виявити та фон. Модель C-V не заснована на функції визначеня країв, щоб зупинити криву, що розвивається на бажаній межі (немає потреби у згладжуванні початкового зображення, навіть якщо воно шумне), при цьому розташування кордонів дуже добре виявляються. Метод може виявляти об'єкти, межі яких є не обов'язково визначаються градієнтом або дуже гладкими межами. Починаючи лише з однієї початкової кривої, ця модель може автоматично виявляти внутрішні контури, при чому не обов'язково необхідно стартувати близько до об'єктів, які підлягають виявленню.

е) Під кластеризацією розуміють розбиття багатовимірної неструктурованої множини даних на ряд підмножин даних зі схожими характеристиками. Кластеризація не використовує етапи навчання, а тренується самостійно використовуючи наявні дані. Кластеризація в основному використовується, коли є наперед визначені клас. Між пікселями визначається критерій подібності, а потім подібні пікселі групуються разом, утворюючи кластери. Групування пікселів у кластери здійснюється за принципом максимізації внутрішньокласової подібності та максимізації міжкласової подібності. Якість результату кластеризації залежить від мір подібності, що використовуються методом, та його реалізації. Виділяють такі алгоритми кластеризації як жорстка кластеризація, k-кластеризація, нечітка кластеризація тощо.

Жорстка кластеризація передбачає чіткі межі між кластерам, а піксель належить одному і лише одному кластеру. Популярний і відомий жорсткий алгоритм кластеризації – це алгоритм кластеризації K-середніх. Алгоритм K-середніх є методом кластеризації для розділення n пікселів на k кластерів, де k < n. Основні переваги цього алгоритму є його простота і низька обчислювальна вартість, які дозволяють це зробити ефективно на досить великих об’ємах даних. Головний недолік полягає в тому, що кількість кластерів необхідно визначити завчасно, це не дає той самий результат кожного разу, коли виконується алгоритм і отримані кластери залежать від початкових визначень центроїдів.

Техніка нечіткої кластеризації класифікує значення пікселів з великою мірою точності і в основному підходить для додатків, орієнтованих на прийняття рішень як класифікація тканин і виявлення пухлин тощо. Нечітка кластеризація розділяє вхідні пікселі на кластери або групи на основі деякого критерію подібності, при якому подібні пікселі належать одному кластеру. Використовуваним критерієм подібності може бути відстань, зв’язок, інтенсивність. Отриманий розділ покращує розуміння людей і допомагає приймати більш обґрунтовані рішення. Перевага нечіткої системи полягає в тому, що їх легко зрозуміти, оскільки функція членства правильно розподіляє простір даних. Алгоритми нечіткої кластеризації включають FCM алгоритм, GK (Gustafson-Kessel), GMD, FCV (різноманітні нечіткі C), AFC, FCS, FCSS, FCQS, FCRS алгоритм тощо.

є) Сегментація на основі нейронної мережі повністю відрізняється від звичайні алгоритми сегментації. У цьому випадку зображення в першу чергу відображається як нейронну мережа. Кожен нейрон інтерпретується як піксель, таким чином задача сегментації зображення перетворюється в задачу мінімізації енергії. Нейронна мережа була навчена з навчальною вибіркою набору для визначення зв'язку і ваги між вузлами. Потім нові зображення можна сегментувати за допомогою навченої нейронної мережі, наприклад, ми зможемо витягти зображення ребер за допомогою динамічних рівнянь, які спрямовують стан кожного нейрона до мінімальної енергії, визначеної нейронною мережею.

Сегментація нейронної мережі включає два важливі кроки – початкове виділення та сегментація зображень на основі нейронної мережі. Виділення ознак є дуже важливим, оскільки воно визначає вхідні дані нейронної мережі, спочатку деякі функції витягуються з зображення, щоб вони стали придатними для сегментації, а потім щоб вони були входом нейронної мережі. Усі вибрані особливості складаються з дуже нелінійного простору ознак кластера межа.

Сегментація на основі нейронної мережі має три основні характеристики:

* висока паралелізація та швидкі обчислювальні можливості, що робить його придатним для застосування в режимі реального часу;
* можливість покращити результати сегментації, коли дані відхиляються від нормалізованої ситуації;
* нечутливість до шуму;
* зменшена потреба у фаховому втручанні під час процес сегментації;

Однак у нейронних мереж є і недоліки сегментації, наприклад:

* повинна бути відома інформація про сегментацію заздалегідь;
* ініціалізація може вплинути на результат зображення сегментації;
* нейронну мережу потрібно навчати за допомогою процесу навчання;
* заздалегідь період навчання може бути дуже тривалим, і при цьому слід уникати перетренування.

ж) Багатооб’єктна сегментація зображення покликана отримати інформацію, включену в зображення. Тому розраховуються різні результати сегментації. Таким чином, сегментація зображень є багатоцільовою задачею оптимізації. Розгляд кількох критеріїв (цілей) починається від розуміння візерунка зображення до його вибору задіяний процес сегментації (вибір/вилучення ознак, міра схожості/неподібності) і, нарешті, оцінка виходу (оцінка обґрунтованості). Раніше проблему сегментації зображень розглядали як монооб’єктивну. Монооб'єктивні зображення розглядають лише один об’єкт на одне зображення сегментації. Такі типи сегментації зображень мають хорошу якість, але не дозволяють проводити процеси високого рівня (сегментація зображення розглядається як процес низького рівня) – розпізнавання образів, відстеження об'єктів і аналіз сцени на високому рівні. Вибір ознак – це процес визначення критеріїв подібності використовується в процесі сегментації, тепер або лише один критерій використовується, тобто інтенсивність пікселів, або зробити його багатоцільовим

У задачі необхідно розглянути кілька критеріїв подібності, щоб сегментувати однаково зображення, яке може бути інтенсивним, кольоровим, текстурним, формовим, просторовим. Наприклад, при сегментації на основі медичного зображення комп’ютерної томографі можна розглянути численні характеристики, такі як інтенсивність, форма та положення. Аналогічно у критеріїв для внутрішнього шаблону подібність групування може бути множинною, просторова узгодженість з ознаками однорідності, зв'язність з компактністю, різноманітність з точністю. Для сегментації зображення можлива комбінація кількох методів для отримання відповідного результату, і може бути проведено багаторазову оптимізацію та процес прийняття рішень буде використовувати оцінку множинної валідності.

Існує два загальних підходи до багатоцільової оптимізації Перший підхід полягає в поєднанні кількох цілей функції в одну складену функцію, а другий полягає в тому визначити набір рішень, які не домінують по відношенню до кожної мети.

Звичайний підхід з зваженою формулою (WFA) де багатоцільова задача перетворюється на проблему з єдиною метою [21], в якій зазвичай виконується присвоєння чисельної ваги кожній цілі, а потім об’єднання значення зважених критеріїв в одне значення шляхом додавання або множення зважених критеріїв. Якість даного кандидата модель задається однією з двох формул які наведені нижче:

Q = w1c1 +w2c2 +…+ wncn,

Q = w1c1 ×w2c2 ×… × wncn,

де *Wi*  – ваги, присвоєні критеріям ci;

 *n* – кількість критеріїв оцінки.

Підхід Парето (PTA) замість трансформації багатоцільової задачі в єдину цільову функцію, з подальшим її розв’язанням, передбачає використання багатооб’єктного алгоритм вирішення проблеми. Формулювання починається з одночасної оптимізація кількох цілей, розумне рішення полягає у дослідженні набору рішень, кожне з яких задовольняє цілі на прийнятному рівні без домінування іншого рішення, такі рішення називаються недомінованими і область цих рішень називається фронтом Парето.

## 1.3 Аналіз програмних засобів опрацювання біомедичних зображень

Розвиток сучасних технологій призводить до значного збільшення застосування засобів комп'ютерного зору і пов’язаного з ним програмного забезпечення в першу чергу в галузі медицини, а також в автоматизації, галузі безпеки, smart-технологіях, та інших. З цим пов'язаний значний розвиток численних платформ, фреймворків, і бібліотек програмного забезпечення.

Розглянемо найпопулярніші засоби комп’ютерного зору.

OpenCV – це бібліотека програмного забезпечення для машинного навчання та комп’ютерного зору що має відкритий вихідний код. Створений з метою забезпечення загальної архітектури для програм комп’ютерного зору, OpenCV надає доступ до понад 2500 класичних і сучасних алгоритмів.

Ці алгоритми можуть бути для виконання різноманітних завдань, включаючи виявлення та розпізнавання обличчя, ідентифікацію об’єктів, вилучення 3D-моделей об’єктів, відстеження рухомих об’єктів та з’єднання кількох кадрів у зображення з високою роздільною здатністю.

OpenCV має кілька інтерфейсів, таких як C++, Python, Java і MATLAB, і він підтримує більшість операційних систем, включаючи Windows, Android, Linux і Mac.

Плюси:

* використання безкоштовне і є доступний вихідний код;
* широке ком’юніті, хороша підтримка;
* доступ до більш ніж 2500 алгоритмів;
* дозволяє гнучко налаштувати код для виконання конкретних задач.

Мінуси:

* він не такий простий у використанні, як інші інструменти, такі як MATLAB.

MATLAB – це платформа програмування, яка використовується для ряду різних додатків, таких як машинне навчання, глибоке навчання, а також обробка зображень, відео та сигналів.

Він поставляється з набором інструментів комп’ютерного зору, який має кілька функцій, програм і алгоритмів, які допоможуть розробити та реалізувати будь-які рішення для завдань, пов’язаних із комп’ютерним зором.

Плюси:

* простий у використанні та навчанні;
* багато безкоштовних ресурсів;
* своя мова програмування, простіше реалізовувати код;
* дозволяє дуже швидко кодувати, відмінно підходить для швидкого прототипування;
* має зручний процес автоматичного налагодження;
* вважається найкращим інструментом для дослідників;
* всі функції дуже добре задокументовані.

Мінуси:

* орієнтований на дослідницьку роботу;
* інструмент не є безкоштовним у використанні;
* працює досить повільно при реалізації багатьох завдань;
* непросто інтегруватися зі сторонніми інструментами.

TensorFlow є однією з найпопулярніших наскрізних платформ машинного навчання з відкритим кодом з повним набором інструментів, ресурсів і бібліотек. TensorFlow особливо корисний для створення та розгортання додатків, пов’язаних із комп’ютерним зором, які працюють на основі машинного навчання.

TensorFlow є одним із найпростіших інструментів комп’ютерного зору, який дозволяє користувачам розробляти пов’язані з комп’ютерним зором моделі машинного навчання для таких завдань, як розпізнавання обличчя, класифікація зображень, виявлення об’єктів тощо. Tensorflow, як і OpenCV, також підтримує різні мови, такі як Python, C, C++, Java або JavaScript.

Для реальних проектів комп’ютерного зору TensorFlow Lite –полегшена реалізація машинного навчання на пристрої за допомогою граничних пристроїв. Як частина TensorFlow, TF Lite значно прискорює реалізацію машинного навчання із зменшеним розміром моделі та високою точністю при набагато вищій ефективності, що дає змогу використовувати машинне навчання скрізь, де в цьому є необхідність.

Плюси:

* платформа з відкритим кодом;
* платформа сумісна з кількома мовами;
* забезпечує постійні оновлення для додаткових функцій та покращень.

Мінуси:

* інструментарій забирає надзвичайно багато ресурсів.

CImg інструментарій з відкрити кодом, простий у використанні, ефективний і призначений для створення дуже інструментів для розробки алгоритмів обробки зображень на C++. Завдяки своїй загальній концепції він може охоплювати широкий спектр додатків обробки зображень.

Плюси:

* простий у використанні.

Мінуси:

* неоптимізований;
* нестpуктурована документація.

Більшість бібліотек надають доступ до стандартного набору алгоритмів низького та середнього рівнів комп’ютерного зору. Перевагою таких інструментів як OpenCV та MATLAB Image Processing Toolbox є ширший набір алгоритмів у порівнянні з іншими та хороший супровід зі сторони ком’юніті та розробників.

CUDA – це програмно-апаратна архітектура паралельних обчислень, що дозволяє суттєво збільшити обчислювальну продуктивність завдяки використанню графічних процесорів NVIDIA.

При використанні цієї технології необхідно знати такі поняття:

* пристрій (device) – сама відеокарта, графічний процесор (GPU) – виконує команди центрального процесора;
* хост (host) – центральний процесор (CPU) – запускає різноманітні завдання, виділяє пам'ять, etc.;
* ядро (kernel) – функція (завдання), яка виконуватиметься на GPU.

CUDA дозволяє розробникам реалізовувати на особливому спрощеному діалекті мови C алгоритми, що використовуються у графічних процесорах NVIDIA, та включати спеціальні функції до тексту програми на C. Архітектура CUDA дозволяє розробнику на власний розсуд організовувати доступ до набору інструкцій та пам'яті GPU.

Основні відмінності між виконанням задач на CPU та GPU закладені на етапі проектування. Ядро CPU розробляється для поетапного запуску та максимально швидкого виконання одиничного потоку послідовних інструкцій, основна увага приділяється при цьому продуктивності системи. CPU оптимізовані для обробки потоку даних який може включати як цілі числа, так і числа з плаваючою точкою. Розробники CPU намагаються також розпаралелити виконання інструкцій шляхом збільшення кількості ядер, однак у такого підходу є обмеження на зростання продуктивності, в першу чергу – обмеження на кількість ядер на кристалі, крім того експоненціально зростають і вартість, і енергоспоживання системи, і це невисоких показниках швидкості обчислень та продуктивності. Крім того, CPU дуже важливо забезпечити швидкий і довільний доступ до будь-якої комірки пам'яті.

GPU, в свою чергу, орієнтовані на максимально швидке виконання великої кількості однотипних паралельних потоків інструкцій. Відеочіп приймає в якості початкових даних групу полігонів, проводить над ними запрограмовані операції і на виході подає пікселі. Обробляти полігони та пікселів можна обробляти в паралелі, незалежно один від одного, їхній стан не залежить один від одного. Тому виконання інструкцій на GPU організовано значно простіше і розпаралелено ще на етапі проектування. Крім того, через початково паралельну організацію роботи в GPU простіше організувати подачу інструкцій на виконання на велику кількість виконавчих блоків, на відміну від складно організованої роботи конвейєра послідовного потоку інструкцій для CPU.

Відрізняється між GPU nf CPU ще й полягає у принципі доступу до пам'яті. У GPU він пов'язаний і легко передбачуваний – якщо з пам'яті читається піксель текстури, через деякий час це відбудеться і у сусідніх пікселів. Та й при записі те саме – піксель записується у фреймбуфер, і через декілька циклів буде записуватиметись розташований поряд. Таким чином організація пам'яті відрізняється від тієї, що використовується у CPU. І відеочіпу, в порівнянні з універсальними процесорами, не потрібна кеш-пам'ять великих об’ємів, а для текстур потрібно невеликі об’єми.

Внутрішня пам’ять чіпу (кеш) у CPU i GPU мають значні відмінності як в принципах роботи, так і безпосередньо в реалізації. Для центрального процесора використання кеш-пам’яті має критичне значення для збільшення продуктивності шляхом зменшення затримок в доступі до основних даних. Відповідно графічні процесори використовують кеш для тимчасового кешування даних з метою збільшення пропускної здатності.

Деякі відмінності є в кешуванні. Універсальні центральні процесори застосовуюють кеш-пам'ять для покращення продуктивності за рахунок зменшення затримок звернень до пам'яті, а GPU використовує кеш чи загальну пам'ять для збільшення діапазону пропускання. При цьому, CPU необхідний кеш великого розміру, а також складний механізм передбачення в коді, щоб завчано завантажувати в кеш необхідні дані. Однак часто трапляються ситуації помилки при передбаченні, вибору неправильної гілки розгалуження коду, і, як наслідок, завантаження в кеш даних, які надалі використовуватись не будуть. Такі ситуації негативно впливають на загальну продуктивність. Збільшення розмірів кеш-пам’яті вимагає збільшення їх площі на поверхні чіпа та відповідно збільшення енергоспоживання. Відеочіпи за рахунок паралельного виконання багатьох потоків не мають таких проблем з доступом до пам’яті – якщо потік очікує на дані з пам’яті, GPU продовжує виконання інших потоків без затримок та очікування.

Організація роботи з пам'яттю у GPU та CPU має серйозні відмінності. На GPU зазвичай встановлюються по кілька контролерів пам'яті. Крім того, на відеокартах застосовується швидша пам'ять, і в що забезпечує в кілька разів кращу пропускну здатність, а це відповідно впливає на швидкість роботи та продуктивність, особливо при роботі з для паралельними потоками інструкцій, які оперують з величезними масивами даних.

В універсальних процесорах велика частина площі чіпу та значна кількість мікроелементів витрачається на буфери для команд, апаратне передбачення розгалуження та кристали кеш-пам'яті великого об’єму. Подібна архітектура розрахована на прискорене виконання нечисленних потоків команд. Відеочіпи витрачають площі чипу на додаткові масиви виконуючих блоків, блоки управління потоками, розподілену пам'ять невеликого об'єму і багатоканальні контролери пам'яті. Хоча подібна будова і не прискорює виконання окремих потоків, але дозволяє відеочіпу досягти хорошого приросту продуктивності при обробці великої кількості паралельних потоків, які вимагають одночасного виконання і високої пропускної здатності пам'яті.

Сучасні відеокарти дуже часто мають свою систему охолодження, що дозволяє збільшувати навантаження та досягати значного підвищення продуктивності при роботі зі складними обчисленнями та великими обсягами даних. Однак слід врахувати що при цьому значно зростає енергоспоживання таких систем.

Є безліч відмінностей і організації багатопоточності. CPU виконує кілька потоків обчислень одним процесорнм ядром, а GPU можуть підтримувати декілька тисяч потоків на кожен з мультипроцесорів, яких у чіпі декілька штук. І якщо перехід з виконання одного потоку на інший для CPU може обійтися в сотні тактів, то для GPU перемикання кількох потоків обходиться в один такт.

Підсумовуючи вищезазначене, можна констатувати, що на відмінну від сучасних універсальних центральних процесорів GPU створені для об’ємних паралельних арифметичних обчислень.

Переваги:

* інтерфейс програмування CUDA базується на стандартній мові програмування С з деякими розширеннями, що полегшує процес вивчення та впровадження архітектури CUDA;
* CUDA надає доступ до розділеної між потоками пам'яті, яка може бути застосована для організації кешу з широкою пропускною здатністю, в порівнянні з текстурними вибірками;
* ефективніша передача інформації між системною та пам’яттю GPU;
* відсутність необхідності в графічних інтерфейсах з надмірністю та накладними витратами;
* лінійна адресація пам'яті, можливість запису за довільними адресами;
* апаратна підтримка цілих і бітових операцій.

Недоліки:

* Енегроспоживання таких систем є досить суттєвим.

## Постановка завдання до виконання кваліфікаційної роботи

На основі проведеного аналізу сучасних систем обробки та налізу біомедичних зображень та алгоритмів опрацювання зображень можна виділити такі завдання для успішного вирішення поставлених задач:

* провести класифікацію існуючих типів біомедичних зображень;
* проаналізувати методи та алгоритми опрацювання зображень;
* провести аналіз програмних засобів опрацювання біомедичних зображень;
* провести аналіз та порівняння алгоритмів одномірної та багатомірної регресії;
* сформувати навчальну вибірку для навчання згорткової нейронної мережі;
* розробити реляційну базу даних системи опрацювання зображень та провести її тестування.

## Висновки до розділу 1

Проведено дослідження та класифікацію сучасних видів біомедичних зображень, що дозволило виділити основні групи біомедичних зображень для проведення подальшої обробки.

Проведено дослідження алгоритмів обробки зображень яке дозволило сформувати послідовність роботи груп алгоритмів в процесі роботи програмної системи.

Проведено дослідження структур та функціональних можливостей програмних засобів обробки біомедичних зображень, що дозволило в подальшому спроектувати власну систему оброблення зображень.

# 2 Методи та алгоритми аналізу цитологічних та гістологічних зображень на основі лінійної регресії

## 2.1 Узагальнений алгоритм опрацювання біомедичних зображень

Результатом мікроскопічного дослідження є цифрове зображення у одному з наступних форматів - .jpg, .png, .bmp. Для використання алгоритмів лінійної регресії потрібно обчислити кількісні характеристики мікрооб’єктів. До кількісних характеристик ядер клітин відносять площу, периметр та окружність. Метод обробки цитологічних зображень включає наступні етапи, що виконуються у чіткій послідовності (рисунок 2.1).

Вхідне зображення

Адаптивна попередня обробка зображень

Фільтрація

Гістограмне вирівнювання

Сегментація цифрових зображень

Виділення мікрооб’єктів та обчислення кількісних характеристик

 Зберігання в Базу Даних

Застосування лінійної регресії

Рисунок 2.1 – Етапи обробки цитологічних зображень

Даний алгоритм дозволяє описати усі етапи обробки цифрового зображення, а також етапів попередньої обробки, що використовуються для зменшення похибки опрацювання зображень під час аналізу отримани даних.

1. Першоий етап починається з завантаження цифрового зображення у пам'ять комп’ютера.
2. Другий етап складається з попередьої обробки зображень, яка включає в себе фільтрацію аткими фільтрами як гаусовий та медіанний для зменшення рівня шумів. Також додатково застосовується гістограмне вирівнювання.
3. На третьому етапі відбувається автоматична сегментація цифрового зображення на основі загорткової нейронної мережі типу U-Net.
4. Четвертий етап здійснює обробку кількісних характеристик ядер клітин, які були виділені після етапу сегментації.
5. На п’ятому етапі кількісні характеристик ядер клітин записуються у сховище даних для подальшої обробки.
6. Останній етап включає застосування набору алгоритмів лінійної регресії для обчислення кількісних характеристик ядер клітин.

Більшість зображень піддається впливу різного роду шумів ще на етапі формування. Тому, першим етапом обробки зображень є фільтрація. Наявність шумів на зображенні може причинити неточності та спотворення на етапі сегментації та розпізнавання. Наприклад, система може сприйняти шуми за окремі об'єкти, що може негативно вплинути на подальші дослідження.

В результаті досліджень можна виділити такі основні види шумів: адитивний Гаусовий та імпульсний. Адитивний Гаусовий шум характеризується додаванням до кожного пікселя зображення значень з відповідного нормального розподілу з нульовим середнім значенням. Такий шум з'являється в пристроях формування цифрових зображень. Імпульсний шум характеризується заміною частини пікселів значеннями фіксованої або випадкової величини. Такий шум пов'язаний з втратами при передачі зображень по каналах зв'язку. Зазвичай, в одному зображенні можна зустріти обидва види шумів, такі шуми називають комбінованими.

Гаусова фільтрація – це фільтр лінійного згладжування, який може видаляти гаусовий шум, і його ефект полягає в зменшенні різких змін у градаціях сірого зображення, тобто зображення являється розмитим. Гаусова фільтрація дуже ефективна для зниження шуму, який піддається нормальному розподілу, за рахунок розмивання зображення. Звичайно, іноді метою згладжування та фільтрації зображення є розмитість зображення.

Вираз для операції згортки гаусового фільтра для пікселя з координатами *x*, *y* такий:

$$G\left(x,y\right)=\frac{1}{2πσ^{2}}e^{{-\left(x^{2}+y^{2}\right)}/{2σ^{2}}}.$$

Значення кожного пікселя отримує середньозважене значення самого себе та значень інших пікселів навколо, а також сам спосіб зваження, заснований на функції ядра функції Гаусса.

Медіанна фільтрація реалізується у вигляді процедури локальної обробки у ковзаючому вікні, яке включає в себе певне число пікселів. Для кожного положення вікна виділені в нього відстеження за значенням. Середній по своєму положенню піксель в ранжованому списку називається медіаною для групи пікселів даного положення вікна. Цим значенням і замінюється центральний піксель у вікні обробки. Медіанний фільтр відноситься до числа нелінійних фільтрів, замінюючих середнім значення аномальні точки і вкиди незалежно від їх амплітудних значень, і є стійким за визначенням, здатним анулювати навіть безкінечно великі відсотки.

Алгоритм медіанної фільтрації має явно виражену вибірковість до елемента маси з немонотонною складовою послідовності чисел у межах апертур і найбільш ефективно виключає із зображення одиночні вкиди, які потрапляють на краї ранжованого списку. З огляду на ранжування в списку медіанні фільтри добре подають шуми та перешкоди, протяжність яких становить менше половини вікна. Стабільною точкою є послідовність (у одномірному випадку) або масова (у двомірному випадку), які не змінюються при середній фільтрації. У одномірному випадку стабільними точковими медіанними фільтрами є "локально-монотонні" послідовності, які медіанні фільтри вносять без змін. Виключення складають деякі періодичні подвійні послідовності.

Завдяки цій особливості медіанні фільтри при оптимально вибраному розмірі вікна можуть зберігати без спотворення різкі границі об’єктів, подавляючи некорельовані та слабокорельовані шуми та малорозмірні деталі.

Медіанна фільтрація зображень найбільш ефективна, якщо шум на зображенні має імпульсний характер і представляє собою обмежений набір пікових значень на фоні нулів. В результаті використання медіанного фільтру границі різких перепадів значень яскравості на зображеннях не змінюються. Це дуже корисна властивість саме для зображень, на яких контури несуть основну інформацію.

При середній фільтрації зашумлених зображень ступінь згладжування контурів об'єктів напряму залежить від розмірів апертур фільтра та форми маски. При малих розмірах апертури краще зберігаються контрастні деталі зображення, але в меншому ступені подавляються імпульсні шуми. При великих розмірах апертур спостерігається зворотна картина. Оптимальний вибір форми згладжуючої апертури залежить від особливостей задачі та форми об'єктів. Особливе значення має для задач збереження перепадів (різких границь яскравості) у зображеннях.

Під зображенням перепаду ми розуміємо зображення, в якому точки на одну сторону від деякої лінії мають одиничне значення a, а всі точки по стороні від цієї лінії – значення *b*, *b*≠*а*. Якщо апертура фільтра симетрична щодо початку координат і містить його в собі, то медіанний фільтр зберігає будь-яке зображення перепадів. Це виконується для всіх апертур з непарним числом елементів (рисунок 2.2).



Рисунок 2.2 – Приклади масок за мінімальною апертурою

Для спрощення обмежимося використанням фільтра з квадратною маскою розміром 3х3.

Таким чином, алгоритм медіанного фільтра буде складатись з таких послідовних кроків:

1. Вибір значень, які потрапили у вікно фільтра *w*.
2. Сортування значень вікна *w*.
3. Вибір медіанного значення *m*(*w*).
4. Запис медіанного значення в точку *х*.

Ковзаючий фільтр переглядає пікселі зображення зліва-направо і зверху-вниз, при цьому вхідну двовимірну послідовність також представимо у вигляді послідовного числового ряду пікселів {{*x*(*n*)}} зліва-направо зверху-вниз. З цієї послідовності в кожній поточній точці маска фільтра виділяє масив *w*(*n*) як n-eлементний вектор, який в даному випадку містить усі елементи з вікна 3×3, центровані навколо *x*(*n*), і сам центральний елемент, якщо це передбачено типом маски:

$$w=\left|x\_{1}(n),x\_{2}(n),...,x\_{w}(n)\right|.$$

У цьому випадку значення xi відповідає відображенню зліва-направо та зверху-вниз вікна 3 × 3 в одновимірний вектор, як показано на рисунку 2.3.



Рисунок 2.3 – Елементи вікна маски 3х3

Елементи даного вектора, як і для одномірного медіанного фільтра, також можуть бути впорядковані в ряд за зростанням або зменшенням своїх значень:

$r=\left|x\_{1}(n),x\_{2}(n),...,x\_{w}(n)\right|$.

Визначено значення медіани *y*(*n*)=*med*(*r*(*n*)), центральний піксель маски замінений значенням медіани. Якщо за типом маски центральний відлік не входить до ряду *r*(*n*), то медіанне значення перебуває у вигляді середнього значення двох центральних пікселів ряду *w*(*n*).

Наведені вирази не пояснюють способу знаходження вихідного сигналу поблизу кінцевих і прикордонних точок кінцевих послідовностях і зображеннях. Один із простих прийомів полягає в тому, що потрібно знаходити медіану лише тих точок усередині зображення, які потрапляють у межі апертури. Тому для точок, розташованих поряд із кордонами, медіани будуть визначені, виходячи з меншої кількості точок.

## 2.2 Алгоритми виділення кількісних характеристик ядер клітин

Для виділення ядер клітин на зображенні спочатку потрібно виділити самі клітини, провівши об’єднання окремих однорідних областей. Проаналізувавши зображення клітин, можна виділити такі закономірності зображень:

* ядро клітини сильно відрізняється по кольору відносно всієї клітини, і має темне забарвлення;
* при цьому ядро клітини яскравість або не змінює, або це відбувається поступово;
* на границі ядра клітини спостерігається різкий перепад яскравості;
* зміна яскравості клітин відбувається поступово;
* на границях клітини спостерігається різкий перепад яскравості, проте не такий значний, як у ядрі.

Введемо позначення: *Cor* – визначене ядро клітини. *RCor* – однорідна область на зображенні, що може бути визнана ядром клітини. *Cell* – клітина на вхідному зображенні. *RCell* – область зображення, яка може бути клітиною. При цьому сусідніми областями ми будемо вважати області контури яких дотикаються.

На основі отриманої моделі зображення з областями клітин та масиву ядер клітин, використаємо наступний алгоритм виділення клітин:

1) знаходимо множину однорідних областей {*RCor*}, які можуть відповідати ядрам клітин вхідного зображення;

2) перевіряємо всі елементи множини {*RCor*} на критерії відповідності однорідних областей ядрам клітин для знаходження множини ядер клітин {*Cor*};

3) приєднуємо всі сусідні однорідні області, що межують з *Cori*- ядром, утворюючи *RCelli* - область клітини;

4) нарощуємо *RCelli* області клітини, для знаходження країв *Celli* клітини;

5) повернення на етап 3 для ядра *Cori*+1 ядра, якщо перевірено всі ядра – завершення роботи алгоритму.

Для знаходження точок, що можуть належати ядрам клітин, проводимо порогову сегментацію монохромного вхідного зображення

$I→ I\_{мон}→Ib$*.*

Значення параметра яскравості порогової сегментації *t* = 90 (встановлене по замовчуванню) може корегуватись. Після операції порогової обробки отримуємо бінаризоване зображення, на якому виділено окремі однорідні групи пікселів, що можуть відповідати областям ядер клітин {*RCor*}. Для відсіювання хибних областей ядер клітин проводимо перевірку інформативних характеристик виділених однорідних областей: площа() *A* *RCori* та колоподібність *Circle* (*RCori*):

$0,9×A(R\_{etal})\leq A(Cor\_{i})\leq 1.1×A(R\_{etal})$*,*

де *A*(*REtal*) – значення площі еталонної області ядра, яке визначається експертом для відповідного типу мікрооб’єктів.

$0,8\leq Circle(Cor\_{i})\leq 1.2$,

де $Circle(Cor\_{i})=\frac{P(RCor\_{i})^{2}}{A(RCor\_{i})}$.

Результатом даного етапу є множина областей ядер клітин {*RCor*}.

Знаходження однорідних областей вхідного зображення, що відповідають областям ядрам клітин. Для цього співставляємо масив однорідних областей ** та маси областей ядер клітин, отриманих на поперепедньому етапі $\{OCor\_{q}\}$. Під час співставлення можуть виникнути наступні випадки:

Випадок 1. Область ядра $OCor\_{i}$ повністю співпадає з однорідною областю $OCor\_{i}=O\_{j}$. В даному випадку однорідна область  визнається ядром клітини $Cor\_{k}$.

Випадок 2. Область ядра $OCor\_{i}$ повністю включається в однорідну область $O\_{j}\in OCor\_{i}$ та площа області ядра займає не менше 50% площі однорідної області $P(OCor\_{i})\geq 0.5⋅P(O\_{j})$. В даному випадку область ядра $ON\_{i}$ визнається неповною, а однорідна область  визнається ядром клітини $Cor\_{k}$.

Випадок 3. Область ядра $OCor\_{i}$ повністю включається в однорідну область$ O\_{j}\in OCor\_{i}$ та площа області ядра займає менше 50% площі однорідної області $P(OCor\_{i})<0.5⋅P(O\_{j})$. В даному випадку область ядра $OCor\_{i}$ визнається неправильно визначеною та видаляється з подальшого розгляду.

Випадок 4. Однорідна область  повністю включається в область ядра $O\_{j}\in OCor\_{i}$ та площа однорідної області займає не менше 90% площі області ядра $P(O\_{j})\geq 0.9⋅P(OCor\_{i})$. В даному випадку область ядра , а однорідна область  визнається ядром клітини $Cor\_{k}$.

Випадок 5. Область ядра $OCor\_{i}$ повністю включається в декілька однорідни областей$O\_{1}\in OCor\_{i},O\_{2}\in OCor\_{i},...,O\_{p}\in OCor\_{i}$ та площа області ядра займає не менше 50% площ кожної з однорідних областей $P(OCor\_{i})\geq 0.5⋅P(O\_{1}),P(OCor\_{i})\geq 0.5⋅P(O\_{2}),...,P(OCor\_{i})\geq 0.5⋅P(O\_{p})$ (рисунку 2.9d). В даному випадку область ядра $OCor\_{i}$ визнається визначеною не повністю. Ядру клітини $Cor\_{k}$ відповідає область утворення об’єднанням однорідних областей .

Випадок 6. Область ядра $OCor\_{i}$ повністю включається в декілька однорідних областей $O\_{1}\in OCor\_{i},O\_{2}\in OCor\_{i},...,O\_{p}\in OCor\_{i}$. Площа області ядра займає не менше 50% площ деякої мнодини з однорідних областей, а в інших площак перетину є меншою $P(OCor\_{i})\geq 0.5⋅P(O\_{1}),P(OCor\_{i})\geq 0.5⋅P(O\_{2}),...,P(OCor\_{i})<0.5⋅P(O\_{p})$. В даному випадку область ядра $OCor\_{i}$ визнається визначеною некоректно. Область утворена об’єднанням однорідних областей, для яких виконується умова 50% визнається ядром клітини $Cor\_{k}$.

Результатом даного етапу є масив ядер клітин $\{Cor\_{k}\}$.

Приєднання сусідніх однорідних областей, що межують з $Cor\_{i}$-им ядром, утворюючи -у область клітини. Базуючись на апріорній інформації про те, що навколо ядра в недеформованих клітинах розміщена цитоплазма, приєднуємо всі однорідні області, що є сусідніми до ядра клітини. При приєднанні однорідних областей проводиться перевірка на дві умови:

* однорідна область є сусідньою до областей, які відповідають ядру клітини;
* яскравість однорідної області є вищою за яскравість ядра, в іншому випадку дана клітина визнається деформованою і дане ядро виключається з подальшого аналізу.

Результатом роботи даного етапу є масив ядер клітин  з частково визначеною цитоплазмою.

Однією з найважливіших завдань аналізу даних є класифікація – віднесення об'єктів предметної області до заздалегідь певних груп, які називають класами. При цьому кожному класу повинні належати об'єкти, близькі за своїми властивостями. Узагальнюючи властивості відомих об'єктів класу нові, віднесені щодо нього об'єкти, можна отримувати знання про них.

Завдання класифікації вирішується за допомогою аналітичних моделей, які називають класифікаторами. Класифікувати об'єкт означає пред'явити набір його ознак (зазвичай представлених у вигляді вектора) на вхід моделі-класифікатора, яка повинна надати йому мітку або номер класу. В даний час розроблено велику кількість різних видів класифікаторів, для побудови яких використовуються як статистичні методи (логістична регресія, дискримінантний аналіз), так і методи машинного навчання (нейронні мережі, дерева рішень та ін.).

Необхідність використання в аналізі даних великої кількості різноманітних методів класифікації, обумовлена тим, що задачі, що вирішуються з її допомогою, можуть мати свої особливості, пов'язані, наприклад, з поданням вихідних даних, їх кількістю і якістю, що вимагає вибору адекватного класифікатора. Тому вибір класифікатора, що відповідає особливостям розв'язуваної задачі аналізу, є важливим фактором отримання правильного рішення.

Віднесення класифікації до інтелектуальним технологіям аналізу даних обумовлено тим, що у повсякденному житті свідомість людини, у якого постійно потрапляють нові об'єкти навколишнього світу, зіставляє їх із відомими об'єктами і оцінює ступінь їх подібності. Потім на основі цієї оцінки об'єкт асоціюється з певною групою (класом). Таким чином, класифікація є найбільш «природним» для людського інтелекту способом отримання знань про процеси та явища, що відбуваються в навколишньому світі.

З огляду на сказане, можна припустити, що це методи класифікації у тому чи іншому вигляді використовуватимуть формалізоване поняття «подібності», міра якого оцінюватиметься з допомогою деякої функції. У статистичних методах аналізу мірою подібності є ймовірність приналежності об'єкта класу, яка оцінюється кожному за класу, після чого вибирається той із них, котрій ця ймовірність найбільша.

У метричних методах мірою подібності є відстань (наприклад, евклідова) у векторному просторі, де кожен об'єкт представлений своїм вектором ознак. Новий об'єкт швидше за все належить до того ж класу, що більшість сусідніх з ним об'єктів. Метричні методи зазвичай використовуються для побудови класифікаторів на основі машинного навчання.

Перевагою статистичних методів є їхня хороша математична обґрунтованість, недоліком – низька пояснювальна здатність. Використання імовірнісних оцінок дозволяє з високою точністю передбачити якого класу належить об'єкт, але не дозволяє сказати чому. Тому результати статистичних методів класифікації можуть бути складними розуміння та інтерпретації результатів.

Недоліком метричних методів є їх евристичний характер - вони можуть дати неточне і неоднозначне рішення, але вважається прийнятним у більшості значущих випадків. Однак при цьому вони мають високу пояснювальну здатність і тому їх результати простіше інтерпретувати.

Оскільки ознаки, на основі яких проводиться класифікація, можуть мати різну фізичну природу і, відповідно, діапазони значень, для покращення результатів класифікації буде корисно виконати нормалізацію навчальних даних.

Нормалізація – це приведення всіх значень ознаки нового діапазону. Це корисно, оскільки значення ознак можуть змінюватися дуже великому діапазоні. Причому значення різних ознак можуть відрізнятися на кілька порядків. А після нормалізації вони всі будуть у вузькому діапазоні.

Найбільш популярним способом нормалізації є нормалізація методом мінімума-максимума. Для того, щоб застосувати цей метод, має бути відоме максимальне та мінімальне значення ознаки.

$x\_{new}=\frac{x\_{old}-x\_{min}}{x\_{max}-x\_{old}}$.

Також досить популярним методом є Z-нормалізація. Діапазон нових значень для Z-нормалізації виглядає так:

$$(-3σ[X],3σ[X]),$$

де $σ[X]$ – середньоквадратичне відхилення ознаки *X.*

Виконується Z-нормалізація за такою формулою:

$$x\_{new}=\frac{x\_{old}-M\left|X\right|}{σ[X]}$$

де *M*[*X*] – математичне сподівання ознаки *X*.

У разі застосування *Z*-нормалізації до кількох ознак діапазон значень для них буде різним.

## 2.3 Класична модель лінійної регресії

Термін "лінійність" в алгебрі відноситься до лінійного зв'язку між двома або більше змінними. Якщо проведемо це співвідношення у двомірному просторі (між двома змінними), то отримаємо пряму.

Лінійна регресія виконує завдання передбачити значення залежної змінної (*y*) на основі даної незалежної змінної (*x*). Цей метод регресії з’ясовує лінійну залежність між *x* (вхід) та *y* (вихід). Звідси і назва - лінійна регресія. Якщо намітити незалежну змінну (*x*) на осі *x* та залежну змінну (*y*) на осі *y*, лінійна регресія дає нам пряму лінію, яка найкраще відповідає точкам даних, як показано на рисунок 2.4.



Рисунок 2.4 – Лінійна регресія

Рівняння наведеної вище прямої виглядає так:

*Y* = *mx* + *b*,

де *b* – перехоплення,

 *m* – нахил лінії.

Отже, в основному алгоритм лінійної регресії дає нам найбільш оптимальне значення для перехоплення та нахилу (у двох вимірах). Змінні *y* та *x* залишаються незмінними, оскільки вони є характеристиками даних і їх не можна змінити. Значення, які ми можемо контролювати – це перехоплення (*b*) та нахил (*m*). Залежно від значень перехоплення та нахилу може бути декілька прямих ліній. В основному алгоритм лінійної регресії робить так, що він поміщає кілька рядків у точках даних і повертає рядок, що призводить до найменшої помилки.

Як вхідні дані дається безліч векторів імовірно незалежних змінних, кожному з яких ставиться у відповідність деяке значення залежної змінної. Ці дані можна подати у вигляді двох матриць:

$\left(\begin{matrix}a\_{11}&a\_{12}&a\_{13}&...&a\_{1n}\\a\_{21}&a\_{22}&a\_{23}&...&a\_{21}\\...& & & & \\a\_{1}&a\_{2}&a\_{3}&...&a\_{mn}\end{matrix}\right)\left(\begin{matrix}b\_{1}\\b\_{2}\\...\\b\_{n}\end{matrix}\right)$.

Якщо передбачається лінійна залежність, наше припущення можна представити у вигляді добутку матриць (для спрощення запису тут і надалі передбачається, що вільний член рівняння представлений як *xn*, а останній стовпець матриці *A* містить одиниці):

$\left(\begin{matrix}a\_{11}&a\_{12}&a\_{13}&...&a\_{1n}\\a\_{21}&a\_{22}&a\_{23}&...&a\_{21}\\...& & & & \\a\_{1}&a\_{2}&a\_{3}&...&a\_{mn}\end{matrix}\right)×\left(\begin{matrix}x\_{1}\\x\_{2}\\...\\x\_{n}\end{matrix}\right)=\left(\begin{matrix}b\_{1}\\b\_{2}\\...\\b\_{n}\end{matrix}\right)$.

Але розв’язання у такої системи швидше за все не буде. Причина – шум або відсутність лінійної залежності. Таким чином, щоб відновити лінійну залежність за реальними даними, зазвичай потрібно ввести ще одне припущення: вхідні дані містять шум і цей шум має нормальний розподіл. Можна робити припущення і про інші типи розподілу шуму, але в переважній більшості випадків розглядають саме нормальний розподіл. Множиною називають лінійну регресію, у моделі якої число незалежних змінних дві чи більше. У цьому випадку залежна змінна (цільова змінна) залежить від кількох незалежних змінних. Регресійну модель, що включає кілька змінних, можна представити у вигляді:

*y* = *b*0 + *m*1*b*1 + *m*2*b*2 + *m*3*b*3 + ... + *mnbn*.

Це рівняння гіперплощини. Модель лінійної регресії у двох вимірах є прямою лінією; у трьох вимірах це площина, а в більш ніж трьох вимірах – гіперплощина. Як і в простій лінійній регресії, параметри моделі обчислюються з допомогою методу найменших квадратів. Відмінність між простою та множинною лінійною регресією полягає в тому, що замість лінії регресії в ній використовується гіперплощина.

Перевага множинної лінійної регресії в порівнянні з простою полягає в тому, що використання в моделі кількох вхідних змінних дозволяє збільшити частку поясненої дисперсії вихідної змінної, і таким чином покращити відповідність моделі даним. Тобто. при додаванні до моделі кожної нової змінної коефіцієнт детермінації зростає.

Однак у множинні лінійної регресії виникають і проблеми, нехарактерні для простої моделі такі як можлива поява мультиколеніарності та необхідність вибирати найкращу модель, у якій мінімальний набір незалежних змінних зможе пояснити найбільшу частину дисперсії залежною.

Подібними властивостями володіють кореляційні підходи в статистичні обробці даних. Кореляція - це поняття, яке можемо проаналізувати зі статистичної точки зору. Враховуючи статистичну природу більшості методів машинного навчання, тому має сенс говорити загалом про кореляцію, але не про причинно-наслідковий зв’язок.

Явища кореляції не впливають на структуру та спосіб роботи класифікатора. Однак вони негативно впливають на індивідуальні прогнози, що може вплинути на якість кінцевих результатів.

Причина криється у деяких загальних обмеженнях, які виникають при аналізі проблеми за допомогою алгоритму. З цих обмежень найважливішим є розмір доступного набору даних.

Кореляція може бути позитивною, якщо збільшення однієї змінної пов’язане зі збільшенням іншої змінної, або негативною, якщо збільшення однієї змінної пов’язане зі зменшенням іншої змінної.

**У статистиці кореляція є мірою ступеня залежності між двома випадковими величинами. Загалом, слово кореляція має на увазі лінійну залежність.**

**Кореляція може бути позитивним явищем, оскільки вона дозволяє робити прогнози щодо однієї змінної, знаючи значення іншої змінної.**

**Кореляція є синонімом залежності, тому дві незалежні випадкові величини зі статистичної точки зору не корелюють.**

Існує кілька мір кореляції, відомих як коефіцієнти кореляції (зазвичай позначаються *ρ* або *r*). Найпоширенішим є коефіцієнт кореляції Пірсона , який вимірює ступінь лінійної залежності між двома змінними. **Це лінійне співвідношення може мати сенс, навіть якщо математична залежність між двома змінними слідує за нелінійною функцією** .

Це інші коефіцієнти кореляції, також чутливі до нелінійних зв'язків, наприклад , рангова кореляція Спірмена та взаємна інформація.

**Коефіцієнт кореляції продукт-момент Пірсона (PPMCC), також відомий як коефіцієнт кореляції Пірсона або просто коефіцієнт кореляції, є мірою якості підгонки методом найменших квадратів до вихідних даних**

При аналізі можна виділити наступну класифікацію хіріктеристик та критеріїв які використовуються при аналізу взаємопов’язаних парметрів.

За ступенем кореляції:

* Позитивна кореляція: вважається позитивною, коли значення двох змінних рухаються в одному напрямку, так що збільшення однієї змінної супроводжується збільшенням іншої змінної або зменшення однієї змінної супроводжується зменшенням. в іншій змінній.
* Негативна кореляція: Вважається негативною, коли значення двох змінних рухаються в протилежному напрямку, так що збільшення однієї змінної супроводжується зменшенням іншої змінної.

На основі зміни пропорції:

* Лінійна: якщо значення величини зміни однієї змінної прагне зберігати постійне відношення до величини зміни інших змінних, то кореляція називається лінійною.
* Нелінійний: якщо значення величини зміни однієї змінної не зберігає постійне відношення до величини зміни інших змінних, то це називається нелінійною кореляцією. Вона також відома як криволінійна кореляція.

На основі кількості досліджуваних змінних:

* Проста кореляція: коли розглядаємо лише дві змінні (двоваріантний аналіз) і перевіряємо кореляцію лише між цими змінними, це називається простою кореляцією.
* Множина кореляція: коли розглядаємо більше трьох або трьох змінних для кореляції одночасно, це називається множинною кореляцією.
* Часткова кореляція: коли одна або кілька змінних залишаються незмінними, а зв'язок вивчається між змінними, що залишилися, це називається частковою кореляцією. Вивчіть зв’язок між двома змінними та припустивши, що інші змінні є постійними.

Як показав проведений аналіз на сьогодні принципи взаємозалежності широко використовуються в різних наукових напрямках, серед яких одне з головних місць посідає машинне навчання для систем з елементами штучного інтелекту.

## 2.4 Алгоритми одномірної та багатомірної регресії

Алгоритм лінійної регресії (Linear Regression) відноситься до алгоритму навчання з контрольованою регресією (Regression). Алгоритм регресії встановлює модель регресії між змінними і виводить кореляцію між змінними та залежними змінними в процесі навчання. Регресійний аналіз може використовуватися для прогнозних або класифікаційних моделей. Розширені алгоритми регресії включають в себе: лінійну регресію, нелінійну регресію, логістичну регресію, поліноміальную регресію, гребінь регресії, регресію лассо і регресію ElasticNet. Серед них найчастіше використовують лінійну регресію, нелінійну регресію та логістичну регресію.

Алгоритм лінійної регресії (лінійної регресії) призначений для знаходження оптимальної прямої лінії для відповідності даними (може бути розширення до багатомірного). Лінійна регресія зазвичай використовує мінімальну суму квадратів різниці між значеннями функцій і прогнозованим значенням моделі в якості функцій втрат і використовує метод найменших квадратів і метод градієнтного спуска для вирахування параметрів остаточної підгонки.

Основна ідея алгоритму лінійної регресії полягає в тому, щоб провести пряму лінію за історичними даними і використовувати цю пряму для прогнозування нових даних. Формула для лінійної регресії виглядає наступним чином:

y = b0 + m1b1 + m2b2 + m3b3 + ... mnbn.

З приведеної вище формули можна зробити висновок, що для лінійної моделі, яка ідеально відповідає цільовому набору даних, задача полягає в тому, щоб вирішити вагу b кожної незалежної від властивостей змінної. Лінійна регресія спочатку будує функцію оптимізації випуклої функції (такої як мінімальна сума квадратів різниці між значенням функцій і значенням прогнозування моделі) та використовує метод найменших квадратів і метод градієнтного спуску для вирахування параметрів остаточної підгонки.

Основні кроки алгоритму класифікації логістичної регресії наступні:

* побудова функції прогнозування, використовуючи n-вимірну лінійну функцію;
* побудова функції втрат, суми квадратів різниці між значенням функції та значенням прогнозування моделі є найменшою;
* використання методу найменших квадратів і методу градієнтного спуску для розрахунку параметрів остаточної підгонки;
* ітеративна оптимізація параметрів остаточної підгонки;
* вивід остаточно підібраних параметрів.

Останнім кроком є оцінка продуктивності алгоритму. Цей крок особливо важливий, щоб порівняти, наскільки добре працюють різні алгоритми на певному наборі даних. Для алгоритмів регресії зазвичай використовуються три показники оцінки:

а) Середня абсолютна помилка (MAE) – середнє абсолютного значення помилок. Він розраховується як:

$$MRE=\frac{1}{n}\sum\_{j=1}^{n}\left|Y\_{j}-y\_{j}\right|.$$

б) Середня квадратична помилка (MSE) є середнім квадратом помилок і обчислюється як:

$$MSE=\frac{1}{N}\sum\_{i}^{n}(Υ\_{i}-y\_{i}).$$

в) Середньоквадратична помилка (RMSE) – це квадратний корінь із середнього квадрата помилки:

$$RMSE=\sqrt{\frac{1}{n}\sum\_{j=1}^{n}\left(Y\_{j}-y\_{j}\right)^{2}}.$$

Основні переваги алгоритму логістичної регресії наступні:

* масштабованість обчислень;
* складність обчислень можна регулювати;
* залежність параметрів;
* універсальність: підходить для неперервних і дискретних наборів даних;
* протишумна здатність, чутливість до відсутності та ненормальних даних;
* хороша інтерпретируємість результатів.

## 2.5 Висновки до розділу 2

В результаті проведених аналітичних досліджень алгоритмів визначення взаємозалежностей між двома змінними можна зробити висновок, що на сьогоднішній день алгоритми та критерії цієї групи знайшли широке застосування в різних галузях.

Здійснений аналіз та порівняння алгоритмів одномірної та багатомірної регресії на основі критеріїв залежності дозволив виділити групу характеристик біомедичних зображень для подальшого дослідження в системі опрацювання зображень.

# 3 Структура системи опрацювання кількісних характеристик біомедичних зображень

## 3.1 Навчальна вибірка для опрацювання даних

Ідеальний набір даних медичних зображень для програми машинного навчання має мати достатній обсяг даних, анотації, достовірність і можливість повторного використання. В основі кожного об’єкта даних медичної візуалізації присутні елементи даних, метадані та ідентифікатор. Ця комбінація являє собою «дослідження зображення». Набір об’єктів даних або набір даних повинен мати достатню кількість обстежень зображення, щоб відповідати заявленим цілям дослідження. Щоб максимізувати розробку алгоритму, як сам набір даних, так і кожне дослідження зображення повинні бути точно описані та позначені. Крім того, ідеальним набором даних є можливість пошуку, доступності, взаємодії та багаторазового використання.

Найважливішими властивостями вибірки для вирішення задач розпізнавання зображень є:

* репрезентативність – характеризує представництво вибірки по відношенню до генеральної сукупності (на практиці дана властивість при невідомих характеристиках генеральної сукупності забезпечується достатністю обсягу та повнотою вибірки);
* повнота вибірки визначається забезпеченістю класів екземплярами;
* розмірність – характеризує, з одного боку, просторову складність вибірки, а з іншого – мінімальну кількість операцій обробки вибірки;
* суперечливість – характеризує кількість однакових об'єктів вибірки, що належать до різних класів;
* рівномірність – показує, наскільки рівномірно розподілені екземпляри вибірки за класами;
* компактність розташування класів у просторі ознак – відображає простоту розв'язання задачі розпізнавання (чим компактніше розташовані екземпляри кожного класу, тим простіше побудувати модель);
* складність – характеризує витрати ресурсів пам'яті (просторова складність) та обчислювальних ресурсів (обчислювальна складність) для обробки вибірки.

Для деяких з цих властивостей можна запропонувати чисельні критерії, якими їх можна охарактеризувати. Однак відомі критерії не відображають всієї повноти властивостей навчальних вибірок, а також придатні не для всіх завдань. В таблиці 3.1 наведено приклади джерел та їх характеристику для формування навчальних вибірок. На основі даних баз вибірок можна сформувати окремі масиви зображень з наперед визначеними характеристиками, що у повній мірі нададуть можливість провести процес навчання. При формуванні навчальних вибірок необхідно врахувати особливості досліджуваних зображень (колірні характеристики, способи кодування та подання), особливості характеристик мікрооб’єктів на них (розмір, тип, взаємне розташування тощо), а також можливість швидкого оновлення або додавання нових зображень.

Існує декілька підходів до формування репрезентативних вибірок. Один з яких полягає у рандомному виборі навчальних прикладів з усієї генеральної сукупності можливих прикладів. Але даний підхід не завжди є доречним, оскільки наявний матеріал не може повністю описати всі об’єкти. Іншим негативним фактором є те, що медичні зображення завжди чітко розподілені під вирішення конкретної задачі. Це спричиняє нерівномірний розподіл прикладів, які в свою чергу суперечать вимогам рівномірності. Під рівномірністю розподілу класів мається на увазі факт того, що приклади різного роду класів будуть зображені у вибірках в однаковому співвідношенні між класами, якщо ймовірність появи зображених об’єктів клісів, що розпізнаються як однакові.

Таблиця 3.1 – Приклади існуючих навчальних вибірок зображень, що застосовуюються для навчання систем розпізнавання зображень

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Назва навчальної вибірки | Кількість прикладів, N | Кількість класів | Розміри зображень, пікселів  | Помилка класифікації | Кількість параметрів мережі, h |
| MNIST | 60000 навчальних, 10000 тестових | 10 | 28х28x1 | 0,23% | 300 000 |
| SVHN | 73257 навчальних, 26036 тестових | 10 | 32х32x3 | 1,94% | 1 860 000 |
| CIFAR10 | 50000 навчальних, 10000 тестових | 10 | 32х32x3 | 3,47% | 600 000 |
| NORB | 24300 навчальних, 24300 тестових | 50 | 32×32x2 | 7.1% | 300 000-1 000 000 |
| Imagenet 2012 | 1 281 167 навчальних, 50000 тестових | 1000 | 256х256х3 | 15,3% | 60 000 000- 138 000 000 |
| CIFAR100 | 50 000 навчальних, 10 000 тестових | 20/100 | 32х32х3 | 31,75% | 1 870 000 |

Іншим підходом може бути автоматична класифікація, при якій відбувається відмова від наявної класифікації об'єктів, виконується кластеризація об'єктів навчальної вибірки і за її результатами кожному об'єкту ставиться у відповідність номер класу, отриманий автоматично. Даний підхід застосовують у тих випадках, коли немає можливості отримані додаткові дані. Відповідно до другого підходу, використовуються дані про класифікацію об'єктів колективом незалежних експертів, і класифікація об'єктів визначається за допомогою обробки результатів класифікацій кожним експертом. Якщо такі дані апріорно не можуть бути отримані, то в якості експертів може виступати множина правил рішення, побудованих на основі вихідної вибірки. З часом збільшується набір досліджуваних даних, що вимагає постійного оновлення навчальної вибірки і, як наслідок, коригування правил рішення класифікації.

Вхідними елементами навчальної вибірки є цифрові біомедичні зображення. Зображення характеризується такими показниками:

* тип зображення (RGB, в градаціях сірого, чорно – біле тощо);
* глибина кольору (16 біт, 8 біт, 1 біт);
* роздільна здатність (640х480, 800х600, 1024х768, 1600х1200).

В якості експертів може виступати множина правил рішення, побудованих на основі вихідної вибірки. З часом збільшується набір досліджуваних даних, що вимагає постійного оновлення навчальної вибірки і, як наслідок, коригування правил рішення класифікації. Коригування може здійснюватись як в автоматичному режимі, коли нові правила будуть формуватись на основі аналізу нових екземплярів в навчальній вибірці так і в ручному режимі, коли користувач вносить правки у існуючі правила.

Формування навчальної вибірки для проведення комп’ютерних експериментів здійснювалось на основі бази даних цифрових гістологічних і цитологічних зображень передракових та ракових станів молочної залози «BPCI2100», що налічує близько 2100 зображень поділених на класи.

Таблиця 3.2 – Параметри навчальної вибірки для класифікації засобами згорткових нейронних мереж

|  |  |
| --- | --- |
| Параметр | Значення |
| Кількість зображень | 50-100 |
| Розміри, пікселі | 200х200, 150х150, 100х100, 50х50 |
| Кількість класів | 4 |
| Кількість епох | 10-30 |
| Відношення навчальної вибірки до тестової | 70:30 |
| Функція активації | ReLU |

Процес формування навчальних вибірок є одоним з головних підготовчих етапів роботи з програмними системами які включають в себе елементи штучного інтелекту на основі згорткових нейронних мереж. Від якості обраних зразкі в навчальну вибірку в кінцевому результаті буде залежати точність отриманих результатів роботи системи в цілому.

## 3.2 Порівняння алгоритмів лінійної регресії

Модель лінійної регресії (linear regression) – часто використовується в статистиці моделі лінійної залежності однієї змінної у від іншої, незалежної змінної х, або декількох змінних. У розумінні машинного навчання у – це відповідь, а незалежні змінні – ознаки. Проста лінійна регресія виглядає як функція згортки з безліччю ознак з ваговими коефіцієнтами:

$$a\left(x\right)=f\left(x,ω\right)=\sum\_{i=1}^{k}ω\_{i}x\_{i}-ω\_{0}.$$

Задача відновлення регресії досить часто зустрічається при обпрацюванні експериментальних даних. Для аналізу даних спочатку необхідно побудувати модель. В найпростішому випадку модель лінійної регресії буде представлена у вигляді:

$y\_{i}=β\_{1}+ β\_{2}x\_{i}+ ε\_{i}$*,*

де *y*, *x* – змінні за якими ведеться спостереження;

 $β\_{1}$,$ β\_{2}$ – невідомі параметри;

 $ε$ – помилка, випадкова складова.

Алгоритм роботи в даному випадку буде:

* вибрати модель (в нашому випадку – передбачається лінійна залежність) – модель лінійної регресії;
* отримати оцінки невідомих параметрів $Β\_{1},Β\_{2}$;
* прогнозувати, використовуючи оціночні коефіцієнти $Β\_{1},Β\_{2}$ для знаходження оціночного значення $Υ\_{i}$=$Β\_{1}+ Β\_{2}x\_{i}+ ε\_{i}$.

На даному етапі вже можна зробити деякі висновки стосовно точності прогнозу моделі. Для цього необхідно знайти $Ε\_{i}=γ\_{i}- Υ\_{i}$, та обчислити суму таких відхилень. Але при знаходженні суми необхідно врахувати, що відхилення можуть бути як додатними, так і від’ємними. І щоб при знаходженні суми відхилення не нейтралізували один одного, потрібно взяти відхилення по модулю, або, більш загальноприйнятий та розповсюджений метод – обчислити суму квадратів відхилень. Таким чином, отримуємо формулу:

$$Q\left(Β\_{1}, Β\_{2}\right)=\sum\_{i=1}^{n}Ε\_{i}^{2}= \sum\_{i=1}^{n}\left(γ\_{i}- Υ\_{i}\right)\_{i}^{2}.$$

Суть методу мінімальних квадратів полягає в підборі таких $Β\_{1}, Β\_{2}$, щоб $Q(Β\_{1}, Β\_{2})$ була мінімальною. Підставивши конкретні дані під час побудови моделі в результаті необхідно буде вирішити задачу мінімізації функції кількох змінних, для визначення $Β\_{1},Β\_{2}$ коефіцієнтів.

Слід пам'ятати, що при зміні навіть одного значення вихідних даних, всі коефіцієнти змінять свої значення, оскільки вони повністю визначаються вихідними даними.

Регресійну модель можна ускладнити, звівши до мінімуму кількість помилок на кожному етапі побудови моделі. Найбільш зрозумілий і простий, разом з тим часто використовуваний спосіб математичного програмування для вирішення таких задач – метод градієнтного спуску (gradient descent). Це ітераційний алгоритм, на кожному кроці якого вектор ваг $θ$ змінюється в напрямку найбільшого зменшення цільового функціоналу.

$θ=θ-η×∇J\_{θ}(θ)$*,*

де *η* – величина, яка називається швидкістю навчання.

Швидкість навчання η визначає розмір кроків, які необхідно зробити, щоб досягти (локального) мінімуму. Обрання параметра швидкості навчання (learning rate) відбувається експериментальним шляхом, залежить від природи даних, для яких відбувається пошук значення даного параметра. Якщо мінімізуючий метод досить гладкий, то параметр η можна обрати досить великим, після чого час виконання алгоритму навчання буде меншим. Але якщо цільова функція має велику кривизну, то у випадку великих значень швидкості процесу навчання можлива ситуація, де вектор ваг $θ$ буде «перескакувати» своє оптимальне значення. Для подібних випадків слід брати малі значення параметра η – швидкість збіжності до (локального) мінімуму функції буде малою, проте менший шанс його пропустити.

Існує три варіанти градієнтного спуску, які відрізняються тим, скільки даних використовується для обчислення градієнта цільової функції. Залежно від обсягу даних потрібно шукати компроміс між точністю оновлення параметрів і часом, необхідним для їх оновлення.

Пакетний (batch) – на кожній ітерації алгоритму навчальна вибірка проглядається цілком і після цього розраховується нове значення. Оскільки нам потрібно розрахувати градієнти для всього набору даних, щоб виконати лише одне оновлення, пакетний градієнтний спуск може бути дуже повільним і непридатним для наборів даних, які не поміщаються в пам’ять. Пакетний градієнтний спуск також не дозволяє оновлювати модель онлайн, тобто за допомогою нових даних.

Алгоритм пакетного методу реалізований досить просто – потрібно виконувати оновлення ваг за допомогою усередненого значення антиградієнта на всіх прикладах з навчальної вибірки до тих пір, поки оцінка функціоналу J не стабілізується, тобто поки на стане змінюватися його значення занадто повільно або не припинить змінюватися взагалі, або до тих пір, поки змінюються ваги *η*.

Cтохастичний градієнтний спуск оновлює параметри для кожного конкретного значення *х* та *y*. Стохастичний градієнтний спуск виконує одне оновлення за раз. Тому зазвичай працює набагато швидше, і його також можна використовувати «на льоту», оновлюючи значення під час виконання. Стохастичний метод виконує часті оновлення з високою дисперсією, що спричиняє значні коливання цільової функції. Метод стохастичного градієнту дає можливість використовувати вибірки значних розмірів для навчання на відмінну від пакетного методу. Також в даному методі можна змінювати саму роботу алгориту навчання – якщо вибірка занадто велика, то після перегляду об'єкту з цієї вибірки об'єкт можна видаляти і тим самим покращувати оптимізацію використання ресурсів комп’ютера і покращити швидкодію.

Водночас метод стохастичного градієнта також має певні недоліки, зокрема:

* можлива надто повільна збіжність алгоритму;
* якщо функціонал містить велику кількість екстремальних точок, то є доволі велика ймовірність «застрягання» алгоритму в локальних мінімумах. Для уникнення цього застосовується метод випадкової модифікації вектору, внаслідок чого одна або декілька координат вектору частково змінюються і тим самим зменшується ймовірність застрягання.

Можливе використання компромісного методу – коли параметри оновлюються кожні n кроків. Таким чином зменшується дисперсія оновлення параметрів, що може призвести до більш стабільної збіжності; і можуть використовуватись матричні оптимізації, звичайні для сучасних бібліотек машинного навчання. Поширені значення n коливаються від 50 до 256, але можуть відрізнятися для різних реалізацій. Компромісний градієнтний спуск, як правило, є алгоритмом вибору під час навчання нейронної мережі, і термін стохастичний градієнтний спуск зазвичай використовується також, коли використовуються міні-пакети (рисунок 3.1).



Рисунок 3.1 – Коливання цільової функції про компромісному градієнтному спуску.

## 3.3 Реляційна база даних системи

У сучасних інформаційних системах для забезпечення роботи з базами даних використовують системи керування базами даних (СКБД). Система керування базами даних – це система, заснована на програмних та технічних засобах, яка забезпечує визначення, створення, маніпулювання, контроль, керування та використання баз даних. У медичних системах бази даних використовуються для зберігання та оброблення інформації про пацієнтів, результатів дослідження тощо. Наявність бази даних дозволяє проводити статистичний аналіз. БД є ключовим елементом більшості інформаційних систем, в тому числі і спеціалізованих.

Основними перевагами використання баз даних є:

* підтримка цілісності даних;
* підвищена безпека і захист від несанкціонованого доступу;
* підвищення ефективності при зростанні навантаження на систему;
* спільне використання даних;
* контроль за надмірністю даних;
* несуперечливість даних;
* розвинені служби резервного копіювання та відновлення даних.

Одночасно враховуючи специфіку даної роботи не всі дані, які використовує програма, будуть знаходитись в базі даних. Для зменшення навантаження на базу даних та збільшення швидкості роботи такі дані як безпосередньо зображення, що використовуються програмою, а особливо – великі за об’ємом навчальні вибірки зображень для нейромережі, а також навчені згорткові нейромережі, які використовуються лише для одноразового завантаження/вивантаження, будуть зберігатись окремо у файловому сховищі.



Рисунок 3.2 – Даталогічна модель бази даних

Таблиці в базі даних можна умовно розділити на кілька груп на основі аналізу даних які в них знаходяться та є згрупованими на основі деяких характеристи. Дані групи є достатньо умовними та можуть бути зміненими під час подальших аналізів роботи або при зміні поставлених задач.

Таблиця auth\_users зберігає інформацію про користувачів системи. Крім загальної інформації про користувача таблиця містить поле user\_group, яке з іншої таблиці user\_groups зчитує інформацію про групу користувача і його дозволи для роботи в системі. Таблиця logs служить для додаткового документування дій користувачів та видачі їх користувачам у форматі звіту.

Таблиця 3.2 – Структура таблиці бази даних auth\_users

|  |  |  |  |
| --- | --- | --- | --- |
| Поле | Тип даних | Опис поля | Ключ |
| iduser | INT | Ідентифікатор користувача | Первинний |
| user\_login | VARCHAR(255) | Логін |  |
| user\_name | VARCHAR(255) | Ім’я користувача |  |
| user\_group | INT |  | Зовнішній |

Варто візначати, що система ідентифікації користувачів має на меті в першу чергу фіксацію дій користувачів в системі для полегшення роботи з програмою, перегляду звітів роботи, можливості створення конкретним користувачем своїх моделей тощо. Задача організації безпеки даних, шифрування на рівні програмного забезпеченя в даній роботі не ставилась. Організація безпеки та забезпечення цілісності даних відбувається засобами бази даних.

Таблиці CNN, train\_dataset та dataset\_files призначені для роботи з нейромережами. В таблиці CNN зберігаються службові дані про нейромережу, таблиця train\_dataset містить інформацію про навчальні вибірки, сформовані користувачем. Безпосередньо самі нейромережі зберігаються на диску у вигляді окремих файлів.

Опрацювання медичних зображень є ключовим етапом в розробленій системі. Після етапу попередньої оброблки та сегментації відбувається етап обчислення кількісних ознак мікрооб’єктів, які зберігаються в базі даних для подальшої класифікації. У таблиці «Сell» зберігається інформація про площу, периметр, висоту, ширину, окружність контуру, координати центру, ширину, висоту обмежуючого прямокутника кожної клітини. Оскільки однією з характеристик цитологічних та гістологічних зображень є мікрооб’єкти, оточені складним за геометричними і оптичними характеристиками фоном, то доцільніше виділяти не самі мікрооб’єкти, а їх ядра. Ядра на зображеннях характеризуються областями з низьким рівнем яскравості. Таблиця “Core” містить інформацію кількісні характеристики по кожному ядру.

Структуру таблиці Core наведено в табл. 3.3.

Таблиця 3.3 – Структура таблиці бази даних Core

|  |  |  |  |
| --- | --- | --- | --- |
| Поле | Тип даних | Опис поля | Ключ |
| idCore | INT | Ідентифікатор ядра | Первинний |
| idCell | INT | Ідентифікатор клітини | Зовнішній |
| OutlineNumb | INT(11) | Контур |  |
| Area | FLOAT(11) | Площа |  |
| Perimeter | FLOAT(11) | Периметр |  |
| Height | FLOAT(11) | Висота |  |
| Width | FLOAT(11) | Ширина |  |
| xO | FLOAT(11) | координата Х |  |
| yO | FLOAT(11) | координата Y |  |
| LargeAxis | FLOAT(11) | Велика вість |  |
| SmallAxis | FLOAT(11) | Мала вісь |  |

Після виділення об’єктів необхідно провести їхній опис. Опис зображення – це процес виділення інформативних ознак об’єкта.

Розглянемо конкретно операції по обчисленню кількісних характеристик описаних мікрооб’єктів.

Загальна площа об’єкта оприділяється як область яку описує контур об’єкта. Дуже часто під площею об’єкта розуміють ту частину зображення яка належить об’єкту. Це твердження може бути невірним якщо в об’єкті присутні «дірки», в такому випадку їхню площу потрібно вираховувати окремо. Однак для підрахунку загальної площі враховується також і площа дірок, а якщо об’єкт є суцільним, то відповідно загальна площа та просто площа будуть рівними. Для знаходження площі ядра використовуємо наступну формулу (площа ядра – кількість точок, що належать ядру клітини):

.

Периметр – це довжина контуру об’єкта та обчислюється за наступною формулою:

.

Однією з важливих ознак об’єкта є велика вісь. Для визначення великої осі необхідно визначити на контурі дві найвіддаленіші точки.

Для подальшого аналізу також слід обчислити компактності, що визначається за наступною формулою.

.

При роботі з медичними зображеннями, в першу чергу при дослідженні формування новоутворень, дуже часто необхідно вирахувати ядерно-цитоплазматичне відношення – відношення площі ядра до площі цитоплазми – при багатьох видах пухлин це співвідношення різко зростає. Завчасна діагностика таких клітин і відповідно пухлин є однією з найбільш розповсюджених задач аналізу медичних зображень.

Тому додатково розглянемо обчислення кількісних характеристик клітин для таблиці Cell.

Обчислення площі цитоплазми – площа клітини без врахування площі ядра:

.

Ядерно-цитоплазматичне відношення – відношення прощі ядра до площі цитоплазми :

.

Ядерна щільність визначається за формулою:

,

де *n*1, *n*2, *n*3 – кількість ядер на  у кожній дослідженій ділянці.

На підставі отриманих даних про площу ядра клітини і ядерної щільності пухлини була розрахована площа цитоплазми клітин досліджуваних новоутворень по формулі:

,

де  – шукана середня площа цитоплазми клітини;

  – середня площа ядра пухлинної клітини цього новоутворення;

  – ядерна щільність цього новоутворення.

З урахуванням отриманих даних проводилася оцінка ядерно-цитоплазматичного відношення в пухлинних клітинах новоутворень по формулі:

,

де  – середня площа цитоплазми клітини даного новоутворення;

  – середня площа ядра пухлинної клітини цього новоутворення.

Для статистичної обробки даних застосовувалися непараметричні методи. Вираховувались середні значення вибірки, а також мінімальні та максимальні значення.

Таблиця “Model” служить для зберігання моделей лінійної регресії. В залежності від постановки задачі моделі з таблиці можна завантажувати, оновлювати, додатково навчати, та використовувати для прогнозування кількісних показників мікрооб’єктів на конкретному зображені.

## 3.4 Тестування результатів роботи

В загальному вигляді структура системи опрацювання кількісних характеристик об’єктів біомедичних зображень представлена на рисунку 3.2.

В цілому роботу програмного комплексу можна умовно розділити на три етапи – етап попередньої обробки, етап опрацювання зображення та етап побудови моделі регресії на основі отриманих даних. Відповідно у реалізації програмного комплексу було передбачено три великих програмних модулі для кожного етапу. Крім того, додатково було реалізовано допоміжні модулі для організації графічного інтерфейсу, діалогу з користувачем, роботи за файловою системою та базою даних.

При умовному поділі модулів на основні та допоміжні потрібно виділити одну особливість – модуль навчання згорткової нейронної мережі за логікою побудови програми повинен відноситись до допоміжних модулів, оскільки етап навчання ЗНМ в проводиться до безпосередньо роботи з зображеннями. Однак якщо врахувати що саме цей модуль є одним з найскладніших в реалізації, а виконання саме цієї частини програми забере найбільше часу та використає найбільше ресурсів, ця частина програми буде розглядатись при описі роботи основного алгоритму як частина модуля по роботі зі згортковою нейронною мережею.

Рисунок 3.2 – Загальна структура роботи програмного комплексу

Завантаження зображення

Попередня обробка

Класифікація зображення

Сегментація

Виділення мікрооб’єктів

Вибір структури ЗНМ

Завантаження ЗНМ

Збереження структури ЗНМ у файл

Завантаження навчальної вибірки

Навчання ЗНМ

ЗНМ

Обчислення кількісних характеристик мікрооб’єктів

База даних

Створення моделі

Навчання

Прогнозування

Вивід/експорт результатів

Лінійна регресія

Модуль preparation містить набір функцій по реалізації попередньої обробки зображення. Реалізовано алгоритми гаусової фільтрації, медіаної фільтрації та додатково гістограмного вирівнювання. Детальніше конкретно алгоритми попередньої обробки вже було розглянуто в попередніх розділах. Цей модуль не працює з базою даних і для зберігання результатів роботи використовує файлову систему. Зберігання зображень в основній базі даних могло мати наслідком зниження продуктивності та надійності останньої. Використання додаткової бази даних для зберігання зображень було визнано недоцільним – це збільшило б навантаження на сервер БД та зменшило надійність програмного комплексу.

Модуль segment містить набір функцій для реалізації процесу сегментації зображень. Умовно роботу модуля можна розділити два великих блоки – блок створення/навчання згорткової нейронної мережі та блок сегментаці зображень. Кожен блок має свій інтерфейс для взаємодії з користувачем.

Створення та навчання згорткової нейронної мережі необхідно провести заздалегідь, до початку роботи над зображеннями. Фактично, виконання цього блоку не входить до основного алгоритму роботи програми, але є необхідним для подальшої роботи.

В користувача є можливість створити нову нейромережу із заданими параметрами, або завантажити з файлу раніше створену і/або навчену мережу. Різні нейромережі можуть бути навчені з різними параметрами і бути орієнтованими на сегментацію певних видів зображень. Вибір нейромережі в даному випадку покладений на користувача і проводиться останнім на основі візуального огляду або опису зображення.

Навчання нейромережі є досить складним і тривалим процесом. Який крім того потребує багато ресурсів, в основному – обчислюваних ресурсів та часу роботи комп’ютера. Час навчання мережі в сильно залежить від кількості класів, на які поділено зображення, типу процесора (CPU чи GPU), швидкості навчання тощо. Як було зазначено вище, використання GPU для проведення обчислень дозволяє значно прискорити процес навчання нейромережі.

Рисунок 3.3 – Порівняльний аналіз результатів навчання нейронної мережі для класифікації цитологічних зображень для GPU та CPU.

Час навчання нейронної мережі значною мірою залежить від кількості класів вихідного зображення, ніж від кількості самих зображень. На швидкість навчання впливає багато факторів

Одним з параметрів, який може встановити користувач, є величина норми навчання, що впливає значення зворотнього поширення помилки та на перерахунок вагових коефіцієнтів нейронів. Чим більша ця норма, тим швидше мережа буде навчатися, чим менша ‒ тим повільніше, однак використання дуже великих або дуже малих значень може сильно погіршити якість результатів. Зазвичай, норму навчання вибирають в межах від 0.01 до 0.000001.

Епоха – це один період дискретизації, що містить процес прямого та зворотнього поширень а також функції втрат і оновлення ваг.

Параметр F1 Score ‒ це число від нуля до одного, що пояснює, наскільки успішними були тренування мережі. Це аналогічно до відсотка, 1 – найкращий результат, а нуль – найгірший. Фактично, F1 Score – це вірогідність того, що результати мережі є правильними.

Таблиця 3.3 – Порівняльний аналіз результатів навчання ЗНМ

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| К-сть зобра-жень | Класів, од. | Епох, од. | Навчаль-ні дані до тестових | Ф-ція актива-ції | Норма навчання | Час навчання (GPU), хв | F1 Score, % |
| 86 | 4 | 50 | 0.7 | ReLU | 0.01 | 21 | 75 |
| 86 | 4 | 60 | 0.8 | ReLU | 0.01 | 24 | 85 |
| 86 | 4 | 50 | 0.7 | ReLU | 0.01 | 22 | 40 |
| 86 | 4 | 60 | 0.8 | ReLU | 0.01 | 32 | 78 |
| 86 | 4 | 55 | 0.7 | ReLU | 0.01 | 26 | 90 |
| 86 | 4 | 50 | 0.8 | ReLU | 0.01 | 17 | 93 |

При навчанні нейромережі також необхідно уникати ситуації «перенавчання». Така ситуація виникає тоді, коли при досягненні високих результатів при подальшому навчанні якість розпізнавання починає погіршуватись. В такому випадку, кількість ітерацій навчальних циклів для досягнення оптимальних результатів необхідно підбирати експериментально.

Модуль regression містить функції для створення та використання моделей лінійної регресії на основі отриманих раніше при обробці результатів сегментації зображення даних. Цей модуль працює безпосередньо з базою даних. Створені та побудовані моделі можна зберегти в базі даних та завантажити при необхідності. Також вже існуючу модель можна «доповнити» новими даних, наприклад, на основі нового зображення. Перерахунок моделі на основі нових даних умовно можна назвати навчанням моделі. Отримані результати роботи моделей можна використовувати для оцінки залежності параметрів та прогнозування.

Для оцінки продуктивності роботи алгоритму побудови моделей лінійної регресії зазвичай середні відхилення конкретних значень від функції моделі. Ми будемо використовувати значення середньоквадратичної помилки $RMSE$ для тестової моделі. Рисунки 3.4-3.6 ілюструють результати дослідження.

На рисунку 3.4 наведено результат застосування лінійної регресії для параметрів «площа, довжина головної осі».



Рисунок 3.4 – Результат застосування лінійної регресії для параметрів «площа, довжина головної осі»

Значення RMSE для параметрів «площа, довжина головної осі» становить 26.15.

На рисунку 3.5 наведено результат застосування лінійної регресії для параметрів «площа, периметр».



Рисунок 3.5. Результат застосування лінійної регресії для параметрів «площа, периметр»

Значення RMSE для параметрів «площа, периметр» становить 38.21.

Результати порівняння параметрів кількісних характеристик мікрооб’єктів засобами лінійної регресії наведено на рисунку 3.7.



Рисунок 3.6 – Результати порівняння параметрів кількісних характеристик мікрооб’єктів на основі лінійної регресії

В результаті порівняння можна зробити висновок, що використання сукупності параметрів «площа, довжина головної осі» має меншу похибку RMSE порівняно із сукупністю параметрів «площа, периметр».

У даній роботі детально проаналізовано та наведено усі етапи опрацювання зображень: фільтрацію, попереднє оброблення, сегментацію на основі згорткових нейронних мереж, виділення кількісних характеристик ядер клітин і застосування лінійної регресії.

## 3.5 Висновки до розділу 3

Сформовано навчальну вибірку для навчання згорткової нейронної мережі на основі обраного набору харіктеристи біомедичних зображень, що дозволило провести процес навчаня згорткової нейронної мережі;

Розроблена реляційна базу даних та реалізована системи опрацювання зображень на основі запропонованих вище алгоритмів та структур, що дозволило провести опрацювання та оцінювання біомедичних зображень.

# Висновки

На основі проведених аналітичних досліджень методів та алгоритмів опрацювання зображень, а також відомих програмних систем аналізу та обробки біомедичних зображень можна зробити такі висновки:

1. Проведено дослідження алгоритмів обробки зображень яке дозволило сформувати послідовність роботи груп алгоритмів в процесі роботи програмної системи.
2. Проведено дослідження структур та функціональних можливостей програмних засобів обробки біомедичних зображень, що дозволило в подальшому спроектувати власну систему оброблення зображень.
3. На основі визначених кількісних характеристик ядер клітин, завдяки використанню методу лінійної регресії, спрогнозовано кількісні показники мікрооб’єктів на цитологічному зображенні.
4. Сформовано навчальну вибірку для навчання згорткової нейронної мережі на основі обраного набору характеристик біомедичних зображень, що дозволило провести процес навчання згорткової нейронної мережі;
5. Розроблена реляційна базу даних та реалізована системи опрацювання зображень на основі запропонованих вище алгоритмів та структур, що дозволило провести опрацювання та оцінювання біомедичних зображень.
6. На основі проведених комп’ютерних експериментів досліджено, що сукупність параметрів «площа, довжина головної осі» має в 1.4 рази меншу похибку RMSE порівняно з сукупністю «площа, периметр».

# Список використаних джерел

1. Березький О.М. Методи, алгоритми та програмні засоби опрацювання біомедичних зображень: монографія / [ О.М. Березький, Ю.М. Батько, К.М, Березька, С.О. Вербовий, Т.В. Дацко, Л.О. Дубчак, І.В. Ігнатєв, Г.М. Мельник, В.Д. Николюк, О.Й. Піцун ]; під наук. ред. Березький О.М., Тернопіль . ТНЕУ «Економічна думка», 2017. - 330 с.
2. Bradski G. Learning OpenCV - Computer Vision with the OpenCV Library / G. Bradski.,2018 – 580 с.
3. Гонсалес Р. Цифровая обработка изображений в среде MATLAB / Р. Гонсалес, Р. Вудс, С. Эддинс. - М.: Техносфера, 2016. – 616 с.
4. Сафонов В. О. Параметризованные типы данных. История, теория, реализация и применение / В.О. Сафонов. – М.: Издательство Санкт-Петербургского университета, 2013. – 116 c.
5. Секунов Н.Ю. Самоучитель Visual C++ 6.0 / Н.Ю. Секунов. – М.: СПб: BHV, 2014. - 960 c.
6. Дороничева А.В. Методы Распознавания Медицинских Изображений Для Задач Компьютерной Автоматизированной Диагностики / А.В. Дороничева.,2014. – 10 – 18 с.
7. Барінов Е. Ф. Гістологія, цитологія та ембріологія. Кн.1: Цитологія і загальна ембріологія: навч. посіб. для студ. вищ. мед. навч. закл. IV рівня акредитації / Е. Ф. Барінов, Ю. Б. Чайковський . – К. : Медицина – 2010. – 216 c.
8. Тюрин Ю.Н. Анализ данных на компьютере. / Ю.Н. Тюрин, 2015. – 384с.
9. Березький О.М. Аналіз алгоритмів обробки зображень. / О.М. Березький, Ю.М. Батько.,2006. –212-216c.
10. Березский О.Н. Алгоритмы анализа и синтеза биомедицинских изображений / О.Н.Березский.,2007. – 134-144c.
11. Березький О.М. Алгоритм проходження контуром об’єкта з використанням зворотного ходу / О.М. Березький, Ю.М. Батькою.,2009 –516-523с.
12. Березький О.М. Методи та алгоритми перетворення контурів зображень в афінному просторі. / О.М.Березький.,2009. – 185-189с.
13. Дуда Р. Распознавание образов и анализ сцен. / Р. Дуда, П. Харт.,1976. – 511 с.
14. Гонсалес Р. Цифровая обработка изображений. / Р. Гонсалес, Р. Вудс.,2005. – 1072 с.
15. Сойфер В.А.. Методы компьютерной обработки изображений. / В.А. Сойфер., 2003.– 784 с.
16. Фурман Я.А. Введение в контурный анализ; приложения к обработке изображений и сигналов. / Я.А.Фурман, А.В Кревецкий, А.К. Передреевю,2003. – 592 с.
17. Chitradevi B. An Overview on Image Processing Techniques / B. Chitradevi, P.Srimathi.,2014. – 6466-6472с.
18. Pablo F. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces / F. Pablo.,2013. – 323с.
19. Jayashree Kalpathy-Cramer. Evaluating performance of biomedical image retrieval systems – an overview of the medical image retrieval task at ImageCLEF 2004– 2013 / Alba García Seco de Herrera,.2015. – 55–61с.
20. Berezsky O. Algorithm of determination of image contours of biological nature / O. Berezsky, Yu. Bat’ko.,2006. –642-644с.
21. Яне Б. Цифровая обработка изображений. / Б. Яне., 2007. – 584 с.
22. Романенко І.О. Аналіз ефективності сучасних методів сегментації цифрових зображень / І.О. Романенко.,2016. – 172-174с.
23. ШапироДж. Компьютерное зрение. / Дж. Шапиро, Дж. Стокман.,2006. – 752с.
24. Страуструп, Б. Язык программирования C++. Специальное издание / Б. Страуструп., 2008. – 1104 с.
25. Шилдт, Г. С++: базовый курс / Г. Шилдт – М.: Издательский дом «Вильямс», 2008. – 624 c.
26. Васильев, А.Н. Самоучитель C++ с примерами и задачами/ А.Н. Васильев. – СПб.: Наука и Техника, 2010. – 480с.
27. Павловская, Т.А. С/ С++. Программирование на языке высокого уровня / Т.А. Павловская. – СПб.: Питер, 2011. – 461 с.
28. Эферган М. С++: справочник / М. Эферган.,2018. 256с.
29. Чен М.С. Программирование на С++:1001 совет: Наиболее полное руководство по Java и Visual J++/ М.С. Чен., 2017. – 640с.
30. Вебер Дж. Технология С++в подлиннике. QUE Corporation / Дж. Вебер,2017. – 256с.
31. Березький О.М. Segmentation algorithms of biomedical images: development and quantitative evaluation / О.М. Березький, Ю.М. Батько, Г.М. Мельник, С.О. Вербовий, О.Й. Піцун // Штучний інтелект. – Київ. – 2016. – №3 (73). – С. 104-116.
32. Березький О.М. Класифікація гістологічних та цитологічних зображень на основі згорткових нейронних мереж / О.М. Березький, О.Й. Піцун, А.Р. Боднар, Т.М. Долинюк // Штучний інтелект. – Київ. – 2017. – №1 (75). – С. 33-42.
33. Hopfild J.J.Neural networks and physical systems with emergent collective computational abilities. / J.J. Hopfild// Proc. Nat. Acad. Sci. USA. – 1982. – 79. – Pp.2554-2558.
34. Изотов П.Ю Технология Реализации Нейросетевого Алгоритма В Среде Cuda На Примере Распознавания Рукописных Цифр / П.Ю. Изотов, С.В.Суханов, Д.Л Головашкин // Компьютерная оптика . – Т 34, №2. – 2010. – C. 243-251 153.
35. Юдин Д.А. Применение метода екстремального обучения нейронной сети для классификации областей изображения / Д.А. Юдин, В.З. Магергут// Научние ведомостиБелГУ, Серия: Информатика. – 2013. - № 8(151). – С. 95-103.
36. Berezsky O. GPU – based biomedical image processing / O. Berezsky, O. Pitsun, L. Dubchak, P. Lyaschynsky, P. Lyaschynsky // Proceedings of XIV International Conference Perspective Technologies and methods in mems design (MEMSTECH 2018) 18-22 April, 2018, Lviv-Polyana, Ukraine, P. 96-99.
37. Тюрин Ю.Н. Анализ данных на компьютере. М.: ИНФРА-М, Финансы и статистика, 2015. 384с.
38. База даних цифрових гістологічних та цитологічних зображень передракових та ракових станів молочної залози «BPCI2100». /О.М. 150 Березький, Г.М. Мельник, С.О. Вербовий, О.Й. Піцун, В.Д. Николюк, Т.В. Дацко // Свідоцтво про реєстрацію авторського права на твір №75359 від 14.12.2017 р.
39. Вирт Н. Алгоритмы и структуры данных. М.: Мир, 2019. 345с.
40. Касаткин В. Н. Информация. Алгоритмы. ЭВМ. М.: Просвещение, 2011. 219с.
41. Hopfild J.J.Neural networks and physical systems with emergent collective computational abilities. / J.J. Hopfild// Proc. Nat. Acad. Sci. USA. – 1982. – 79. – Pp.2554-2558.
42. Березький О.М. Системи автоматизованої мікроскопії: стан та перспективи розвитку / О.М. Березький, С.О. Вербовий, О.Й. Піцун // Вісник Хмельницького національного університету. – 2016. – №2 (235). – С. 61-68
43. Марков А. С. «Базы данных. Введение в теорию и методологию. Финансы и статистика». 2016. С.24-35.
44. Березький О.М. Розроблення реляційної бази даних інтелектуальної системи автоматизованої мікроскопії / О.М. Березький, О.Й. Піцун, С.О. Вербовий, Т.В. Дацко // Науковий вісник НЛТУ України : збірник науковотехнічних праць. Львів : РВВ НЛТУ України. – 2017. – №. 27(5). – С.125-129.
45. Голованов Е.А. Основы корреляционного и регрессионного анализа. - М.: Наука, 1991.
46. Приходько А. И. Практикум по эконометрике. Регрессионный анализ средствами Excel : Учеб. пособие для студентов вузов, обучающихся по спец. 061800 "Математические методы в экономике" / А. И.Приходько, 2004 - 179 с.
47. Выгодчикова, И. Ю. Алгоритм оценки параметров линейной множественной модели регрессии по минимаксному критерию / И.Ю. Выгодчикова., 2019. - 216 c.
48. Кремер, Н.Ш. Эконометрика: учебник для вузов / Н.Ш. Кремер, Б.А. Путко.,2010. – 328 с
49. Batko Y. Graphical interface of hybrid intelligent systems for biomedical imaging analysis / Y. Batko, G. Melnyk, O. Pitsun.,2016 - 121-124с.
50. Методичні рекомендації до виконання дипломної роботи з освітньо-кваліфікаційного рівня “Магістр”. Спеціальність „Комп’ютерні системи та мережі”/О.М. Березький, Л.О. Дубчак, Г.М. Мельник /Під ред. О.М. Березького Тернопіль: ТНЕУ, 2018. 41с.